LDOW2017: 10th Workshop on Linked Data on the Web

Jens Lehmann
University of Bonn, DE
lehmann@uni-bonn.de

Sören Auer
University of Bonn, DE
auer@cs.uni-bonn.de

Sarven Capadisli
University of Bonn, DE
info@csarven.ca

Krzysztof Janowicz
UC Santa Barbara, USA
jano@geog.ucsb.edu

Christian Bizer
University of Mannheim, DE
chris@bizer.de

Tom Heath
Arup, UK
tom.heath@arup.com

Aidan Hogan
DCC, Universidad de Chile
ahogan@dcc.uchile.cl

Tim Berners-Lee
W3C/MIT, USA
timbl@w3.org

ABSTRACT

The 10th Linked Data on the Web workshop (LDOW2017)[1] was held in Perth, Western Australia on April 3rd, 2017, co-located with the 26th International World Wide Web Conference (WWW2017). In its 10th anniversary edition, the LDOW workshop aims to stimulate discussion and further research into the challenges of publishing, consuming, and integrating structured data on the Web as well as mining knowledge from said data.

1. INTRODUCTION

On the 10th anniversary of the LDOW workshop series, we look back on a decade of Linked Data research, and the broader technical and commercial impact of Linked Data principles and technologies. The Linked Data ethos has always reflected a pragmatic interpretation of the Semantic Web vision; one where the role of the Web architecture has been at least as important as the knowledge representation formalisms by which the data is described. In fact the publication of the Linked Data Principles[2] in 2006, and the community initiatives[3] that sprang from them, represented something of a watershed moment, beyond which the *Web* began to receive as much attention as the *semantics* part from those seeking to build a Semantic Web.

The LDOW workshop series has always been the focal point for researchers seeking to absorb, share, and develop new ideas related to Linked Data. In its 10 year history to date the workshop has been instrumental in bringing Linked Data into the mainstream of research in many aspects of computing and related disciplines. This is reflected in the adoption of Linked Data research into the core topics of top-tier conferences and journals over the last decade.

This period has also seen dramatic changes in the availability and profile of, and reliance on, data in general, not just to the research community, but to society and industry as a whole. Major trends in open data, big data, data science, and the Internet of Things have profoundly changed our awareness of data and the importance we place on them, as well as the challenges they bring. While these trends may come and go, the underlying direction of travel continues to be towards greater availability, coverage, integration, and exploitation of data.

In the same period, the Web has become even more deeply enmeshed in the fabric of society, from commerce and leisure to politics and social discourse. Linked Data sits at the intersection of these trends in data and the Web, and as both grow increasingly ubiquitous. we expect to see the continued influence of Linked Data principles and technologies.

2. A DECADE OF LINKED DATA

The impact of Linked Data research in the last decade has been both extensive and varied. Here we look at some of the milestones of the past 10 years.

LD Knowledge Bases and Vocabularies. One of the early driving forces in the LD initiative was the emergence of knowledge-bases such as *DBpedia*, extracting structured data from Wikipedia; *GeoNames*, describing detailed geographical information; and *MusicBrainz*, describing musical artists and works. Also key to early LD was the re-use of core vocabularies such as *FOAF*, for describing personal information; *DC* for describing basic document meta-data; and *SIOC*, for describing online communities. Together, these formed the core of the earliest LOD cloud. The past ten years have seen the emergence of hundreds of further such datasets and vocabularies. Some key datasets to emerge include *YAGO*, another dataset extracted from Wikipedia with focus on class hierarchies and temporal information; *Wikidata*, a collaborative knowledge-base designed to complement Wikipedia; *LinkedGeoData*, providing a structured-data export from OpenStreetMap; and *Web Data Commons*, collecting embedded meta-data extracted from billions of webpages found in the *Common Crawl*. Key vocabularies that have emerged include *schema.org*, for describing entities mentioned in webpages; *SKOS*, for defining taxonomic hierarchies; *VoID*, for publishing meta-data about datasets; *PROV*, for tracing provenance of datasets, *SSN* for sensors and their observations. New datasets and vocabularies continue to emerge while existing initiatives grow and mature.

Data Portals. As new datasets and vocabularies emerged, the potential of using LD principles and tools to standardise the publication – and enable the subsequent integration – of related datasets within a specific domain began to be re-

[1]http://events.linkeddata.org/ldow2017/
[2]http://www.w3.org/DesignIssues/LinkedData.html
[3]https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3055510

alised. Key to this was the publication of the *5⋆'s of Linked Open Data*, which outlined how applying the Linked Data principles can enhance the reusability of structured data published on the Web. This led to the emergence of Linked Data Portals in various domains. Governments – spurred on by the growing political demand for transparency – were a major adopter, with *data.gov.uk* and *data.gov* publishing multitudinous datasets in LD for the British and U.S. governments, respectively; other public organisations have since followed suit, with, e.g., the *EU Data Portal* adopting LD technology. Other portals emerged focussing on scientific data, with *Bio2RDF* and *OpenPhacts* offering an entry-point onto various key biomedical datasets; the *RKBexplorer* portal aggregating bibliographical meta-data from various sources; and so forth. Other initiatives emerged from the cultural heritage domain, where the *Europeana* portal offers multi-lingual meta-data on items found in museums, libraries and other cultural institutions. Other Data Portals emerged to catalogue datasets available across domains, including *DataHub*, *LODStats*, and *LOD Laundromat*.

Standardisation. As LD began to achieve more and more adoption, new requirements were emerging for the Semantic Web standards, leading to the recommendation of new versions. An important achievement was the standardisation of *SPARQL* in 2008: the first standard query language for RDF; this was extended with a variety of features in 2013, very much informed by emerging requirements for LD. *OWL* received an update in 2009, with new features and new tractable profiles emerging to deal with larger-scale applications. *RDF* later received a major update in 2014, key to which was the standardisation of various syntaxes – such as *Turtle* – that had found favour in the LD community; various flavours of the *JSON-LD* and *RDFa* syntaxes were also standardised. Another major development was the 2015 standardisation of the *Linked Data Platform* (*LDP*), which paved the cowpaths worn by the LD community. Other standardisation efforts included the recommendation of *RDB2RDF* methods and languages for exporting relational data to RDF; *SKOS* for describing taxonomic information; *PROV* for capturing provenance; and others besides. Standardisation continues, with (just for example) the *RDF Data Shapes* Working Group soon to recommend a language for specifying structural constraints on RDF graphs.

LD Tools and Applications. As new standards emerged, so too did implementations thereof, with a wide range of parsers, query engines, reasoners, extractors, etc., being released (mostly open source). Aside from standards, however, other core challenges quickly became clear in LD adoption. In particular, the question of how to automatically link datasets – how to realise the fourth LD principle at large-scale – emerged as key for the LD vision; various proposals and tools for linking datasets thus emerged, with *SILK* and *LIMES* becoming popular frameworks amongst LD publishers. On the flip side, another key question was how to consume LD: how could end-users search and browse the collective data that the Web of Data provides? A number of data search engines emerged, key amongst which was *Sindice* and the related *Sig.ma* interface, which allowed for semantic search. Other applications, such as *Tabulator* or *Haystack*, allow for live browsing of Linked Data. The use of LD technologies – albeit to various degrees of adoption – likewise found use within major companies; with the BBC building LD-powered websites for various major sporting events; *IBM Watson* incorporating LD knowledge-bases for its Jeopardy-playing engine; Facebook proposing its *Open Graph* protocol to decentralise certain social features; and Google offering novel semantic search features powered by *schema.org* and the *Google Knowledge Graph*.

Research. A variety of key issues have been tackled in the LD literature, relating to linking, decentralisation, data quality, amongst others. The resulting literature has been key to realising the aforementioned achievements and others besides. When LDOW began, LD research was a niche topic, but a decade on, such research can be found published not only in Semantic Web and Web venues, but also in areas such as AI, Databases, Data Mining, etc. There are still many open questions left to tackle, and the LD literature is sure to continue to expand in the next decade and onward.

3. ONWARD TO THE NEXT DECADE

Aside from the presentation of eight technical papers on a variety of emerging Linked Data research topics, this year's workshop will also include an open session to discuss the trends that are currently observed concerning the publication and consumption of structured data from the Web as well as the challenges that arise from these trends. We expect the discussion to touch upon points such as:

- the roles of the different data publication mechanisms in the context of the Web of Data, ranging from data portals publishing CSV files, over websites annotating simple data structures using schema.org, to networks of data providers setting RDF links in order to ease the integration of complex structured data.
- the relation between corporate knowledge graphs and the effort that is invested into cleansing them in order to present users with a single truth; and the decentralized nature of the Web of Data, which allows everybody to make statements about everything leading to higher heterogeneity and a more complex concept of truth on the one side, but multiple perspectives on the other side.
- mechanisms and incentives for increasing the degree of interlinking in structured data on the Web, especially, e.g., *schema.org* data.
- an increased focus on the role of, and support for, graph-based data (such as RDF) in popular big data processing frameworks.
- the role of Linked Data in enabling data discovery on the Web, irrespective of data model or serialisation.

Within the open session, we will also discuss the next steps for the LDOW workshop series to better understand its role and influence as part of the broader Web Science community, as well as initiatives that emerge around the workshop. One of the emerging initiatives is proposed as "Pioneering the Linked Open Research Cloud", which is intended to encourage researchers to apply techniques and best practices in Linked Data to scholarly communication. The intention of this initiative is to ensure that research articles as well as peer-reviews for example are better discoverable, accessible, interactive, and more social through the use of native Web technologies and standards.