Automatically Generating Wikipedia Info-boxes from Wikidata

Tomás Sáez
DCC, Universidad de Chile
tsaez@dcc.uchile.cl

Aidan Hogan
Institute for the Foundations of Data
DCC, Universidad de Chile
ahogan@dcc.uchile.cl

ABSTRACT

Info-boxes provide a summary of the most important meta-data relating to a particular entity described by a Wikipedia article. However, many articles have no info-box or have info-boxes with only minimal information; furthermore, there is a huge disparity between the level of detail available for info-boxes in English articles and those for other languages. Wikidata has been proposed as a central repository of facts to try to address such disparities, and has been used as a source of information to generate info-boxes. However, current processes still rely on human intervention either to create generic templates for entities of a given type or to create a specific info-box for a specific article in a specific language. As such, there are still many articles of Wikipedia without info-boxes but where relevant data are provided by Wikidata. In this paper, we investigate fully automatic methods to generate info-boxes for Wikipedia from the Wikidata knowledge graph. The primary challenge is to create ranking mechanisms that provide an intuitive prioritisation of the facts associated with an entity. We discuss this challenge, propose several straightforward metrics to prioritise information in info-boxes, and present an initial user evaluation to compare the quality of info-boxes generated by various metrics.

CCS CONCEPTS

• **Information systems** → Wikis;

KEYWORDS

Wikipedia, Wikidata, Info-box, Ranking

ACM Reference Format:
Tomás Sáez and Aidan Hogan. 2018. Automatically Generating Wikipedia Info-boxes from Wikidata. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3184558.3191647

1 INTRODUCTION

As one of the largest collections of human knowledge – collaboratively edited by hundreds of thousands of active users – Wikipedia hardly needs introduction here. However, Wikipedia is, by its nature, always a work in progress. While new articles are added to reflect new entities, and old articles are edited to improve their quality and accuracy, other work is ongoing to improve the Wikipedia infrastructure itself. One major development along these lines has been the creation of the complementary Wikidata knowledge graph [12]. A core premise behind this knowledge graph is to allow users to curate structured data directly, in a central location, in as language-agnostic and interoperable a manner as possible.

Before Wikidata, most (semi-)structured data associated with Wikipedia was embedded directly into articles in the form of info-boxes, tables, lists, categories, and so forth (where such data was extracted and integrated by various mechanisms to form rich and popularly-used datasets such as DBpedia [5] or YAGO [1]). However, managing data on Wikipedia in this form is far from ideal. For example, when a prolific football player scores a goal in an international match, that goal may necessitate manual edits to many different articles: the total goals of that player in their respective info-box, a table with the top scorers for that tournament, the all-time top scorers for that national team, and so forth; considering that there are 288 actively edited Wikipedias[1] corresponding to different languages, one can see that a single goal scored by a player could potentially require hundreds or thousands of manual edits to maintain the structured data of Wikipedia up-to-date and consistent across different languages. Clearly this situation leads to huge inefficiencies in terms of the use of human effort. Further given the disparity in active editors available for different languages, this leads to many articles not having an info-box provided, inconsistent information across different language versions, and so forth [6, 7].

Recognising such deficiencies in how Wikipedia has managed its structured content, Wikidata was thus proposed to instead gather such content in a central location. Being structured, the underlying data uses language-independent identifiers, where multilingual labels and descriptions can be assigned to individual entities and properties; thereafter, facts are given as tuples of these entities and properties that can be surfaced in any language for which the labels of the constituent entities and properties are available. This feature of Wikidata minimises the effort required to generate information in various languages.[2] Furthermore, being designed from scratch with structured data in mind, Wikidata allows various permutations of the underlying data to be generated with a single query; e.g., rather than human editors having to manually maintain a list of top scorers in a tournament, such a list can be generated and/or refreshed by a query to the underlying dataset as needed.

Since its inception, Wikidata has experienced significant growth and development, where descriptions for 42.6 million items (entities) are now available, with over 18 thousand active users helping to extend and curate the knowledge graph.[3] Thus Wikidata has become a rich source of structured data that can compliment Wikipedia in non-trivial ways. As part of so called "Phase II" of the Wikidata

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191647

[1] According to https://en.wikipedia.org/wiki/List_of_Wikipedias; retr. 2018/01/27.
[2] We refer the reader to a recent survey by Kaffee et al. [2] on the availability of Wikidata labels in various languages.
[3] Statistics from https://www.wikidata.org/wiki/Wikidata:Statistics; retr. 2018/01/27

project, an important goal is to start to generate Wikipedia info-boxes from Wikidata.[4] In fact, a number of Wikipedia info-boxes are already based on information managed by Wikidata; Figure 1 offers an example of a highly-detailed such info-box. In the source markup of the article is the concise instruction "`{{Infobox telescope}}`", which is all that is needed for the info-box to be generated based on the type of the entity: the entity type is associated with a particular template, which indicates how the structured data available in Wikidata for that entity should be rendered as an info-box on the Wikipedia article, allowing users to manually override particular attributes if required.[5] Close to one thousand info-boxes are presently generated in this manner for English Wikipedia.[6]

Looking in more depth at how editors can generate Wikipedia info-boxes from Wikidata, we note that as per the previous example, such generation is guided by templates associated to a particular type. Indeed, there are currently 367 such info-box templates defined for Wikipedia using Wikidata as a source; 22 of these are for creating complete info-boxes, while the remaining 345 are for filling a value in an existing info-box from a Wikidata attribute.[7] Of the 22 complete templates available, these include a variety of types including `telescope` (see Figure 1), `person`, `sumo wrestler`, `South African town`, etc. As per Figure 1, the info-boxes produced for some entities appear to be of high quality.

However, there are some major obstacles to be overcome with this type-centric template-based approach. First and foremost, a suitable template needs to be defined for each type with the particular attributes and their order hard-coded. While this might be straightforward for an entity such as `telescope` (assuming some defeasibility allowing to override attributes in a particular case), it would seem somewhat more difficult to hardcode attributes for a type such as `person`, where a variety of attributes might be of interest depending on their notability, occupation, when they lived, etc. While there is a generic template for entities of type `person`, this template only covers the most essential information that one might consider applicable. We provide an example of an info-box generated from the generic `person` template from Wikidata in Figure 2, which is notably more sparse than the example provided for the telescope; investigating further, the entity `Samuel Argall` has much more information available in Wikidata (`Q1640499`[8]) than displayed in the info-box, such as `country of citizenship`, `place of birth`, `occupation`, `military branch`, etc. (some of which provide external references). To address this, a variety of other templates have been defined for sub-types of people, such as `sumo wrestlers`, `scientist`, `squash player`, etc.; these then allow for creating more detailed info-boxes for such entities.

Still, only a few of the relevant entity types (22 in total) currently have templates: the creation of such templates for all relevant types would require a huge amount of manual labour and coordination. Aside from this, there are a number of further issues to consider with such a template-based approach. First, a particular attribute (e.g.,

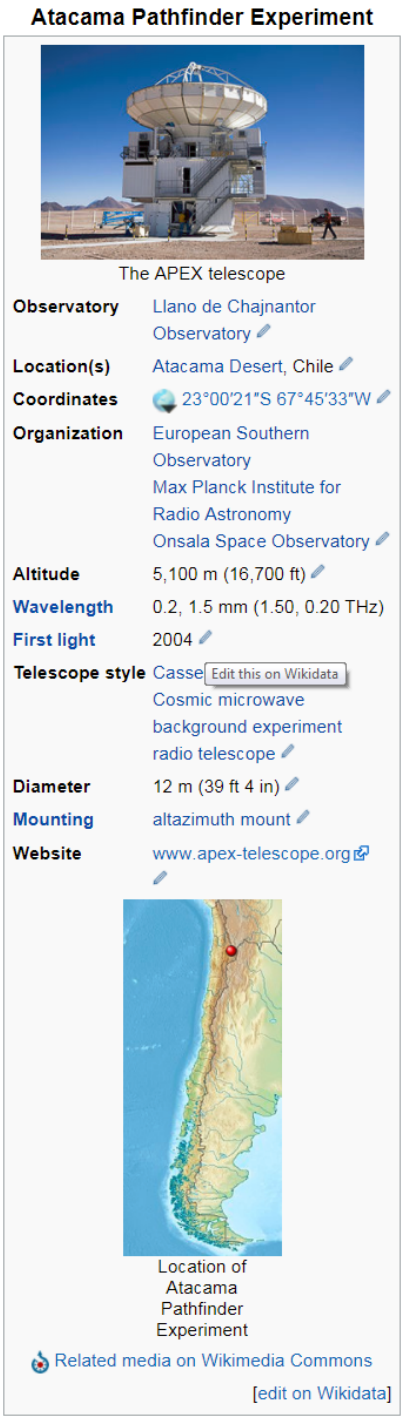

Figure 1: Example of a Wikipedia info-box generated from Wikidata using the `telescope` template; example taken from https://en.wikipedia.org/wiki/Atacama_Pathfinder_Experiment under CC-BY-SA 3.0.

[4] See https://www.wikidata.org/wiki/Wikidata:WikiProject_Infoboxes; retr. 2018/01/27.
[5] See https://en.wikipedia.org/wiki/Template:Infobox_telescope; retr. 2018/01/27.
[6] A list is available at https://en.wikipedia.org/wiki/Category:Articles_with_infoboxes_completely_from_Wikidata; retr. 2018/01/27
[7] These are listed at https://en.wikipedia.org/wiki/Category:Templates_using_data_from_Wikidata; retr. 2018/01/27. The complete info-box templates are named `Template:Infobox xxx`.
[8] See https://www.wikidata.org/wiki/Q1640499; retr. 2018/01/27.

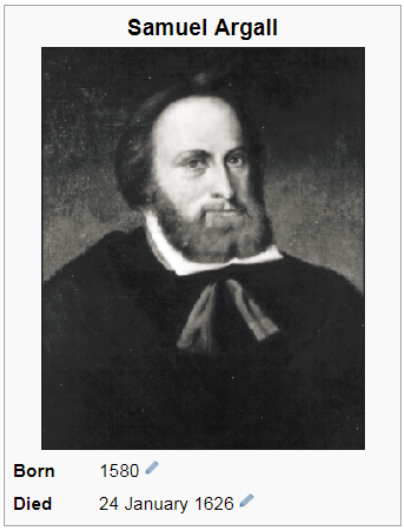

Figure 2: Example of a Wikipedia info-box generated from Wikidata using the `person` template; example taken from https://en.wikipedia.org/wiki/Samuel_Argall under CC-BY-SA 3.0.

award, *occupation*, etc.) may take multiple values, where only some may be sufficiently notable to warrant placement in the info-box. Second, the appropriate attributes to display may sometimes depend on the entity in question, not just the type of entity; for example, there may be cases where an entity crosses types (e.g., a sumo wrestler that is also a notable scientist); furthermore, for example, an attribute such as *sibling* may not be considered noteworthy for a particular type of entity, but when the value for that attribute is a person as famous as *Barack Obama*, it may warrant inclusion.

Along these lines, we investigate fully-automatic techniques by which info-boxes in a specified language can be generated for a particular entity without any information other than provided by Wikidata, meaning no manually-specified templates, no assumptions of existing training data for info-boxes, etc. Our hypothesis is that we can derive statistics from the structure of Wikidata itself that can, in a fully generic manner, be used to prioritise the attribute–value pairs for an entity in Wikidata that are likely to be interesting/relevant to a user in the context of a Wikipedia info-box.

We see this as preliminary research, and may perhaps compliment the existing mechanisms by which info-boxes are generated from Wikidata. In particular, our proposed method could be used as a type-agnostic default in the case that a suitable template is not available (in a given language); a user would simply have to add a generic command `"{{Infobox Wikidata}}"` and the article would be populated with a fully-automatic info-box based on the language of the current article. Furthermore, our method may be useful to prioritise values for a given attribute in an existing template where ranks are not explicitly provided.

2 RELATED WORK

Even aside from the internal efforts to generate info-boxes from Wikidata using templates (as discussed previously), we are not the first work to consider automatically generating or enriching Wikipedia info-boxes with (semi-)structured information.

A number of approaches have been proposed to extract structured information from Wikipedia in order to generate or otherwise enhance info-boxes. These include systems – such as Kylin [14], iPopulator [4], Wikitology [10] – that analyse Wikipedia text looking for sentences from which to extract a value for a particular attribute. Other works have looked at using information extraction techniques over sources external to Wikipedia to generate facts that can be used to improve info-boxes [11, 13]. While these works propose methods to automatically extract information relevant to info-boxes from various sources, we instead assume Wikidata as a source of information and focus primarily on the problem of ranking attributes and values when generating an info-box.

Like us, other works have rather proposed to use existing sources of structured information to enhance Wikipedia. Yus et al [15] propose the Infoboxer system, which uses DBpedia [5] to help users create info-boxes by suggesting popular (missing) attributes and validating the range of attributes (e.g., checking that the value for *place of birth* is a *location*); though some ideas overlap with this work, their focus is rather on semi-automated info-box generation. Kaffee [3] proposes a method to automatically generate placeholders for Wikipedia articles based on Wikidata statements; however, in her approach it is proposed that Wikipedia administrators will manually generate an appropriate ordering of attributes for display, whereas a main objective of this work is to develop and evaluate automated ranking schemes for presenting information in info-boxes. We see our work – more focused on ranking of information – as being complementary to these previous proposals.

3 PROTOTYPE FOR GENERATING INFO-BOXES FROM WIKIDATA

We have created a prototype service for generating info-boxes from Wikidata entities that takes as input: (1) a particular `Q` code identifying a Wikidata entity, and (2) a language code. From this information, the service creates an info-box for that entity in that language, prioritising the attribute–value pairs that it deems most relevant for that entity based on a particular ranking methodology (using statistics compiled offline from a Wikidata dump). In particular, given a `Q`-code and a language, the steps are as follows:

(1) The `Q`-code and language are filled into a SPARQL query template that will retrieve all attribute–value pairs associated with that entity from the Wikidata Query Service.[9] The query template is given by Listing 1: it retrieves not only the properties and values for the given entity, but also their primary labels (given by `rdfs:label`) in the given language; it retrieves no further information (in this initial prototype, no consideration is given to qualifiers or references for example). Values that are datatypes (e.g., dates, numbers, etc.) will not have a label associated with them, but the Wikidata Label Service will leave this value `UNBOUND`, allowing the info-box service to select the label where available; otherwise the value itself is used. Special consideration must be given to retrieve the labels of the given property, which involves some indirection as shown in the query.

[9] https://query.wikidata.org/

Listing 1: SPARQL Query to retrieve meta-data from Wikidata for entity with ID 'yy' for language-tag 'xx'

```
PREFIX rdfs:     <http://www.w3.org/2000/01/rdf-schema#>
PREFIX wikibase: <http://wikiba.se/ontology#>
PREFIX wd:       <http://www.wikidata.org/entity/>
PREFIX bd:       <http://www.bigdata.com/rdf#>
PREFIX wdt:      <http://www.wikidata.org/prop/direct/>

SELECT ?pLabel ?prop ?val ?valLabel
WHERE {
 wd:Qyy ?prop ?val .
 ?ps wikibase:directClaim ?prop .
 ?ps rdfs:label ?pLabel .
 SERVICE wikibase:label { bd:serviceParam wikibase:language 'xx'. }
 FILTER((LANG(?pLabel)) = 'xx' && (?prop != wdt:P18))
}
```

(2) Once all attribute–value pairs have been retrieved, the primary label of the entity (`rdfs:label`) will be used to generate the title of the info-box, while the property `P18` (`image`) will be used to display an image of the entity, if available (in case that multiple images are available, we currently choose one arbitrarily). Thereafter, we are left to decide which of the remaining attribute–value pairs to display, and in what order. This prioritisation of information – based on statistics that are compiled from Wikidata offline – is the focus of this preliminary research, and will be described presently.
(3) With all attribute–value pairs prioritised, the info-box can be previewed in HTML or output as Wikicode.

We provide an online demo for generating info-boxes provided an appropriate Wikidata `Q` code.[10] We highlight that this is a prototype of a system; a (hypothetically) deployed version would rather be tightly integrated with Wikipedia, where given a command `"{{Infobox Wikidata}}"`, a deployed version would automatically detect the language version of Wikipedia, find the correct `Q` code in Wikidata, use the service to compile the info-box, and then display it accordingly in the final HTML page.

More generally, the focus of this research is on point (2) above: prioritising the attribute–value pairs returned, selecting those to present in the generated info-box and applying an appropriate ordering. We now discuss this process in more detail.

4 RANKING ATTRIBUTE–VALUE PAIRS

Info-boxes are intended to be a succinct summary of structured data available about a particular entity, where information is prioritised by its potential relevance to a user, starting with the most important information first. In designing an automatic info-box generation service from Wikidata, an important and non-trivial aspect is then deciding which attribute–value pairs are the most important to display in the info-box. This is necessary to decide which pairs to display (in some cases Wikidata offers more information than required for a concise info-box) and in what order. To apply such a ranking, we require information beyond that provided by the query in Listing 1. Hence we will take a dump of Wikidata and extract some (rather straightforward) statistics from it.

Before we describe these statistics, we give some very brief preliminaries. We currently consider the "truthy version" of Wikidata, which gives direct triples of the form (s, p, o) without qualifiers, references, etc.; this version selects the best non-deprecated value amongst competing values, which will, for example, include the most recent population reading for a city.[11] An example triple might be (`wd:Q42`, `wdt:P19`, `wd:Q350`), where `wd:Q42` refers to *Douglas Adams*, `wdt:P19` refers to *place of birth*, and `wd:Q350` refers to *Cambridge*. Another example is (`wd:Q42`, `wdt:P570`, `11 May 2001`), where `wdt:P570` refers to *date of death* and the object value of the triple is a datatype: a date. We then consider a Wikidata dump as forming a graph G comprised of a set of triples (as can be represented, for example, in RDF).

First we consider the relative importance of the attributes (e.g., `born`, `died`, `observatory`, `altitude`, etc.) independently of the particular value. A straightforward idea is to consider the FREQUENCY of the attribute as being an interesting metric to determine its importance, with the intuition that more frequently used attributes are more important. For example, we may consider that an attribute `country of citizenship` might be used more often in the data than `blood type`, and hence the former attribute should be prioritised over the latter. More specifically, we can define the frequency of an attribute p from a graph G as simply:

$$\text{freq}(p, G) = |\{(s, o) : (s, p, o) \in G\}|$$

In other words, the frequency of an attribute is simply the number of triples in the graph where that attribute is defined.

Though easy to compute, this frequency measure has some limitations. Firstly, the frequency of an attribute may not correlate well with its relevance to a user; this will have to be validated empirically through, e.g., a user evaluation. Secondly, such a measure can only rank attributes, and will not help us to rank the values associated with a given attribute. This can be problematic for entities that have a lot of values defined for a given attribute. For example, Barack Obama is stated in Wikidata as having won 11 awards, all of which are associated with the same "normal rank" by editors (each being an equally valid value); however, these awards are not of equal prominence or fame, and range from the `Nobel Peace Prize` to `Order of Sikatuna` (a diplomatic merit known in the Philippines). In the info-box, we might like to present only the most important values for this property, ordered by said importance; however, the frequency measure is too coarse-grained for such a feature.

Hence the second measure we consider is PAGERANK [8], which is a popular centrality-based measure used to estimate the importance of nodes in a graph. This measure was originally defined for directed graphs, where we construct such a graph from the Wikidata dump where we consider each triple (s, p, o) as a directed edge $s \to o$, thus not offering any special consideration to the edge-label and not considering triples where o is a datatype value (such as a date,

[10] https://s3-us-west-2.amazonaws.com/infobox-coloro/index.html

[11] It is important to note that these ranks do not represent the importance of a particular value for a multi-valued attribute, but rather are intended to represent a preference amongst competing values, such as selecting a more recent population reading, or the current mayor; etc. Such ranks would not be used, for example, to rank the awards won by Barack Obama since all values are equally valid (though not equally prominent).

RANDOM

Douglas Adams	
Description:	English writer and humorist
People Australia ID:	847711
Munzinger IBA:	00000020676
spouse:	Jane Belson
AlloCiné person ID:	97049
educated at:	St John's College
date of birth:	1952-03-11T00:00:00Z
Runeberg author ID:	adamsdou
openMLOL author ID:	140290
native language:	British English
PORT person ID:	208947
place of birth:	Cambridge
UNZ author identifier:	AdamsDouglas
IMDb ID:	nm0010930
National Library of Israel ID:	000163846
occupation:	comedian, novelist
Goodreads author ID:	4
child:	Polly Jane Rocket Adams
NDL Auth ID:	00430962
topic's main category:	Category:Douglas Adams
manner of death:	natural causes
Discogs artist ID:	134923
NKCR AUT ID:	jn19990000029
Google Doodle:	douglas-adams-61st-birthday
Freebase ID:	/m/0282x

FREQUENCY

Douglas Adams	
Description:	English writer and humorist
instance of:	human
sex or gender:	male
occupation:	screenwriter, playwright, comedian, dramaturge, children's writer, novelist, science fiction writer
date of birth:	1952-03-11T00:00:00Z
given name:	Douglas
country of citizenship:	United Kingdom
Commons category:	Douglas Adams
place of birth:	Cambridge
date of death:	2001-05-11T00:00:00Z
Freebase ID:	/m/0282x
VIAF ID:	113230702
official website:	http://douglasadams.com/
place of death:	Santa Barbara
genre:	science fiction, comedy, satire
languages spoken, written or signed:	English, British English
educated at:	University of Cambridge

PAGERANK

Douglas Adams	
Description:	English writer and humorist
instance of:	human
sex or gender:	male
country of citizenship:	United Kingdom
native language:	English, British English
languages spoken, written or signed:	English, British English
residence:	London
religion:	atheism
occupation:	screenwriter, novelist, playwright, science fiction writer, children's writer, comedian, dramaturge
manner of death:	natural causes
instrument:	guitar
place of birth:	Cambridge
genre:	comedy, science fiction
cause of death:	myocardial infarction
given name:	Douglas
educated at:	University of Cambridge
employer:	BBC

COMBINED^{+}

Douglas Adams	
Description:	English writer and humorist
instance of:	human
sex or gender:	male
country of citizenship:	United Kingdom
languages spoken, written or signed:	English, British English
native language:	English
occupation:	screenwriter, novelist, playwright, science fiction writer, children's writer, comedian, dramaturge
residence:	London
date of birth:	1952-03-11T00:00:00Z
given name:	Douglas
place of birth:	Cambridge
Commons category:	Douglas Adams
date of death:	2001-05-11T00:00:00Z
Freebase ID:	/m/0282x
religion:	atheism
VIAF ID:	113230702
manner of death:	natural causes
genre:	comedy, science fiction

COMBINED$^{\times}$

Douglas Adams	
Description:	English writer and humorist
instance of:	human
sex or gender:	male
country of citizenship:	United Kingdom
languages spoken, written or signed:	English, British English
occupation:	screenwriter, novelist, playwright, science fiction writer, children's writer, comedian, dramaturge
native language:	English, British English
place of birth:	Cambridge
given name:	Douglas
residence:	London
genre:	comedy, science fiction, satire
religion:	atheism
educated at:	University of Cambridge
manner of death:	natural causes
place of death:	Santa Barbara
instrument:	guitar

Figure 3: Info-boxes generated under our five ranking strategies for Douglas Adams (Q42); each info-box contains 25 attribute–value pairs (not counting the first hard-coded Description pair); we highlight that the PAGERANK and COMBINED$^{\times}$ info-boxes contain no datatype values since PageRank is set to 0 for such values.

number, etc.).[12] Upon applying the PageRank algorithm over this graph, we derive a score for each s and o value in the graph.

Unlike the frequency measure which applies to Wikidata properties (with `P*` identifiers, such as `award received`), PageRank thus rather applies to Wikidata entities (with `Q*` identifiers, such as `Barack Obama`, `Nobel Peace Prize`, `Order of Sikatuna`, etc.); hence frequency is a measure that can be used to rank attributes, while PageRank can be used to rank values. As such, the frequency and PageRank measures can be considered complementary and can be combined to rank attribute–value pairs. Thus for a given pair (p, o) and a graph G, we consider two straightforward ways to combine the frequency score of p and the PageRank score of o, the first based on a summation/mean of the terms:

$$\text{rank}^{+}(p, o, G) = \frac{\text{norm}(\text{freq}(p, G)) + \text{norm}(\text{prank}(o, G))}{2}$$

and the second based on a product of the terms:

$$\text{rank}^{\times}(p, o, G) = \text{norm}(\text{freq}(p, G)) \times \text{norm}(\text{prank}(o, G))$$

where $\text{prank}(o, G)$ denotes the PageRank score of o in the graph G, and norm defines a normalisation function that linearly maps the

[12] One may question whether or not direction plays an important role in the Wikidata graph since, for example, one can define a triple (s, child, o) equivalently as (o, parent, s) with an inverse edge label. However, we argue that direction does play an important role in Wikidata since the out-degree of nodes tends to be bounded, whereas in-degree is not; for example, citizens link to their countries but countries do not link to their (potentially too numerous) citizens. Furthermore, in the original formulation of PageRank for assigning importance to web-pages based on links, we note that Wikidata offers hyperlinks from the webpages of s to o, but not vice-versa.

values into an interval $[0, 1]$, with 0 denoting the minimum value of the measure for all p (in the case of frequency) or $s|o$ (in the case of PageRank), and 1 denoting the analogous maximum value.[13]

The intuition of these combinations is that the summation measure should act as a form of "disjunction", where an attribute–value pair (p, o) can get a high rank from either having a high p rank, or a high o rank (but in the best case, both ranks are high). On the other hand, the product measure acts as a form of "conjunction", where an attribute–value pair (p, o) must have a high rank for both values to have a high overall rank. For example, if $\text{norm}(\text{freq}(p, G)) = 1$ but $\text{norm}(\text{prank}(o, G)) = 0$, then $\text{rank}^{+}(p, o, G) = \frac{1}{2}$, whereas $\text{rank}^{\times}(p, o, G) = 0$. It is important to highlight that since we do not have PageRank scores for datatype values[14], the rankings of pairs (p, o) where o is a datatype value will be at most $\frac{1}{2}$ for the summation and 0 for the product.

The frequency and PageRank measures are computed off-line and loaded into a local index within the info-box generation service. Such values can be updated periodically (though in general we would not consider these values to be sensitive to short-term changes relatively speaking, in terms of overall orderings).

5 USER EVALUATION

In order to evaluate the relative quality of the info-boxes generated by the different measures, we performed an initial user evaluation where users were presented the info-boxes generated for various entities by the following ranking variants:

Random baseline (RAND) Attribute–value pairs are ordered randomly; this strategy is intended as a baseline.

Attribute frequency (FREQ) Attribute–value pairs are ordered by the frequency of their attributes (within each attribute, values are ordered alphabetically).

Value PageRank (PR) Attribute–value pairs are ordered by their value's PageRank.

Combined (+) (COM^{+}) Attribute–value pairs are ordered by the mean of the attribute's frequency and the value's PageRank (rank^{+}).

Combined (×) (COM$^{\times}$) Attribute–value pairs are ordered by the product of the attribute's frequency and the value's PageRank ($\text{rank}^{\times}$).

To avoid creating overly-long info-boxes for users to review, we selected a threshold of 25 attribute–value pairs to display in each info-box.[15] For a given entity, the info-box is then constructed with the primary label of the entity used as the title; for the purposes of the evaluation, we do not display images as they should not vary across the different orderings. The top-25 attribute–value pairs are then grouped by attribute: the attributes are ordered based on their top ranked value, and values within each attribute are then listed according to the order of their corresponding pair.

[13]This normalisation process is necessary for the summation version since the maximum PageRank value is less than zero (being based on a probabilistic measure), while the maximum frequency value is in the millions.

[14]Technically there is no problem ranking such values as any other node in the graph, but the results would not make much sense: for example, the PageRank scores of two boolean values would be incomparable with an arbitrary number of date values.

[15]We selected this threshold based on an informal survey of the number of such values in the info-boxes of featured articles in Wikipedia.

We select 15 entities for evaluation based on the most common types on Wikidata; these entities are as follows:

- **3 People**
  - `Douglas Adams` [`Q42`]
  - `Michelle Bachelet` [`Q320`]
  - `David Lynch` [`Q2071`]
- **2 Countries**
  - `Chile` [`Q298`]
  - `Zimbabwe` [`Q954`]
- **2 Chemical elements**
  - `Gold` [`Q879`]
  - `Water` [`Q283`]
- **2 Species**
  - `Dog` [`Q144`]
  - `Platypus` [`Q15343`]
- **2 Intellectual works**
  - `The Bible` [`Q1845`]
  - `12 Angry Men` [`Q2345`]
- **2 Astronomic bodies**
  - `Mars` [`Q111`]
  - `Betelgeuse` [`Q12124`]
- **2 Buildings**
  - `Eiffel Tower` [`Q243`]
  - `Guggenheim Museum` [`Q179199`]

We constructed five info-boxes for each entity, corresponding to the five ranking strategies previously outlined. Figure 3 provides an example of the five info-boxes generated for Douglas Adams (Q42) under each of the investigated strategies.

To evaluate the generated info-boxes, we gather 12 evaluators (mostly students of a Semantic Web course). Since these evaluators were all native Spanish speakers, the info-boxes were generated in Spanish. For each entity, the five info-boxes were printed on a sheet side-by-side. No explicit indication was given as to which info-box corresponded to which ranking and for different entities, the order of presentation of the info-boxes was randomised to avoid, e.g., a first or last info-box evaluation bias.

Each evaluator was provided a form where they were instructed to provide an identifier for the sheet (entity) and a score on a 7-level Likert scale from 1 (very poor) to 7 (very good) for each of the five info-boxes. Before the marking began, the evaluators were briefly instructed on the criteria for evaluation: that the presented options should be evaluated as potential info-boxes for Wikipedia, that they should look out for the relevance of presented data, and that they should consider the ordering of attributes and also values for each attribute in their assessment of the info-boxes.

6 RESULTS

The evaluators were given 20 minutes to provide their scores. A total of 145 complete evaluations (each evaluation giving five scores for each strategy) and 1 incomplete evaluation were collected in this time. The mean inter-rater standard deviation was ~1.23 (from a rating scheme with an interval of 6) for corresponding evaluations.

In Figure 4, we present the overall results for each ranking measure taking the mean of all evaluations; the error bars here denote the standard deviation. All four strategies outperform the baseline

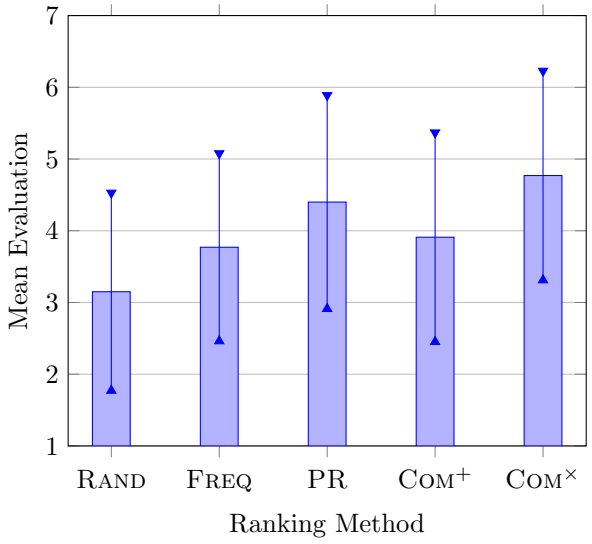

Figure 4: Overall mean evaluation per ranking measure; error bars indicate standard deviation

with statistical significance ($p < 1.3 \times 10^{-5}$).[16] In fact, the results for all pairs of methods are significantly different from each other ($p < 7.8 \times 10^{-5}$), except Freq and Com^{+} ($p = 0.25$). The measures based on PageRank-generated info-boxes were, on average, evaluated better than their counterparts. The best measure overall was the combined measure using multiplication. This result was quite surprising: it indicates that (a) the ranking of values is considered more important than the ranking of attributes, and (b) that the evaluators did not put a strong emphasis on the presence of datatype values in the info-boxes. Rather the evaluators valued the presence of important values in the info-box (PageRank) more than common attributes (frequency); furthermore, property–value pairs with a high attribute frequency and high PageRank value (Com$^{\times}$) had better evaluations than just considering PageRank (PR).

In Figure 5, we provide more detailed results with mean evaluations per entity and ranking measure. Though in general this plot reveals the same trends when comparing the ranking measures (for example, Com$^{\times}$ performs best on average for almost all entities), we can now see that the absolute evaluations vary quite noticeably across different entities, and indeed, different types of entities. In particular, info-boxes describing people were highly rated, while entities relating to chemical compounds or species were much lower rated on average. This perhaps suggests that the quality of info-box generated is sensitive to the type of entity; as a general trend, we can observe that more specific types of entities (i.e., people, works, buildings) tended to be evaluated better than very generic types, possibly because it is unclear what would be the important attributes for something as generic as `Water`.

[16] Statistical significance results are based on a paired t-test over 145 complete responses; all p-values are two-tailed.

7 CONCLUSIONS

In this paper, we have investigated a fully automatic method for generating Wikipedia info-boxes from corresponding Wikidata descriptions. The method can be applied to generate info-boxes for the languages supported by Wikidata and without requiring any specific manual input, such as type-specific templates. The core of the method relates to the prioritisation of property–value pairs taken from Wikidata, where we looked at four core measures: the first based on the frequency of an attribute, the second based on the PageRank of the value, the third based on the average of both, and the fourth based on a product of both. We argued that frequency and PageRank should be considered complimentary since one is useful to rank attributes while the other is useful to rank values. We then conducted an initial user evaluation comparing the info-boxes generated by these four methods and a random baseline for a selection of 15 entities from 7 popular entity types. The results showed that although all ranking schemes outperformed the random baseline, users put a higher emphasis on the PageRank of the values when evaluating the quality of an info-box than on the frequency of the attribute. Furthermore, evaluations varied across different types of entities, where we saw an initial trend that users tended to more strictly evaluate "general" entities such as `Water`, `Dog`, etc., when compared with specific persons, works, etc.

The results presented herein should be considered preliminary; indeed, we have not compared, for example, the results obtained by template-based methods.[17] In general, we see these fully automated methods as a possible way to complement the existing type-specific template approach currently used to generate info-boxes from Wikidata in situations, for example, where such templates are not available. However, the trend of users preferring info-box rankings based on PageRank scores for values is potentially quite important since the current template-based approach only considers an ordering of attributes; the results here suggest that considering a ranking of values is also important, helping to select important values for multi-valued attributes such as *award received*, or helping to boost the rank of an attribute–value pair when the value is a prominent one, such as *sibling: Barack Obama*.

The measures we present here explore initial ideas on how such a ranking could be applied for automatic info-box generation from Wikidata. A benefit of the proposed approach is that it makes minimal assumptions and is applicable to entities of any type without requiring training sets or other manual inputs. But there are other possible directions that could be explored, such as to use machine learning methods to identify important properties for specific types of entities [9], or perhaps rather relying on semi-automated methods as proposed by Yus et al [15] leveraging the DBpedia dataset.

Another crucial aspect to be considered is that of the community of editors involved with Wikipedia: how they would perceive such a tool and how it could be made more usable for them. For example, editors may wish to change some of the automated attribute–value pairs, or to restrict information to that for which references exist. Another open issue is the provision of links, images and other complex values. A final limitation is that the info-boxes automatically generated by our methods, particularly those based on PageRank,

[17] It is not entirely clear how this could be done fairly given the varying level of detail present in such info-boxes, as evidenced by Figures 1 and 2.

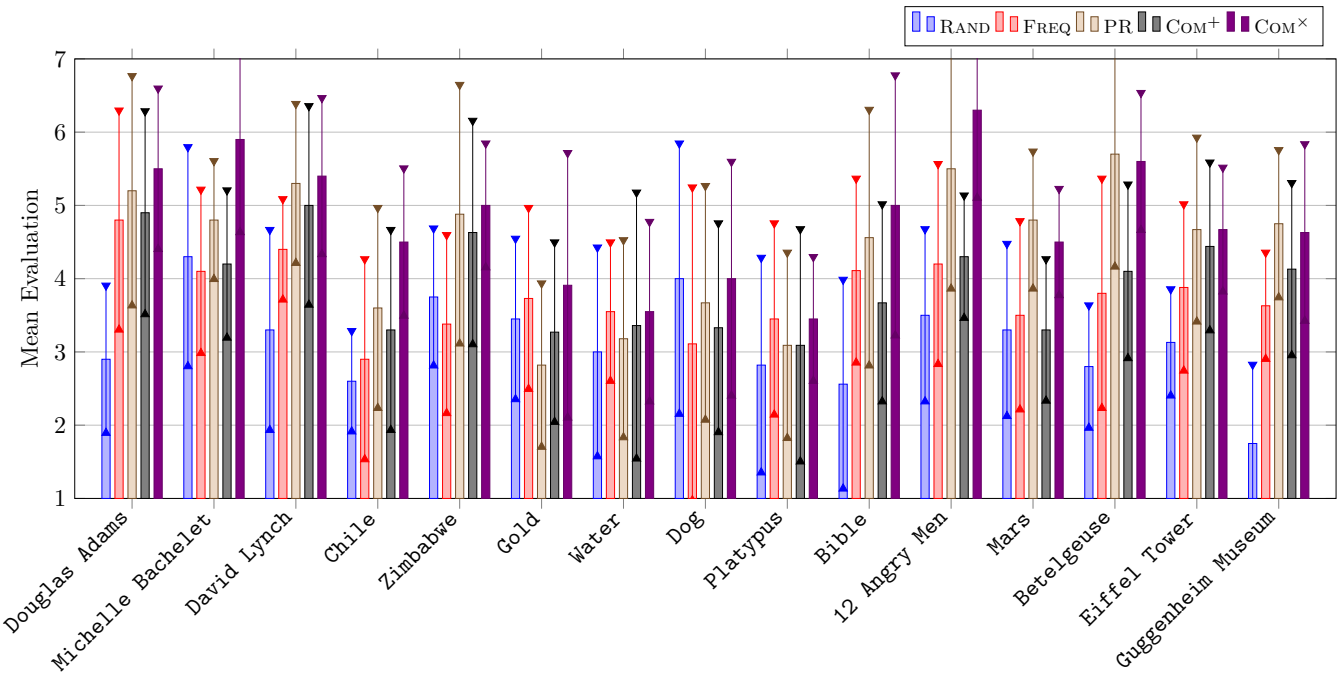

Figure 5: Detailed mean results per entity and ranking measure; error bars indicate standard deviation

may not follow a consistent style for entities of similar types; a potential future direction may thus be to consider class-specific rankings of attributes and/or values.

In any case, we believe that more research on the generation of info-boxes could obviate the need for type-specific templates and could accelerate Wikidata's impact on the structured-data views of Wikipedia across several languages. Furthermore, our results suggest that contrary to many of the currently-proposed methods, the importance of values – not just attributes – should play an important role when populating an info-box.

Acknowledgements. *This work was supported by the Millennium Institute for the Foundations of Data and by Fondecyt Grant No. 1181896. We would like to thank all of the students that participated in our study for their contribution.*

REFERENCES

[1] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. 2013. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artif. Intell.* 194 (2013), 28–61. https://doi.org/10.1016/j.artint.2012.06.001

[2] Lucie-Aimée Kaffee, Alessandro Piscopo, Pavlos Vougiouklis, Elena Simperl, Leslie Carr, and Lydia Pintscher. 2017. A Glimpse into Babel: An Analysis of Multilinguality in Wikidata. In *International Symposium on Open Collaboration (OpenSym)*. 14:1–14:5. https://doi.org/10.1145/3125433.3125465

[3] Lucie-Aimée Kaffee. 2016. *Generating Article Placeholders from Wikidata for Wikipedia: Increasing Access to Free and Open Knowledge.* Bachelor Thesis. HTW Berlin University of Applied Sciences.

[4] Dustin Lange, Christoph Böhm, and Felix Naumann. 2010. Extracting structured information from Wikipedia articles to populate infoboxes. In *ACM Conference on Information and Knowledge Management (CIKM)*. 1661–1664. https://doi.org/10.1145/1871437.1871698

[5] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web Journal* 6, 2 (2015), 167–195.

[6] Wlodzimierz Lewoniewski. 2017. Completeness and Reliability of Wikipedia Infoboxes in Various Languages. In *Business Information Systems Workshops (BIS)*. 295–305. https://doi.org/10.1007/978-3-319-69023-0_25

[7] Wlodzimierz Lewoniewski, Krzysztof Wecel, and Witold Abramowicz. 2017. Relative Quality and Popularity Evaluation of Multilingual Wikipedia Articles. *Informatics* 4, 4 (2017), 43. https://doi.org/10.3390/informatics4040043

[8] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 1999-66. Stanford InfoLab.

[9] Simon Razniewski, Vevake Balaraman, and Werner Nutt. 2017. Doctoral Advisor or Medical Condition: Towards Entity-Specific Rankings of Knowledge Base Properties. In *Advanced Data Mining and Applications*. 526–540. https://doi.org/10.1007/978-3-319-69179-4_37

[10] Zareen Saba Syed and Tim Finin. 2010. Approaches for Automatically Enriching Wikipedia. In *Collaboratively-Built Knowledge Sources and Artificial Intelligence (AAAI Workshop)*.

[11] Thong Tran and Tru H. Cao. 2013. Automatic Detection of Outdated Information in Wikipedia Infoboxes. *Research in Computing Science* 70 (2013), 211–222.

[12] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.

[13] Fei Wu, Raphael Hoffmann, and Daniel S. Weld. 2008. Information extraction from Wikipedia: moving down the long tail. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 731–739. https://doi.org/10.1145/1401890.1401978

[14] Fei Wu and Daniel S. Weld. 2007. Autonomously semantifying Wikipedia. In *ACM Conference on Information and Knowledge Management (CIKM)*. 41–50. https://doi.org/10.1145/1321440.1321449

[15] Roberto Yus, Varish Mulwad, Tim Finin, and Eduardo Mena. 2014. Infoboxer: Using Statistical and Semantic Knowledge to Help Create Wikipedia Infoboxes. In *ISWC 2014 Posters & Demonstrations Track*. 405–408. http://ceur-ws.org/Vol-1272/paper_123.pdf