The Web of Data: Decentralized, collaborative, interlinked and interoperable

Sören Auer and Sebastian Hellmann

Universität Leipzig, Institut für Informatik, AKSW,
Postfach 100920, D-04009 Leipzig, Germany,
{auer|hellmann}@informatik.uni-leipzig.de
http://aksw.org

Abstract

Recently the publishing and integration of structured data on the Web gained traction with initiatives such as Linked Data, RDFa and schema.org. In this article we outline some fundamental principles and aspects of the emerging Web of Data. We stress the importance of open licenses as an enabler for collaboration, sharing and reuse of structured data on the Web. We discuss some features of the RDF data model and its suitability for integrating structured data on the Web. Two particularly crucial aspects are performance and scalability as well as conceptual interoperability, when using the Web as a medium for data integration. Last but not least we outline our vision of a Web of interlinked linguistic resources, which includes the establishment of a distributed ecosystem of heterogeneous NLP tools and services by means of structural, conceptual and access interoperability employing background knowledge from the Web of Data.

Keywords: Linked Data, Interoperability, Data Web

1. Introduction

Tim Berners-Lee conceived the vision of the Giant Global Graph[1] connecting all data on the Web and allowing to discover new relations between the data. This vision has been pursued by the Linked Open Data community, where the Linked Open Data (LOD) cloud now comprises 295 repositories and more than 30 billion RDF triples[2]. Although it is difficult to precisely identify the reasons for the success of the LOD effort, advocates generally argue that open licenses as well as open access are key enablers for the growth of such a network as they provide a strong incentive for collaboration and contribution by third parties. (Bizer, 2011) argues that with RDF the overall data integration effort can be "split between data publishers, third parties, and the data consumer", a claim that can be substantiated by looking at the evolution of many large datasets constituting the LOD cloud. We outline some stages of the linked data publication and refinement (cf. (Auer and Lehmann, 2010; Berners-Lee, 2006; Bizer, 2011)) in Figure 1 and will discuss these in more detail throughout this article.

In this overview article accompanying a presentation at the LREC 2012 conference we discuss some crucial aspects of the emerging Web of interlinked Open Data: The importance of open licenses and open access as an enabler for collaboration, the ability to interlink data on the Web as a key feature of RDF as well as scalability and decentralization. We elaborate on how conceptual interoperability can be achieved by (1) re-using vocabularies and (2) agile ontology development (3) meetings to refine and adapt ontologies (4) tool support to enrich ontologies and match schemata. Finally, we introduce our vision of a Web of tightly interlinked linguistic resources.

[1] http://dig.csail.mit.edu/breadcrumbs/node/215

[2] http://www4.wiwiss.fu-berlin.de/lodcloud/state/

2. Open licenses, open access and collaboration

DBpedia, FlickrWrappr, 2000 U.S. Census, LinkedGeoData, LinkedMDB are some prominent examples of LOD datasets, where the conversion, interlinking, as well as the hosting of the links and the converted RDF data has been completely provided by third parties with no effort and cost for the original data providers[3]. DBpedia, for example, was initially converted to RDF solely from the open data dumps provided by Wikipedia. With Openlink Software a company supported the project by providing hosting infrastructure and a community evolved, which created links and applications. Although it is difficult to determine whether open licenses are a necessary or sufficient condition for the collaborative evolution of a data set, the opposite is quite obvious: *Closed* licenses or *unclearly licensed* data are an impediment to an architecture which is focused on (re-)publishing and linking of data. Several datasets, which were converted to RDF could not be re-published due to licensing issues. Especially, these include the Leipzig Corpora Collection (LCC) (Quasthoff et al., 2009) and the RDF data used in the TIGER Corpus Navigator (Hellmann et al., 2010). Very often (as is the case in the previous two examples), the reason for closed licenses is the strict copyright of the primary data (such as newspaper texts) and researcher thus being unable to publish their annotations and derived data. The open part of the American National Corpus (OANC[4]) on the other hand has been converted to RDF and was re-published successfully using POWLA (Chiarcos, 2012). Thus, the work contributed to OANC was directly reusable by other scientist and likewise the same accounts for the RDF conversion.

Note that the *Open* in Linked Open Data refers still mainly to *open access*, i.e. retrievable by HTTP.Only around 18% of the datasets of the LOD cloud provide clear licensing

[3] More datasets are available here: http://thedatahub.org/tag/published-by-third-party

[4] http://www.anc.org/OANC/

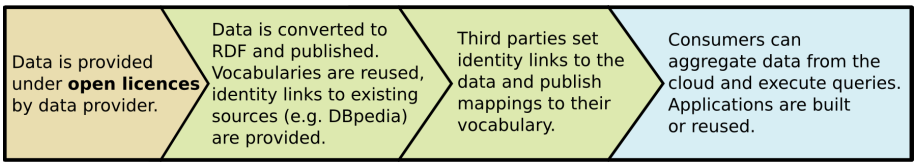

Figure 1: Summary of several methodologies for publishing and exploiting linked data (Chiarcos et al.,). The data provider is only required to make data available under an open license (left-most step). The remaining steps for data integration can be contributed by third parties and data consumers.

information at all[5]. Of these 18% an even smaller amount is considered *open* in the spirit of the open definition[6] coined by the Open Knowledge Foundation.

3. RDF as a data model

The RDF data model is very simple yet powerful. Inspired by linguistic categories, the RDF data model is based just one single elementary structure – RDF statements (or triples) consisting of a subject, predicate and object. Each of these components is essentially a worldwide (or in the case of blank nodes locally) unique identifier – IRIs. For objects also data values (called literals) together with a datatype or language tag are allowed. RDF as a data model has distinctive features, when compared to its alternatives. Conceptually, RDF is close to the widely used Entity-Relationship Diagrams (ERD) or the Unified Modeling Language (UML) and allows to model entities and their relationships. XML is a serialization format, that is useful to (de-)serialize data models such as RDF. Major drawbacks of XML and relational databases are the lack of (1) global identifiers such as IRIs, (2) standardized formalisms to explicitly express links and mappings between these entities and (3) mechanisms to publicly access, query and aggregate data. Note that (2) can not be supplemented by transformations such as XSLT, because the linking and mappings are implicit. All three aspects are important to enable ad-hoc collaboration. The resulting technology mix provided by RDF allows any collaborator to join her data into the decentralized data network employing the HTTP protocol with immediate benefits herself and others. In addition, features of OWL can be used for inferencing and consistency checking. Inferencing allows, for example, to model transitive properties, which can be queried on demand, without expanding the size of the data. While XML can only check for validity, i.e. the occurrence and order of data items (elements and attributes), consistency checking allows to verify, whether a dataset adheres to the semantics given by the formal definitions of the used ontologies.

4. Performance and scalability

RDF, its query language SPARQL and its logical extension OWL provide features and expressivity that go beyond relational databases and simple graph-based representation strategies. This expressivity poses a performance challenge to query answering by RDF triples stores, inferencing by OWL reasoners and of course the combination thereof. Although the scalability is a constant focus of RDF data management research[7], the primary strength of RDF is its flexibility and suitability for data integration and not superior performance for specific use cases. Many RDF-based systems are designed to be deployed in parallel to existing high performance systems and not as a replacement. An overview over approaches that provide Linked Data and SPARQL on top of relational database systems, for example, can be found in (Auer et al., 2009). The NLP Interchange Format (cf. section 6.) allows to express the output of highly optimized NLP systems (e.g. UIMA) as RDF/OWL. The architecture of the Data Web, however, is able to scale in the same manner as the traditional WWW as the nodes are kept in a de-centralized way and new nodes can join the network any time and establish links to existing data. Data Web search engines such as Swoogle[8] or Sindice[9] index the available structured data in a similar way as Google does with the text documents on the Web and provide keyword-based query interfaces.

5. Conceptual interoperability

While RDF provides structural (or syntactical) interoperability, conceptual interoperability is achieved by globally unique identifiers (i.e. IRIs) for entities, classes and properties, that have a defined meaning. These unique identifiers can be interlinked via `owl:sameAs` links on the entity-level, re-used as properties on the vocabulary level and extended or set equivalent via `rdfs:subClassOf` or `owl:equivalentClass` on the schema-level. Following the ontology definition of Gruber (Gruber, 1993), the aspect that ontologies represent a "shared conceptualization" stresses the need to collaborate in order to achieve a shared understanding. On the class and property level RDF and OWL give users the freedom to reuse, extend and relate other work within their own conceptualization. Very often, however, it is the case that groups of stakeholders actively discuss and collaborate to form some kind of agreement on the meaning of identifiers (as e.g. described in (Hepp et al., 2006)). In the following, we outline some examples on how conceptual interoperability can be achieved:

- In a knowledge extraction process (e.g. when converting relational databases to RDF) vocabulary identifiers can be re-used during the extraction process. Especially community-curated vocabularies such as FOAF,

[5] `http://www4.wiwiss.fu-berlin.de/lodcloud/state/#license`

[6] `http://opendefinition.org/`

[7] `http://factforge.net` or `http://lod.openlinksw.com` provide SPARQL interfaces to query billions of aggregated facts.

[8] `htpp://swoogle.umbc.edu`

[9] `http://sinidce.com`

SIOC, Dublin Core and the DBpedia Ontology are suitable candidates for reuse as this leads to conceptual interoperability with all applications and knowledge bases that also use the same vocabularies. This aspect has been the rationale for designing Triplify (Auer et al., 2009), where the SQL query syntax was slightly extended to map query results to existing RDF vocabularies.

- During the creation process of ontologies, direct collaboration can be facilitated with tools that allow agile ontology development such as OntoWiki, Semantic Mediawiki or the DBpedia Mappings Wiki[10]. In this way, conceptual interoperability is achieved by a decentralized group of stakeholders, who work together over the Internet. The created ontology can be published and new collaborators can get involved to further improve the ontology and tailor it to their needs.
- In some cases, real life meetings are established, e.g. in the form of Vo(cabulary)-Camps, where interested people meet to discuss and refine vocabularies. VoCamps can be found and registered on `http://vocamp.org`.

6. Towards a Web of interlinked linguistic resources

In recent years, the interoperability of linguistic resources and NLP tools has become a major topic in the fields of computational linguistics and Natural Language Processing (Ide and Pustejovsky, 2010). The technologies developed by the Semantic Web initiative during the last decade have produced formalisms and methods that enable the publication and linking of comprehensive knowledge bases, while still providing implementations that scale for large data. Some current projects in the NLP domain seem to follow the same approach such as the graph-based formalism *GrAF* developed in the ISO TC37/SC4 group (Ide and Suderman, 2007) and the *ISOcat data registry* (Windhouwer and Wright, accepted), which can benefit directly from the widely available tool support, once resources were converted to RDF. It is the declared goal of GrAF to be a pivot format for supporting conversion between other formats and it was not primarily designed to be used directly. Also, the ISOcat already offers a Linked Data interface. In addition, other datasets have already converted to RDF such as the typological data in Glottolog/Langdoc (Chiarcos et al.,) or Wiktionary[11]. An overview of such approaches can be found in (Chiarcos et al., 2012).

An important factor for improving the quality of the output generated by NLP tools is the availability of large quantities of qualitative background knowledge, such as on the currently emerging Web of Linked Data (Auer and Lehmann, 2010). Many NLP tasks can greatly benefit from making use of this wealth of knowledge being available on the Web in structured form as *Linked Open Data* (LOD). The precision and recall of Named Entity Recognition, for example, can be boosted when using background knowledge from DBpedia, Geonames or other LOD

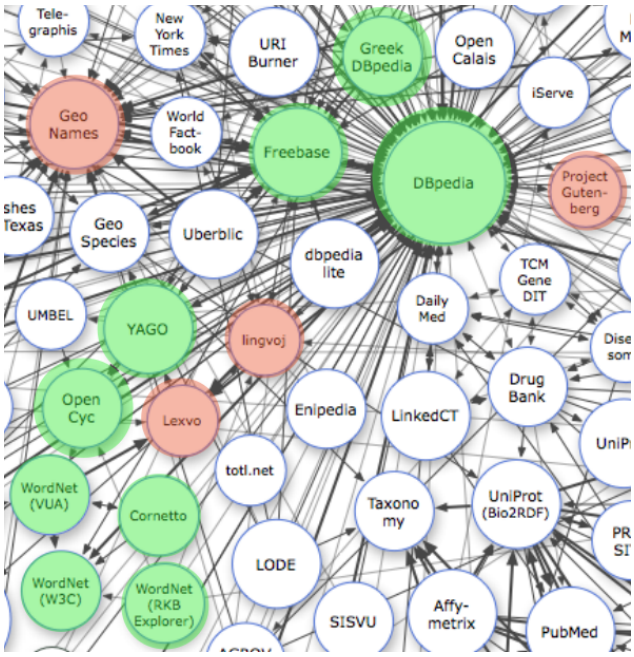

Figure 2: Language resources in the current Linked Open Data cloud. Lexical-semantic resources are colored green and linguistic metadata red.

sources as crowdsourced and community-reviewed and timely-updated gazetteers. Of course the use of gazetteers is a common practice in NLP. However, before the arrival of large amounts of Linked Open Data their creation and maintenance in particular for multi-domain NLP applications was often impractical. Figure 2 shows a snapshot of the LOD cloud with highlighted language resources that are particularly relevant for NLP.

The use of LOD background knowledge in NLP applications poses some particular challenges. These include: *identification* – uniquely identifying and reusing identifiers for (parts of) text, entities, relationships, NLP concepts and annotations etc.; *provenance* – tracking the lineage of text and annotations across tools, domains and applications; *semantic alignment* – tackle the semantic heterogeneity of background knowledge as well as concepts used by different NLP tools and tasks.

Besides the availability of Linked Data, we are currently observing a plethora of *Natural Language Processing* (NLP) tools and services being freely available and new ones appearing frequently. Especially relevant for the Semantic Web are tools and web services, that provide *Named Entity Recognition* (NER) as well as reusable identifiers (IRIs) for entities found in the Linked Data Cloud. The recently published NLP Interchange Format (NIF)[12] aims to improve interoperability for the output of such NLP tools as well as for linguistic data in RDF, documents and structured data published on the Web.

NIF addresses the interoperability problem on three layers: the *structural*, *conceptual* and *access* layer. NIF is based on a Linked Data enabled IRI scheme for identifying elements in (hyper-)texts (structural layer) and a compre-

[10] `http://mappings.dbpedia.org`

[11] `http://dbpedia.org/Wiktionary`

[12] Specification: `http://nlp2rdf.org/nif-1-0`

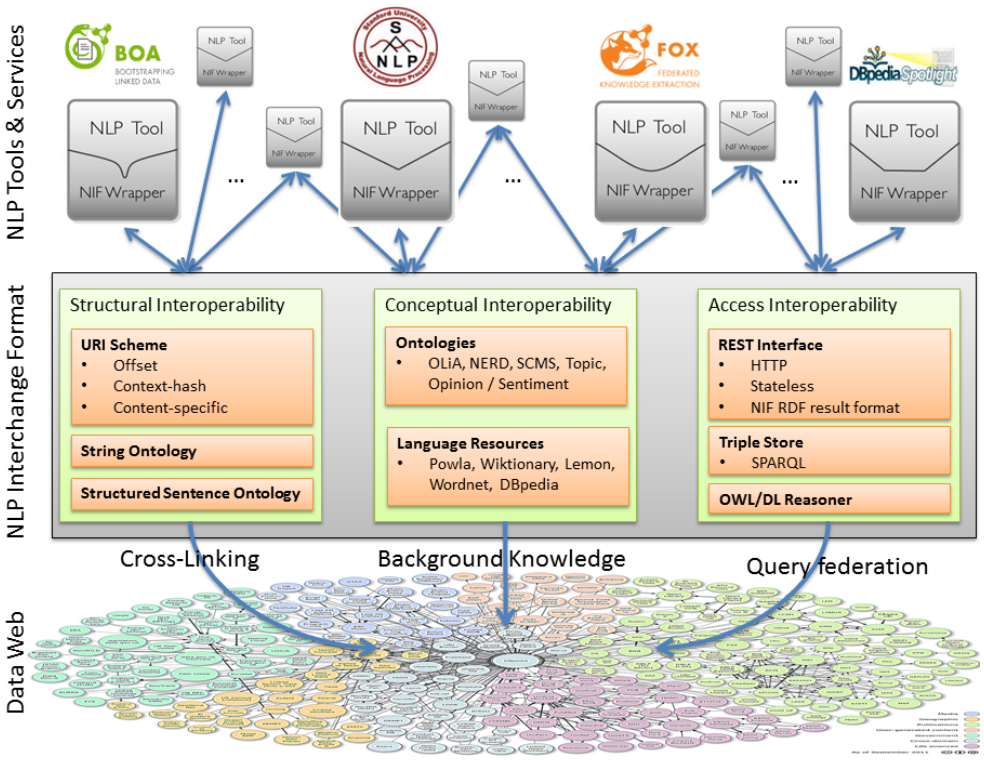

Figure 3: NIF architecture aiming at establishing a distributed ecosystem of heterogeneous NLP tools and services by means of structural, conceptual and access interoperability employing background knowledge from the Web of Data.

hensive ontology for describing common NLP terms and concepts (conceptual layer). NIF-aware applications will produce output (and possibly also consume input) adhering to the NIF ontology as REST services (access layer). Other than more centralized solutions such as UIMA and GATE, NIF enables the creation of heterogeneous, distributed and loosely coupled NLP applications, which use the Web as an integration platform. Another benefit is, that a NIF wrapper has to be only created once for a particular tool, but enables the tool to interoperate with a potentially large number of other tools without additional adaptations. NIF can be partly compared to LAF and its extension GrAF(Ide and Pustejovsky, 2010) as LAF is similar to the proposed IRI schemes and the String ontology[13], while other (already existing) ontologies are re-used for the different annotation layers of NLP[14]. Furthermore, NIF utilizes the advantages of RDF and uses the Web as an integration and collaboration platform. Extensions for NIF can be created in a decentralized and agile process, as has been done in the NERD extension for NIF (Rizzo et al., 2012). Named Entity Recognition and Disambiguation (NERD)[15] provides an ontology, which maps the types used by web services such as *Zemanta*, *OpenCalais*, *Ontos*, *Evri*, *Extractiv*, *Alchemy API* and *DBpedia Spotlight* to a common taxonomy. Ultimately, we envision an ecosystem of NLP tools and services to emerge using NIF for exchanging and integrating rich annotations. Figure 3 gives an overview on the architecture of NIF connecting tools, language resources and the Web of Data.

[13] `http://nlp2rdf.lod2.eu/schema/string/`

[14] examples for such ontologies are OLiA, NERD and lemon

[15] `http://nerd.eurecom.fr`

7. References

Sören Auer and Jens Lehmann. 2010. Making the web a data washing machine - creating knowledge out of interlinked data. *Semantic Web Journal*.

Sören Auer, Sebastian Dietzold, Jens Lehmann, Sebastian Hellmann, and David Aumueller. 2009. Triplify: Lightweight linked data publication from relational databases. In Juan Quemada, Gonzalo León, Yoëlle S. Maarek, and Wolfgang Nejdl, editors, *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, pages 621–630. ACM.

Tim Berners-Lee. 2006. Design issues: Linked data. `http://www.w3.org/DesignIssues/LinkedData.html`.

Chris Bizer. 2011. Evolving the web into a global data space. `http://www.wiwiss.fu-berlin.de/en/institute/pwo/bizer/research/publications/Bizer-GlobalDataSpace-Talk-BNCOD2011.pdf`. Keynote at 28th British National Conference on Databases (BNCOD2011).

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. Towards a linguistic linked open data cloud: The open linguistics working group. *Traitement automatique des langues*, to appear.

Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors. 2012. *Linked Data in Linguistics. Representing Language Data and Metadata*. Springer. companion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with

the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany.

Christian Chiarcos. 2012. Powla: Modeling linguistic corpora in owl/dl. In *Proceedings of 9th Extended Semantic Web Conference (ESWC2012)*.

Thomas R. Gruber. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220.

Sebastian Hellmann, Jörg Unbehauen, Christian Chiarcos, and Axel-Cyrille Ngonga Ngomo. 2010. The TIGER Corpus Navigator. In *9th International Workshop on Treebanks and Linguistic Theories (TLT-9)*, pages 91–102, Tartu, Estonia.

Martin Hepp, Daniel Bachlechner, and Katharina Siorpaes. 2006. Harvesting wiki consensus - using wikipedia entries as ontology elements. In Max Völkel and Sebastian Schaffert, editors, *Proceedings of the First Workshop on Semantic Wikis – From Wiki to Semantics, co-located with the 3rd Annual European Semantic Web Conference (ESWC 2006)*, Workshop on Semantic Wikis. ESWC2006, June.

N. Ide and J. Pustejovsky. 2010. What does interoperability mean, anyway? Toward an operational definition of interoperability. In *Proc. Second International Conference on Global Interoperability for Language Resources (ICGL 2010)*, Hong Kong, China.

Nancy Ide and Keith Suderman. 2007. GrAF: A graph-based format for linguistic annotations. In *Proc. Linguistic Annotation Workshop (LAW 2007)*, pages 1–8, Prague, Czech Republic.

Matthias Quasthoff, Sebastian Hellmann, and Konrad Höffner. 2009. Standardized multilingual language resources for the web of data: http://corpora.uni-leipzig.de/rdf. In *3rd prize at the LOD Triplification Challenge, Graz, 2009*.

Giuseppe Rizzo, Raphaël Troncy, Sebastian Hellmann, and Martin Brümmer. 2012. NERD meets NIF: Lifting NLP Extraction Results to the LinkedData Cloud. In *Proceedings of Linked Data on the Web Workshop (WWW)*.

M. Windhouwer and S. E. Wright. accepted. Linking to linguistic data categories in ISOcat. In *Linked Data in Linguistics (LDL 2012)*, Frankfurt/M., Germany, Mar.