The Workshops of the Tenth International AAAI Conference on Web and Social Media
Wiki: Technical Report WS-16-17

Literature, Geolocation and Wikidata (Extended Abstract)

Finn Årup Nielsen
DTU Compute, Technical University of Denmark
Richard Petersens Plads
DK-2800 Kongens Lyngby, Denmark

Abstract

Littar was the second-prize winning entry in an app competition. It implemented a system for visualizing places mentioned in individual literary works. Wikidata acted as the backend for the system. Here I describe the Littar system and also some of the issues I encountered while developing the system: How locations and literature can be related, what types of location-literature relations are possible within Wikidata, what limitations there are and what questions we may ask once we have enough data in Wikidata.

Introduction

The geographical position datatype enables geotagging of items in Wikidata. Besides real-world objects in the physical world, Wikidata editors can also geotag art and literature that refer to places either directly through geographical properties or indirectly by linking to items which feature a geocoordinate property. Among the interesting applications are maps pinpointing depictions of paintings as implemented in the Crotos search and display engine: The Callisto service of Crotos features a zoomable world map with Wikidata items associated with art, either through geotagged depictions, geotagged collections (usually museums) or direct geocoordinate (the latter is mostly applied on statues).

For literature, Wikidata users can tag literary work items with the narrative location property and this formed the basis for the Littar system that queried Wikidata and created a web page with a map pinpointing narrative locations.

Literature and location

The term "spatial turn" has been used to denote a recent increased interest in places and spaces in humanities and social sciences (Rosiek 2015). In literary science, it means a focus on places and spaces away from a person-centric or chronological description of literature. In their studies, literary historians may consider named real-life geolocatable places but also more generic sites. An example that Rosiek gives on the former is *Gurre*, a Danish castle ruin in Northern Zealand referenced in a number of works, and for the latter the example is *the beach* (Rosiek 2015). The use of specific places may vary over time and Rosiek cites the French historian Alain Corbin who argued that *the beach* (or the more general French word: *rivage*) was "discovered" between 1750 and 1840. Rosiek notes that Corbin's theory is supported for Danish creative works, where it is difficult to find *rivage* paintings or texts before 1750, while there exists a multitude of examples after 1840.

Note that named geographical places does not necessarily associate with a geographical position, e.g., take Agatha Christie's fictional St. Mary Mead. Fictional places is, on the other hand, not necessarily not geolocatable, e.g., Johannes V. Jensen's Graabølle may be taken to be a fictionalized version of his childhood village Farsø.

Apart from the narrative location, literary historians have also consider where a text has been composed, collected, used and "canonized" (Mai 2010, volume 1, page 9). Mai structures her work of Danish literature both with respect to time and place with the earliest places of relevance being the cathedral, the manor, the academy; later, the salon, the vicarage; and in more modern times, the newspaper, the metropolis and the Internet.

It varies considerably how authors choose to make their work geolocatable. For example, the tales of Hans Christian Andersen may be devoid of geolocatable places, while Patrick Modiano has plenty. The first four sentences in his *Fleurs de ruine* has five geolocatable places: streets, an institute, a church, a movie theater, — all in a specific part of Paris. Sometimes the places in the narrative are not just separate points. The movement of the characters may be detailed to such an extent that it is possible to create a geographical path of the narrative. One example is Thomas E. Kennedy's novel *Kerrigan in Copenhagen*, where the main character walks and drives on named streets between named bars in Copenhagen in his effort to write a bar guide book.

Literary historians seem mostly to have been concerned with works of fiction when studying literature and places, but non-fiction works may also relate to places. Places are very often named in non-fiction narrative works that recount events in past times, e.g., historical works, biographies and literature in the true crime genre.

Geolocated literature in Wikidata

The primary property in Wikidata for descriping geolocation in literature is the *narrative location* (P840, 6154).[1] For the property suggestion, the primary domain seems to have been for films, but the domain is now set to general "work (literature, film, music, everything with a narrative)". One even finds some paintings (perhaps controversially) using the property. Usually paintings would use another property, the *depicts* (P180, 580) to describe the location of the content. Surprisingly, one finds several hundreds of applications of this property in literature items. Over a dozen literary works use *location of final assembly* (P1071, 15) to indicate were the work was created. This property is also used by paintings that furthermore can use the *coordinates of the point of view* property (P1259, 0). For films, there is also the *filming location* (P915, 29).[2]

For paintings, *collection* (P195, 334) is often used to describe in which entity the original work appears, e.g., which museum collection. The physical location of statues may be indicated with *coordinate location* (P625, 283) and *location* (P276, 189). For the modern mass-manufactured book, it rarely makes sense to speak about collection or location of the work, but notable individual incunables, such as the *Book of Kells*, has the *collection* property set. Handwritten manuscripts of modern authors may be organized in library collections. Such manuscripts could use the *collection* property if they are described in Wikidata.

Many non-fiction books may be about one specific place: a building, a street, a castle, a town, etc. In this case the *main subject* (P921, 59512) property can be used. *Inspired by* (P941, 73) could be used to describe works that are inspired by locations. In Wikidata, I find only one work set to be inspired from a geolocatable item: The novel *Les toiles de Sidi Moumen* inspired by the 2003 Casablanca bombing. Location related to the publishing of a literary work may be indicated with *place of publication* (P291, 2878) or *publisher* (P123, 27525).

If we focus on the author, then there are several properties that can be used to indicate geolocatable values: *place of birth* (P19), *place of death* (P20), *place of burial* (P119), *educated at* (P69), *residence* (P551) and possibly *employer* (P108). These properties can also be associated with date information. Travelling may be an important part of the life of an author. If the author stays long enough at a place then *residence* could possibly come into use. However, I know of no present Wikidata property to record more temporary stays. For instance, it is unclear how one should record that Johannes V. Jensen was delayed for several hours at Brinkley Station during his travel to U.S.A. (an incident that has relevance, as it inspired him to his landmark poem *At Memphis Station*).

[1] Here P840 denotes the Wikidata property identifier and 6154 is the number of property values for items that are instances of *book* or *literary work* or their subclasses.

[2] Note that the subclass hierarchy of literary works may be fairly broad, e.g., at the time of writing, instances of *superhero film* are regarded as a literary work.

Littar

Littar ("litteratur radar") is a system for visualizing the narrative location of literary works. Presently, it is essentially just a webpage with Javascript referencing a KML file and a tile server, either Google Maps or OpenStreetMap. The demonstration is available from fnielsen.github.io/littar/. It was created in response to an app competition set up by DBC, the Danish national library service, that posed the question "How can data science be used to provide library users with new and better experiences?"

It quickly became clear to me, that the way Wikidata users typically edit the narrative location property, would make it difficult to create an "interesting" map. The narrative location is typically linked to geographically broad items, such as cities and countries. I was interested in geographically more narrow places—streets, buildings, small towns—to get a densely annotated map. Because of the Danish nature of the competition I wanted to focus on Danish narrative locations and as there were few—and the few there were, were tagged on broad levels—I spend a considerable time entering data myself. I created a script that could extract named geographical entities. The gazetteer was established via a Wikidata query, and the returned names were added to a long regular expression, so it could be applied on digitized texts. However, most of the data was added by skimming, reading or rereading works for mentioning of location. I added fictional as well as non-fictional work, poetry, novels, short stories, biographies including autobiographies, true crime, etc.

Apart from the location item itself, I also added an excerpt from the work via the *quote* property (P1683) as a reference for the claim that the narrative takes place at the location. It could be controversial due to copyright. However, since the added excerpts for copyrighted works were typically just a single sentence, it could be argued that it is well within fair use. For public domain works, one may be safe to add longer quotes, but still restricted by the 400 character limit imposed by Wikidata. Most of the entered excerpts was in Danish, a few in English. In Wikidata contexts, it is somewhat unusual to use the reference as a field for "content"-like data.

Once the data existed in Wikidata, I used one large SPARQL query to get work, location, genre, geographical coordinate, author and reference excerpt in one call. From the returned data I used a Python program to construct a KML file, which was then referenced in Javascript code on the webpage. Each narrative location was associated with a KML Placemark tag and the description tag under the Placemark was set to show location and the list of location-work relations including the excerpt for that location. Narrative locations set to Copenhagen and Denmark were left out. I controlled the color of the Placemark with the genre of the work. For each listed work, a link was also created to the DBC search engine at bibliotek.dk based on the title of the work. The map is shown in Figure 1.

Discussion

A considerable number of places mentioned in the literature, that I used as the basis for Littar, were not initially existing

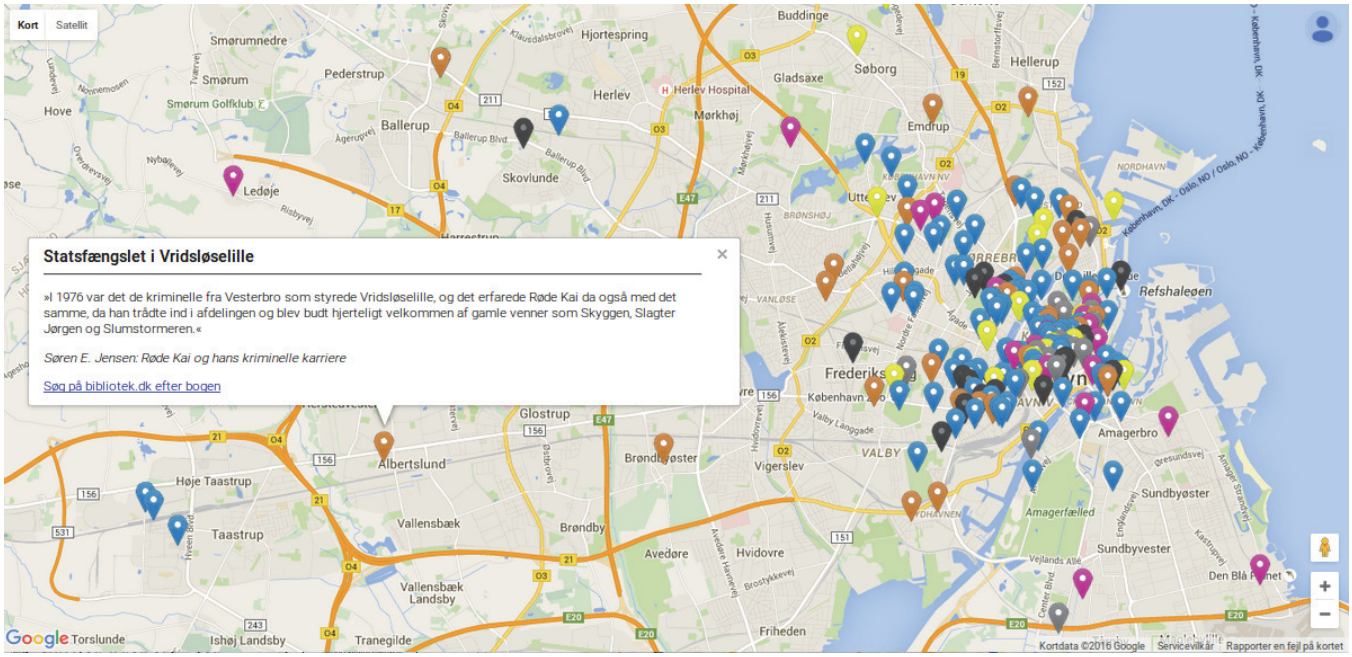

Figure 1: Littar map with Google Map and a quote from the book *Røde Kai og hans kriminelle karriere* by Søren E. Jensen. Here the user has navigated to the Copenhagen suburb Albertslund and pressed the placemark for The State Prison in Vridsløselille, so the popup with information about the location and work appears. Note the concentration of placemarks at the center of Copenhagen.

in Wikidata and had to be created. Such places will not be detected by a named entity recognizer based on Wikidata data.

In some works, the narrative takes place in named organizations. Wikidata users have a tendency to avoid geotagging organizations on the main claim level, and instead move them to a qualifier below the *headquarters location* property. My SPARQL query missed a few geotagged places on that account, —a more complicated UNION SPARQL query could possibly identify these "hidden" geographical coordinates.

The non-geolocatable[3] locations that have been used the most as narrative locations are fictional towns, such as Öreskoga, Glimmerdagg, Santa Teresa and Orbajosa. They are far outnumbered by the locations which are geolocatable. New York City, London and Paris are the most popular tags, and American states, different cities, countries and counties are those that appear most often as narrative locations. Appalachian Mountains and Christiansborg Palace are among the few examples that do not fit into the group of common types of frequent narrative locations. The "lack" of narrative locations means that Wikidata yet is less relevant for "computational literary science". Lets say that we want to determine how the *beach* is used through time to investigate the theory of Corbin. I find that Wikidata describe 203 painting as depicting the beach, see Figure 2, while Wikidata has no works set with the beach as a narrative location: Considerable work needs to be done before Wikidata becomes relevant for this mode of analysis in computational literary science. Indeed, having spend some time on editing depictions of artworks, I have found it considerably faster to tag paintings, than to tag literary works with narrative location. A considerable effort lies ahead. Natural language processing (NLP) has a long history of establishing annotated text corpora, and many resources are readily available for NLP tools, see, e.g., (Bird, Klein, and Loper 2009). With the advent Amazon's Mechanical Turk, crowd-sourcing technologies have become popular in this domain, see, e.g., (Callison-Burch and Dredze 2010). Following in the line of such efforts, a semi-automated system with automated extraction of place names from literary works and subsequent feed into a gamified interface, like Magnus Manske's *Wikidata Game*, could help.

One limitation in Wikidata is the scope of the narrative location property. In some literary texts, the place is merely described with no "action" going on. This can particularly be the case for poems, e.g., in the Danish national anthem we are told that broad beaches stand near salty *eastern beach*. Here the beach can hardly be said to a narrative location and it is neither the main subject. There is a need for a Wikidata property that can do to literary work what the depict property does for artworks. A property that can be used in a wider sense than the narrative location and main subject properties.

How can we record what function a place has? Home, work, murder site, travel destination. For instance, when crime writer Jussi Adler-Olsen let his protagonist has his home in Allerød we can easily indicate the narrative location as Allerød, but it is unclear how we should indicate that it is

[3] which do not have P625 property

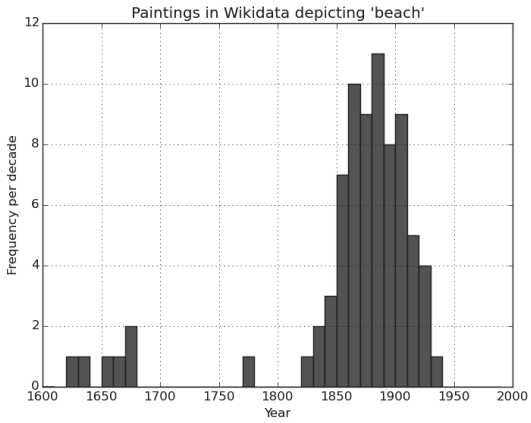

Figure 2: Paintings in Wikidata depicting "beach" as a function of decade.

his home, — perhaps with the generic qualifier *as* (P794)?

Some newer Wikidata visualization systems use on-the-fly queries to the SPARQL endpoint and create the visualization on-the-fly with Javascript libraries. The examples are Aron Ambrosiani's "Museum on Wikidata" from aron.ambrosiani.se/museums/ and angryloki's "Wikidata graph builder" from angryloki.github.io/wikidata-graph-builder/. Most recently map generation has been added as a result type in the Wikidata query service. The present Littar SPARQL query is too slow for on-the-fly queries, but may possibly be modified to a query that does not fetch all aspects of the data needed in the visualization in the first call.

The number of different places in the present map of the Littar application is low. I count 6154 narrative locations for literary works globally and 514 in Denmark. If the number of distinct narrative location grows considerably then there is a need for a system that adaptively loads data, like presently done in Callisto.

As I entered data, I saw a concentration of narrative locations in the center of Copenhagen compared to the rest of Denmark, see Figure 1. Recent research has addressed possible geographical biases in peer-produced works (Johnson et al. 2016; Straumann and Graham 2014). Is there are literary geographical bias? Can the Copenhagen concentration be explained by a per capita effect? Is there an effect for other cities? Or is it only related to a biased Wikidata entry? These are some of the questions I hope to address in future work.

References

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural Language Processing with Python*. Sebastopol, California: O'Reilly.

Callison-Burch, C., and Dredze, M., eds. 2010. *Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk: Proceedings of the Workshop*. Stroudsburg, PA, USA: Association for Computational Linguistics.

Johnson, I. L.; Lin, Y.; Li, T. J.-J.; Hall, A.; Halfaker, A.; Schning, J.; and Hecht, B. 2016. Not at home on the range: Peer production and the urban/rural divide. In *CHI'16*. ACM.

Mai, A.-M. 2010. *Hvor litteraturen finder sted*. Gyldendal.

Rosiek, J. 2015. *Danmark, Gurre, stranden. Steder i dansk litteratur*. U Press.

Straumann, R., and Graham, M. 2014. The geographically uneven coverage of Wikipedia. Information Geographies at the Oxford Internet Institute.