Chaudron: Extending DBpedia with measurement

Julien Subercaze

Chaudron: Extending DBpedia with measurement

Julien Subercaze[1]

Univ Lyon, UJM-Saint-Etienne, CNRS
Laboratoire Hubert Curien
UMR 5516, F-42023, SAINT-ETIENNE, France
julien.subercaze@univ-st-etienne.fr

Abstract. Wikipedia is the largest collaborative encyclopedia and is used as the source for DBpedia, a central dataset of the LOD cloud. Wikipedia contains numerous numerical measures on the entities it describes, as per the general character of the data it encompasses. The DBpedia Information Extraction Framework transforms semi-structured data from Wikipedia into structured RDF. However this extraction framework offers a limited support to handle measurement in Wikipedia.
In this paper, we describe the automated process that enables the creation of the Chaudron dataset. We propose an alternative extraction to the traditional mapping creation from Wikipedia dump, by also using the rendered HTML to avoid the template transclusion issue.
This dataset extends DBpedia with more than 3.9 million triples and 949.000 measurements on every domain covered by DBpedia. We define a multi-level approach powered by a formal grammar that proves very robust on the extraction of measurement. An extensive evaluation against DBpedia and Wikidata shows that our approach largely surpasses its competitors for measurement extraction on Wikipedia Infoboxes. Chaudron exhibits a F1-score of .89 while DBpedia and Wikidata respectively reach 0.38 and 0.10 on this extraction task.

Keywords: Wikipedia, Extraction,DBpedia, Measurement, RDF, Formal Grammar

1 Introduction

Wikipedia is a free content internet encyclopedia, currently the largest encyclopedia with more than five million articles in its English version; overall 38 milllion articles in over 250 languages. Wikipedia is currently in the top ten of the most viewed websites in the world, with, on average 18 billion page views and nearly 500 million unique visitors per month.

As a central knowledge repository on the Web, Wikipedia serves as a seed for the creation of knowledge bases such as Google's Knowledge Graph or Wolfram Alpha. The Intelligence in Wikipedia project [23] used Wikipedia to derive an ontology [26] and fostered open information extraction [27]. The linked open data movement has long understood the central role of Wikipedia in the web of

knowledge. The DBpedia effort is, since 2007, an open source project that aims at translating entries from Wikipedia into RDF and making it publicly available [2]. Another extraction of Wikipedia into RDF was made in the Yago project [19] using a heuristic based approach [6]. After a tentative to create a semantic backend for Wikipedia [7], the same authors created a new Wikimedia project, entitled Wikidata [21,22], that aims at structuring Wikipedia data in order make them machine readable.

Triples in the DBpedia dataset are mainly extracted from Wikipedia dumps through the DBpedia Information Extraction Framework (DIEF for the rest of the paper). This framework uses manually defined mappings to transform Wikipedia semi-structured content – called Infoboxes – into structured content. These mappings encompass, among others, the structuration of physical measurements intro RDF triples. A measurement is the assignment of a number to a property of an object or an event. For instance, the fact that Lionel Messi is 1.70 meter tall is written as follows in DBpedia:

```
@prefix dbr: <http://dbpedia.org/resource/> .
@prefix dbo: <http://dbpedia.org/ontology/> .
@prefix dbd: <http://dbpedia.org/datatype/> .

dbr:Lionel_Messi dbo:Person/Height  "170"^^dbd:centimeter .
dbr:Lionel_Messi dbo:Height  "1.70"^^xsd:double .
```

The DIEF, while providing one the most valuable asset to the LOD cloud [18], is not perfect: 40% of the property occurences are unmapped, leaving room for focused automated mappings. Some mappings do not include units [9,24] or some are incorrect[1] and transcluded Infoboxes are not properly managed, especially in chemistry[2] [12].

In this paper, we tackle the issue of structuring measures from Wikipedia Infoboxes into RDF, where the property is a physical quantity. We take into account the above mentioned issues and design a novel approach that aims for robustness. The contributions of this paper are threefold. *First*, we extract and make public Chaudron, a dataset of over 900.000 measurements of the utmost practical interest that complements DBpedia. *Second*, to create this dataset, we devised a novel robust approach based on a formal grammar for units detection. *Third*, we show that processing rendered HTML is a viable approach for extracting triples from Wikipedia.

The paper is organized as follows: Section 2 presents measurements and how measurements are managed in Wikipedia's Infoboxes. Section 3 describes Chaudron's extraction techniques. Section 4 presents the Chaudron dataset and its characteristics. Section 5 describes the dataset availability and its potential applications. The quality evaluation of the dataset is presented in Section 6. Section 7 provides the concluding remarks.

[1] The Bowatenna Dam has a `dbp:plantCapacity` of 40 W instead of 40 MW

[2] See for instance DBpedia resources corresponding to chemical elements, compounds and drugs, e.g. `Iron` and `Nicotine`

2 Measurement

Measurement is, the assignment of a number to a characteristic or event, which can be compared with other objects or events [13]. To facilitate comparisons in various fields, measurement systems have been developed. Altough large progress have been made towards unification, there still exists various measurement systems used to describe similar physical dimensions. The metrication – the process of converting to the metric system of units of measurement, also known as the International System of Units (ISU) – began in France after the revolution and spread across the globe. Nowadays, except the metric system, the US customary units and the burmese units of measurement are heavily used in the United Sates and Burma respectively. Due to the large volume of trade between the United States and Canada, the US customary is also in use in some limited contexts in Canada, mainly agriculture and engineering. Altough United Kingdom officially uses the metric system, imperial units is widespread among the public; commonwealth countries are currently in similar situations. Furthermore, non standard units are also widely used in some special contexts or areas. Computer scientists, for instance, are common users of the *rack unit* – one rack (U) being 44.45 *mm* – that is used to measure rack-mountable computer equipment such as servers or network elements. In conclusion, there exists a large body of custom units, thus they should be catered or converted accordingly.

This diversity of units system, each of them used by large number of people on earth, is reflected in Wikipedia's articles. In its english version, height, length and the weight are usually displayed in both ISU and US customary units. Non-standard units also appear in Wikipedia articles. This leads to the first challenge in this study, i.e., to identify the set of units that are possibly used in Wikipedia Infoboxes. We adress this issue in Section 3.2. We first start with how measurement are described within these Infoboxes.

2.1 Measurement in Wikipedia's Infoboxes

Wikipedia contains a large number of measures, most of them are stored in Infoboxes. An Infobox is a Wikipedia template that is used to represent a summary of the article, as stated in Wikipedia's documentation. It presents a summary of the most relevant facts in form of key-value pairs that are displayed as a table. Units are mainly from the three most common systems of units (ISU, US customary & Imperial) and may include other customary systems. In some rare cases, other unit systems may be used. These cases are sufficiently rare to be ignored. As noted by the authors of DBpedia [10], most of the editors do not follow the recommendations and good practices for formatting. As a consequence, the units chosen for display are not necessarily standards and may belong either to the three main systems, as well as isolated units of measure that are specialized in some domains. We investigated how measurements are stored and displayed in Wikipedia's Infoboxes. From our experience, they are pigeonholed in three categories:

1) Definition

Templates calls the convert module to convert (ft in) to m

```
{{infobox [...]

| label9   = Listed height

|  data9   = {{#if:{{{height ft|}}}
     |{{convert|{{{height ft|0|}}}|ft|
     {{#if:{{{height in|}}}|{{{height in}}}|0}}
     |in|m|2|abbr=on|order={{{height order|}}}}}}
     {{{height footnote|}}}
```

2) Instance

Template Infobox basketball biography filled with the values of Kobe Bryant

```
{{Infobox basketball biography
| name          = Kobe Bryant
| image         = Kobe Bryant 2014.jpg
| caption       = Bryant in 2014
| position      = [[Shooting guard]]<!-
| height ft     = 6
| height in     = 6
| weight lb     = 212
| league        =
| team          =
| number        = 8, 24
```

3) Display

Formatted text is displayed with both U.S. customary and SI units.

Personal information	
Born	August 23, 1978 (age 38) Philadelphia, Pennsylvania
Nationality	American
Listed height	6 ft 6 in (1.98 m)[a]
Listed weight	212 lb (96 kg)
Career information	
High school	Lower Merion (Ardmore, Pennsylvania)
NBA draft	1996 / Round: 1 / Pick: 13th overall
Selected by the Charlotte Hornets	
Playing career	1996–2016
Position	Shooting guard
Number	8, 24
Career history	
1996–2016	Los Angeles Lakers

Fig. 1: Example of Infobox that uses the {{Convert}} template to format the height of a person. Here the basketball player Kobe Bryant.

Conversion In this case, units are displayed through the template {{Convert}}. This template allows to display a measurement with two units. It is mainly used to display measurement in both metric and Imperial or US customary systems. The main syntax is the following : {{Convert|val|$unit_{source}$|$unit_{dest}$}}. For instance the following expression {{Convert|1|lb|kg}} gives "1 pound

(0.45 kg)". The {{Convert}} template supports a closed list of units suitable for conversion.

Formatted display The formatted display is used when one does not require a unit conversion. The {{Val}} template formats a measure with units to a readable form. It is also used to display uncertainty in the measurement. The main syntax is the following: {{Val|number|ul=unit code}}. Similarly to the {{Convert}} template, {{Val}} supports a predefined list of units[3], however it also supports arbitrary units that may be customised by the editor.

Free riding When none of the above described templates are used, editors use custom formatting to describe the measurements. This is the less formatted case of all three. Whereas common practice lead to output that are formatted similarly to {{Val}}, edge cases happen where the output does not resemble to any given format.

Some Infoboxes automatically format certain values, i.e., when an editor fills up the template instance with a key-value pair, the template will process the value with one of the two above described templates or formats it along a predefined pattern. Figure 1 depicts this process: the height is processed by the {{Convert}} templates. Other Infoboxes do not support automatic formatting, this is either due to the impractibility of using a single prefix – range of values maybe too large – or due to the lazyness of the Infobox template editors.

3 Measures Extraction

To extract measures from Infobox, we are facing the following problem: we must determine whether a key/value pair is of the form: <physicalQuantity / numericalValue Unit>. Our approach is divided in three parts. In the first part, we filter the key/value pairs to discard irrelevant entries. This process is described in Section 3.1. In the second part, we apply patterns to the values of key-value pairs for the single unit case. This pattern matching approach requires a list of units to be matched. We detail in Section 3.2 how we obtained such a list of units. In the third part, we develop a novel technique based on formal grammar to match complex units: this robust technique is presented in Section 3.3.

3.1 Infobox parsing & Filtering

As stated in the previous section, measures are often displayed by the mean of templates. These latter may be combined with other templates, making the extraction from the markup text a complex procedure. The DBpedia Information Extraction Framework extracts triples directly from the markup. There also exist very convenient tools to programatically access Wikipedia raw data, as well as its edit history [3].

[3] https://en.wikipedia.org/wiki/Template:Val/list

For our process, such an approach based on wiki markup is not suitable. Templates, such as {{Convert}} or {{Val}}, are complex pieces of code from different languages (Lua, PHP) and are subject to changes. Infoboxes may be aliased (e.g. {{chembox}},{{drugbox}}), or transcluded[4] as for instance chemical elements whose Infoboxes are templates from a template [5]. Therefore trying to reproduce the code stack from Mediawiki[6] in another language is overly complex. This is why the DIEF is unable to extract triples from the Infoboxes of chemical elements. To give an insight into this complexity, the Infobox settlement, one of the most used Infobox in Wikipedia, transcludes more than 50 different templates and more than 40 modules. Moreover, these modules and templates may also themselves transclude other modules and templates.

We chose to parse HTML generated code using a local copy of Wikipedia. This method, while less computationally performant, is more robust. To reduce the performance overhead, we first filter the content from the Wikipedia dump using a custom MWDumper[7] filter. Thus, for a page to pass the filter, its Infobox must either contains digits and units from the list (see Sec 3.2) or the template of this Infobox must contains units or refer the templates for conversion or formatted display. The filter only retains the Infobox text and discards the rest of the page. This allows us to drastically reduce the size and the number of entries on our local Wikipedia copy, i.e., from 16 million entries in the original dump to 1.4 million entries. The size of our local database dump is 2GB against 50GB for its original counterpart. Consequently, this speeds up the overall process as compared to processing the whole set of Wikipedia articles.

While processing HTML from the local Wikipedia pages, a second filter is applied on each key/value pair. This filter is activated before the pair is sent to the units extractor. To avoid useless computation, we verify that the value of the key/value pair contains digits, but not only digits: ensuring that a unit is potentially present after the digits. Once this filter is succesfully passed, we try to match the value against the following pattern: numericalValue Unit, where the list of all unit is thereafter defined.

3.2 List of Units

In order to apply the pattern matching approach, we need a list of units used in Wikipedia. For this purpose, we extracted the pages contained in the category Units of measurement and its subcategories. These pages are not necessarily only units and outliers have to be discarded. To ensure that we retain only units, we determine if the first sentence of the page validates certain textual patterns (e.g the first sentence contains "is a unit of") or contains a link to unit pages (Unit of measurement, SI units, Derived SI units, etc). Not every unit is linked

[4] https://www.mediawiki.org/wiki/Transclusion

[5] See for instance the element Iron and its Infobox which is itself an instance of the Infobox element.

[6] Mediawiki is the software that powers Wikipedia

[7] https://www.mediawiki.org/wiki/Manual:MWDumper

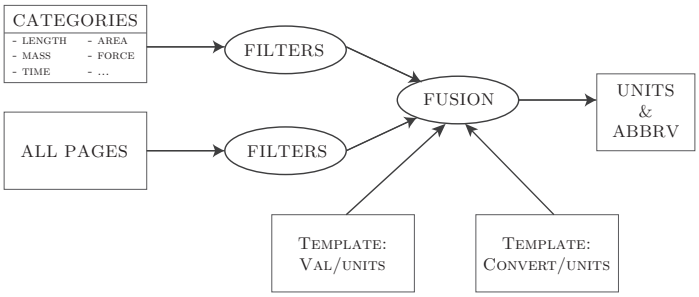


Fig. 2: Extracting the list of units from Wikipedia

to its respective category, we therefore apply more restrictive patterns on every page to extract a second list of units.

In Section 2, we described the usage of the {{Convert}} and {{Val}} templates. Each of these templates has an associated list of units, that we also process to extract our third and fourth lists. Finally the four lists are merged, and in total, 464 atomic units and their abbreviations are extracted. When available, we also retain the URL of the Wikipedia article that describes the unit. Figure 2 summarizes this process and depicts the fusion process from the different sources.

This list of units allows us to apply a simple pattern matching technique to extract measurements. This technique consists of detecting numerical values followed by one of the unit from the list. This approach covers simple cases where the measurement can be expressed using a single physical dimension like height or weight. However it falls short when the quantity measured is described by a combination of several units. For example the molar mass expressed in $g \cdot mol^{-1}$ cannot be extracted using this technique.

3.3 Extraction with formal grammar

The above described approach only covers simple cases but falls short on multiple complex cases, i.e., when the dimension of the physical quantities are given using a combination of units. For instance the molar heat capacity of a substance is given in $kJ/(mol \cdot K)$. The formula that describes the unit of such a measure follows, by its very nature, a formal grammar. Formal grammars precisely define the structure of valid sentences in a language. Formal grammars are one of the common underlying tools of formal languages such as computer programming languages, but are not expressive enough to describe natural language where ambiguity arise. However there exists some approach to describe natural language using formal grammar such as the Attempto Controlled English [5] that can be used for knowledge representation purposes [4].

Herein we develop a formal grammar that validates whether or not a string is a valid formula. A valid a formula is made up of a single unit or of an arith-

metic expression of valid prefix and units couple. The prefixes are the 26 valid prefixes from the ISU. The abbreviations of the seven core units from the ISU as from its derived units[8] are incorporated in the list. Since Wikipedia makes a very large use of conversion to display measurement in both SI and imperial, it is sufficient to extract the data.

$$
\begin{aligned}
\langle \text{UNIT} \rangle &\models \text{m} \mid \text{g} \mid \text{K} \mid \text{A} \mid \text{J} \mid \text{F} \mid \text{V} \mid \ldots \\
\langle \text{PREFIX} \rangle &\models \text{Y} \mid \text{Z} \mid \text{E} \mid \text{P} \mid \text{T} \mid \text{G} \mid \text{M} \mid \ldots \\
\langle \text{OP} \rangle &\models \cdot \mid / \\
\langle \text{SIGN} \rangle &\models + \mid - \\
\langle \text{NUM} \rangle &\models (1 \ldots 9)\texttt{+} \\
\langle \text{EXP} \rangle &\models \langle \text{EXPT} \rangle \, (\langle \text{OP} \rangle \, \langle \text{EXPT} \rangle)\texttt{*} \\
\langle \text{EXPT} \rangle &\models \langle \text{PREFIX} \rangle \texttt{?} \, \langle \text{UNIT} \rangle \, (\langle \text{POW} \rangle)\texttt{?} \mid \text{“(”} \, \langle \text{EXP} \rangle \, \text{“)”} \\
\langle \text{POW} \rangle &\models \text{“*”}(\langle \text{SIGN} \rangle)\texttt{?}\langle \text{NUM} \rangle
\end{aligned}
$$

Fig. 3: Simplified BNF Grammar of the unit formula checker. `*`,`+` and `?` are the standard Kleene operators. The exponentiation is denoted “*”, the multiplication $\cdot$ and the division `/`.

We defined valid formula to be standard arithmetic formula including parentheses, along with the operators of multiplication and division between (prefix/units) and exponentiation on the units. We give a simplified version of the grammar in Figure 3. This version omits the complicated details required for units that are not made of a single terminal symbol. For example the mole which is abbreviated `mol` is one of these cases, since its first letter `m` collides with the letter used for meter. This requires some special treatment in the grammar than do not carry a particular interest to be precisely described here. The complete grammar is available on Chaudron's website[9].

An input formula is described as valid, if it can syntactically be parsed by the grammar. An example is given in Figure 4 for the ISU definition of the Volt.

3.4 Extraction & discussion

The extraction process is run against the rendered HTML pages obtained from the local Wikipedia copy. The HTML is filtered using CSS selectors in order to retain only the Infobox. Afterwards, the Infobox – which is concretely a `table` in HTML – is parsed to extract keys and values. Values, if not matched by the

[8] `https://en.wikipedia.org/wiki/International_System_of_Units#Derived_units`

[9] `http://w3id.org/chaudron/`

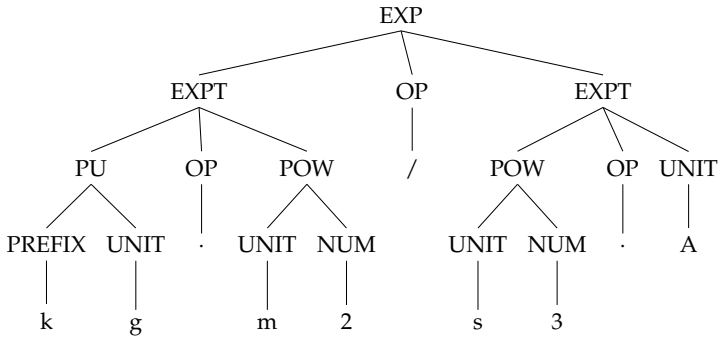


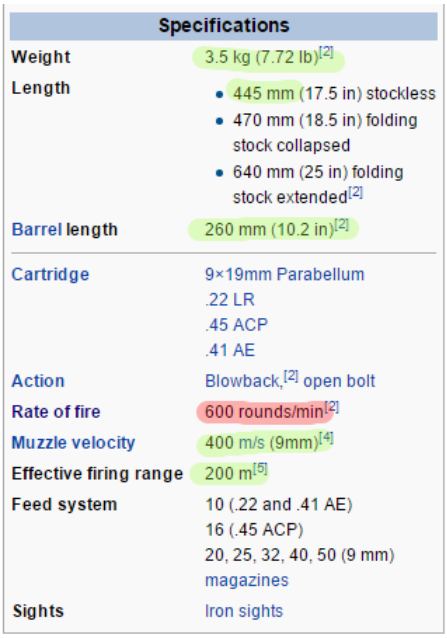
Fig. 4: Simplified parse tree for the SI equivalent of the Volt : $(kg \cdot m^2)/(s^3 \cdot A)$ using the grammar defined in Figure 3. Note that the grammar also parses the equivalent formulation: $kg \cdot m^2 \cdot s^{-3} \cdot A^{-1}$.

simple pattern matching described above, are analyzed by the parser obtained from our grammar. If a value is match ced by the parser, the formula is parsed to extract the unit and to determine whether the numerical value represents an interval or not. The representation of intervals is discussed in the next section.

As we stated in Section 2, we aim at identifiying the largest possible set of physical measurement in Wikipedia Infoboxes. This excludes any measurement that uses non standard units. Figure 5 depicts such an example. Fields highlighted in green are candidates to be extracted since they contain standard units. The field highlighted in red contains a measurement using a non-standard unit *rounds per minute*, that in fact represents a frequency. These measurements could be interesting to extract but they raise some issues. First of all, being non standard, these units do not connect to any existing system and are of limited use to their very context. Second, the notation of these units is also non-standard, i.e. it does not follow a rigorous notation. For instance *rounds per minute* appears in some articles under the abbreviation *rpm*; this, under a different context, means *rotation per minute* which is also a frequency, but of a very different type of events. Therefore these non-standard units – that do not measure physical quantities but are characteristics of some entities – are edge cases that are difficult to handle.

4 Dataset

The dataset that we obtain by extracting measures from the Infoboxes in the English Wikipedia contains over 900.000 measurements represented in 3.9 million triples. These measures extend existing knowledge for more than 450,000 DBpedia resources. More than 3.000 properties are measured with more than 600 units. Tables 1 and 2 present the most encountered units and properties in the dataset.



Specifications	
Weight	3.5 kg (7.72 lb)[2]
Length	• 445 mm (17.5 in) stockless • 470 mm (18.5 in) folding stock collapsed • 640 mm (25 in) folding stock extended[2]
Barrel length	260 mm (10.2 in)[2]
Cartridge	9×19mm Parabellum .22 LR .45 ACP .41 AE
Action	Blowback,[2] open bolt
Rate of fire	600 rounds/min[2]
Muzzle velocity	400 m/s (9mm)[4]
Effective firing range	200 m[5]
Feed system	10 (.22 and .41 AE) 16 (.45 ACP) 20, 25, 32, 40, 50 (9 mm) magazines
Sights	Iron sights

Fig. 5: Example of Infobox that contains standard and non standard units

Representation. Representing units of measures and measurement has been largely discussed within the community and several approaches have been proposed [14,8,15,11,17]. We refer the reader to [16] for an overview and interesting discussions. The most common issue with these approaches is that they come with closed lists (often large, but closed) of supported units that do not contain the whole set of units of our dataset. Our approach is an open formalism that resembles those used to represent measures using reification to describe the various properties of the measure. For instance, to describe the fact that the Three Gorges Dam has a Nameplate Capacity of 22,500 Megawatts, we obtain the following representation:

```
@prefix dbr: <http://DBpedia.org/resource/> .
@prefix cha: <http://w3id.org/chaudron/ontology/> .

dbr:Three_Gorges_Dam
 cha:measure [
  cha:value ''22,500''^^xsd:decimal;
  cha:unit ''MW''^^xsd:string;
  cha:physicalProperty ''NameplateCapacity''^^xsd:string;
  cha:DBpediaResource dbr:Nameplate_Capacity
  ] .
```

We are able to link the physical quantity measured to the DBpedia resource using `cha:DBpediaResource dbr:Nameplate_Capacity`. This is possible since the key in the key/value pair in the Infobox links to the Wikipedia page `Nameplate Capacity`.

Our extraction framework also supports the extraction of interval of values. Intervals are commonly encountered in Wikipedia to describe measure that are variable under the context (dimension of soccer field) or could not be precisely measured. To represent the bounds of the intervals in our dataset, we use the following properties: `cha:minValue` and `cha:maxValue`.

We also propose an alternative formalism suitable to process custom datatypes, based on [9]. Using custom datatypes, one could also integrate conversion and calculus at the triple store level. Dedicated libraries to process units are now widely available. For instance, the Java Specification Request 363 led to an implementation[10] of the units and measure API, that is a perfect candidate to be integrated with Jena in order to realize the vision of [9]. To our opinion this would provide a more adequate solution for units management than extensible SPARQL queries with Javascript [25].

Units	Count	Percentage
m	352,506	37.124
km^2	242,423	25.531
kg	61,202	6.445
mm	45,401	4.781
km	44,763	4.714

Table 1: Top units

Measure	Count	Percentage
height	138,985	14.637
elevation	125,345	13.201
areaTotal	110,351	11.622
weight	70,081	7.381
areaLand	44,765	4.714

Table 2: Top measures

5 Availability & Applications

The Chaudron dataset is fully integrated into the Linked Open Data Web. The resources that are qualified with measurements refer to Wikipedia articles and therefore could be directly binded to DBpedia resources. As a linked dataset, our goal is to ensure the best availability and description of Chaudron to ensure its wide adoption. We follow the current practices [1] to describe and publish our dataset.

License Chaudron is available under the terms of the Attribution-ShareAlike 3.0 license at the following persistent adress : `http://www.w3id.org/chaudron/`. The dataset is registered at Datahub[11] and described using VoID[12]. We also provide a public SPARQL endpoint, that is available from Chaudron's homepage.

[10] `https://github.com/unitsofmeasurement/`

[11] `https://datahub.io/dataset/chaudron`

[12] `http://w3id.org/chaudron/voID.ttl`

Applications. The Chaudron dataset extends DBpedia in a way that allows new applications to be developed as well as to enhance existing ones. Since measurement concern universal values, the triples do not only extend the english version of DBpedia but the complete set of localized DBpedia versions. Chaudron also fills the gap with Chemistry [12] data and leads the way to the development of new applications in this field. DBpedia has been succesfully used as the primary knowledge source for Question answering Systems, especially within the Question Answering over Linked Data workshop [20]. With the integration of Chaudron, existing Q&A systems will see their performance increased, since Chaudron provides them with triples from the utmost practical interest. Notably, Question Answering systems would be able to answer a new class of queries, that is, queries comparing entities by an attribute, including top-k queries. "What is the heaviest element between Oxygen and Gallium ?", or "What are the ten fastest roller coasters in the world ?" are examples of such questions. More complex questions such as "In which country, the biggest dam of the 19th century was built?", combine knowledge from both DBpedia and Chaudron.

6 Evaluation

To assess the quality of the dataset, we conducted a thorough evaluation to determine the precision and recall of our approach and compare with other datasets extracted from Wikipedia. We compare our extraction with the one of DBpedia [2], and Wikidata [22]. We previously discussed the construction of DBpedia; Wikidata follows a similar process as Wikipedia: the approach is to crowdsource the data acquisition.

In order to evaluate the quality of the dataset, we manually annotated over 300 statements from Infoboxes over 40 Wikipedia articles of different nature with the objective to maximize the coverage of categories: our evaluation includes vehicles, cities, persons, weapons, celestial objects, regions, buildings, drugs, rockets and others types of articles containing measurement of different nature. These measurements encompass height, weight, force, diameter, boiling point, torque, volume. The goal of this manual evaluation is to cover many types of categories and measurements. The detailed evaluation is open and can be consulted online [13].

The evaluation was conducted as follows: articles categories (buildings, drugs, weapons, ...) were identified and among them 40 articles were chosen. For each article, the fields in the Infobox have been manually identified. For each dataset, we manually identified one of the following situations: the measurement – including its units – is correctly extracted; the measurement is missing; the measurement is incorrectly extracted. For DBpedia, the latter case is very common. That is, measurement is partially extracted, mostly only the numerical value, but the units are missing. Units are sometimes present in the rela-

[13] `https://docs.google.com/spreadsheets/d/1yKFU1MMakEsKF08b3jMNBe4m91SRVXlsG6WXtNTPzlw/edit?usp=sharing`

	Chaudron	DBpedia	Wikidata
Precision	**.976** ($\frac{248}{254}$)	.555 ($\frac{50}{90}$)	.941 ($\frac{16}{17}$)
Recall	**.817** ($\frac{248}{311}$)	.289 ($\frac{90}{311}$)	.055 ($\frac{17}{311}$)
F1	**.889**	.381	.103

Table 3: Evaluation of the measurement extraction task for Chaudron, DBpedia and Wikidata.

tionship between the subject and the numerical values (height, weight, length) for instance, but in the majority of the cases this crucial information is missing, making the measurement valueless. As stated in the definition given in Section 2, measurement are meant to be compared and thus required units. Without units, a numerical value assigned to a unitless property does not constitue a measurement.

The results of this evaluation are presented in Table 3. Wikidata, with its crowdsourced approach offers a high precision, but the crowdsourced approach shows its limit on the recall. Out of the 311 measures in this evaluation, Wikidata contains only 17, out of which 16 are correct. We believe that this result to the lack of incentive of people to translate measurement into Wikidata. DBpedia uses also a manual approach, that defines translation for given Infobox templates. Therefore the recall is much higher than Wikidata (.289 vs 0.055) but it indicates that there is room for improvement in the manually defined templates. The main issue with DBpedia and measurement is how units are handled. While some properties contain the unit as the `rdf:label`, most of properties do not include this information. This is the main reason of the low precision (.55) exhibited by DBpedia on this task. Our approach outperforms its competitors on both precision (.976) and recall (.817), naturally F1-score follows. The formal grammar approach proves very robust on the extraction task. The issues encountered are mainly due to the lack of support for measurement precision (using symbol $\pm$) and to multiple values in the same cell. For instance, when an attribute gets different value depending on context: the Tesla S has different electric range depending on the model. This remains an open issue that could be fixed in the next version of Chaudron.

An important remark is that this evaluation has been conducted to cover the largest body of domains containing measurement. However data is not uniformly distributed as shown in the Infobox usage list [14]. For instance, the template `Infobox:settlement` reaches by far the first place. More precisely, at the time of writing, out the 3.7 millions use of Infoboxes in Wikipedia articles, the top 10 Infoboxes cover more than the half of use in articles. The top 100 covers 84 percents of the total usage. Would we have conducted an evaluation that follows the distribution, we would have had mainly settlements, taxobox

[14] `https://en.wikipedia.org/wiki/Wikipedia:List_of_infoboxes`

and person articles. However, such an evaluation would leave aside complete domains that are of the utmost practical interest.

7 Conclusion & Future Work

We presented Chaudron, to our knowledge, the first dataset of measurement that extends DBpedia with more than 3.9 million valuable triples. We described the different techniques used to filter and structure measures from Wikipedia Infoboxes into RDF, including pattern matching and formal grammar. The evaluation shows that our automated approach proves very robust on this extraction task and largely outperforms DBpedia and Wikidata. Our representation of measurement, in its n-ary version, offers the opportunity to support custom datatype processing using recent research results [9].

A side result of this work, nevertheless interesting for other practitioners is to demonstrate the soundness of extracting Wikipedia data from the rendered HTML instead of using Wiki markup. Future work will include the extraction of measurement from Wikipedia's plain text and will benefit from the Chaudron data for validation. Since Wikipedia's text contains way more information than the Infoboxes, we therefore hope to lift a larger body of measurements.

References

1. Keith Alexander, Richard Cyganiak, Michael Hausenblas, and Jun Zhao. Describing linked datasets. In *LDOW*, 2009.
2. Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007.
3. Oliver Ferschke, Torsten Zesch, and Iryna Gurevych. Wikipedia revision toolkit: efficiently accessing Wikipedia's edit history. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 97–102, 2011.
4. Norbert E Fuchs, Kaarel Kaljurand, and Tobias Kuhn. Attempto controlled english for knowledge representation. In *Reasoning Web*, pages 104–124. Springer, 2008.
5. Norbert E Fuchs, Uta Schwertel, and Rolf Schwitter. Attempto controlled englishnot just another logic specification language. In *International Workshop on Logic Programming Synthesis and Transformation*, pages 1–20. Springer, 1998.
6. Gjergji Kasneci, Maya Ramanath, Fabian Suchanek, and Gerhard Weikum. The yago-naga approach to knowledge discovery. *ACM SIGMOD Record*, 37(4):41–47, 2009.
7. Markus Krötzsch, Denny Vrandečić, and Max Völkel. Semantic mediawiki. In *International semantic web conference*, pages 935–942. Springer Berlin Heidelberg, 2006.
8. David Leal and Andrea Schröder. Rdf vocabulary for physical properties, quantities and units. Technical report, Technical report, ScadaOn-Web. Available at http://www. s-ten. eu/scadaonweb/NOTE-units/2002-08-05/NOTE-units. html, 2002.
9. Maxime Lefrançois and Antoine Zimmermann. *Supporting Arbitrary Custom Datatypes in RDF and SPARQL*, pages 371–386. Springer International Publishing, Cham, 2016.

10. Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195, 2015.
11. Claudio Masolo, Stefano Borgo, Aldo Gangemi, Nicola Guarino, and Alessandro Oltramari. Wonderweb deliverable d18, ontology library (final). *ICT project*, 33052, 2003.
12. Peter Murray-Rust. Chemistry for everyone. *Nature*, 451(7179):648–651, 2008.
13. Elazar J Pedhazur and Liora Pedhazur Schmelkin. *Measurement, design, and analysis: An integrated approach*. Psychology Press, 2013.
14. H Sofia Pinto and J Martins. Revising and extending the units of measure" subontology. In *Proceedings of IJCAI2001s workshop on IEEE standard upper ontology, Seattle, WA*. Citeseer, 2001.
15. Florian Probst. Observations, measurements and semantic reference spaces. *Applied Ontology*, 3(1-2):63–89, 2008.
16. Hajo Rijgersberg, Mark van Assem, and Jan Top. Ontology of units of measure and related concepts. *Semantic Web*, 4(1):3–13, 2013.
17. Hajo Rijgersberg, Mari Wigham, and Jan L Top. How semantics can improve engineering processes: A case of units of measure and quantities. *Advanced Engineering Informatics*, 25(2):276–287, 2011.
18. Max Schmachtenberg, Christian Bizer, and Heiko Paulheim. Adoption of the linked data best practices in different topical domains. In *International Semantic Web Conference*, pages 245–260. Springer, 2014.
19. Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM, 2007.
20. Christina Unger, Corina Forascu, Vanessa Lopez, Axel-Cyrille Ngonga Ngomo, Elena Cabrio, Philipp Cimiano, and Sebastian Walter. Question answering over linked data (QALD-5). In *Working Notes of CLEF 2015*, 2015.
21. Denny Vrandečić. Wikidata: A new platform for collaborative data collection. In *Proceedings of the 21st International Conference on World Wide Web*, pages 1063–1064. ACM, 2012.
22. Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.
23. Daniel S Weld, Fei Wu, Eytan Adar, Saleema Amershi, James Fogarty, Raphael Hoffmann, Kayur Patel, and Michael Skinner. Intelligence in wikipedia. In *AAAI*, volume 8, pages 1609–1614, 2008.
24. Dominik Wienand and Heiko Paulheim. Detecting incorrect numerical data in dbpedia. In *European Semantic Web Conference*, pages 504–518. Springer, 2014.
25. Greg Williams. Extensible SPARQL functions with embedded javascript. In Sören Auer, Christian Bizer, Tom Heath, and Gunnar Aastrand Grimnes, editors, *Proceedings of the ESWC'07 Workshop on Scripting for the Semantic Web, SFSW 2007, Innsbruck, Austria, May 30, 2007*, volume 248 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2007.
26. Fei Wu and Daniel S Weld. Autonomously semantifying wikipedia. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 41–50. ACM, 2007.
27. Fei Wu and Daniel S Weld. Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 118–127. Association for Computational Linguistics, 2010.