Towards Green Linked Data

Julia Hoxha[1], Anisa Rula[2], and Basil Ell[1]

[1] Institute AIFB, Karlsruhe Institute of Technology, {julia.hoxha, basil.ell}@kit.edu,
[2] Dipartimento di Informatica Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca, anisa.rula@disco.unimib.it

Abstract. We here present a vision of what needs to be addressed when designing and publishing linked data on the Web. Our approach aims at reducing the amount of incorrect, irrelevant, or redundant content – which can also be seen as *pollution* in the Web of Data – when publishing linked data. At the foundation lie the design principles adapted from green engineering.We envision a holistic framework that evaluates, along these principles and their respective assessment metrics, datasets from publishers and allows configuration of new validation tools.

1 Introduction

The rapid growth of the Web of Data has contributed to the creation of large amounts of linked data that often results in low quality content. For this reason, it is important to investigate the problem of *pollution* from which the *linked data environment* may suffer. Pollution refers in our case to incorrect, irrelevant, or redundant content, which aggregates low value to users and services that consume these data. Examples include broken links, ambiguous use of owl:sameAs, redundant definition of vocabularies, multiple URIs for the same resource in a dataset, complex vocabularies that cannot be efficiently reused, uncomprehensible data, unaccessible data, non-maintained data, etc.

We approach the field of green engineering, since it has long been involved with quality assurance when designing materials, processes and systems that are benign to the environment. Moreover, this field offers an ecological perspective when discussing the problems encountered in linked data publishing. At the foundation of our approach lie the fundamental principles of green engineering [2], which we adapt for the linked data setting. We aim at providing a vision of what needs to be addressed when designing and publishing linked data, in order to minimize pollution in the Web of Data, increase reuse and achieve sustainability. To concretize this vision, we introduce a framework that applies the principles with measurable aspects to evaluate how green the datasets from publishers are.

The issue of quality on the Web of Data has been addressed along aspects such as syntax errors and inconsistencies in datasets [6], link discovery and maintainance [8], quality and trustworthiness assessment based on provenance information [5], etc. Our approach is not complementary to these works, rather encompasses and aligns them to our principles. The framework that we introduce is holistic and based on green engineering aspects, which can be extended with new measures and validation tools for higher quality of the published data.

2 Green Linked Data Principles

We introduce each principle with a short description and assessment measures, which are partly contribution of Web community [1]. Basic resources related to the Web of Data include vocabularies, datasets, RDF links, and URIs. The principles are non-orthogonal, therefore some measures occur in different principles.

Principle 1. Inherent rather than circumstantial

Ensure that data are as inherently benign as possible

Benign refers to data that maximize the qualities in which the publishers and consumers are interested. Publishers are interested that their data is consumed, i.e. data is 1)accessible 2)understandable by consumers and 3)meet their demand.

Dimension	**Measures**
Accessibility	server accessibility; accessibility of a SPARQL-endpoint; dereferenceable URIs; accessibility of the RDF dumps
Reliability	usage of a dedicated provenance vocabulary [5]; basic provenance information; usage of digital signatures
Comprehensibility	labeling and readable description of classes, properties and entities; indication of exemplary URIs; exemplary SPARQL queries

Table 1. Dimensions and Measures of Principle 1

Principle 2. Prevention Instead of Treatment

It is better to prevent waste than to treat or clean up after it is formed.

Publishers should strive to produce data with "zero-waste", which in the Web of Data results from the lack of use or consumption, i.e. consumers (human and machines) are unable to effectively exploit published data for beneficial use.

Dimension	**Measures**
Visibility	listing in linked data catalogues
Consistency	valid entity definition as members of disjoint classes inconsistent values for properties; usage of uniform datatypes valid usage of inverse-functional properties
Comprehensibility	see Table. 1

Table 2. Dimensions and Measures of Principle 2

Principle 3. Maximize Reuse

Reuse existing resources: vocabularies, URIs, links

Publishers should strive to maximize usage of provenance information, usage of established vocabularies, and referencing of prominent URIs.

Dimension	**Measures**
Provenance	existence of provenance information usage of established provenance ontologies
Uniformity	usage of an established representation format usage of established vocabularies;referencing of established URIs
Redundancy	multiple URIs for same entity; identity resolution

Table 3. Dimensions and Measures of Principle 3

Principle 4. Design for Separation
Modularization operations should be a component of the design process
Engineering large monolithic ontologies leads to artifacts that can rarely be reused, due to fitting to the design requirements. Modularization helps solve this challenge using instead a set of micro-ontologies, therefore increasing opportunities for the reuse of the developed artifacts.

Dimension	**Measures**
Provenance	usage of related provenance ontologies metadata on derivation history, data engineering/generation process
Partitionability	usage of micro-ontologies; metadata on micro-ontologies

Table 4. Dimensions and Measures of Principle 4

Principle 5. Maximize Efficiency
Design datasets in order to maximize efficient exploitation
Published linked data should allow consumers to search, query and browse them achieving required results with minimum effort and time.

Dimension	**Measures**
Validity	no syntax errors; proper datatypes no deprecated classes and properties; usage of proper datatypes
Size	number of triples, number of internal and external links scope and level of detail in the dataset
Performance	reasoning performance scalability; browsing efficiency
Consistency	see Table. 2

Table 5. Dimensions and Measures of Principle 5

Principle 6. Output-Pulled Versus Input-Pushed
Bringing content and publishing rate in line with demand
Publishers should have possible consumers in mind when designing their data.To this aim, they should cover user needs providing only the necessary resources.

Dimension	**Measures**
Consumer requirements	existence of concrete user requirements; existing tools or applications consuming such/similar data
Semantic Gap	identification of semantic data and vocabulary gap via query logs [7] identification of sparse result sets and near matches [4]

Table 6. Dimensions and Measures of Principle 6

Principle 7. Conserve Complexity
When making design choices, publishers should strive to reuse a complex ontology or dataset as it is, instead of *recycling* i.e. extracting parts of it and modifying them for further use. Complexity should be viewed as an investment for reuse.

Dimension	Measures
Complexity	size of the vocabulary (classes, properties, instances, derived properties, etc.); entropy distance-based structure complexity
Partitionability	see Table. 4

Table 7. Dimensions and Measures of Principle 7

Principle 8. Meet Need, Minimize Excess
Design for unnecessary capability or capacity solutions should be considered a design flaw
Publishers should try to provide datasets that meet the necessary capabilities, with no excessive details, while "'one size fits all"' solutions are a design flaw.

Dimension	Measures
Scope	conformance to user requirements (e.g. competency questions); level of detail of dataset
Granularity	no one-size-fits-all but domain-specific micro-ontologies
Size	see Table. 5

Table 8. Dimensions and Measures of Principle 8

Principle 9. Design for Afterlife
Design for performance in a commercial afterlife
It is necessary to provide updates and maintainance after the planned end of life of the data. To reduce waste, components that remain functional and valuable can be recovered for reuse and/or reconfiguration.

Dimension	Measures
Validity	see Table. 5
Accessibility	see Table. 1
Timeliness	indication of the most recent data validation (update) frequency of validation; exclusion of outdated data;inclusion of recent data; appropriate metadata to indicate outdated or deprecated dataset/URIs
Targeted Lifetime	indication of maintainance period usage of proper vocabularies on maintainance data

Table 9. Dimensions and Measures of Principle 9

3 Green Linked Data Framework

In this section, we introduce an envisioned framework, which is a Web platform addressing three main groups of visitors 1) those who want to learn about linked data and the green approach, 2) publishers that wish to check their linked data before publishing them online, and 3) software developers who can contribute with validators that check particular measures pertaining to the principles.

We have initiated the implementation of this framework online[1], aiming to make it a future point of reference for the users of the Web of Data. Through the

[1] http://www.greenlinkeddata.org

introduction of the green principles and the dimensions in which they expand, as well as via the further enrichment of the website with materials and related links, we aim to raise the concern among these users about the importance of the quality of linked data published online.

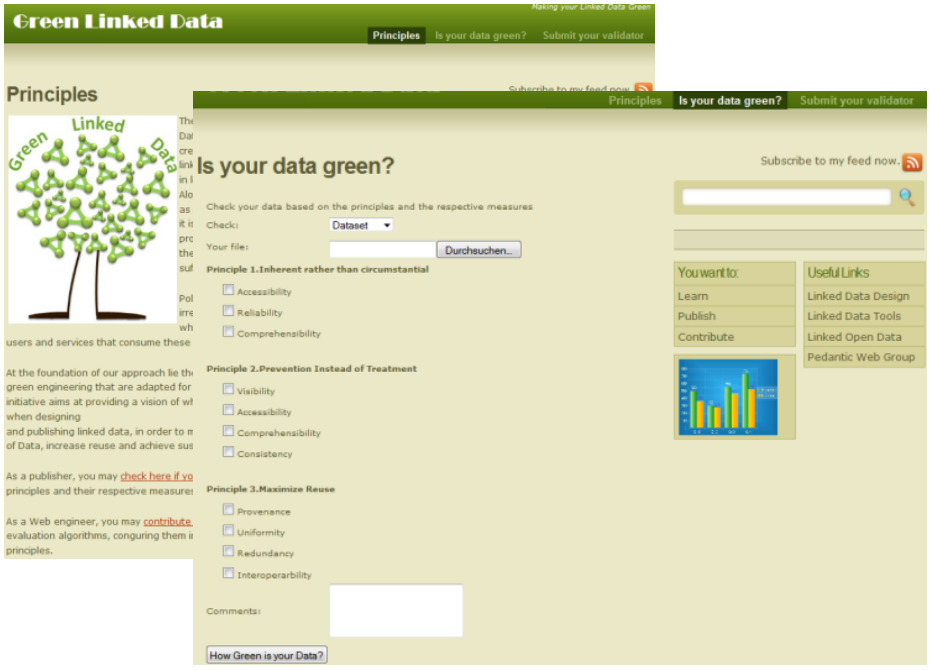

Fig. 1. greenlinkeddata.org Framework

Besides its informative nature, the platform aims at enabling users to make concious decisions about the data they need to publish, and most importantly help them evaluate how these data conform to the green principles. Therefore, for the publishers the framework offers the possibility to automatically check the datasets or vocabularies they want to publish based on the measures defined. A publisher may choose to check its data towards one or several principles and dimensions (Fig. 1).

The evaluation of the data will be done through validators which will consist of open source or off-the-shelf algorithms offered in the Web of Data community, as well as new validators (e.g. to check comprehensibility) that we are implementing. A more interesting feature of the framework is the possibility provided to software developers to submit their validators, for example as Web services. For example one measure for the comprehensability of a dataset is the labeling completeness metric LC_{lp} where lp is a set of labeling properties such as `rdsf:label`. This metric evaluates the ratio of non-information resources for which at least one label is defined [3].

There is also the possibility to suggest new dimensions and respectively contribute with appropriate validators. Thus, our goal is to provide an open framework, where Web users not only contribute with validators, but also with new

ideas, materials and tools. Furthermore, the platform allows adding to each principle in the website new comments that may consist of, but are not restricted to, best practices, benefits or even difficulties they have had when dealing with those aspects. They may also contribute with suggestions on how to extend dimensions and measures of that principle.

4 Discussion and Conclusion

At the foundation of this approach lie green engineering principles, which we have transfered to linked data publishing. In contrast to the physical artifacts addressed in the original approach, we deal with data that represent immaterial artifacts. The fundamental differences between these two types of artifacts have necessarily been taken into account.

Physical artifacts are subject to decay and abrasion in consequence of usage. They cannot easily be duplicated or distributed, and possess the property of excludability. Since material goods are naturally scarce, this can lead to rivalry. In contrast, data are immaterial, thus can be easily duplicated and distributed, without being subject to decay. While porting the principles to the linked data setting, we have extraced only 9 of the original 12 principles, excluding e.g. those dealing with renewable type of resources, infrastructure used to create and provide the data, or discussion on green energy consumption.

In our future work, we will focus on extending the principles with other measures and bringing to life via further development the envisioned framework.

References

1. *Quality criteria for linked data sources, http://sourceforge.net/apps/mediawiki/trdf/index.php?title=quality-criteria-for-linked-data-sources*, 2010.
2. P. Anastas and J. Zimmerman, *Design through the 12 principles of green engineering*, Engineering Management Review, IEEE, 35 (2007), p. 16.
3. B. Ell, D. Vrandečić, and E. Simperl, *Labels in the Web of Data*, in Proceedings of the 10th International Semantic Web Conference (ISWC2011), Lecture Notes in Computer Science, Berlin / Heidelberg, 2011, Springer.
4. H.-J. Happel, *Semantic need: guiding metadata annotations by questions people #ask*, in Proceedings of ISWC'10- Volume Part I, Berlin, Heidelberg, 2010, Springer-Verlag, pp. 321–336.
5. O. Hartig, *Provenance Information in the Web of Data*, 2009.
6. A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres, *Weaving the pedantic web*, in 3rd International Workshop on Linked Data on the Web (LDOW2010), in conjunction with 19th International World Wide Web Conference, CEUR, 2010.
7. P. Mika, E. Meij, and H. Zaragoza, *Investigating the semantic gap through query log analysis*, in Proceedings of the 8th International Semantic Web Conference, ISWC '09, Berlin, Heidelberg, 2009, Springer-Verlag, pp. 441–455.
8. J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov, *Discovering and Maintaining Links on the Web of Data*, in The Semantic Web - ISWC 2009, A. Bernstein, D. R. Karger, T. Heath, L. Feigenbaum, D. Maynard, E. Motta, and K. Thirunarayan, eds., vol. 5823, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, ch. 41, pp. 650–665.