Using YAGO for the Humanities

Thomas Rebele[1], Arash Nekoei[2], and Fabian M. Suchanek[1]

[1] Telecom ParisTech, France
[2] University of Stockholm, Sweden

Abstract. In this paper, we study how data from the Semantic Web can be used for case studies in the Humanities. We conduct a proof of concept, using the YAGO knowledge base to study life expectancy, birth rates, and the age at childbirth over time. We also discuss the information extraction methods that we used to make YAGO sufficiently complete for these analyses to work.

1 Introduction

Recent advances in the Semantic Web have led to more and more data about the real world in structured form. This data is stored in knowledge bases (KBs). In parallel to this development, the Digital Humanities have established themselves as a research field in their own right, using computational methods to support researchers in their quest to understand history, language, or the arts. Recently, these fields of research have found common ground [1, 9], with the insight that the data of the Semantic Web can help deliver answers to questions in the Digital Humanities. In this paper, we want conduct a proof of concept, and investigate in how far data from YAGO [11], one of the largest general-purpose KBs on the Semantic Web, could help study questions of history and society.

Our first observation is that the data in YAGO is very incomplete. For example, only half of the people in YAGO have a place of residence.[3] Therefore, our first challenge is to make the data more complete. YAGO extracts its information by automated means from the online encyclopedia Wikipedia and other sources. A large part of the incompleteness in YAGO stems from the fact that these extraction mechanisms miss important pieces of information in this process. In this paper, we describe how these extraction mechanisms can be improved so that they distill even more information from the source. We show that our techniques improve the coverage of YAGO on certain attributes by up to 60%.

To show the usefulness of this new data, we then proceed to several case studies: First, we use the data in YAGO to trace the evolution of life expectancy across history, by gender. Then, we use the data to refute the myth that full moon days see more births than other days. Finally, we trace the age at child birth over time. With these analyses, we show that the data from the Semantic Web can help shed light on questions of the humanities.

[3] Incompleteness is a general problem on the Semantic Web: In DBpedia [7], only 0.2% of people have a gender, and in Wikidata [15], only 3% of people have father [10].

This paper is structured as follows: Section 2 discussed related work. Section 3 presents our improved methods for information extraction, which we evaluate in Section 4. Section 5 presents our case studies, before Section 6 concludes.

2 Related Work

Several projects have recently started to create large knowledge bases. Among the most visible ones are YAGO [13], DBpedia [7], NELL [3], BabelNet [8], WikiData [15], and Google's Knowledge Vault [5]. In this paper, we compare the coverage of our improved YAGO to DBpedia, which was also derived from Wikipedia.

The idea of using data from the Semantic Web to support the Digital Humanities is in its infancy [1, 9]. Schich et al. [12] use Freebase to trace the birth and death locations of intellectuals. However, their study was limited to only 150,000 people, while we aim an order of magnitude higher. De La Croix et al. [4] trace the longevity of famous people across history. Likewise, their study was limited to the 300,000 people in the "Index bio-bibliographicus notorum hominum", while we aim to show the value of Semantic Web data, which is much more ample. Gergaud et al. [6] come closest to our approach: They build a database of 1,1m people from Wikipedia and study the economic impact of these individuals. In this paper, we show how to build a database that contains twice as many people from Wikipedia.

3 Methods

3.1 Information Extraction in YAGO

In accordance with the RDF standard, YAGO stores information in the form of triples. Each triple consists of a subject, a relation name, and an object – as, e.g., in ⟨Cleopatra, wasBornin, Egypt⟩. In total, YAGO contains 10 million entities (like persons, organizations, cities, etc.), and more than 120 million triples about these entities. The KB knows 100 relation names, which have been defined manually. YAGO assigns each entity to one or several classes. Cleopatra, e.g., is in the class *People from Alexandria*. This class is a subclass of the class *Person*, which is in turn a subclass of *Organism*, and so on.

The main part of the information in YAGO stems from Wikipedia. Every article in Wikipedia becomes an entity in YAGO. The triples are extracted from the infoboxes in Wikipedia and the category names. This works with a modular architecture [2], in which small Java programs (called *extractors*) produce sets of triples (called *themes*). These themes are then post-processed by other extractors: they are cleaned, deduplicated, and checked for consistency. This results in a sequence of themes of ever cleaner data, of which the final themes constitute the YAGO KB. The improvements that we propose follow this schema: we propose to add new extractors and link them into this process.

One particularity of YAGO is that it has a manually evaluated precision of 95% with respect to Wikipedia. This means that, statistically, only 1 triple out of 20 does not correspond to the facts in Wikipedia. All of our improvements have to respect this quality constraint.

3.2 Gender

The infoboxes of Wikipedia do not mention the gender of a person. Therefore, earlier versions of YAGO did not have gender information. Gender was added only in YAGO3, and it was extracted from the occurrence of pronouns on the page. The *GenderPronounExtractor* counts the number of occurrences of "he" and "she" in the articles. If the number of "he" is at least twice the number of occurrences of the word "she", and if the number of occurrences is at least 10, the gender is assumed to be male (and vice versa). This worked well, but it had a rather low coverage. Only 61% of people had a gender.

We improved this method by making use of the gender-specific categories in Wikipedia. For example, Cleopatra is in the category *Female people from Alexandria*. We wrote the *GenderCategoryExtractor*, which considers every article x, and produces the fact $\langle x$, hasGender, female$\rangle$ if the article is in a category that contains the substring *female* (analogously for the male categories). This works well, but it still has a low coverage.

We improve upon this as follows: We wrote the *GenderNameExtractor*, which collects the first names of all people with a known gender. The same first name ν may be associated to people of different gender – either because the name is used for both males and females, or because of errors in Wikipedia or the extraction process. Our goal is to determine whether ν is associated to one gender in the majority of cases. Let us say that our sample for name ν has a proportion of $p\%$ males. We want to know what is the proportion of male people with the name ν in the real world. For this purpose, we run a statistical test: We assume that our set of people with first name ν is a sample from the real world. Then we generalize the proportion of males in our sample to the proportion of males in the real world. We use the Wilson estimator for this purpose, with $\alpha = 5\%$. Only if the lower bound of the Wilson interval is at least 95%, we assign the name to the male gender (analogously for female). The values of α and 95% are those that the YAGO evaluation [13] used.

All of these three extractors produce sets of facts in the YAGO framework. These are then collected by a fourth extractor, which assigns at most one gender to each person, giving priority to the *GenderCategoryExtractor*, followed by the *GenderNameExtractor* and the *GenderPronounExtractor*.

3.3 Dates

YAGO harvests the birth dates and death dates of people from two sources: From the infoboxes of Wikipedia and from the categories of Wikipedia. The infoboxes contain a list of attribute-value-pairs, of the form *date-of-birth = Jan 8, 1935*. As for the categories, there exists one category per birth year (e.g., *1935 births*).

YAGO uses the date from the infobox where available, and defaults to the year from the category otherwise.

The problem with this approach is that the extraction quality is higher for the categories than for the infoboxes. This is due to more varied date formats in the infoboxes (*01/08/1935, 08/01/1935, 8 Jan 1935*, etc.). Therefore, we modified the respective extractor as follows: If only one birth date is available, that date is used. If the year from the infobox and the category coincide, the date from the infobox is used. Otherwise, we use the year from the category (and abandon the date from the infobox). We proceed analogously for death dates.

3.4 Locations

YAGO extracts the birth place, death place, and place of residence from the infoboxes of Wikipedia. The problem with this approach is that the data in the infoboxes is very sparse. However, some categories give away the nationality or the place of residence of a person, as in *Egyptian queens regnant*.

Therefore, we have written a new extractor (*CategoryLocationExtractor*) that harvests also the categories of Wikipedia. We first compiled a list of demonyms – in part from the Wikipedia list of demonyms, and in part from the lists of empires on Wikipedia. Our extractor then scans the categories of a person x, and counts the number of times each country's demonym appears. For every country y that appears most often, we create a fact $\langle x, \text{livedIn}, y\rangle$. Our rationale is that the categories can neither determine the birth place nor the nationality reliably, but that they can at least indicate a place of residence.

4 Results

In this section, we study the effect that our new extraction methods have. We consider two axes: Precision and coverage. For precision, we took a sample of 100 facts per method, and checked them manually against Wikipedia, as we usually do for YAGO [13]. For coverage, we counted the number of unique people who have a certain attribute. We use the Wikipedia dumps from 2017-02-20 to generate YAGO. We compared our coverage to the previously implemented methods of YAGO, run on the same Wikipedia dumps.

We also report the coverage in DBpedia, the other big knowledge base extracted from Wikipedia. These numbers have to be taken with a grain of salt: First, there exist different versions of DBpedia, one for each Wikipedia language. We take here the English one (which is the largest). YAGO, in contrast, extracts from 10 Wikipedias. Second, the precision of DBpedia is not known. We did not want to make statements about the quality of DBpedia in place of its authors. Thus, our results should not be understood as a direct comparison.

Table 1 shows our results. YAGO contains 2,284,927 people. Our precision values are very good. The 2% wrong results for genders were exclusively due to type errors, not to the extraction itself. The extraction of residences also suffered from that problem. In addition, it produced 6% of anachronistic residencies (such

as *German Empire* instead of *Germany*). We counted them as correct for the purposes of our study. The precision of birth and death dates is extraordinary.

In addition, our methods have significantly increased the number of data points for all attributes that we consider. Genders, e.g., increased by 35%. Almost three times as many people as before have a place of residence.

Extraction	YAGO before	YAGO now	Precision	DBpedia (en)
Birth Dates	1,566,227	1,688,282	100%	819,371[4]
Death Dates	745,411	822,234	100%	322,212[4]
Place of Residence	692,770	2,085,110	97%	683,854[5]
Gender	1,471,120	1,983,734	98%	4,419[6]

Table 1. Coverage of people and precision of our methods

5 Case Studies

In this section, we aim to show the usefulness of our data in case studies.

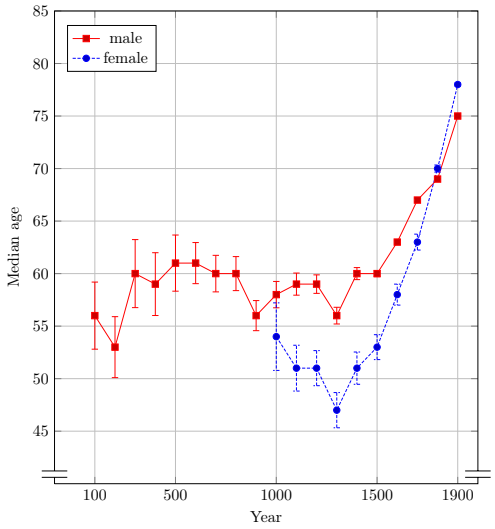

Fig. 1. Median age over time, by year of birth (with the Student's t confidence interval at $\alpha = 95\%$).

[4] from the DBpedia statistics Web page
[5] union of birth-place, death-place, and residence from "Mapping based objects"
[6] "Gender" dataset

5.1 Life expectancy over time

Our first study (Figure 1) investigates the pattern of life spans across history. It does so by plotting the average life span of males and females against time. We restricted the study to centuries where we had more than 100 men and women, respectively. Interestingly, there is no trend in the data until the 15th century, with the average age fluctuating around 53 and 60 years for females and males, respectively. The effects of the Black Death are clearly mirrored in our data: life expectancy decreases in the 13th century. Beyond that, there is an steady increase in the life span across genders during the last 500 years. We also see that women generally had a shorter life time in our data. This changes, however, in the 19th century: Women live longer than men. As our data is quite dense from the 19th century onwards, this fact is statistically significant in our data.

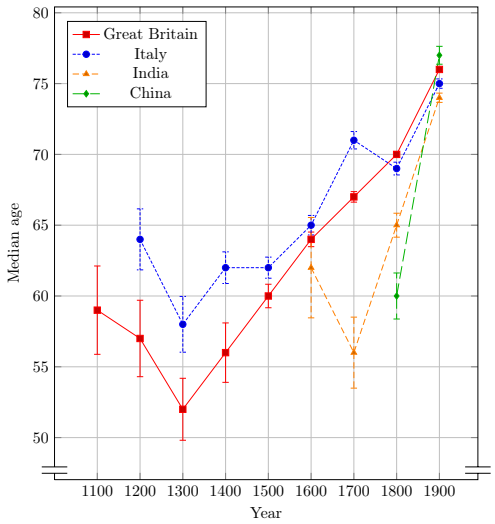

Fig. 2. Median age over time, by year of birth (with the Student's t confidence interval at $\alpha = 95\%$).

We can also drill down into the historical life expectancy per country. Figure 2 shows 4 countries that existed continuously over the past 1000 years, together with the median age of their population (only for centuries with more than 100 data points). The figure shows that the life span has been increasing since the Black Death in the two developed countries in the sample, Italy and Great Britain. We also see a catch-up effect: India and China have been experiencing a much larger increase in life span during the last 200 years. Another interesting observation is the take-over of Italy by Great Britain. This is mainly due to a deceleration of Italy rather that an acceleration of Great Britain, and it is thus difficult to argue that is related to industrial revolution. Today, all 4 countries

have a comparable life expectancy in our data. This result has to be taken with a grain of salt: Wikipedia (and hence YAGO) contains mainly the elite population. Results may thus not generalize to the full population of a country. Our next study will shed light on that divergence.

5.2 Births per month

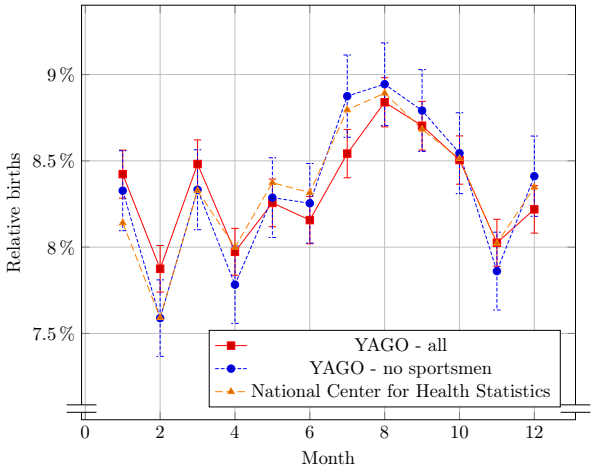

Fig. 3. Births per month (with the Student's t confidence interval at $\alpha = 95\%$).

Our next study tests whether the likelihood of making it to Wikipedia/YAGO is associated with the month of birth (Figure 3). For this purpose, we plotted the number of births in the US per month in Figure 3, and compared it to the number of births per month according to the U.S. birth registry[7]. We find that generally, both graphs show the same peaks. However, we find that people born in January are slightly more likely to be in Wikipedia than expected.

One potential explanation for this pattern is that our data comprises many sportsmen, who benefit from the relative age effect[8]: Sportsmen born early in the year are slightly older and thus slightly more mature than sportsmen born later in the year with whom they usually compete. This makes them more successful, and hence slightly more likely to appear in Wikipedia. Indeed, if sportsmen are removed from the graph ("YAGO other"), our curve becomes more similar to the US census data.

[7] From the National Center for Health Statistics: `http://abcnews.go.com/Health/Science/story?id=990641`, [14]

[8] `https://en.wikipedia.org/wiki/Relative_age_effect`

5.3 Full Moon Myth

A popular myth has it that there is an increase in births on full moon days[9]. We wanted to analyze this conjecture with our data. For this purpose, we computed the time points of the full moon over the past centuries. We restricted our analysis to 1600 to 2000, as our full moon calculation bases on the Gregorian calendar introduced in 1582. We also did point checks with known historical full moon dates to make sure our computation is correct.

We consider the two days that precede, and the two days that follow a full moon day as "full moon days". Among the 146,463 days between 1600 and 2000, we classified 24,793 days as full moon days. This gives a total ratio of full moon days of 16.9%. During that time period, 691,616 people were born, of which 117,081 fall on a full moon day. This gives a total ratio of full moon births of 16.9% – exactly as expected. The Wilson score interval at $\alpha = 95\%$ is 16.9%±0.1%. Thus, the full moon has no influence on the birth rates in our data.

Additionally, we plotted the proportion of people born on or around a full moon day between 1600 and 2000 per century. Figure 4 shows our results. The great fluctuation seems to indicate a relationship between the period of the moon and the frequency of births. However, this fluctuation is mainly due to sparse data. As the confidence interval shrinks, the proportion of full moon births converges to the expected value of 16.9%.

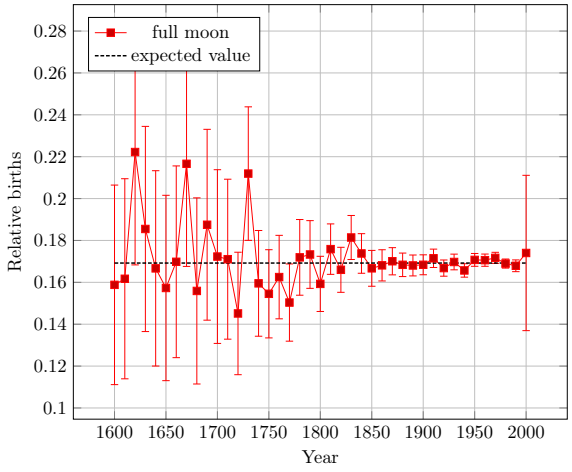

Fig. 4. Proportion of births on a full moon day (±2.5 days, with a Wilson score interval at $\alpha = 95\%$).

[9] `https://www.google.com/#q=full+moon+birth`

5.4 Age at first child birth

We now turn towards the age at which people become parents. Figure 5 shows the age at which people have their first and last child during the last millennium. Again, we restricted our analysis to centuries where we had at least 100 parent-child pairs. Males show an increase in the age at which they have their first child, moving from around 28 to around 32. For females, we see similar story, with one important difference: for females, the increase is concentrated during the last 2 centuries. However, the age at which males have their last child is almost unchanged until 1700s, and declining since then. For females, the age at the birth of last child is unchanged. We speculate that two demographic phenomena are creating these patterns. First, women having less children by mainly postponing the age at which they have their first child. Second, the age difference between fathers and mothers is declining. These two forces together can create an increasing first-child age of mothers and a declining last-child of fathers with constant last-child age of mothers and first-child of fathers.

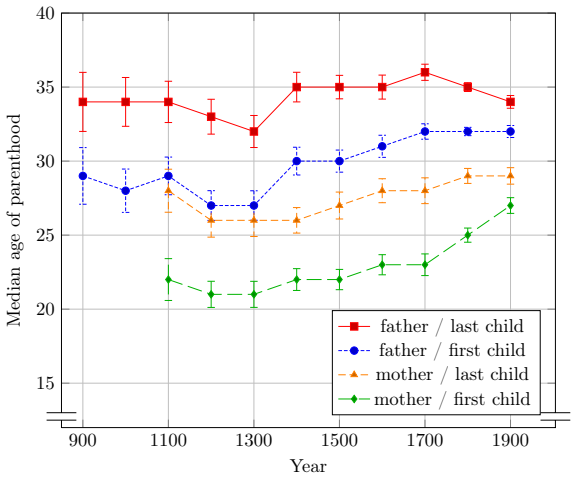

Fig. 5. Age of parenthood, by year of birth (with the Student's t confidence interval at $\alpha = 95\%$)

6 Conclusion

In this paper, we have investigated how data from the Semantic Web can help research in the Digital Humanities. As a proof of concept, we have used the YAGO knowledge base, one of the largest general-purpose knowledge bases on the Semantic Web, to study the life expectancy of people across different times; the age at first childbirth; and the myth that full moon days see more births. We have also presented methods to improve the coverage of YAGO.

Our methods are integrated in the YAGO infrastructure, and the data we generate will be included in the next release of YAGO. The data of YAGO can be downloaded freely at `http://yago-knowledge.org`. A SPARQL endpoint is available at `https://w3id.org/yago/sparql`. We also provide the source code used for this study. The code for generating the previous YAGO is available at `https://github.com/yago-naga/yago3/tree/whise2017-YAGO-before`. Our new algorithms are available at `https://github.com/yago-naga/yago3/tree/whise2017`. The data generated for this study is available at `https://www.thomasrebele.org/projects/whise2017`. The entire YAGO project is available as open source at `https://github.com/yago-naga/yago3`.

Acknowledgments This research was supported by the grants ANR-11-LABEX-0045-DIGICOSME and ANR-16-CE23-0007-01 ("DICOS").

References

1. Alessandro Adamou, Enrico Daga, and Leif Isaksen, editors. *Workshop on Humanities in the Semantic Web*, volume 1608 of *CEUR Workshop Proceedings*, 2016.
2. Joanna Asia Biega, Erdal Kuzey, and Fabian M. Suchanek. Inside YAGO2s: A Transparent Information Extraction Architecture . In *WWW demo track*, 2013.
3. A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. Hruschka Jr., and T. Mitchell. Toward an architecture for never-ending language learning. In *AAAI*, 2010.
4. David de la Croix and Omar Licandro. The longevity of famous people from Hammurabi to Einstein. *Journal of Economic Growth*, 2015.
5. X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In *SIGKDD*, 2014.
6. Olivier Gergaud, Morgane Laouenan, and Etienne Wasmer. A brief history of human time. exploring a database of notable people. *Sciences Po Economics Discussion Papers*, 2017.
7. J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. v. Kleef, S. Auer, and C. Bizer. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 6(2), 2015.
8. Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.
9. Angeliki Rapti, Dimitrios Tsolis, Spyros Sioutas, and Athanasios Tsakalidis. A survey: Mining linked cultural heritage data. In *EANN*, 2015.
10. S. Razniewski, F. M. Suchanek, and W. Nutt. But what do we actually know? *AKBC workshop*, 2016.
11. Thomas Rebele, Fabian M. Suchanek, Johannes Hoffart, Joanna Asia Biega, Erdal Kuzey, and Gerhard Weikum. YAGO: a multilingual knowledge base from Wikipedia, Wordnet, and Geonames . In *ISWC*, 2016.
12. Maximilian Schich, Chaoming Song, Yong-Yeol Ahn, Alexander Mirsky, Mauro Martino, Albert-Lszl Barabsi, and Dirk Helbing. A network framework of cultural history. *Science*, 2014.

13. F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *WWW*, 2007.
14. United States Department of Health and Human Services (US DHHS), Centers for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS), Division of Vital Statistics. Natality public-use data 2003-2006, and 2007-2015, on CDC WONDER Online Database, February 2017. Accessed at `https://wonder.cdc.gov/natality.html` on Jul 24, 2017 9:36:18 AM.
15. D. Vrandečić and M. Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 2014.