Station to Station: Linking and Enriching Historical British Railway Data

Mariona Coll Ardanuy[1,4], Kaspar Beelen[1,4], Jon Lawrence[3], Katherine McDonough[1,4], Federico Nanni[1], Joshua Rhodes[1], Giorgia Tolfo[2] and Daniel C.S. Wilson[1,4]

[1] *The Alan Turing Institute, London, United Kingdom*
[2] *The British Library, London, United Kingdom*
[3] *The University of Exeter, Exeter, United Kingdom*
[4] *Queen Mary University of London, London, United Kingdom*

Abstract

The transformative impact of the railway on nineteenth-century British society has been widely recognized, but understanding that process at scale remains challenging because the Victorian rail network was both vast and in a state of constant flux. Michael Quick's reference work *Railway Passenger Stations in Great Britain: a Chronology* offers a uniquely rich and detailed account of Britain's changing railway infrastructure. Its listing of over 12,000 stations allows us to reconstruct the coming of rail at both micro- and macro-scales; however, being published originally as a book, this resource was not well suited for systematic linking to other geographical data. This paper shows how such a minimally-structured historical directory can be transformed into an openly available structured and linked dataset, named `StopsGB` (Structured Timeline of Passenger Stations in Great Britain), which will be of widespread interest across the historical, digital library and semantic web communities. To achieve this, we use traditional parsing techniques to convert the original document into a structured dataset of railway stations, with attributes containing information such as operating companies and opening and closing dates. We then identify a set of potential Wikidata candidates for each station using DeezyMatch, a deep neural approach to fuzzy string matching, and use a supervised classification approach to determine the best matching entity.

Keywords

entity linking, digital humanities, open science, toponym resolution, railway stations

1. Introduction

The transformative impact of the railway on nineteenth-century British society has been widely recognized, but understanding that process at scale remains challenging because the Victorian rail network was both vast and in a state of constant flux. Several machine-readable resources exist that include information on the British railway system. However, those that are openly available lack both coverage as well as historical specificity. In contrast, Michael Quick's

CHR 2021: Computational Humanities Research Conference, November 17–19, 2021, Amsterdam, The Netherlands

mcollardanuy@turing.ac.uk (M. Coll Ardanuy); kbeelen@turing.ac.uk (K. Beelen); j.lawrence3@exeter.ac.uk (J. Lawrence); kmcdonough@turing.ac.uk (K. McDonough); fnanni@turing.ac.uk (F. Nanni); jrhodes@turing.ac.uk (J. Rhodes); giorgia.tolfo@bl.uk (G. Tolfo); dwilson@turing.ac.uk (D.C.S. Wilson)
0000-0001-8455-7196 (M. Coll Ardanuy); 0000-0001-7331-1174 (K. Beelen); 0000-0001-6561-6381 (J. Lawrence); 0000-0001-7506-1025 (K. McDonough); 0000-0003-2484-4331 (F. Nanni); 0000-0002-4017-2777 (J. Rhodes); 0000-0001-6886-775X (D.C.S. Wilson)

reference work *Railway Passenger Stations in Great Britain: a Chronology*[1] offers a uniquely rich and detailed account of Britain's changing railway station infrastructure. It includes over 12,000 stations with information such as their opening and closing dates and operating companies.

Quick's *Chronology* has been an important resource for railway enthusiasts and historians. However, being published originally as a book (with detailed station information in the form of free text), this resource was not well suited for systematic linking to other geographical data. In this paper, we turn the text of the *Chronology* into a structured dataset, linked to Wikidata and georeferenced. In this process, we distinguish two main steps. First, we use traditional parsing techniques to convert the minimally structured Word document into a structured dataset. Then, we link each of the identified stations to the corresponding referent entry in Wikidata or, if missing, the closest most suitable entry. To achieve this, we use DeezyMatch[2] [14], a deep neural approach to fuzzy string matching, to identify the set of potential Wikidata candidates for each station, and use a supervised classification approach to determine the best matching entity. While the data processing step is dataset-specific, the linking process is largely generalizable to other structured datasets with metadata fields containing place information in plain text.

Charting the growth of Britain's rail network in relation to other geographically rich data sources will allow us to reconstruct the coming of rail at both micro- and macro-scales, and understand the railway in fuller context than has been previously possible. We are making the resulting linked dataset openly available for download, thereby opening new possibilities for data-driven research on the history of the railway network and its profound impact on society at large.[3]

2. Related Work

2.1. Linked Open Data, the Semantic Web, and Digital Humanities

Applications of linked open data and semantic web technologies to cultural heritage have grown substantially and the last decade has seen the appearance of many projects dedicated to creating and publishing linked historical data sets.[4] The fruits of this labour have been intensely explored by digital humanities (DH) scholars—for whom new types of access have created novel ways of studying culture and history—but also by libraries, museums and archives. For research at the interface of humanities and data science, the advantages of applying semantic technologies are manifold: the interconnected nature of the data lends itself well to qualitative exploration (facilitating serendipity and storytelling with data[5]), but also, for quantitative approaches, it is possible to leverage linked data for more refined modeling of historical and

[1] https://rchs.org.uk/railway-passenger-stations-in-great-britain-a-chronology/, version 5.02 released September 2020 by the Railway and Canal Historical Society [last accessed 14 September 2021].

[2] https://pypi.org/project/DeezyMatch/ [last accessed 14 September 2021].

[3] The data and code necessary to reproduce the linking experiments reported in this paper are available on Github via https://github.com/Living-with-machines/station-to-station. The `StopsGB` dataset is available on the British Library research repository via https://doi.org/10.23636/wvva-3d67.

[4] See for example, Linked Infrastructure for Networked Cultural Scholarship https://lincsproject.ca/, Digging into Linked Parliamentary Data https://blog.history.ac.uk/tag/digging-into-linked-parliamentary-data/, and Golden Agents https://www.goldenagents.org/ [last accessed 14 September 2021].

[5] See for example the DIVE project within the CLARIAH MediaSuite https://mediasuite.clariah.nl/documentation/glossary/dive [last accessed 14 September 2021].

cultural phenomena [20].

2.2. Candidate Selection and Resolution on Historical Sources

Successfully linking entities in cultural heritage data to a given knowledge base (KB) depends on many prior decisions. The choice of KB has the most evident impact on the linking performance: if knowledge contained in the chosen resource is incomplete or faulty, then this is likely to be reflected in the linking process. The openly available GeoNames geographical database[6] is one of the largest and most commonly-used resources for linking geographical entities [15, 28]. GeoNames integrates geographical data from many different sources and its records are complemented with volunteered information, resulting in a resource that contains over 11 million unique locations with a total of over 25 million associated geographical names. Resources based on Wikipedia and other Wikimedia projects have steadily become the most popular for generic entity linking approaches [3, 7], partly due to the fact that they contain encyclopedic knowledge formulated in natural language. Among these, Wikidata, as the central storage for the structured data of Wikimedia projects, has in recent years emerged as an exceedingly valuable resource for linking data across sources from different domains [25, 9].

While it has traditionally received little attention in the research community, candidate selection and ranking (the task of identifying and ranking potential matching entities from a knowledge base) has been shown to also have a significant impact on the downstream task of entity linking (see [5] for an overview). Established entity linking systems such as DBpedia Spotlight [19] and TagMe! [10] employ very basic candidate selection strategies, which perform sufficiently well on contemporary sources in English, but fail to address the many challenges of working with historical documents (such as diachronic and spelling variations, OCR errors, etc. [18, 22, 24]). Recent research in DH [26, 14] has focused on developing deep learning approaches. In particular, Hosseini et al. [14] recently introduced DeezyMatch, a Python open-source library for fuzzy string matching and candidate ranking, that we have employed in our work.

After having identified a set of potential entity candidates based on a string mention, multiple strategies have been presented to resolve the mention to the correct KB entry [27], such as deriving relatedness and relevance measures between co-occurring entities from the networked structure of the knowledge base (starting from [29]) or modeling the similarity of textual content, when this is available in the KB (see for instance how Wikipedia content could be used for the task [11]). Given the specificity of our setting, where we have entity mentions with minimal textual content describing them, we in part follow recent studies in the field [23] by relying on Transformers-based pre-trained models such as BERT [8] to derive a measure of text similarity between the mention and the candidate's description in Wikidata, and we combine this with more geographically-motivated strategies for entity resolution [1].

2.3. British Railway Station Data

Several resources exist that contain information about historical or modern stations in England, Wales, Scotland, Northern Ireland, and sometimes also Ireland. However, of those that are openly available, none compares to the rich detail (in terms of additional descriptors) or extensive coverage for England, Wales, and Scotland found in the *Chronology*.

[6] https://www.geonames.org [last accessed 14 September 2021].

Martí-Henneberg et al. [17, 12, 13] released snapshots of railway station data for 1851, 1861, and 1881 as part of their research with the Cambridge Group for the History of Population and Social Structure (CAMPOP). These three datasets, henceforth referred to collectively as `Campop`, are based on the content of a historical atlas that maps railway tracks and stations active between 1807-1998 on 1-inch Ordnance Survey maps [4]. The snapshots available from the UK Data Service are exports from a time-dynamic GIS of stations and tracks. Each record contains a unique object ID and point data for each station, but no other attributes such as names, opening or closing dates, or operators.

Another key resource is a subset of the GB1900 gazetteer created through a crowdsourcing project to transcribe labels on the second edition of the 6-inch-to-one-mile Ordnance Survey maps for England, Wales, and Scotland [2], which we henceforth refer to as `GB1900`.[7] By filtering only those labels containing 'station' type labels, we created a useful dataset for comparison with the *Chronology* entries. Labels represent stations on map sheets that were printed between 1888 and 1913. Because GB1900 labels were geolocated using a point in the bottom left-hand corner of the first character of the label text, this is often not the same as a station location. GB1900 does not provide the name of the station, as labels were often only '*Sta.*' or '*Station*'.

Wikidata contains records for both modern and historical railway stations. Station entries are geolocated and often situated within spatial hierarchies (city, county, state) and time-framed. They may include details like the 'operator' (railway company), and often provide links to domain-specific knowledge bases (such as the UK Railway Station code from National Rail). Other interesting properties indicate where a station is located in relation to other stations on the line, opening and closing dates, connecting lines, number of tracks, and additional external identifiers.[8] However, overall coverage of rail-specific information in Wikidata is sparse.

Although other richly documented resources exist online, few of these are amenable to computational research: the 'Disused Stations' website was created to 'build up a comprehensive database' of closed British railway stations (currently 2230 passenger stations and 14 goods stations);[9] 'RailScot' and 'RAIL MAP online: Historic railways, railroads and canals', and the 'Register of Closed Railways' (since 1901) do not currently have mechanisms for sharing their underlying data.[10]

3. The *Railway Passenger Stations* dataset

3.1. The source material

The Railway and Canal Historical Society (R&CHS) *Railway Passenger Stations in Great Britain: A Chronology* was first published privately by Michael Quick in 1996 as a by-product of his work mapping Britain's historical railway network. Now in its fifth edition, much expanded and online only, the *Chronology* has benefited greatly from the input of local and railway historians over the past quarter-of-a-century. The Quick et al. *Chronology* is a directory of every known passenger railway station in England, Scotland and Wales, past and present.

[7] GB1900 is available from https://data.nls.uk/data/map-spatial-data/gb1900/ [last accessed 14 September 2021].

[8] For example 'Stevenage railway station', https://www.wikidata.org/wiki/Q19970.

[9] See http://www.disused-stations.org.uk/ [last accessed 14 September 2021].

[10] https://www.railscot.co.uk/ and https://www.railmaponline.com/ and https://www.registerofclosedrailways.co.uk/ [last accessed 14 September 2021].

Importantly, it seeks to understand the railway system 'from the point of view of the traveller in times past', rather than 'from the companies' standpoint', and therefore includes informal stops used by landowners, workmen, sports enthusiasts and holiday-makers, as well as stations identified in the railway companies' public timetables (*Chronology*, 6).

The *Chronology* began as a document listing the opening dates of British railway stations. The content expanded significantly and now includes a range of details, such as the principal service providers, type of station (passenger, goods, worker, private, etc.), disambiguation cues to help locate the station if more than one station with the same name exists (e.g. 'Ashton, near Bristol'), opening and (where applicable) closing dates, station name at opening and any changes, any additional notes about the station, and a shorthand reference to finding the station on an OS map. Source information for the above is provided with meticulous detail and is derived mainly from contemporary, primary sources including company timetables, company reports and local newspapers, and supplemented with information from secondary works deemed authoritative.

The *Chronology* therefore offers a uniquely rich insight into the ebb and flow of the British rail system from its inception to the present day. The Society has established a Railway Chronology Group co-ordinated by Ted Cheers to collate revisions to the *Chronology*, which is available to download as a pdf from its website, but is maintained as an MS Word document. This latter version was kindly shared with us as part of our data sharing agreement with the Society, and was used to construct a structured dataset for linking. The Word document maintains a (mostly) regular structure from station to station, which made it a good candidate for parsing and transforming into (explicitly) structured data.

3.2. Processing

Railway stations share certain formatting features in the MS Word document: they always appear at the beginning of a new paragraph, in bold and upper case, and have the same font size. When more than one station exists in a town, the *Chronology* groups them together under a heading of that town name, underlined and of a larger font size than that of the comprised stations. For example, the first reference to 'Aberavon' in Figure 1 is not a station, but rather a kind of generic or phantom header name which sometimes lists attributes that all stations in that place share (in this example, the operating company and a map reference). The entries listed beneath place headings are railway stations, often with names abbreviated to their initials when they match the place name. For example, the place *Aberavon* has the following stations: *A Sea Side* and *A Town*, which should be read as *Aberavon Sea Side* and *Aberavon Town*. The entry 'Aberayron' in the same figure, on the other hand, is the only railway station in the eponymous town and, therefore, appears as a sole entry, and is preceded by no heading.[11]

The regular formatting of the document meant that we could define `xpath` expressions to identify both generic places and concrete railway stations, and therefore transform the Word document into tabular data. Were these not styled in the document, identifying them correctly would have been extremely laborious, and probably required strong supervision in the form of human annotations. We used regular expressions to expand the abbreviated names to their full names, by matching initials to the corresponding tokens of the generic place. These operations resulted in a structured dataset of 12,676 railway station entries in 9,667 places, each with

[11] Text in red indicates updates to the document since it was first shared online.

ABERAVON [RSB] {map 85}.
A SEA SIDE op 1 March 1899 (station at Sandfields, Aberavon op on Wednesday, *S Wales Daily Post, Thursday 2nd)*; clo 3 December 1962 *(RM January 1963)*. A Jubilee Road pre-opening *(RAC)*.
A TOWN op 25 June 1885 *(Cambrian, 26th)* as A; became A PORT TALBOT 1 December 1891 *(RCG)*; P T 1895 tt *(Cl; RCG ref April)*; A TOWN 1 July 1924 *(GW cirular 18 June)*; clo 3 December 1962 *(RM January 1963)*.
ABERAYRON [GW] op 12 May 1911 *(co n Lampeter)*; clo 12 February 1951 *(Cambrian News, 16 February 1951, cited by Cozens)* – see 1951**.

Figure 1: Snapshot of the MS Word document version of the *Chronology*.

Table 1
Railway stations in Aberavon and Aberayron in `StopsGB`. Aberavon has two stations (Aberavon Sea Side and Aberavon Town), Aberayron has only one. Column *Content* contains information about the railway station. The first 'Aberavon' mention (ID 25–27) does not correspond to a station, but an abstraction whose features are shared among all railway stations in this place. `StopsGB` also includes other fields, such as the abbreviated station name, operating companies, alternate names, referenced stations, first opening date, and last closing date (not shown due to space limitations).

Id	Place	Station	Content
25-27	ABERAVON	ABERAVON	[RSB] {map 85}.
25-28	ABERAVON	ABERAVON SEA SIDE	op 1 March 1899 (station at Sandfields, Aberavon op on Wednesday, S Wales Daily Post, Thursday 2nd); clo 3 December 1962 (RM January 1963). A Jubilee Road pre-opening (RAC).
25-29	ABERAVON	ABERAVON TOWN	op 25 June 1885 (Cambrian, 26th) as A; became A PORT TALBOT 1 December 1891 (RCG); P T 1895 tt (Cl; RCG ref April); A TOWN 1 July 1924 (GW cirular 18 June); clo 3 December 1962 (RM January 1963).
26-30	ABERAYRON	ABERAYRON	[GW] op 12 May 1911 (co n Lampeter); clo 12 February 1951 (Cambrian News, 16 February 1951, cited by Cozens) – see 1951**.

a unique place–station identifier pair. We set apart 491 items to annotate for the linking experiments (see section 3.3), of which only eight had some parsing error, due to existing, but rare, formatting inconsistencies in the MS Word document. Table 1 shows the entries in the newly structured dataset (henceforth `StopsGB`, for 'Structured Timeline of Passenger Stations in Great Britain') corresponding to those in Figure 1.

The content of the *Chronology* entries is rigorously formatted, despite being in free text form. With the help of punctuation (e.g. squared parentheses for companies and curly brackets for map information) and other types of markers (e.g. *op/clo* preceding opening and closing dates) or formatting options (e.g. capitalized full words indicating alternate station names), we were able to parse the content with regular expressions. We extracted opening and closing dates, operating companies, alternate names (names by which the railway station has been known at different moments in time), referenced stations, disambiguators (additional information on where the station is located), and a reference to an OS map location.[12]

[12]The following scores represent precision and recall respectively, on 219 entries that were manually annotated to evaluate the parsing: alternate station names: 0.91/0.85; companies: 1.0/1.0; first opening dates: 0.98/0.98; and last closing dates: 0.97/0.98. Alternatively, we experimented using a deep learning sequential LSTM tagging approach, which interestingly worked significantly worse (given the limited amount of training data) than the approach based on regular expressions, which greatly benefited from the very regular formatting of the text content.

3.3. Annotation

We manually linked 491 randomly selected entries from the *Chronology* to Wikidata, of which 217 were used for method development, 219 were used for testing, and the rest were discarded because they were cross-references or contained parsing errors. Wikidata has substantial records for current and historical railway stations, even for those long in disuse. Therefore, a large proportion of these cases could be matched directly to a Wikidata entry. Where the *Chronology* entry contained a place header for major settlements rather than a specific railway station (e.g. 'Aberavon' above) we signaled this by prefixing the Wikidata identifier with `ppl` for 'populated place'. The same procedure was followed for small settlements where a Wikidata identifier could be found only for the town or village, and not for the station.

There were also a small number of cases where the location of a station with no Wikidata match could be identified with enough certainty from its name and description to find a nearby, alternative Wikidata identifier. In these cases the identifier code was prefixed with `opl`, for 'other place,' to indicate that it was a proximate rather than direct link. For instance, there was no match for Newcastle's Moor Edge station, but we were able to make a proximate link with the city's Town Moor (`Q11898308`) since we know that this temporary station served race meetings that were held on Town Moor.[13]

4. Linking experiments and evaluation

We describe the Wikidata-based resource that we use for linking in Section 4.1. The linking is performed in two steps. First, given a query (a railway station, a place, or a station alternate name), we narrow the full set of Wikidata candidates down to those that may potentially be referred to by this query. This is called *candidate selection* and is described in Section 4.2. The next step is to determine the correct entity given the candidates selected in the previous step. This step is called *entity resolution* and is addressed in Section 4.3.

For reference, Figure 2 provides a simplified overview of the linking process that is described throughout this section, using one entry in the *Chronology* as an example.

4.1. Linking resource

We extracted all locations in Wikidata[14] by filtering the entries that have a *coordinate location* property (`P625`), i.e. entries that can be located on the Earth's surface through their geographical latitude and longitude. For each entry we kept a series of features that describe the entry (geographically, historically, politically). This resulted in 8,094,093 entries, which we narrowed down to those located in Great Britain, filtering them by their location within a polygon of coordinates enclosing the island.[15] The resulting dataset (henceforth *WikiPlaces* gazetteer) is composed of 671,320 entries. Next, we created a further subset composed of those entries from the *WikiPlaces* gazetteer that are either instances of station-related classes or their English label has the words '*station*', '*stop*', or '*halt*', not preceded by '*police*', '*signal*', '*power*', '*lifeboat*',

[13] In total, 55 entries were annotated as populated places and 19 as other places. There were 4 entries for which no Wikidata match could be provided.

[14] We used the 20200925 Wikidata dump from https://dumps.wikimedia.org/wikidatawiki/entities/ and followed the approach described in https://akbaritabar.netlify.app/how_to_use_a_wikidata_dump to parse the entities [last accessed 14 September 2021].

[15] We have used the Ordnance Survey OpenData Boundary-Line™ ESRI Shapefile from https://osdatahub.os.uk/downloads/open/BoundaryLine [last accessed 14 September 2021].

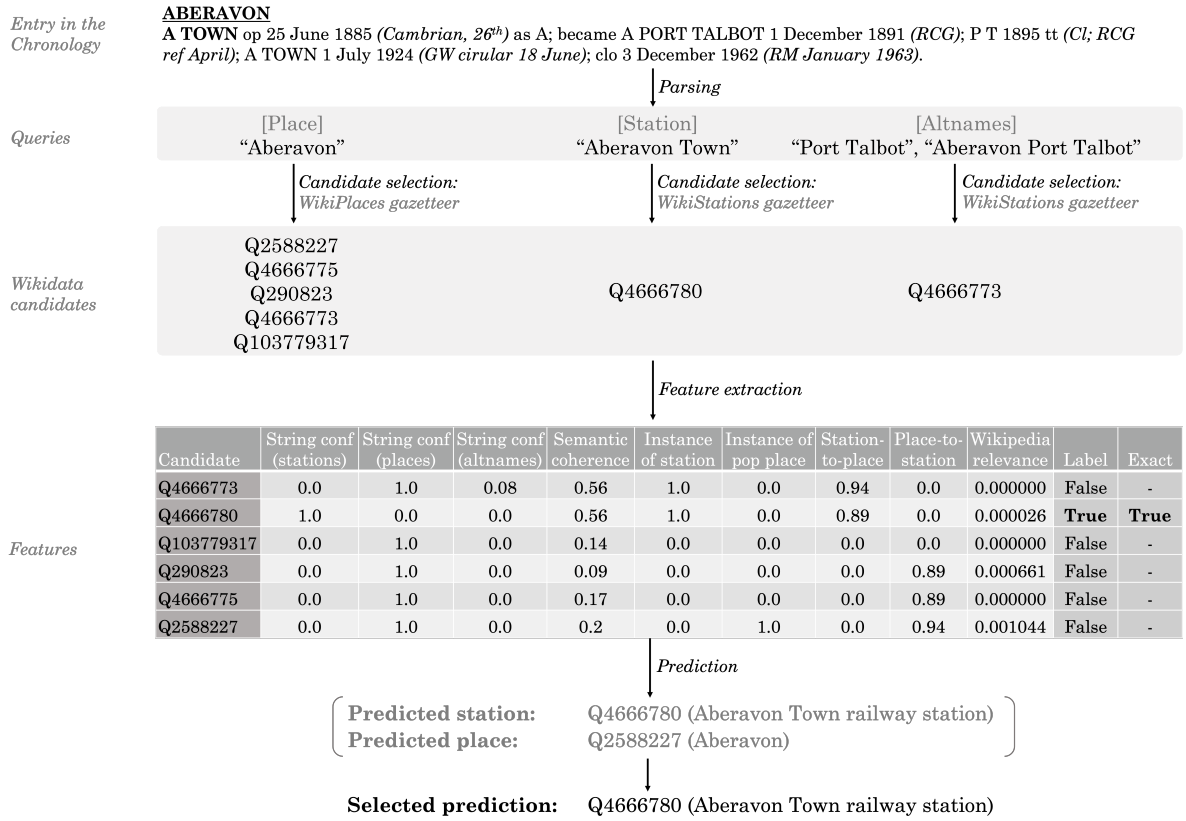

Figure 2: Overview of the linking steps, using the Aberavon Town railway station entry as an example. First, as described in section 3.2, we identify different queries: place, station, and station alternate names). Then, candidates are found for each type of query in either the *WikiPlaces* gazetteer (for places) or *WikiStations* gazetteer (for stations and alternate names). We use one name variation ($nv = 1$) in this example (described in Section 4.2). We then extract several features for each candidate (see Section 4.3.1). Columns 'Label' and 'Exact' are provided through the annotations, and are available only for entries that are in the development set (used for training and development) or in the test set (used for evaluation). Column 'Label' indicates the most appropriate Wikidata match, and column 'Exact' indicates whether the Wikidata entry is an exact match to the railway station, or whether it is a proximate (prefixed `ppl` or `opl` during the annotation process, as described in Section 3.3). Finally, given the set of candidates and their features, the resolution method will predict a Wikidata match, in this case `Q4666780`, correctly corresponding to the Wikidata entry for Aberavon Town railway station. The different resolution baselines and methods are described in Section 4.3. While most methods predict just the final Wikidata entry, the `SVM refined` method predicts one entry for station and one for place, and selects the best match based on the confidence of these predictions.

'*pumping*', or '*transmitting*'. This procedure leads to the retrieval of many false positives but at this point we are interested in maximizing recall at the expense of precision: we maximize precision during the subsequent linking steps (described in sections 4.2 and 4.3). The resulting dataset is composed of 9,361 entries, henceforth referred to as the *WikiStations* gazetteer.

We improved the Wikidata-based gazetteers in two ways. First, Wikidata provides structured and curated sets of alternate names in terms of labels and aliases in different languages, but which are relatively limited when compared to other resources such as Wikipedia or Geonames. We therefore use the links between Wikidata and Wikipedia[16] and between Wikidata

[16]The Wikipedia link structure has been largely exploited in the past in order to expand the alternate names

and Geonames to expand our gazetteers with alternate names from these resources. Secondly, we make use of the linking between Wikidata and Wikipedia to obtain—for each Wikidata entry in our gazetteers—the number of incoming links of the corresponding Wikipedia page, if available. This measure is traditionally used as a proxy for relevance in entity linking systems (see for instance [29]).[17] The final *WikiPlaces* has 670,325 entities (after filtering out unlabelled entries) with 823,304 alternate names; the final *WikiStations* gazetteer has 9,361 entries with 33,156 alternate names.

4.2. Candidate selection

As discussed in Section 3.2, each entry in `StopsGB` has a *station name* and a *place name* field and, when available, also a list of *alternate names* for the station. Because one of the aims of linking is to geolocate the entries, we decided that, in those cases in which the railway station is not present in Wikidata (as in the case of New Tredegar Colliery railway station), we provide an approximated location (i.e. New Tredegar, the location of this station for miners). Therefore, in this step we aim to retrieve Wikidata entries that are potentially referred to by one of the query fields (station, place, or alternate names). Both the *station* and *alternate names* fields refer to stations, whereas the *place* field refers to more generic place names. Therefore, we retrieve Wikidata candidates for both the station and the alternate names fields by querying them against the *WikiStations* gazetteer; and retrieve Wikidata candidates for the generic place field from the *WikiPlaces* gazetteer.

4.2.1. Approaches

We have experimented with three different approaches for candidate selection: (1) ***exact match***: Wikidata candidates are retrieved if one of the alternate names of the Wikidata entry is identical to the query; (2) ***partial match***: candidates are retrieved if the query is contained in one of their alternate names (i.e. there is a string overlap), and are ranked according to amount of overlap; and (3) ***deezy match***: candidates are retrieved and ranked using DeezyMatch [14], an open-source software library for fuzzy string matching and candidate ranking using neural networks. Both *partial* and *deezy* match allow for fuzzy string matching.[18] To have a more extended overview of the impact of this step, we tested candidate selection considering the set of candidates corresponding to the top ranked one, three and five candidate name variations of a query (henceforth *nv*).[19]

of entities in knowledge bases [3, 7]. We use the Wikipedia-based gazetteer described in [6].

[17] We have employed the 20200920 English Wikipedia dump and processed it using WikiExtractor (https://github.com/attardi/wikiextractor [last accessed 14 September 2021]), to extract single pages and their structure in sections, as in [21].

[18] See [5] for an extensive comparison between DeezyMatch and traditional string similarity measures for candidate selection.

[19] To show this with an example, consider the scenario in which we choose to retrieve three name variations (`nv = 3`) per query: given the query 'PARKGATE', DeezyMatch returns the following three most similar candidate strings from Wikidata (scores in parentheses represent cosine distance): 'Parkgate' (0.0), 'Park Gate' (0.0152), and 'Parkergate' (0.0162), which are then expanded to all Wikidata candidate entries that have this alternate name, i.e. 7 candidate entities for 'Parkgate' (such as `Q7138469`, a village in Cheshire, and `Q7138470`, a village in Scotland), 4 candidate entities for 'Park Gate', and one for 'Parkergate'.

4.2.2. Metrics

Given a mention, we assess the performance of each method in generating a ranked list of name variations of potential entity candidates by reporting **precision** at *nv* (either 1, 3 or 5), meaning how many times a name variation of the correct entity appears in the retrieved results. Note that increasing the number of potential name variations will consequently impact the precision of the retrieved ranking, which can be taken as a measure of difficulty of the following resolution step. In addition, we report the **mean average precision**[20] at the same *nv*: this will offer a glance on the quality of the ranking. Finally, we report **binary retrieval** to highlight how many times at least one name variation of the correct entity is retrieved at *nv*—this will set the skyline for the following resolution step (meaning that if the correct entity is not retrieved at the selection stage, the mention cannot be resolved correctly).

4.2.3. Evaluation

We report a comparison of the different approaches to select and rank potential candidates for given query inputs in Table 2. We compare two evaluation settings: (1) *strict*, which assesses the performance only on those queries for which there exists a Wikidata entry corresponding to the station (i.e. not preceded by neither `ppl` nor `opl` in the annotations), which we use on queries from the *station* and *alternate* name fields of the structured dataset; and (2) *appr*, which assesses the performance on all queries, in which case a true positive is not whether the correct railway station is found, but whether the best possible match on Wikidata (according to the annotators) has been retrieved.

The results in Table 2 provide an interesting portrayal of the forthcoming entity resolution task, described in section 4.3. We see that the gain of allowing more name variations than just the most similar one is very low (the increase of *retr* is minimal) compared to the increase in difficulty of the task (shown by a decrease in precision). MAP, however, stays high, indicating the importance of string similarity confidence, especially using DeezyMatch. The retrieved candidates and their confidence score are therefore passed on to the next step, which will resolve each entry in `StopsGB` to the best matching Wikidata entity.

4.3. Entity Resolution

At this point, for each entry in `StopsGB` we have up to three sets of candidates: a set of candidates for the station name, one for the place name, and one for possible alternate names. The final step of the pipeline, entity resolution, takes the retrieved candidate entities and returns only one best match per entry. We performed our experiments on candidates selected with DeezyMatch, because this is the approach that had the highest MAP score overall, and the largest variation in precision depending on number of retrieved candidates. We performed experiments with $nv = 1$ and $nv = 5$.

4.3.1. Features and baselines

We defined several features for each candidate to quantify the compatibility between the Wikidata candidate and the *Chronology* entry. The features we used are the following:

[20]Mean average precision (MAP) is a popular metric in information retrieval that highlights how well the ranking (overlap score in the case of perfect and partial match, and confidence score in the case of DeezyMatch) correlates with the labels.

Table 2
Performance of the candidate selection approaches (*exact*, *partial*, and *deezy* match) for different query inputs ('stns' for 'stations', 'alts' for alternate station names, and 'places' for generic places), in terms of precision (p), mean average precision (map), and binary retrieval ($retr$), in either a *strict* or approximate (*appr*) evaluating scenario, on Wikidata candidates matching up to 1, 3, and 5 string variations (*nv*) of the original mention.

		$nv = 1$			$nv = 3$			$nv = 5$		
Eval	Approach	p	map	retr	p	map	retr	p	map	retr
Strict	exact:stns	0.66	0.68	0.71	–	–	–	–	–	–
Strict	partial:stns	0.66	0.68	0.71	**0.6**	0.68	**0.72**	**0.59**	**0.69**	**0.72**
Strict	deezy:stns	**0.67**	**0.69**	**0.72**	0.56	**0.69**	**0.72**	0.55	**0.69**	**0.72**
Strict	exact:stns+alts	**0.64**	0.68	0.72	–	–	–	–	–	–
Strict	partial:stns+alts	**0.64**	**0.69**	0.72	**0.57**	0.67	**0.73**	**0.56**	0.68	**0.73**
Strict	deezy:stns+alts	0.63	**0.69**	**0.73**	0.52	**0.69**	**0.73**	0.51	**0.69**	**0.73**
Appr	exact:stns+alts+plcs	**0.33**	0.72	0.79	–	–	–	–	–	–
Appr	partial:stns+alts+plcs	0.32	**0.73**	**0.80**	**0.21**	0.61	**0.81**	**0.18**	0.49	**0.82**
Appr	deezy:stns+alts+plcs	0.29	0.71	**0.80**	0.19	**0.71**	0.8	**0.18**	**0.71**	0.8

- **String confidence:** DeezyMatch confidence score between the mention and the candidate alternate name for (a) stations, (b) places, and (c) station alternate names. We generated one feature for each.
- **Semantic coherence:** The semantic similarity between the Wikidata candidate and the entry in `StopsGB`, using transformer-based sentence embeddings [23].[21]
- **Wikipedia relevance:** Number of incoming links a Wikidata candidate has on Wikipedia (as a proxy for entity popularity), normalized against the maximum number of incoming links in the set of candidates.
- **Wikidata class:** (a) whether the candidate is an instance of a railway station class, and (b) whether the candidate is an instance of a populated place.
- **Station-to-place and place-to-station geographical compatibility:** If the candidate is a railway station, normalized geographical closeness to the closest place candidate; if the candidate is a generic place, normalized geographical closeness to the closest station candidate.

Each candidate is therefore represented as a vector of features, followed by its label (true if it is the correct entity for a given entry, false otherwise), and whether it is an exact match (i.e. the railway station) or an approximate match (i.e. the best possible match given that the exact match does not exist). We use three of these features (string confidence, semantic coherence, and relevance in Wikipedia) as **baseline methods** for the task, by selecting the candidate that has the highest score from the pool of overall retrieved candidates. In the case of the *string confidence* baseline, we select the top match amongst railway stations and, only if none

[21] We use the description, the historical county and administrative region information for the Wikidata candidate; and the place, disambiguation cues, maps description, alternate names, and references for the `StopsGB` entry. We have used the default pre-trained model: `paraphrase-distilroberta-base-v1`, which is trained on large scale paraphrase data.

has been retrieved, the top match amongst places. We also compute a **skyline**, which is the highest possible score reachable, given the available set of candidates.

4.3.2. Supervised resolution approaches

We propose a supervised approach that trains a Support Vector Machine (SVM) on the development set (i.e. one SVM trained on all query/candidate combinations at once) and learns whether a candidate is a correct match for a given query or not. We then apply the resulting classifier on a query basis (i.e. on the set of possible candidates per query only, as in the baseline methods[22]), return the probability score instead of returning a label, and select the most confident match from the subset of possible candidates. We propose two different SVM variations:[23] (1) *SVM simple* trains the SVM on the development set using all features, without distinguishing between *strict* and *approximate* instances; whereas (2) *SVM refined* is a dual classification system: it trains an SVM classifier using all features on the subset of queries for which there is a strict match, and an additional classifier using all features on the subset of queries for which there is *not* a strict match. The idea behind *SVM refined* is that the learning objective is different if the goal is to predict entities of the type 'station' or generic places. We combine the two based on the confidence score of the first classifier (i.e. the *station* classifier): if the confidence of a prediction is lower than a certain threshold (found based on experiments on the development set), we will apply the second classifier (i.e. the *generic place* classifier).

As a comparison, we employed the same features in a Learning to Rank (L2R) [16] pipeline, using *RankLib*.[24] The weight parameter is learned by optimizing for the precision at 1 (P@1) using coordinate ascent with linear normalization.

4.3.3. Metrics and evaluation

Table 3 summarizes the results of our experiments. As in the previous step, we also provide two evaluation scenarios: *strict* only accepts exact entities as true (only entities referring to the correct railway station), whereas *approximate* accepts place entities if the station does not exist as an entity in Wikidata. We present the results for the resolution task in terms of precision (how many times the mention is correctly matched with the correct entity) as well as approximate accuracy at 1, 5, and 10 km (*Acc@km*) (i.e. how many times the mention is correctly geo-located within 1, 5, and 10 km from the gold standard coordinates).

An analysis of the most indicative features for both classifiers proves our assumption that predicting stations and generic places are two different learning tasks. The most indicative features for the *stations* classifier in the *strict* scenario, where $nv = 1$, are (ranked from higher to lower prominence) station name string confidence, station-to-place compatibility, Wikidata class if candidate is a railway station, and semantic coherence. The most indicative features of the *generic places* classifier in the *approximate* scenario, where $nv = 1$, are Wikipedia relevance, place-to-station compatibility, and place string confidence. This distinction between station and place favours *SVM refined* especially when $nv = 5$ and in the *approximate* setting. The results in Table 3 confirm that using a larger nv does not compensate for the resulting increased difficulty of the task. Nevertheless, the good performance of *SVM refined* when $nv = 5$ suggests that it is a robust resolution system, which does not suffer from a higher number of

[22]Note that in all cases the queries will be different between the development and the test set.
[23]Both are linear SVMs, where the `C` parameter is tuned on the development set.
[24]https://sourceforge.net/p/lemur/wiki/RankLib/ [last accessed 14 September 2021].

Table 3
Performance of the resolution methods in terms of precision, and accuracy at 1, 5, and 10 km ($Acc@km$) for the two evaluation settings: *strict* and *approximate*. Each approach is evaluated from candidates extracted with DeezyMatch, with number of string variations specified by nv.

	Strict	*Approximate*			
Approach	Precision	Precision	Acc@1km	Acc@5km	Acc@10km
skyline (deezy, nv=1)	*0.73*	-	-	-	-
string confidence (deezy, nv=1)	0.66	0.69	0.77	0.84	0.85
wikipedia relevance (deezy, nv=1)	0.10	0.16	0.54	0.8	0.81
semantic coherence (deezy, nv=1)	0.30	0.32	0.58	0.78	0.79
RankLib (deezy, nv=1)	**0.68**	0.7	0.79	0.85	**0.86**
SVM simple (deezy, nv=1)	**0.68**	**0.71**	**0.8**	**0.86**	**0.86**
SVM refined (deezy, nv=1)	0.67	0.7	0.79	**0.86**	**0.86**
skyline (deezy, nv=5)	*0.73*	-	-	-	-
string confidence (deezy, nv=5)	0.66	0.68	0.77	0.85	0.85
wikipedia relevance (deezy, nv=5)	0.06	0.11	0.42	0.65	0.65
semantic coherence (deezy, nv=5)	0.25	0.26	0.45	0.61	0.62
RankLib (deezy, nv=5)	0.68	0.71	0.79	**0.86**	**0.87**
SVM simple (deezy, nv=5)	0.67	0.68	0.76	0.82	0.82
SVM refined (deezy, nv=5)	**0.69**	**0.72**	**0.8**	**0.86**	**0.87**

candidates, in particular in comparison with *SVM simple* and the Wikipedia relevance and semantic coherence baselines.[25]

Based on the results of our experiments, we applied *SVM refined* on the full `StopsGB` dataset (i.e. 12,676 rows), using $nv = 1$. For each entry, we provide predictions of Wikidata entries both for station and place, together with the confidence score of these predictions. We also provide the Wikidata ID of the selected entity (i.e. the predicted station if the confidence score is above a certain threshold; the predicted place if not) and its latitude and longitude.

5. Discussion

Linking information on railway stations serves the larger aim of enabling historical research based on heterogeneous, interconnected data sources. This section offers a quick comparison with the publicly available `Campop` data and also showcases some novel research avenues that emerge from the enrichment and linking of historical information. The goal, here, is to sketch these opportunities and more elaborate analyses will appear in future work.

Compared to existing datasets, `StopsGB` expands our knowledge of historical stations in many ways. Not only does it fill gaps in the current record, it also extends the time frame, spanning almost two centuries. To compare the differences visually, we can map `StopsGB` and `Campop` data. Figure 3 includes all stations opened up to 1999 and compares them to the combined `Campop` stations (e.g. the union of 1851, 1861, and 1881 stations): it appears that `StopsGB` provides a more complete picture of the station landscape (e.g. red points that are

[25]The *string confidence* baseline is a very strong baseline, especially in the strict evaluation scenario, and indicates that most station names are quite unique. It is worth mentioning that both the *string confidence* baseline and RankLib produce different results at each run. For this reason, the results reported are averaged over 5 runs to present a more reliable overview.

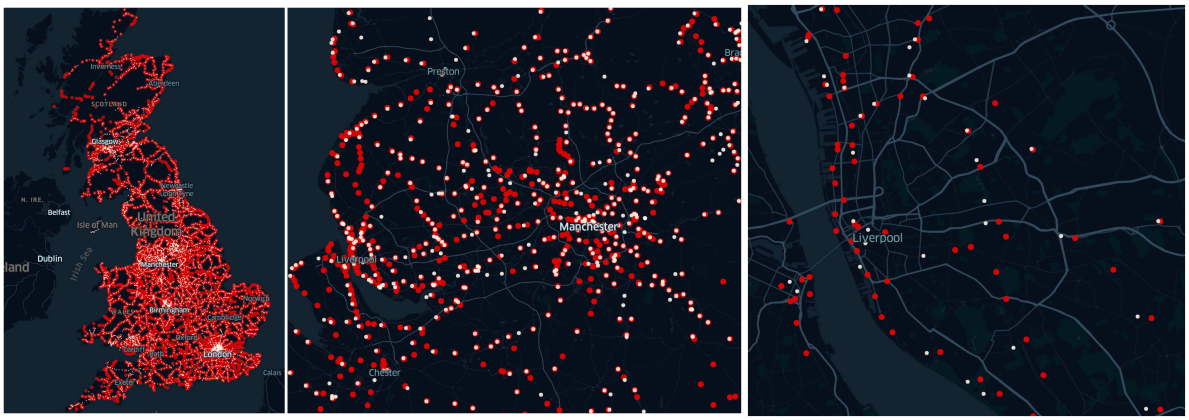

Figure 3: Station locations for Great Britain, northwest England and Merseyside as derived from `StopsGB` (red) and `Campop` (white).

Figure 4: This image zooms in on Bolton, showing stations (red dots) obtained from `StopsGB`, and industrial buildings (purple), churches (yellow) and schools (green) obtained from `GB1900` labels. Map images from the 2nd edition of the 6-inch-to-one-mile Ordnance Survey sheets are courtesy of the National Library of Scotland.

not paired with an overlapping or adjacent white point). However, this also points to some complexities, as neither data set is complete, nor do they overlap in clear ways. Scrutinizing these differences and overlaps between `Campop` and `GB1900` (as well as with modern data) is part of future work.

To highlight some novel research approaches made possible by `StopsGB`, we sketch out two case studies that exploit links between data to understand the place of rail in industrializing communities. Figure 4 shows how rail is embedded in the urban landscape. Focusing on Bolton, it plots stations (red) in relation to industrial buildings, churches and schools.[26] Blending linked data with visual information (in this case, historical maps) provides new means to explore the context of station and rail, both quantitatively and qualitatively. This approach allows us to explore (using more abstract measures) the spatial distribution of stations, but we can also zoom in on specific areas for a 'close reading' of the spatial context of the rail. Moreover, by exploiting information on the opening and closing of stations, we can obtain a dynamic and detailed image of the evolution of the British rail network. Figure 5 shows the spread of the railway during the nineteenth century.

[26]These labels are obtained by matching entries in `GB1900`. Industrial terms are '*works*', '*mill*', '*mills*', '*factories*', '*factory*', '*workshop*', '*wks*', '*manufactory*'. '*Schools*' and '*sch*' are used for plotting schools. Religious buildings were captured by '*church*', '*ch*', '*chap*', '*chapel*' and '*cathedral*'.

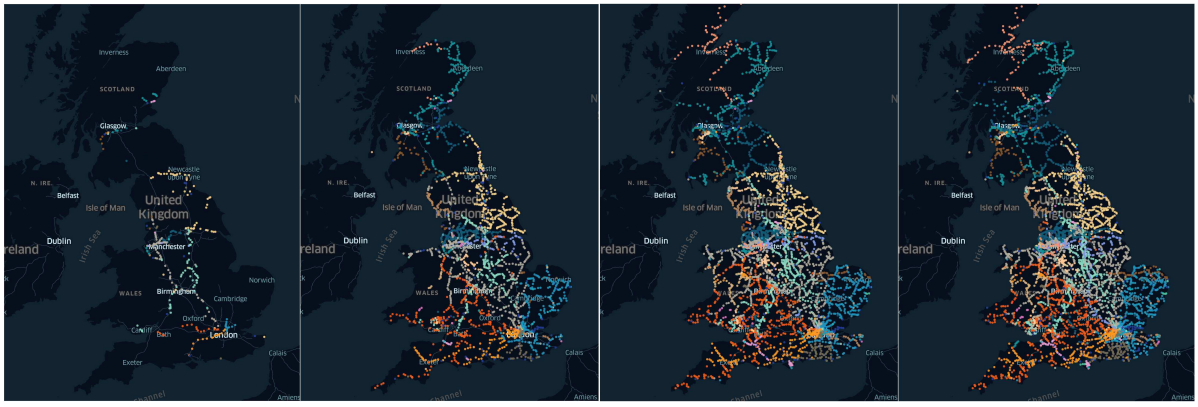

Figure 5: Evolution of stations between 1840, 1860, 1880 and 1900. Stations are colored by the company operating them in 1922, or at the date they closed (if earlier).

6. Conclusion

Leveraging the links between Wikidata and the *Chronology* station descriptions in these examples demonstrates the power of a station dataset that can be queried not only by location, but also by date or any other attribute so carefully collected by Quick and other contributors from the Railway and Canal Historical Society. Our work to translate this exceptional community-curated resource into a geolocated dataset is an early step that will allow history and geography researchers to craft new narratives about the railway, and the process of industrialisation it accompanied.

Author contributions

After the first author, authors are listed in alphabetical order. The names in the following roles are sorted by amount of contribution and, if equal, alphabetically: *Conceptualization:* KM, JL, DW; *Methodology:* MCA, FN, KB; *Implementation:* MCA, FN, KB, GT; *Reproducibility:* FN, MCA; *Historical Analysis:* KB, KM, JL, JR, DW; *Data Acquisition and Curation:* DW, MCA, GT, FN; *Annotation:* JL, KM; *Project Management:* MCA; *Writing and Editing:* all.

Acknowledgments

We thank the Railway and Canal Historical Society for sharing the Microsoft Word version of *Railway Passenger Stations in Great Britain: a Chronology* by Michael Quick. Work for this paper was produced as part of *Living with Machines*. This project, funded by the UK Research and Innovation (UKRI) Strategic Priority Fund, is a multidisciplinary collaboration delivered by the Arts and Humanities Research Council (AHRC), with The Alan Turing Institute, the British Library and the Universities of Cambridge, East Anglia, Exeter, and Queen Mary University of London.

References

[1] E. Acheson, M. Volpi, and R. S. Purves. "Machine learning for cross-gazetteer matching of natural features". In: *Ijgis* (2020).

[2] P. Aucott and H. Southall. "Locating past places in Britain: creating and evaluating the GB1900 Gazetteer". In: *International Journal of Humanities and Arts Computing* 13.1 (2019), pp. 69–94.

[3] R. Bunescu and M. Paşca. "Using encyclopedic knowledge for named entity disambiguation". In: *11th Conference of the European Chapter of the Association for Computational Linguistics*. Trento, Italy: Association for Computational Linguistics, 2006. URL: https://www.aclweb.org/anthology/E06-1002.

[4] M. H. Cobb. *The railways of Great Britain, a historical atlas at the scale of 1 inch to 1 mile*. Shepperton, Surrey: Ian Allan Pub., 2005.

[5] M. Coll Ardanuy, K. Hosseini, K. McDonough, A. Krause, D. van Strien, and F. Nanni. "A deep learning approach to geographical candidate selection through toponym matching". In: *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*. 2020, pp. 385–388.

[6] M. Coll Ardanuy, K. McDonough, A. Krause, D. C. Wilson, K. Hosseini, and D. van Strien. "Resolving places, past and present: toponym resolution in historical British newspapers using multiple resources". In: *Proc. of GIR*. 2019.

[7] S. Cucerzan. "Large-scale named entity disambiguation based on Wikipedia data". In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Prague, Czech Republic: Association for Computational Linguistics, 2007, pp. 708–716. URL: https://www.aclweb.org/anthology/D07-1074.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv:1810.04805* (2018).

[9] M. Ehrmann, M. Romanello, A. Flückiger, and S. Clematide. "Extended overview of CLEF HIPE 2020: named entity processing on historical newspapers". In: *CLEF 2020 Working Notes. Conference and Labs of the Evaluation Forum*. Vol. 2696. Conf. Ceur. 2020.

[10] P. Ferragina and U. Scaiella. "TagMe: on-the-fly annotation of short text fragments". In: *Proc. of CIKM*. 2010.

[11] B. Hachey, W. Radford, J. Nothman, M. Honnibal, and J. R. Curran. "Evaluating entity linking with Wikipedia". In: *Artificial intelligence* 194 (2013), pp. 130–150.

[12] J. Henneberg, M. Satchell, X. You, L. M. W. Shaw-Taylor, E. A. Wrigley, and M. Cobb. *1861 England, Wales and Scotland railway stations*. 2018. DOI: 10.5255/ukda-sn-852995.

[13] J. Henneberg, M. Satchell, X. You, L. M. W. Shaw-Taylor, E. A. Wrigley, and M. Cobb. *1881 England, Wales and Scotland railway stations*. 2018. DOI: 10.5255/ukda-sn-852996.

[14] K. Hosseini, F. Nanni, and M. Coll Ardanuy. "DeezyMatch: A flexible deep learning approach to fuzzy string matching". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2020, pp. 62–69.

[15] M. D. Lieberman, H. Samet, and J. Sankaranarayanan. "Geotagging with local lexicons to build indexes for textually-specified spatial data". In: *2010 IEEE 26th international conference on data engineering (ICDE 2010)*. Ieee. 2010, pp. 201–212.

[16] T.-Y. Liu. *Learning to rank for information retrieval*. Springer, 2011.

[17] J. Marti-Henneberg, M. Satchell, X. You, L. M. W. Shaw-Taylor, and E. A. Wrigley. *1851 England, Wales and Scotland railway stations*. 2018. DOI: 10.5255/ukda-sn-852994.

[18] K. McDonough, L. Moncla, and M. van de Camp. "Named entity recognition goes to Old Regime France: geographic text analysis for early modern French corpora". In: *International Journal of Geographical Information Science* 33.12 (2019), pp. 2498–2522.

[19] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. "DBpedia Spotlight: shedding light on the web of documents". In: *Proc. of SEMANTiCS*. 2011.

[20] A. Meroño-Peñuela, A. Ashkpour, M. Van Erp, K. Mandemakers, L. Breure, A. Scharnhorst, S. Schlobach, and F. Van Harmelen. "Semantic technologies for historical research: A survey". In: *Semantic Web* 6.6 (2015), pp. 539–564.

[21] F. Nanni, S. P. Ponzetto, and L. Dietz. "Entity-aspect linking: providing fine-grained semantics of entities in context". In: *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*. 2018, pp. 49–58.

[22] A. Olieman, K. Beelen, M. van Lange, J. Kamps, and M. Marx. "Good applications for crummy entity linkers? The case of corpus selection in digital humanities". In: *Proc. of SEMANTiCS*. 2017.

[23] N. Reimers and I. Gurevych. "Sentence-BERT: Sentence embeddings using siamese BERT-networks". In: *Proc. of EMNLP* (2019).

[24] M. Rovera, F. Nanni, S. P. Ponzetto, and A. Goy. "Domain-specific named entity disambiguation in historical memoirs". In: *Proc. of CLIC* (2017).

[25] A. Sakor, K. Singh, A. Patel, and M.-E. Vidal. "Falcon 2.0: An entity and relation linking tool over Wikidata". In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2020, pp. 3141–3148.

[26] R. Santos, P. Murrieta-Flores, and B. Martins. "Learning to combine multiple string similarity metrics for effective toponym matching". In: *International journal of digital earth* (2018).

[27] W. Shen, J. Wang, and J. Han. "Entity linking with a knowledge base: issues, techniques, and solutions". In: *IEEE Transactions on Knowledge and Data Eng.* (2014).

[28] R. Simon, E. Barker, L. Isaksen, and P. de Soto Cañamares. "Linking early geospatial documents, one place at a time: annotation of geographic documents with Recogito". In: *e-Perimetron* 10.2 (2015), pp. 49–59.

[29] I. H. Witten and D. N. Milne. "An effective, low-cost measure of semantic relatedness obtained from Wikipedia links". In: (2008).