Universität Leipzig
Fakultät für Mathematik und Informatik
Institut für Informatik

Master Thesis

DataID
Towards Semantically Rich Metadata for Complex Datasets

Abstract: <Gegenstand und Resultate der Arbeit. Was ist neu? Warum sollte man die Arbeit lesen?>

Leipzig, Oktober 2016

vorgelegt von
Markus Freudenberg
Studiengang Informatik

Betreuender Hochschullehrer:

Dr. Sebastian Hellmann
Fakultät für Mathematik und Informatik
Betriebliche Informationssysteme, Semantic Web

Acknowledgements

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Schreibhilfen

Das folgende Kapitel dient der Themenabgrenzung und soll gibt Orientierung bei der Arbeit. Es ist in der fertigen Arbeit nicht zu finden.

Die Kapitel Mindmap, Exposé und Zeitplan sollten VOR Beginn der Arbeit erstell werden und beim Betreuer abgegeben werden; das ist nicht direkter Bestandteil der schriftlichen Ausarbeitung der Bachelor/Master/Diplomarbeit.

1.1 Einfache Worte

Hier wird mit einfachen Worten, also ohne Fachwörter zu benutzen, das Thema in einem Satz beschrieben.

1.2 Wissenschaftlicher Dreisatz

Nachfolgend der wissenschaftliche Dreisatz, um das Thema und das Vorgehen näher zu beschreiben.

Thema Hier wird das Thema kurz und bündig wiedergegeben, z.B.: Ich forsche an / arbeite an / untersuche / beschäftige mich mit etc.

Erkenntnisinteresse Nachfolgend wird die Fragestellung herausgehoben, z.B.: Weil ich herausfinden möchte was / wie / warum / ob etc.

Absicht und Ziele Hier werden die Ziele und Absichten dargestellt, also : Um zu zeigen warum, wie, weshalb etc. Es folgt eine Auflistung der Ziele und Teilziele:

1. Darstellung der Methode.
2. Darstellung der Funktionsweise.

3. Unterlegen mit Messerwerten.
4. etc...

1.3 Mindmap

Die nachfolgende Mindmap stellt die Zusammenhänge der Arbeit dar. Dabei sollten die Wissensgebiete sowie noch zu klärende Themen abgesteckt werden. Das Vorgehen sollte dabei ersichtlich sein.

1.4 Exposé

Das nachfolgende Exposé gibt einen Abriss der geplanten Arbeit. Behandelt werden muss das Thema, die Motivation, die Methoden, das Vorgehen, eine Literaturliste sowie das zu erwartende Ergebnis.

1.5 Zeitplan

Nach dem Exposé folgt ein grober Zeitplan bzw. Projektplan. Dieser sollte die wesentlichen Arbeitsschritte, Meilensteine und Konsultationstermine enthalten. Beispiele sind etwa die Literaturrecherche, das Design, die Implementierung sowie ein oder mehrere Prototypen-Meilensteine. Auch die Schreibarbeit sollte geplant werden. Beispiele ist der Abschluss des Inhaltsverzeichnisses, der Abschluss einer Stichpunktfassung, die erste Grobfassung etc. Auch eine gewisse Zeit für Korrekturen sollte eingeplant werden.

2 Introduction

2.1 Motivation

In 2006, Clive Humby coined the phrase "the new oil" for (digital) data[1], heralding the ever-expanding realm of what is now summarised as: Big Data. Attributed with the same transformative and wealth-producing abilities, once connected to crude oil bursting out of the earth, data has become a cornerstone of economical and societal visions. In fact, the amount of data generated around the world has increased dramatically over the last years, begging the question if those visions have already come to pass.

The steep increase in data produced can be ascribed to multiple factors. To name just a few:

- The growth in content and reach of the World Wide Web.
- The digitalising of former analogue data.
- The realisation of what is called the Internet of Things (IoT)[2].
- The shift of classic fields of research and industry to computer-aided processes and digital resource management (e.g. digital humanities, industry 4.0).
- Huge data collections about protein sequences or human disease taxonomies are established in the life sciences.
- Research areas like natural language processing or machine learning are generating and refining data.
- In addition, open data initiatives like the Open Knowledge Foundation are following the call for 'Raw data, Now!'[3] of Tim Berners-Lee, demanding open data from governments and organisations.

[1] https://www.theguardian.com/technology/2013/aug/23/tech-giants-data
[2] http://siliconangle.com/blog/2015/10/28/page/3#post-254300
[3] http://www.wired.co.uk/news/archive/2012-11/09/raw-data

As a new discipline, data engineering is dealing with the fallout of this trend, namely with issues of how to extract, aggregate, store, refine, combine and distribute data of different sources in ways which give equal consideration to the four V's of Big Data: Volume, Velocity, Variety and Veracity[4].

Datasets are the building blocks of these endeavours. They are the combination of multiple datums bundled together by at least one dimension of distinction (such as source, topic or category). When working with these bricks of information, additional data about datasets (or metadata) is needed. Dataset metadata enables users to discover, understand and (automatically) process the data it holds, as well as providing provenance on how a dataset came into existence. This metadata is often created, maintained and stored in diverse data repositories featuring disparate data models that are often unable to provide the metadata necessary to automatically process the datasets described. In addition, many use cases for dataset metadata call for more specific information than provided by most available metadata vocabularies, depending on the use case at hand. Extending existing metadata models to fit these scenarios is a cumbersome process resulting often in non-reusable solutions.

One vocabulary for dataset metadata is breaking this trend. Since its introduction in 2013, the Data Catalog Vocabulary [DCAT] vocabulary, a W3C recommendation, has been widely adopted as a foundation for dataset metadata in research, government and industry [needs at least one link to back this up]. The very general approach adopted by the authors of DCAT allows for portraying any given (digital) object with this ontology. Extending DCAT is very easy and mappings to other metadata formats are not difficult to achieve.

Conversely, the general approach of DCAT is often too imprecise where specificity is needed. A short list of the more pressing issues resulting from this imprecisness:

- Insufficient provenance information
- Missing relations between Datasets
- Relations to agents are too cursory
- Technical description of web-resources is lacking, restricting the accessibility of the data
- General lack of specificity, inviting non-machine-readable expressions of resources

[4] http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Similar findings were concluded at the W3C/VRE4EIC workshop Smart Descriptions & Smarter Vocabularies (SDSVoc) in 2016 [reference to the proceedings (not yet published)]. As a result of lacking specificity, current representations of datasets with DCAT are often not contributing to the main benefits for publishing data on the web [DWBP] : Reuse, Comprehension, Linkability, Discoverability, Trust, Access, Interoperability and Processability. This, in turn, amplifies broader problems with published datasets, especially in the open data community, reflected by the Open Data Strategy[needs link], defining the following six barriers for "open public data" [needs link], proposed by the European Commission in 2011:

1. a lack of information that certain data actually exists and is available,
2. a lack of clarity of which public authority holds the data,
3. a lack of clarity about the terms of re-use,
4. data made available in formats that are difficult or expensive to use,
5. complicated licensing procedures or prohibitive fees,
6. exclusive re-use agreements with one commercial actor or re-use restricted to a government-owned company.

Many issues with DCAT itself or their manifestation in reality can be solved by existing ontologies, even when restricted only to W3C recommended ontologies. For example, the PROV Ontology [PROV-O], deals with questions on how to record provenance information on a very granular level. While the Open Digital Rights Language [ODRL] provides machine readable descriptions of licenses and other policies. The existence of problems, like those listed above, despite these offered solutions, speaks to a larger problem of missing organisational structures for landscaping of vocabularies (offering recommendations on combining, revising and usage of ontologies). A study of 91 commonly used vocabularies concluded[**?**]:

> "Our validation detected a total of 6 typos, 14 missing or unavailable ontologies, 73 language level errors, 310 instances of ontology namespace violations and 2 class cycles which we believe to be errors."

These errors accumulate when strong interdependencies exist between vocabularies, adding logical and practical problems, which make model unification inconsistent.

2.2 Objectives

In this thesis I will present the metadata model of DataID, a multi-layered metadata ecosystem, which, in its core, describes complex datasets and their different manifestations, as well as relations to agents like persons or organisations, in regard to their rights and responsibilities.

Improving the portrayal of PROVENANCE, LICENSING and ACCESS, while maintaining the easy EXTENSIBILITY and INTEROPERABILITY of DCAT, are the linchpin objectives in our effort to present a comprehensive, extensible and interoperable metadata vocabulary. Multiple well established ontologies (such as PROV-O, VOID and FOAF [needs links]) are reused for maximum compatibility to establish a uniform and accepted way to describe and deliver dataset metadata for arbitrary datasets and to put existing standards into practice.

The DataID Ecosystem is a suite of ontologies comprised of DataID core and multiple extension ontologies, clustered around DataID core. It is the result of a modularisation process, which was necessary to preserve EXTENSIBILITY and INTEROPERABILITY of the DCAT vocabulary, on which all ontologies are based.

I want to present my solution for most of the current problems with dataset metadata in general and DCAT in particular, following these objectives:

1. Provide sufficient support for extensive and machine-readable representations for PROVENANCE, LICENSING and data ACCESS.
2. Achieve this by extending DCAT with well established ontologies to solve as many issues as possible (favour W3C recommended ontologies).
3. Show, that by modularising into a landscape of ontologies, DataID preserves the general character of DCAT, supporting EXTENSIBILITY and INTEROPERABILITY.
4. Prove that the resulting ecosystem is capable of serving for complex demands on dataset metadata (proving EXTENSIBILITY).
5. Demonstrate the INTEROPERABILITY with other metadata formats.
6. Evaluate the universal applicability of DataID for datasets in any given domain or scenario.

link

DataID was developed under the sponsorship of the ALIGNED project, following its main goals:

- to be part of a unified software and data engineering process
- describing the complete data lifecycle and domain model
- with an emphasis on data quality and integrity

2.3 Structure

The remainder of this document is structured as follows:

add structure of thesis here

3 Grundlagen

Das Grundlagenkapitel, kurz und knapp. Hier sollten die Themen auftauchen, die später wieder aufgegriffen werden und zum Verständnis der Arbeit nötig sind (mehr nicht). Hier stehen meist Referenzen/ Zitate von Lehrbüchern und anderen Publikationen. Das Kapitel sollte ca. 5 - 10, max. 15 Seiten umfassen.

4 Related Work

The Data Catalog Vocabulary (DCAT) is a W3C Recommendation [?] and serves as a foundation for many available dataset vocabularies and application profiles. In [?] the authors introduce a standardised interchange format for machine-readable representations of government data catalogues. The DCAT vocabulary includes the special class Distribution for the representation of the available materialisations of a dataset (e.g. CSV file, an API or RSS feed). These distributions cannot be described further within DCAT (e.g. the type of data, or access procedures). Applications which utilise the DCAT vocabulary (e.g. datahub.io[5]) provide no standardised means for describing more complex datasets either. Yet, the basic class structure of DCAT (Catalog, CatalogRecord, Dataset, Distribution) has prevailed. Range definitions of properties provided for these classes are general enough to make this vocabulary easy to extend.

DCAT, as opposed to PROV-O, expresses provenance in a limited way using a few basic properties such as `dct:source` or `dct:creator`, thus it does not relate semantically to persons or organisations involved in the publishing, maintenance etc. of the dataset. There is no support or incentive to describe source datasets or conversion activities of transformations responsible for the dataset at hand. This lack is crucial, especially in a scientific contexts, as it omits the processes necessary to replicate a specific dataset, a feature easily obtainable by the use of PROV-O.

Metadata models vary and most of them do not offer enough granularity to sufficiently describe complex datasets in a semantically rich way. For example, CKAN[6](Comprehensive Knowledge Archive Network), which is used as a metadata schema in data portals like datahub.io, partially implements the DCAT vocabulary, but only describes resources associated with a dataset superficially. Additional properties are simple key-value pairs which themselves are linked by `dct:relation` properties. This data model is semantically poor and inadequate for most use cases wanting to automatically consume the data of a dataset.

[5] http://datahub.io/
[6] http://ckan.org/

While not implementing the DCAT vocabulary, META-SHARE [?] does provide an almost complete mapping to DCAT, providing an extensive description of language resources, based on a XSD schema. In addition it offers an exemplary way of describing licenses and terms of reuse. Yet, META-SHARE is specialised on language resources, thus lacking generality and extensibility for other use cases.

Likewise the Asset Description Metadata Schema[7] (ADMS) is a profile of DCAT, which only describes a specialised class of datasets: so-called Semantic Assets. Highly reusable metadata (e.g. code lists, XML schemata, taxonomies, vocabularies etc.), which is comprised of relatively small text files.

DCAT-AP (DCAT Application Profile for data portals in Europe[8]) is a profile, extending DCAT with some ADMS properties. It has been endorsed by the ISA Committee in January of 2016[9]. Due to the stringent cardinality restrictions, extending DCAT-AP to serve more elaborate purposes will prove difficult. As remarked in section 7 the representation of different agent roles is lacking in the current version of DCAT-AP. Neither DCAT-AP nor ADMS give any consideration to defining responsibilities of agents, extending provenance or providing thorough machine-readable licensing information.

Similar problems afflicted the previous version of the DataID ontology[?]. Rooted in the Linked Open Data world, it neglected important information or provided properties (e.g. `dataid:graphName`) which are orphans outside this domain. While already importing the PROV-O ontology, it was lacking a specific management of rights and responsibilities.

Aufführung anderer bzw. ähnlichen Arbeiten zum Thema. Hier stehen die meisten Zitate zu *wissenschaftlichen* Publikationen. Dies müssen nicht nur aktuelle Paper sein, sondern können auch Publikationen älteren Datums sein. Die Prägnanz, die Generalisierung und der Vergleiche stehen hierbei im Vordergrund. Auch kann eine Diskussion enthalten sein.

Das Kapitel sollte so viele Seiten wie nötig umfassen, um eine Einordnung des Sachverhalts gegenüber anderen Gebieten zu erreichen. Der Richtwert sind 5 bis max. 10 Seiten.

[7] https://www.w3.org/TR/vocab-adms/

[8] https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/dcat-ap-v11

[9] https://joinup.ec.europa.eu/community/semic/news/dcat-ap-v11-endorsed-isa-committee

5 Entwurf

In diesem Kapitel wird der Entwurf des System beschrieben. Dies soll kein Lasten / Pflichtenheft sein oder die Anforderungen bis ins kleinste Details aufzählen. Vielmehr sollen wichtige Design Entscheidungen erklärt werden. Hierzu gehören das Design, Konzepte und Modelle.

Wichtig ist, dass die eigenen Ideen und Beiträge als solche hervorgehoben werden. In Bachelorarbeiten sollte hierbei das gelernte Wissen geschickt angewandt werden. In Master und Diplom sollten hierbei wissenschaftliche Innovationen oder die kreative Anwendung anklingen. In einer Dissertation muss hier klar der wissenschaftliche Fortschritt (eine Erfindung, eine Neuerung) sowie der Nutzen ersichtlich sein.

Dies ist das Hauptkapitel der Arbeit und sollte so viele Seiten wie nötig enthalten.

6 Implementierung

In diesem Kapitel werden Implementierungsdetails behandelt. Dabei wird der Aufbau des System, die eigentliche Umsetzung, Konfigurationen und die Handhabung betrachtet. Das Kapitel sollte sich nicht in Details verlieren aber dennoch die Anwendung behandeln.

Das Kapitel sollte soviel wie nötige Seiten umfassen, dabei aber knapp gehalten werden!

7 Auswertung

In diesem Kapitel folgt die Auswertung der Arbeit. Hier werden Messerreihen wie z.B. die Latenzzeit, die Performanz, der Speicherverbrauch, die Erkennungsrate etc. aufgeführt. Ferner werden die Methoden erläutert, die Ergebnisse interpretiert und diskutiert.

Auch hier gibt es keinen Richtwert, das Kapitel sollte umfassend wie nötig sein.

8 Zusammenfassung und Ausblick

Am Ende der Arbeit steht die Zusammenfassung, die alle wichtigen Punkte und Ergebnisse der Arbeit in einfachen Worten wiedergibt. Anschließend folgt ein Ausblick auf anschließende Arbeiten und Themenvorschläge.

Das Kapitel sollte zwischen einer und drei Seiten umfassen.

Glossar

Vocabulary On the Semantic Web, vocabularies define the concepts and relationships (also referred to as "terms") used to describe and represent an area of concern. Vocabularies are used to classify the terms that can be used in a particular application, characterize possible relationships, and define possible constraints on using those terms. In practice, vocabularies can be very complex (with several thousands of terms) or very simple (describing one or two concepts only)[**?**].

Ontology There is no clear division between what is referred to as "vocabularies" (see Vocabulary) and "ontologies". The trend is to use the word "ontology" for more complex, and possibly quite formal collection of terms, whereas "vocabulary" is used when such strict formalism is not necessarily used or only in a very loose sense. Vocabularies are the basic building blocks for inference techniques on the Semantic Web[**?**].

Abkürzungsverzeichniss

<falls viele Abkürzungen vorkommen>

TLA Three Letter Acronym

TLB Translation Lookaside Buffer

RTFM ...

Literaturverzeichnis

Erklärung

"'Ich versichere, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe, insbesondere sind wörtliche oder sinngemäße Zitate als solche gekennzeichnet. Mir ist bekannt, dass Zuwiderhandlung auch nachträglich zur Aberkennung des Abschlusses führen kann"'.

Ort	Datum	Unterschrift