Leveraging Non-Lexical Knowledge for the Linked Open Data Web

Denny Vrandečić, Markus Krötzsch, Sebastian Rudolph, Uta Lösch

Institut AIFB - Karlsruhe Institute of Technology
Karlsruhe, Germany
{firstname.lastname}@kit.edu

Abstract. The Linked Data paradigm introduces the possibility to share machine-readable data across numerous Web resources, thus enabling applications that are traditionally only possible in corporate intranets to be realized on a Web scale. Due to the creation of an increasing number of publicly available Linked Open Data resources, the Web of Data has become a major application area for semantic technologies. This work introduces a recently published data set LON of non-lexical entities (NLEs) that can be used for numerous tasks of quantitative modeling on the Semantic Web. The size of the published data increases the magnitude of the public Linked Data significantly, yet we show how it can be seamlessly integrated into current application architectures for the Web of Data.

1 Introduction

As of today, the Semantic Web has matured to become an essential enabling factor of the ways in which we use the Web in our daily life, work, and business. Given that the principle mechanisms and architecture of the Semantic Web had already been clearly specified in 2001 [2], it may come as a surprise that the implementation of these ideas took almost a decade. Partial successes had, of course, been accomplished earlier – the most prominent example is the problem of ontological modeling which has been solved in 2004 by Smith, Welty, and McGuiness [12] – yet the main breakthrough of semantic technologies on the Web happened only more recently with the appearance of *Linked Open Data* (LOD) [3].

Linked Open Data refers to the practice of publishing structured data on the Web, and interconnecting various such data sources with links that describe their respective relationship. A more rigorous definition is provided in the section "What is Linked Data?" in [3]: "Linked Data is simply about using the Web to create typed links between data from different sources." Many practitioners, including industry and governmental organizations, have followed this astonishingly simple paradigm for publishing data. A particularly important subset of

the published resources has been grouped into the so-called *Linked Data Cloud* that is regularly updated as new resources become available.[1]

In spite of this success, there are still many areas of application for which Linked Data does not provide sufficient coverage. The contribution of this work is to extend the Linked Data Cloud with openly available information that is essential for quantitative modeling and inferencing on the Semantic Web. To this end, we provide public access to a large-scale numerical ontology that has been developed and successfully used at our institution for various years. We argue that this new data set *Linked Open Numbers* lays the foundation for hitherto unrealized applications for semantic technologies on the Web, and opens an avenue for a range of future research topics.

One reason why such important information has not been published on the Semantic Web yet may be the fact that there is a persistent skepticism toward the notion of endowing numbers with an individual identity on the Web. Nonetheless, we have found that there is significant evidence that this approach is justified both on philosophical and on practical grounds. Starting from Plato, it has long been argued that mathematical objects exist in reality and renowned experts, among them Frege [5] and Goedel [6], support this view. In particular, this argument applies to mathematical entities as foundational as numbers. The least doubt about an independent existence is left about the positive integers; according to Leopold Kronecker, "God made the integers; all else is the work of man" [1]. In the light of this overwhelming evidence, the today's Semantic Web practice of expressing numbers as literals seems highly inappropriate at best. Following [8], "URI references are used for naming all kinds of things in RDF." Not providing URIs for numbers therefore constitutes a de-facto denial of their thingness. Therefore, our work of establishing identifiers that do justice to the nature of numbers is a significant first step towards an emancipation of numbers.[2]

Clearly, the resources to store and exhibit information about numbers are limited. This might – at the first glance – lead to the impression that a dataset about the positive integers can never be complete. However, an empirical proof of the infinity of numbers has not been established: in fact, the set of all numbers ever given explicitly turns out to be finite. Moreover, the widely acknowledged school of *ultrafinitism* (see, e.g. [14, 9, 11]) convincingly argues that there are only finitely many positive integers overall. This substantiates the hope that the desirable goal of making the positive integers available to the public in their entirety can be achieved, given adequate support by national and international research funding organizations.

The structure of this paper is as follows. In Section 2, we give an overview of the modeling approach and discuss our technical realization. Section 3 then

[1] A machine-readable description (PNG) of the linked data cloud is available online at `http://linkeddata.org/`.

[2] Still, we refrain from supporting the even more radical approach "all numbers are equal", as this might lead to some counterintuitive mathematical consequences.

explains the significance of our work in the broader context of the Semantic Web, and Section 4 discusses some initial application scenarios. We provide an outlook and offer our conclusions in Section 5. The Linked Open Numbers data set can be publically accessed online at `http://km.aifb.kit.edu/projects/numbers/`.

2 Formalizing Non-Lexical Data in RDF

Following the idea of publishing numbers as RDF, we developed an ontology for describing numbers. However, we think that a similar approach would be applicable to formalizing any other kind of non-lexical data. Therefore, in the remainder of this paper we use the term non-lexical entities (NLE) when referring to numbers, thereby reflecting the broader applicability of our approach.

After a thorough investigation of the most important and most needed information concerning non-lexical data, we identified the following requirements for the ontology:

- The integer value of a NLE should be related to its resource.
- NLEs are ordered; this order should be reflected by the formalization. It should therefore be possible to identify the *predecessor* and the *successor* of a given NLE.
- To support one of the most important paradigms of computer science, Divide-and-conquer, the *prime factorization* of each non-lexical resource should be available in the ontology.
- NLEs can either be described by a series of digits or as a word (which is language-dependent). Labels should therefore be available in digits-form and in various languages.
- For reasons of backward-compatibility, Roman literals should be supported by our approach.
- Due to its high usefulness in various domains such as biology, psychology, and music, and due to the high complexity of its computation, the natural logarithm of each non-lexical resource should be available.

These above-defined requirements are the basis for the definition of the *Numbers* ontology. The ontology defines a taxonomy of NLE types: namely, `Number` is a superclass of `Integer` which itself subsumes `NaturalNumber`, the most specific class is `Prime` as subclass of `NaturalNumber`. As the meaning should be obvious to the educated reader, we will not detail here what the semantics of each of the defined classes is.

The properties `lessThan` and `greaterThan` are defined as inverse properties relating two instances of class `Number`. The properties `next` and `previous` define the predecessor and successor relation on the class `Integer`. Additional properties where introduced for defining the prime factors of a number (`prime`) and its logarithm (`log`).

Although the ontology itself is very small (it consists of less than 100 triples), it was rather complex and hard to obtain as a number of difficult design choices had to be made: which types of NLEs should be included in the ontology? Which

relations between NLEs are best suited for addressing the requirements? We believe that the formalization we propose is the best possible solution to this modeling task. This was evaluated by conducting a user study. In an expert survey, we asked a number (2) of experienced users of semantic technologies whether they saw any weak spot in the formalization. As this was denied by 100% of the participants in the study, we are confident that no further improvements are possible.

The resulting *Numbers* ontology is published adhering to all relevant Semantic Web standards, and was pedantically checked by every available validator.[3] The website is available `http://km.aifb.kit.edu/projects/numbers/`

We use the following URI scheme (as an example we take the NLE 7):

- `http://km.aifb.kit.edu/projects/numbers/n7` is the identifier for the entity, i.e. the URI identifying the abstract number 7. Since we cannot transport the actual number 7 via HTTP a GET on the URI will return a redirect to a simple, electronically transferable representation of the abstract concept, either the representation using RDF or the representation using HTML (see next two bullet points).
- `http://km.aifb.kit.edu/projects/numbers/web/n7` offers a representation of the number 7 in HTML, which most browsers are able to display for human consumption. Note that the Web page and the abstract resource are not the same and therefore have different identifiers.
- `http://km.aifb.kit.edu/projects/numbers/data/n7` is the document that contains some RDF data about the number 7. It also states the connection between the abstract number 7 and the web site describing it. The RDF resource is not meant for human consumption, but only for the machine. Its machine-understandable syntax allows the data to be repurposed and redisplayed automatically, as witnessed by the number of data browsers that are all linked from the HTML page.

We expect that this URI scheme resolves all ambiguities in this field.

3 Impact on the Semantic Web and the Linked Data Cloud

Naturally, a data set of the given size has an immeasurable impact on the Linked Open Data landscape but also on the structure of the Semantic Web as a whole. Since 2007, the former has been investigated and maintained chiefly by the *Linking Open Data project* which propagates linking of existing publicly available datasets. The goal of the initiative is bootstrapping the Web of Data, and participants have identified many available datasets and converted them to Linked Data formats. Combining this approach with the skillful persuasion of stakeholders in the data hosting community, the Linked Data cloud has grown rapidly over the last years: a total of 4.7 billion published triples has been reported in May

[3] To the best of our knowledge

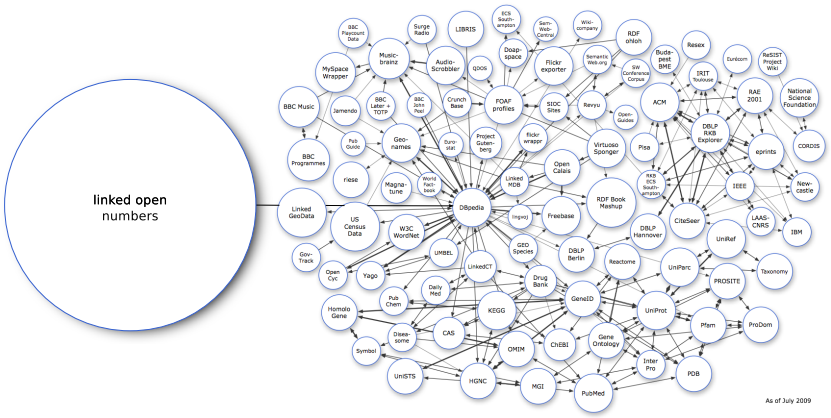

Fig. 1. Linked Open Numbers in the context of the linked open data cloud of July 2009; some sizes are estimated

2009 [3], the total was reported to have increased to 16.0 billion triples in March 2010 [13].

Our newly published RDF data set Linked Open Numbers considerably contributes to the amount of data that is available on the Web. Given the format we use to encode each NLE, a minimum of 14 RDF triples per NLE is stored in our database. Note that there are additional triples for the NLEs linking to Wikipedia, or for having the Roman encoding and the prime factorization attached. As we have encoded 1,000,000,000 NLEs, a conservative estimation of the numbers of triples we have published is 14 billion. We thereby increase the size of the Linked Data Cloud by roughly 87.5%. The significance of this increase can be seen in Fig. 1 which also visualizes the tight integration of Linked Open Numbers into the LOD cloud.

Besides this obvious impact of Linked Open Numbers on the Web of Data, the fundamental relevance of this area of application also suggests a tighter integration into the Semantic Web architecture. Indeed, it is important to ensure maximal interoperability with other semantic technology standards on a conceptual level. Due to restrictions of space, we only outline the most important aspects of this integration. A detailed architectural specification is visualized in Fig. 2 where we show how LON and NLEs are to be integrated into the W3C's Semantic Web technology stack. We point out that the combined LON/NLE module in the stack is not convex since it features two inward corners: this reflects some of the underlying design decisions of our current approach which can be seen as shortcomings in this context. The question whether a non-convex solution is possible without sacrificing practically important characteristics of our solution is left to future research.

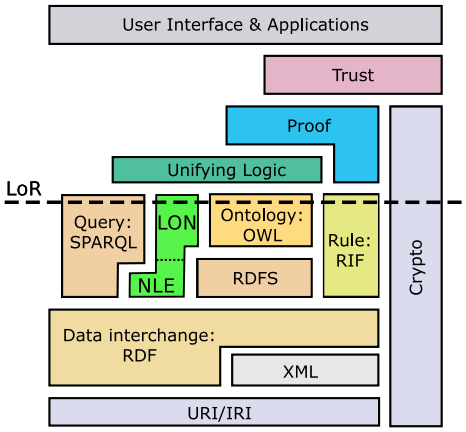

Fig. 2. Linked Open Numbers (LON) and non-lexical entities (NLE) in the context of the Semantic Web architecture where LoR represents the *line of research/realization*

Another question that is closely related to any proposed change in the Semantic Web architecture is whether downwards compatibility to existing technological standards can be ensured. It is well known that around $76\% - 92\%$ of working group activity in any W3C driven standardization effort are related to the problem that has traditionally been called *backwards compatibility*. With the advent of two-dimensional architectural diagrams, however, this problem has grown to include issues of sidewards, downwards, and, occasionally, upwards compatibility. Thus, it is important to ensure that any proposed change to the architecture imposes as little burden as possible for the users of existing technologies.

Yet, there are already a great number of Linked Open Data resources that have been created before the availability of LON as a unified resource for NLEs. Re-modeling these data sets based on the new technologies would be a tedious task that may not always be practical.[4] Fortunately, it is possible to use semantic technologies to solve this problem. Namely, the following SPARQL query creates all RDF statements that use the new approach based on the existing data:

```
CONSTRUCT ?s ?p ?nle
WHERE {?s ?p ?literal . ?nle rdf:value ?literal}
```

Using this query, it is possible to create updated linked data sets with only little human intervention. A remaining challenge is that the query can only be executed on an RDF store that contains the data that is to be transformed

[4] Initial experiments on that issue have been conducted with a group of students who were asked to do the according refactoring in Protégé. Unfortunately, experiments had to be terminated without conclusive results when loading times exceeded 16 hours.

together with the complete LON data set. Initial tests with Jena[5] have not been successful, even with all additional indexing disabled.

4 Application Scenarios

The added value of the paradigm shift initiated by our work cannot be underestimated. By endowing numbers with an own identity, the linked open data cloud will become treasure trove for a variety of disciplines. By using elaborate data mining techniques, groundbreaking insights about deep mathematical correspondences can be obtained. As an example, using our sample dataset, we were able to discover that there are significantly more odd primes than even ones, and – even more excitingly – a number contains 2 as a prime factor exactly if its successor does not. We conjecture these findings to also hold for numbers not yet contained in our dataset.

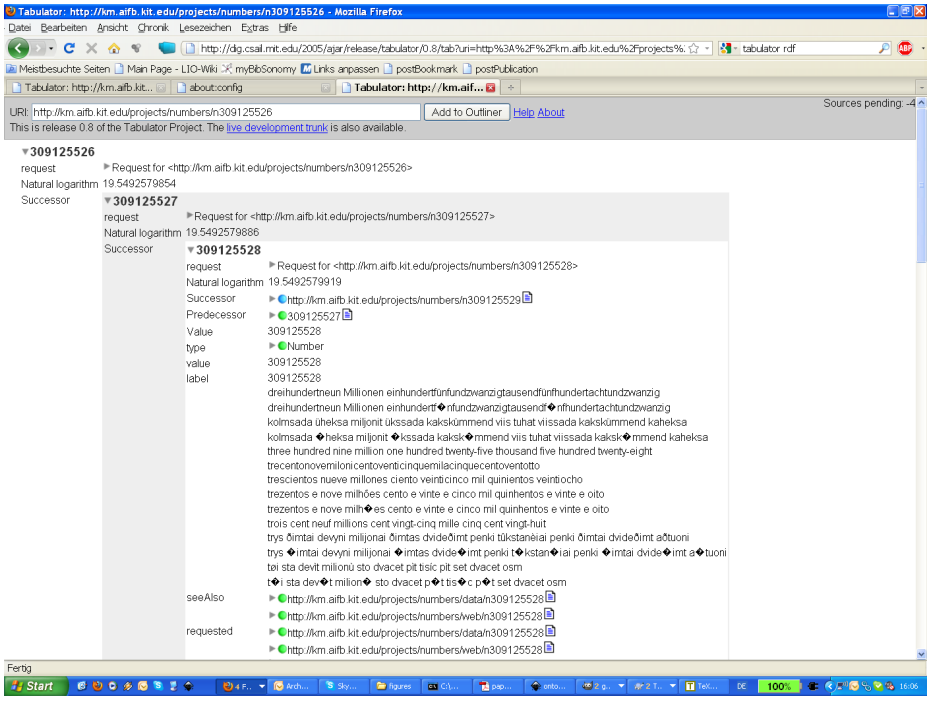

Fig. 3. Browsing the LON in Tabulator

Another prominent application of our data is the use of a Linked Data browser such as Tabulator,[6] as shown in Fig. 3. The figure illustrates a typi-

[5] http://jena.sourceforge.net/
[6] http://www.w3.org/2005/ajar/tab

cal exploration of the data set by a human user, starting from the NLE `http://km.aifb.kit.edu/projects/numbers/n309125526`. The graphical interface of Tabulator provides a concise yet informative overview of the available data, based on which the user can refine the view according to her information need. In the given example, the user has explored entities that are related via the successor property. While this is only a very simple application scenario that does not address a praticular use case, we found that many users are enthralled by the richness and depth of the LON data set. This observation hints at the huge potential of such user friendly interfaces to the Web of Data.

5 Conclusion and Outlook

In this work, we have introduced a recently published data set Linked Open Numbers (LON) of non-lexical entities (NLEs) that can be used for numerous tasks of quantitative modeling on the Semantic Web. Based on the observation that quantity is an essential quality metric for resources on the Web of Data, it is expected that LON will have a major impact on the development of the Linked Open Data web and its potential applications. Yet, we have been able to show how to seamlessly integrate the LON data into the existing Web of Data, and into the Semantic Web architecture.

Considering the impact of our work on the Semantic Web as a whole, it is interesting to note the relationship to the seminal work of Fensel and van Harmelen [4]. A major insight of this work is articulated as follows: "Given that describing the natural numbers already requires countably many axioms, the Web is quite unlikely to require much less." The visionary range of this statement is certainly astonishing, yet it may also be one of the reasons why a rigorous formalization of NLEs has not been attempted until now. Another important result of this work is that "it would take 10,000 triples just to describe each human, which gives us 100 trillion." Based on this calculation of the ultimate size of the Web of Data, we can conclude that our work contributes about 0.014% of the total amount of data on the Semantic Web. While this may appear to be little, it must be kept in mind that the final magnitude of the Web of Data is not going to be reached in the near future.

In spite of the immediate benefits that LON offers to practitioners already, there are also numerous open challenges that should be addressed in future works on the topic. Maybe most obvious is the current limitation of our data set to 10^9 NLEs. While it can be argued that these entities include many of the most important non-lexical entities that are referenced in applications, they do not cover all entities of practical interest yet.[7] The limitations of our current approach mirror the fundamental trade-off between expressive power and computational demands that is typical to knowledge representation. Future technological advances will certainly allow the border to be increased, but ultimately this endeavor will also require additional funding dedicated to that task.

[7] See, e.g., `http://en.wikipedia.org/wiki/9814072356` for a counterexample.

Other perspectives for future research include the extension of the NLE formalism to cover decimal, rational, real, or even complex numbers. Each of these extensions requires the underlying conceptual model to be augmented with suitable constructs. For example, an ontological class for algebraic reals could be practically exploited by implementations since these entities can be stored much more memory-efficient than non-algebraic reals. Moreover, the introduction of probability and vagueness into the data set may have a large practical utility: random numbers are employed in a variety of usage scenarios but as of today are only available in traditional media formats [10].

Another type of future work concerns the export formats in which LON is made available to the public. Currently, only RDF/XML format is supported, but we consider the possibility to provide at least partial exports in formats such as JSON and OWL 2 Manchester Syntax [7]. Another possible candidate is the export of LON using Microformats[8] but it is currently open whether the `value` property of *hCard* is suitable for encoding numerical values of NLEs. An alternative would be to use the *hRecipe* microformat. Creating a new NLE microformat is also an option, but this would contradict the microformat philosophy of having a small canonical set of basic formats that cover a maximal amount of application areas with a minimal amount of vocabulary.

In summary, we consider our work to be a starting point rather than as a conclusive scientific contribution, and we are confident that it will be a suitable beginning for the next decade of Semantic Web research.

6 Acknowledgements

This work would not have been possible without the great support of our colleagues: we thank Philipp Sorg for very valuable technical support, Andreas Harth for being pedantic, and Basil Ell and Irene Schick for their support in making HTML look nice.

References

1. E. T. Bell. *Men of Mathematics.* Simon and Schuster, New York, 1986.
2. T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, pages 96–101, May 2001.
3. C. Bizer, T. Heath, and T. Berners-Lee. Linked data – The story so far. *International Journal on Semantic Web & Information Systems*, 5:1–22, 2009.
4. D. Fensel and F. van Harmelen. Unifying reasoning and search to web scale. *IEEE Internet Computing*, 11(2):94–96, 2007.
5. G. Frege. *Die Grundlagen der Arithmetik. Eine logisch mathematische Untersuchung über den Begriff der Zahl.* Wilhelm Koebner, Breslau, 1884.
6. K. Gödel. What is Cantor's continuum problem? *The American Mathematical Monthly*, 54(9):515–525, 1947.

[8] `http://microformats.org`

7. M. Horridge and P. F. Patel-Schneider, editors. *OWL 2 Web Ontology Language: Manchester Syntax*. W3C Working Group Note, 27 October 2009. Available at `http://www.w3.org/TR/owl2-manchester-syntax/`.
8. G. Klyne and J. J. Carroll, editors. *Resource Description Framework (RDF): Concepts and Abstract Syntax*. W3C Recommendation, 10 February 2004. Available at `http://www.w3.org/TR/rdf-concepts/`.
9. S. Lavine. *Understanding the Infinite*. Harvard, 1994.
10. RAND Corporation. *A Million Random Digits with 100,000 Normal Deviates*. American Book Publishers, 2001.
11. P. K. Rashevsky. On the dogma of the natural numbers. *Russ.Math.Surveys*, 29, 1975.
12. M. K. Smith, C. A. Welty, and D. L. McGuinness, editors. *OWL Web Ontology Language Guide*. W3C Recommendation, 10 February 2004. Available at `http://www.w3.org/TR/owl-guide/`.
13. D. T. Tran. Personal communication, March 2010.
14. P. Vopenka. *Mathematics in the Alternative Set Theory*. Teubner, 1979.