# Catching the Red Priest: Using Historical Editions of Encyclopaedia Britannica to Track the Evolution of Reputations

**Yen-Fu Luo**[†], **Anna Rumshisky**[†], **Mikhail Gronas**[*]
[†]Dept. of Computer Science, University of Massachusetts Lowell, Lowell, MA, USA
[*]Dept. of Russian, Dartmouth College, Hanover, NH, USA
{yluo,arum}@cs.uml.edu, mikhail.gronas@dartmouth.edu

## Abstract

In this paper, we investigate the feasibility of using the chronology of changes in historical editions of Encyclopaedia Britannica (EB) to track the changes in the landscape of cultural knowledge, and specifically, the rise and fall in reputations of historical figures. We describe the data-processing pipeline we developed in order to identify the matching articles about historical figures in Wikipedia, the current electronic edition of Encyclopaedia Britannica (edition 15), and several digitized historical editions, namely, editions 3, 9, 11. We evaluate our results on the tasks of article segmentation and cross-edition matching using a manually annotated subset of 1000 articles from each edition. As a case study for the validity of discovered trends, we use the Wikipedia category of 18th century classical composers. We demonstrate that our data-driven method allows us to identify cases where a historical figure's reputation experiences a drastic fall or a dramatic recovery which would allow scholars to further investigate previously overlooked instances of such change.

## 1 Introduction

Histories of nations are reflected in their shifting borders. But the histories of things immaterial, yet no less interesting–concepts, ideologies, reputations of historical personalities–are mapless. This paper describes the progress of the Knowledge Evolution Project (KnowEvo), which investigates the possibility of using historical digitized text to track and map long-range historical changes in the conceptual landscape, and specifically, the history of intellectual networks and reputations.

One of the ways to investigate the change in how ideas and personalities are represented is to use mention statistics from books written at different historical periods. Google Ngram Viewer is a tool that plots occurrence statistics using Google Books, the largest online repository of digitized books. But while Google Books in its entirety certainly has quantity, it lacks structure. However, the history of knowledge (or culture) is, to a large extent, the history of structures: hierarchies, taxonomies, domains, subdomains.

In the present project, our goal was to focus on sources that endeavor to capture such structures. One such source is particularly fitting for the task; and it has been in existence at least for the last three centuries, in the form of changing editions of authoritative encyclopedias, and specifically, Encyclopaedia Britannica. Throughout their existence, encyclopedias have claimed to be well-organized (i.e., structured) representations of knowledge and have effectively served as its (obviously imperfect) mirrors. Each edition of Encyclopaedia Britannica reflected a collective editorial decision, based on a scholarly consensus, regarding the importance of each subject that has to be included and the relative volume dedicated to it. As such, it can be thought of as a proxy of sorts for the state of contemporary knowledge. Of course, institutions such as Britannica, their claims to universality notwithstanding, throughout their histories have been necessarily western-centric and reflected the prejudices of their time. A note of caution is therefore in order here: what this data allows us to reconstruct is the evolution of knowledge representation, rather than of the knowledge itself.

In this paper, we investigate the feasibility of using historical Encyclopaedia Britannica editions to develop tools that can be used in scholarship and in pedagogy to illustrate and analyze known historical changes and to facilitate the discovery of overlooked trends and processes. Specifically, we focus on the history of intellectual reputations. We are interested in whether certain categories of

people that form an intellectual landscape of a culture can be tracked through time using Britannica's historical editions. We suggest that by measuring changes in the relative importance assigned to a particular figure in successive editions of Britannica we can reconstruct the history of his or her reputation. Thus, each edition can be thought of as a proxy for the contemporary state of knowledge (and reputations in particular), with the history of editions reflecting the history of such states. Continuing previous work (Gronas et al., 2012), we develop a set of tools for cleaning noisy digitally scanned text, identifying articles and subjects, normalizing their mentions across editions, and measuring their relative importance. The data about historical figures and their reputations, based on their representation in different editions of Encyclopaedia Britannica, is available for browsing and visualization through the KnowEvo Facebook of the Past search interface.[1]

We examine the plausibility of using these tools to track the change in people's reputations in various domains of culture. In the current work, we verify the accuracy of our cross-edition normalization methods and conduct a case study to examine whether the discovered trends are valid. As a case study, we look at the reputations of 18th century classical composers. Many of 18th century classical composers had utmost importance for western classical music and exerted far-reaching influence during the following centuries. We looked at the reputation changes between the 11th edition (1911) and the 15th edition (1985–2000), thus covering most of the 20th century.

The results of our case study suggest that our methods provide a valid way to examine the trends in the rise and fall of reputations. For example, the case study revealed that in the course of the 20th century, among the major composers, Handel's reputation underwent the biggest change, as he dropped from being second most important composer (after Johann Sebastian Bach) in the beginning of the century to the fifth position, well behind Gluck and Haydn. Meanwhile, Mozart dethroned Bach, who moved from the first to the second place. Some of the lesser composers (Lotti and Gaensbacher) disappeared from encyclopedia-curated cultural memory altogether; whereas the familiar name of Telemann owes its familiarity to a recent revival. Another notable shift, empirically revealed during the case study, was the change in Vivaldi's legacy. The author of "The Four Seasons", known to his contemporaries as the red priest (due to his hair and profession, respectively) was completely forgotten towards the beginning of the 20th century and then rediscovered and joined the canon in the second part of the century. For a student of musical history these facts are not surprising. However, they have been obtained through an automatic method which can be used in other, less well known areas of cultural history and on a large scale.

[1] http://knowevo.cs.uml.edu

## 2 Related work

A big data analysis of large textual datasets in humanities has been gaining momentum in recent years, as evidenced by the success of Culturomics (Michel et al., 2011), a method based on n-gram frequency analysis of the Google Books corpus, available via the Google Ngram Viewer.

Skiena and Ward (2013) recently applied similar quantitative analysis to empirical cultural history, assessing the relative importance of historical figures by examining Wikipedia people articles. They supplemented word frequency analysis with several Wikipedia-based measures, such as PageRank (Page et al., 1999), page size, the number of page views and page edits. Their approach is complementary to ours: whereas they are interested in the reputations as they exist today, we seek to quantify the dynamics of cultural change, i.e. the historical dimension of reputations, rather than a contemporary snapshot.

A culturomics-like approach applied to large structured datasets (knowledge bases) is advocated in Suchanek and Preda (2014). Our approach is somewhat similar in that the corpus of historical editions of Britannica can be considered a knowledge base, with an important difference being a chronological dimension, absent from such knowledge bases as YAGO or DBpedia. An example of mining a historical corpus for trends using the frequentist approaches to vocabulary shifts as well as normalization to structured sources can be found in the recent work on newspaper and journal historical editions such as Kestemont et al. (2014) and Huet et al. (2013).

Disambiguation of named entities to structured sources such as Wikipedia has been an active area of research in recent years (Bunescu and Pasca, 2006; Cornolti et al., 2013; Cucerzan, 2007; Hof-

fart et al., 2011; Kulkarni et al., 2009; Liao and Veeramachaneni, 2009; Ratinov et al., 2011). Our approach to diachronic normalization between different editions opens the door to time-specific entity disambiguation, which would link the mentions of a particular person in a historical text to the time-appropriate knowledge base, which in this case would be the encyclopedic edition from the same time period.

## 3 Methods

In order to track the change over time, we collected several historical editions of Encyclopaedia Britannica, including the 3rd, the 9th, the 11th, and the 15th editions. The first three editions are OCR-scanned version but the 11th edition is partially proofread by Project Gutenberg.[2] Encyclopaedia Britannica, Inc. gave us the authorization to use the electronic text of the current 15th edition for research.

Our text-processing pipeline includes article segmentation, people article extraction, and article matching. For the case study presented in this paper, we rely on our automated matching of articles between the 11th and the 15th edition. Published in 1911, the 11th edition was a fully reworked version of the encyclopedia which represented a substantial change in the state of knowledge from the last 19th century edition, and which remained mostly unchanged over the next several editions. We use edition 15 (the last paper edition of Encyclopaedia Britannica, converted to electronic form, 1985–2000) to represent the state of encyclopedic knowledge at the end of the 20th century. We normalize the 15th edition Britannica articles to their Wikipedia counterparts, and use Wikipedia categories as the proxy for different domains of culture.

Our framework relies on identifying the corresponding articles about the same historical figure in different editions, which are then used to form the representation for the stream of history. In the following subsections, we describe in detail the approach we used to extract the matching people-related articles, as well as the obtained estimates for system performance on different subtasks of the pipeline.

### 3.1 Article Segmentation

We developed a set of simple title heuristics to identify article titles in the historical editions. We look for uppercase words at the beginning of a line preceded by an empty line and followed by at least one non-empty line; the first word should be at least two characters long, and excludes frequent words such as "OCR", "BIBLIOGRAPHY", "FIG.", and Roman numerals.

For example, the following are the first sentences of the articles for Giorgio Baglivi in the 3rd and 9th editions, respectively:

```
BAGL1V1 (George), a most illustrious physician

BAGLIVI, GIORGIO, an illustrious Italian physician,
```

Note that OCR errors in the first word of article are quite common, as seen here in the 3rd edition title.

In addition, we used metadata regarding the titles present in each volume. For example, articles in the first volume are from A to Androphagi. Therefore, for the first volume, we use the regular expression that extracts potential titles that begin with "A" to "AN". We developed the heuristics for article segmentation in an iterative process which used the fact that article titles in the encyclopedia are sorted alphabetically. Article titles that appeared out of order were examined to refine the heuristics at each step.

The 11th edition of Encyclopaedia Britannica contained the total of 29 volumes, including the index volume. We obtained a digitized copy from Project Gutenberg. The errors caused by the OCR process were corrected manually by distributed proofreaders on the first seventeen volumes available from Project Gutenberg. However, the 4th, 6th, 14th, and 15th volumes are not complete. Therefore, we performed the article segmentation on the original fourteen OCR-Scanned and thirteen revised volumes. We also collected article titles from Project Gutenberg for segmentation and evaluation.

### 3.2 People Article Extraction

We use the 15th edition gender metadata to identify and extract articles about people from the current edition. In order to identify people articles in the historical editions, we use the Stanford CoreNLP named entity recognizer (NER) with pre-trained models,[3] on the first sentence of the article. The common format of a person name is "last name, first name" in the 9th and 11th edition and "last name (first name)" in the 3rd edition. The first token always serves as the article title and is prone to OCR errors, since it is usually all-capitalized, and in some editions, uses a special font.

[2] http://www.gutenberg.org

[3] http://nlp.stanford.edu/software/corenlp.shtml

For most historical figures, the encyclopedia gives both the first name and the last name. Typically, the first name which is not a part of the title is much more accurately recognized by the OCR. We therefore first check if the third token (which corresponds to the first name) was recognized as a person entity by Stanford NER. In cases when it is not recognized as such, we also check the first token. This is done in order to identify historical figures that do not have last name emperors, royal family members, mythological figures, ancient philosophers, etc. However, we observed that Stanford NER often mis-identifies locations as people in the first position. We therefore employ several heuristics to filter out the non-person articles, including checking for the presence of keywords such as 'he', 'his', 'she', 'her', 'born', and date or time mentions in the full text of the article.

### 3.3 Article Matching

We use two complementary strategies to match the articles that refer to the same person across different editions. The first matching method, pairwise matching, relies on the assumption that it is easier to match articles between consecutive editions, since for most articles, the text is likely to have undergone fewer changes. For each person article in a given edition of Encyclopaedia Britannica, we try to find a matching article about the same person in the next edition. If we fail to identify a matching article, we back-off to matching the same article directly to Wikipedia.

The pairwise matching results are concatenated to produce "chains" of matching articles between the four editions of Encyclopaedia Britannica. The last article of each chain is linked to the corresponding Wikipedia article. If the pairwise matching strategy fails to link together the matching articles in two adjacent editions, multiple incomplete chains may be generated. If several incomplete chains are linked to the same Wikipedia article, they are merged.

Figure 1 illustrates the article matching process using Dante Alighieri as an example. In this case, matching from the 3rd edition to the 9th edition fails, but the back-off strategy finds a matching article in Wikipedia. At the same time, pairwise matching between editions 9 and 11 and between editions 11 and 15 succeeds, and the article in edition 15 is successfully matched to Wikipedia. Since the article from the 3rd edition and the article from

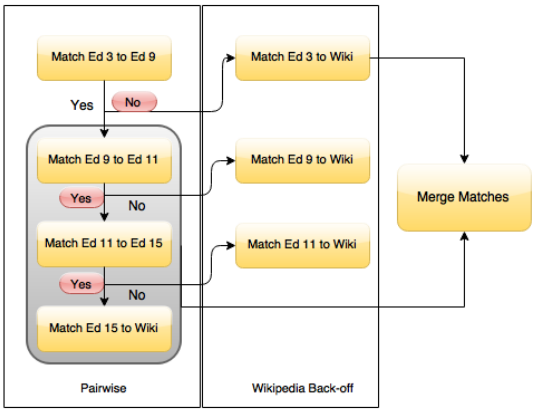


Figure 1: Article matching for *Dante Alighieri*.

the 15th edition are matched to the same Wikipedia article, the two incomplete chains ("Ed. 3–..." and "...–Ed. 9–Ed. 11–Ed. 15") are merged.

Both matching strategies first identify a set of possible matches (a *confusion set*), and then select the best matching candidate using a set of thresholds which were selected using the matching precision obtained on the development set. Two development sets were manually created by one of the authors: (1) 50 randomly selected people articles from 9th edition were matched to the 11th edition, (2) 50 randomly selected people articles from 15th edition were matched to Wikipedia. We describe the two strategies below.

#### 3.3.1 Pairwise Matching of Historical Editions

Note that some people featured in the older edition may not appear in the newer edition at all. Also, some of the people in the later Britannica editions may not have been alive and/or sufficiently known to be included in the encyclopedia when the previous edition was published. Therefore, the pairwise matching process proceeds from the earlier editions to the later editions. We first match the 15th (current) edition to Wikipedia, then the 11th edition to the 15th, the 9th edition to 11th, and finally, the 3rd edition to the 9th edition.

The matching methods are similar for each pair of historical editions. To use matching between the 11th and 15th editions as an example, we match the people articles from the 11th edition to the corresponding articles in the 15th edition by first identifying a set of potential matches (the confusion set) using a heuristic-based search on article titles. We then find the best matching article by computing the cosine similarity (Baeza-Yates et al.,

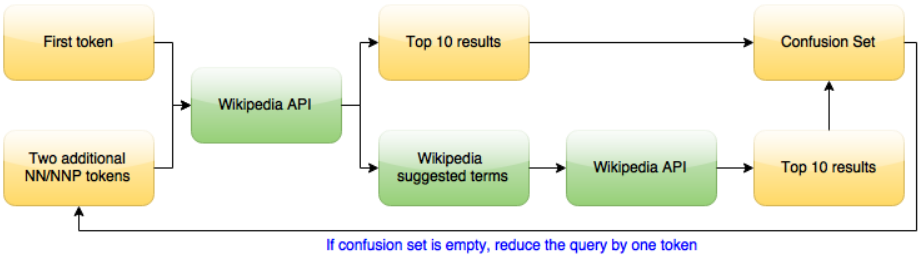


Figure 2: Deriving the confusion set with Wikipedia back-off strategy.

1999) between the original 11th edition article and the candidate article from the 15th edition. We use "bag-of-words" Boolean features on the full text of the article to compute cosine similarity. We then filter out the articles that do not have matches by applying a 0.2 threshold for minimum similarity.

In the present implementation, the confusion set is obtained by searching for all the 15th edition people articles that have the same first word. The resulting set of candidate matches contains all the articles about people with the same last name. We found that for people articles that do not contain last names (such as royalty, ancient writers and philosophers, etc.) searching on the first title word still produces a reasonable confusion set.

However, the first title word may also contain OCR errors. We are currently working on an OCR-correction system specifically tailored to the encyclopedic text. In the present implementation, we use the following solution. If no people with the same last name (first title word) are found, we take all the people with the longest matching prefix of the first title word. For example, due to the longest matching prefix, "Elme", "Elmes, James" in the 11th edition is compared to "Elmen, Gustav Waldemar" in the 15th edition. In this example, the cosine similarity between Elmen and Elmes is 0.07 and our method reports that no corresponding article exists in the 15th edition. However, if an article with the longest matching prefix is a correct match, the cosine similarity measure is likely to be above the selected threshold.

#### 3.3.2 Matching to Wikipedia as a back-off strategy

Using the articles with the longest matching prefix allows us to identify the correct match in case when the OCR error occurs far enough from the beginning of the word. If the misspelling occurs in the very beginning of the first word, the best match for the resulting confusion set will be filtered out by the similarity threshold. For those cases, we use a back-off strategy that attempts to reprocess the articles with no matches by obtaining a new confusion set from Wikipedia API. In order to query the Wikipedia API, we use the first token and two additional tokens with NN, NNP, NNS, or NNPS part-of-speech tags (if any), identified using the CoreNLP part-of-speech module. We use this query to retrieve the top 10 search results from Wikipedia. Wikipedia API also suggests a possible correction to the query. We use the suggested query to retrieve the top 10 search results (if any), which are then used to expand the candidate set. If no results are retrieved using the original and the suggested query constructed from three keywords as described above, the first two keywords are used to repeat the above steps. The process of obtaining the confusion set is illustrated in Figure 2.

In order to reduce processing time, we set up Java Wikipedia Library (JWPL)[4] to access all information in Wikipedia locally. Wikipedia page titles of the candidate set are used to retrieve plain text of Wikipedia article from JWPL. Cosine similarity is then calculated for each candidate article to find the best match. We use "bag-of-words" TF-IDF (Salton and Yang, 1973) scores on the full text of the article to compute cosine similarity. We then filter out the articles that do not have matches by applying a 0.13 threshold for minimum similarity. As mentioned above, the threshold was selected based on the matching precision for the development set. Note that using boolean features for pairwise matching between historical editions effectively reduces the noise caused by the OCR errors. The back-off matching strategy uses TF-IDF features for cosine similarity calculation, since the clean electronic text is available for Wikipedia.

Since the 15th (current) edition is available in electronic form, correct and complete names

[4]https://code.google.com/p/jwpl/

can almost always be retrieved from the metadata. We therefore use the complete names, rather than the first three noun tokens, to retrieve the Wikipedia articles with the same title. The candidates are retrieved using both JWPL functionality and Wikipedia API. If several namesakes are present in Wikipedia, the best match is selected using cosine similarity.

### 3.4 Importance Measure

In the current implementation, we use a simple z-score as an importance measure, with the following formula:

$$importance(a) = (L(a) - average(L))/stddev(L)$$

where $a$ is a particular person article, $L$ is the article length (i.e. the number of words in that article), and average and standard deviation are computed for all articles in a given edition (Gabrovski, 2012).

Note that importance can be measured in a number of ways, for example, using the number of times a person is mentioned in other articles, or using a PageRank on an article graph constructed for each edition. An article graph can be constructed by treating person mentions or "see also" references in a historical edition as edges between the article nodes, making a historical edition more similar to Wikipedia, in which hyperlinks added by the users serve as connecting edges. However, OCR errors make any methods relying on person mentions less robust.

### 3.5 Gold Standard Data

System performance on article segmentation, people article extraction, and article matching was evaluated on a gold standard data created by an independent annotator. An evaluation set of 1000 randomly selected articles was created for each historical edition separately using the article segmentation produced by the system. The articles were divided into 20 equal-size bins, and 50 consecutive articles were picked from each bin. The annotator was asked to go through the 1000 articles for each edition, and perform the following tasks for each article: (1) check if the article segmentation is correct, (2) check if the subject of the article is a person, and (3) for person articles, find the matching articles a) in the next historical edition and b) in Wikipedia. For the 15th edition, 1000 articles were selected using the segmentation provided by the electronic edition, and the annotator performed only the tasks of people extraction and matching to Wikipedia.

Our preliminary pilot annotation experiments conducted during annotator training indicated that annotator error was highly unlikely for these tasks. We therefore created the gold standard using a single annotator whose work was spot-checked for correctness by one of the authors. Table 1 shows the results of annotation for people article extraction and matching. System segmentation accuracy is shown in Table 2.

| | Ed. 3 | Ed. 9 | Ed. 11 | Ed. 15 |
|---|---|---|---|---|
| Total # of person articles in evaluation set | 137 | 368 | 407 | 335 |
| Person articles with matches in the next edition | 75 | 337 | 232 | n/a |
| Person articles with matches in Wikipedia | 124 | 364 | 403 | 327 |

Table 1: Person articles in gold standard data.

## 4 Results

Table 2 shows the results of evaluation for article segmentation, extraction of articles about people, and the matching of corresponding articles across different editions.

### 4.1 Article Segmentation

Segmentation accuracy is the percentage of articles the system segmented correctly. According to the annotation results, the segmentation accuracy for the 3rd, the 9th, and the 11th editions are 92.2%, 96.5%, and 99.9%, respectively. Since the 15th edition is available in XML format, it is excluded from segmentation evaluation.

### 4.2 Person Article Extraction

We estimated the number of person articles in each of the historical editions using the number of articles about people identified in the 1000 articles reviewed manually by the annotator. Table 2 shows the estimate for number of person articles in each edition, as well as the number of articles identified by the system.

The recall and precision for person article extraction for each edition are computed as the ratio of the number of person articles identified correctly by the system to the total number of person articles identified by the annotator (for recall), and the total number of person articles extracted by the system (for precision). Person article recall for all historic editions is around 70%, so there are about 30% of

person articles not recovered. This can likely be addressed by developing additional name patterns or annotating Britannica to retrain CoreNLP NER models. For the 15th edition, the articles that contain gender information in the metadata are identified as person articles by the system, which fails to recover approximately 10% of person articles.

### 4.3 Article Matching

*Person article pairwise precision* is the percentage of article pairs between adjacent historical editions that are identified correctly, relative to the total number of matches identified by the system. Since the 15th edition is matched directly to Wikipedia, it is excluded from this evaluation. *Person article matching precision* is the percentage of person articles for which the matching articles in Wikipedia were identified correctly, relative to the total number of matches identified by the system. The *pairwise recall* and *matching recall* are computed accordingly, with the percentages reported relative to the total number of matches identified by the annotator.

| | Ed. 3 | Ed. 9 | Ed. 11 | Ed. 15 |
|---|---|---|---|---|
| Segmentation Accuracy | 92.2% | 96.5% | 99.9% | n/a |
| Estimated # of Person Articles | 2654 | 5910 | 14823 | 27465 |
| System-detected # of Person Articles | 2089 | 4600 | 10702 | 26230 |
| Person Article Precision | 80.0% | 94.7% | 93.8% | 100.0% |
| Person Article Recall | 67.2% | 72.6% | 74.0% | 91.3% |
| Person Article Pairwise Precision | 88.2% | 99.6% | 96.2% | n/a |
| Person Article Pairwise Recall | 57.7% | 89.6% | 92.6% | n/a |
| Person Article Matching Precision | 81.0% | 96.1% | 93.3% | 96.5% |
| Person Article Matching Recall | 40.0% | 82.6% | 83.6% | 91.3% |

Table 2: Evaluation results for segmentation, person article extraction, and matching.

Note that the last two rows in Table 2 show the matching precision and recall obtained by two complementary strategies described in Section 3, giving the estimates of the overall quality of the matching algorithm. Note that precision and recall improve progressively for the later editions, and with exception of edition 3, we obtain the precision above 90% and recall above 80%. Edition 3 recall is substantially lower due to the diminished quality of the OCR scan, and the differences in the fonts and the formatting conventions. One should also keep in mind that the matching precision and recall are computed over the articles that have been recognized as person articles, therefore in order to get the estimates for the actual number of articles matched correctly, one should factor in person article recall.

## 5 Use Case Study

We applied our approach to the Wikipedia category of the 18th century classical composers in order to investigate whether the output of our algorithm can be used to identify valid trends in the rise and fall in reputations of historical figures. Wikipedia uses collaboratively created categories to group articles based on a variety of classificatory principles.

We investigated the change in the reputations of the 18th century classical composers using the corresponding Wikipedia category. Currently, there are 109 composers in this category. We evaluated manually the matching accuracy for the articles in this category, obtaining 95.5% and 89.1% matching accuracy for the 15th edition and 11th edition respectively, with the lower matching accuracy for 11th edition mainly caused by segmentation errors.

We used Web-based Analysis and Visualization Environment (WEAVE)[5] to visualize and analyze the relative importance, rank, and its change over time for the historical figures in this category. Figure 3 illustrates the change in importance. The legend on the left lists the composers alphabetically. The top two bar charts are their rank in the 11th and 15th editions, respectively. The bottom bar chart shows their reputation change from 11th to 15th edition, sorted on its absolute value.

## 6 Discussion

Table 3 shows the relative ranking for the most important 18th century composers in the 11th and 15th editions of Britannica. Each composer's importance score is shown in parentheses, with the higher relative importance score corresponding to a higher rank. Interestingly, while the top five composers remained the same, the order of importance underwent a significant change. In the 15th edition, Mozart replaced Bach at the top of the hierarchy, a change potentially brought on by the era of sound recording which led to classical music reaching a wider audience; this may have proved detrimental to Bach's difficult polyphonies, while Mozart's light melody lines with suitable harmonic accompaniment rose in popularity.

[5] http://www.oicweave.org


Figure 3: Biggest "movers and shakers" among the 18th century composers.

| 1911 (11th edition rank) | 1985-2000 (15th edition rank) |
|---|---|
| 1. Johann Sebastian Bach ($2.2 \times 10^{-4}$) | 1. Wolfgang Amadeus Mozart ($1.9 \times 10^{-4}$) |
| 2. George Frideric Handel ($2.0 \times 10^{-4}$) | 2. Johann Sebastian Bach ($1.4 \times 10^{-4}$) |
| 3. Wolfgang Amadeus Mozart ($1.5 \times 10^{-4}$) | 3. Joseph Haydn ($7.5 \times 10^{-5}$) |
| 4. Christoph Willibald Gluck ($9.2 \times 10^{-5}$) | 4. Christoph Willibald Gluck ($6.6 \times 10^{-5}$) |
| 5. Joseph Haydn ($5.6 \times 10^{-5}$) | 5. George Frideric Handel ($5.3 \times 10^{-5}$) |

Table 3: The rank of top five 18th century composers. The importance score is shown in parentheses.

The most drastic change within the top 5 composers was Handel's drop from the second to the fifth place. A possible explanation may lie in the history of genres: in the 20th century, the genres of the archaic Italian opera and oratorio that defined Handel's oeuvre lost their popularity and were, in general, less frequently performed and recorded.

These trends seem to be confirmed by the frequency plots for the names of these composers obtained from the Google Ngram Viewer (Figure 4). For the first decade of the 20th century Bach is the most frequently mentioned composer, with Handel and Mozart sharing the second position; towards the end of the century, the mention frequency for Mozart approaches and sometimes surpasses Bach, while the mention frequency for Handel falls.

Two composers that did not even have a dedicated article in the 11th edition, but ranked quite high in the 15th edition are Georg Philipp Telemann and Antonio Vivaldi, aka "the red priest". Their 20th century rediscovery is a well known fact. Importantly, our algorithm has been able to "catch" these two comebacks automatically.

The reverse case is the Venetian Antonio Lotti (1667-1740), a composer who according to our algorithm, was considered rather important in the beginning of the 20th century but lost his stature towards its end. Lotti's rare fans should not be discouraged; he may well be due for rediscovery in the 21st.

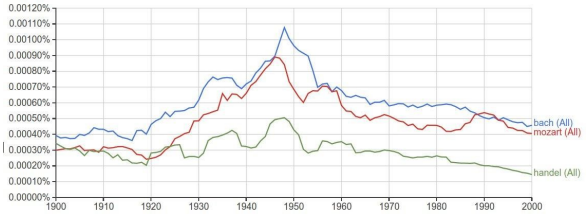


Figure 4: Google Books mention frequency for Mozart, Bach, and Handel.

## 7 Conclusion and Future Work

We have developed a method for matching and comparison of articles about people in historical editions of EB, as well as mapping category information from Wikipedia to EB. Our analysis has shown that the automated comparison between the historical editions of EB can be used to detect and track the historical changes within selected domains of culture. In the future, we plan to extend the pipeline to other editions of EB, thus widening the chronological scope of our research, and scale

up from a few selected categories to a wider range of categories encompassing different domains of cultural and political history.

## References


Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. 1999. *Modern information retrieval*, volume 463. ACM press New York.

Razvan C Bunescu and Marius Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *EACL*, volume 6, pages 9–16.

Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. 2013. A framework for benchmarking entity-annotation systems. In *Proceedings of the 22nd international conference on World Wide Web*, pages 249–260. International World Wide Web Conferences Steering Committee.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL*, volume 7, pages 708–716.

Aleksandar R. Gabrovski. 2012. *Knowevo and Gravebook: Tracking the History of Knowledge*. An undergraduate thesis, Dartmouth College, Hanover, NH.

M Gronas, A Rumshisky, A Gabrovski, S Kovaka, and H Chen. 2012. Tracking the history of knowledge using historical editions of encyclopedia britannica. In *Workshop on Adaptation of Language Resources and Tools for Processing Cultural Heritage Objects. LREC*.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 782–792. Association for Computational Linguistics.

Thomas Huet, Joanna Biega, and Fabian M Suchanek. 2013. Mining history with le monde. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 49–54. ACM.

Mike Kestemont, Folgert Karsdorp, and Folgert Karsdorp. 2014. Mining the twentieth centurys history from the time magazine corpus. *EACL 2014*, page 62.

Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Collective annotation of wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–466. ACM.

Wenhui Liao and Sriharsha Veeramachaneni. 2009. A simple semi-supervised algorithm for named entity recognition. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 58–65. Association for Computational Linguistics.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. 2011. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: bringing order to the web.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1375–1384. Association for Computational Linguistics.

Gerard Salton and Chung-Shu Yang. 1973. On the specification of term values in automatic indexing. *Journal of documentation*, 29(4):351–372.

Steven S Skiena and Charles B Ward. 2013. *Who's Bigger?: Where Historical Figures Really Rank*. Cambridge University Press.

Fabian M Suchanek and Nicoleta Preda. 2014. Semantic culturomics. *Proceedings of the VLDB Endowment*, 7(12):1215–1218.