HDL - Towards a Harmonized Dataset Model for Open Data Portals

Ahmad Assaf[1,2], Raphaël Troncy[1] and Aline Senart[2]

[1] EURECOM, Sophia Antipolis, France, <firstName.lastName@eurecom.fr>
[2] SAP Labs France, <firstName.lastName@sap.com>

Abstract. The Open Data movement triggered an unprecedented amount of data published in a wide range of domains. Governments and corporations around the world are encouraged to publish, share, use and integrate Open Data. There are many areas where one can see the added value of Open Data, from transparency and self-empowerment to improving efficiency, effectiveness and decision making. This growing amount of data requires rich metadata in order to reach its full potential. This metadata enables dataset discovery, understanding, integration and maintenance. Data portals, which are considered to be datasets' access points, offer metadata represented in different and heterogenous models. In this paper, we first conduct a unique and comprehensive survey of seven metadata models: CKAN, DKAT, Public Open Data, Socrata, VoID, DCAT and Schema.org. Next, we propose HDL, an harmonized dataset model based on this survey. We describe use cases that show the benefits of providing rich metadata to enable dataset discovery, search and spam detection.

Keywords: Dataset Metadata, Dataset Profile, Dataset Model, Data Quality

1 Introduction

Open data is the data that can be easily discovered, reused and redistributed by anyone. It can include anything from statistics, geographical data, meteorological data to digitized books from libraries. Open data should have both legal and technical dimensions. It should be placed in the public domain under liberal terms of use with minimal restrictions and should be available in electronic formats that are non-proprietary and machine readable. Open Data has major benefits for citizens, businesses, society and governments: it increases transparency and enables self-empowerment by improving the visibility of previously inaccessible information; it allows citizens to be better informed about policies, public spending and activities in the law making processes. Moreover, it is still considered as a gold mine for organizations which are trying to leverage external data sources in order to produce more informed business decisions [22], despite the legal issues surrounding Linked Data licenses [71].

The Linked Data publishing best practices [25] specifies that datasets should contain metadata needed to effectively understand and use them. *Metadata* is

structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource [118]. Having rich metadata helps in enabling:

- **Data discovery, exploration and reuse**: In [136], it was found that users are facing difficulties finding and reusing publicly available datasets. Metadata provides an overview of datasets making them more searchable and accessible. High quality metadata can be at times more important than the actual raw data especially when the costs of publishing and maintaining such data is high.
- **Organization and identification**: The increasing number of datasets being published makes it hard to track, organize and present them to users efficiently. Attached metadata helps in bringing similar resources together and distinguish useful links.
- **Archiving and preservation**: There is a growing concern that digital resources will not survive in usable forms to the future [118]. Metadata can ensure resources survival and continuous accessibility by providing clear provenance information to track the lineage of digital resources and detail their physical characteristics.

The value of Open Data is recognized when it is used. To ensure that, publishers need to enable people to find datasets easily. Data portals are specifically designed for this purpose. They make it easy for individuals and organizations to store, publish and discover datasets. The data portals can be public like Datahub[3] and the Europe's Public Data portal[4] or private like Quandl[5] and Engima[6]. The data available in private portals is of higher quality as it is manually curated but in lesser quantity compared to what is available in public portals. Similarly, in some public data portals, administrators manually review datasets information, validate, correct and attach suitable metadata information.

Data models vary across data portals. While exhaustively surveying the range of data models, we did not find any that offers enough granularity to completely describe complex datasets facilitating search, discovery and recommendation. For example, the Datahub uses an extension of the Data Catalog Vocabulary (DCAT) [48] which prohibits a semantically rich representation of complex datasets like DBpedia[7] that has multiple endpoints and thousands of dump files with content in several languages [98]. Moreover, to properly integrate Open Data into business, a dataset should include the following information:

- *Access information*: a dataset is useless if it does not contain accessible data dumps or query-able endpoints;
- *License information*: businesses are always concerned with the legal implications of using external content. As a result, datasets should include both

[3] `http://datahub.io`
[4] `http://publicdata.eu`
[5] `https://quandl.com/`
[6] `http://enigma.io/`
[7] `http://dbpedia.org`

machine and human readable license information that indicates permissions, copyrights and attributions;
- *Provenance information*: depending on the dataset license, the data might not be legally usable if there are no information describing its authoritative and versioning information. Current models under-specify these aspects limiting the usability of many datasets.

In this paper, we perform a comprehensive survey of the main data portals and dataset models, that is: CKAN, DKAT, Public Open Data, Socrata, VoID, DCAT and Schema.org. We further analyze these models and suggest a classification for metadata information. Based on this classification, we propose HDL, an harmonized dataset model that addresses the shortcomings of existing dataset models. The remainder of the paper is structured as follows. In Section 2, we present the existing dataset models used by various data portals. In Section 3, we present our classification for the different metadata information. In Section 4, we describe our proposed model and suggest a set of best practices to ensure proper metadata presentation and we finally conclude and outline some future work in Section 5.

2 Data Portals and Dataset Models

There are many data portals that host a large number of private and public datasets. Each portal present the data based on a model used by the underlying software. In this section, we present the results of our landscape survey of the most common data portals and dataset models.

2.1 DCAT

The Data Catalog Vocabulary (DCAT) is a W3C recommendation that has been designed to facilitate interoperability between data catalogs published on the Web [48]. The goal behind DCAT is to increase datasets discoverability enabling applications to easily consume metadata coming from multiple sources. Moreover, the authors foresee that aggregated DCAT metadata can facilitate digital preservation and enable decentralized publishing and federated search.

DCAT is an RDF vocabulary defining three main classes: `dcat:Catalog`, `dcat:Dataset` and `dcat:Distribution`. We are interested in both the `dcat:Dataset` class which is a collection of data that can be available for download in one or more formats and the `dcat:Distribution` class which describes the method with which one can access a dataset (e.g. an RSS feed, a REST API or a SPARQL endpoint).

2.2 DCAT-AP

The DCAT application profile for data portals in Europe (DCAT-AP)[8] is a specialization of DCAT to describe public section datasets in Europe. It defines a

[8] `https://joinup.ec.europa.eu/asset/dcat\_application\_profile/description`

minimal set of properties that should be included in a dataset profile by specifying mandatory and optional properties. The main goal behind it is to enable cross-portal search and enhance discoverability. DCAT-AP has been promoted by the Open Data Support[9] to be the standard for describing datasets and catalogs in Europe.

2.3 ADMS

The Asset Description Metadata Schema (ADMS) [114] is also a profile of DCAT. It is used to semantically describe assets. An asset is broadly defined as something that can be opened and read using familiar desktop software (e.g. code lists, taxonomies, dictionaries, vocabularies) as opposed to something that needs to be processed like raw data. While DCAT is designed to facilitate interoperability between data catalogs, ADMS is focused on the assets within a catalog.

2.4 VoID

VoID [29] is another RDF vocabulary designed specifically to describe linked RDF datasets and to bridge the gap between data publishers and data consumers. In addition to dataset metadata, VoID describes the links between datasets. VoID defines three main classes: `void:Dataset`, `void:Linkset` and `void:subset`. We are specifically interested in the `void:Dataset` concept. VoID conceptualizes a dataset with a social dimension. A VoID dataset is a collection of raw data, talking about one or more topics, originates from a certain source or process and accessible on the web.

2.5 CKAN

CKAN[10] is the world's leading open-source data management system (DMS). It helps users from different domains (national and regional governments, companies and organizations) to easily publish their data through a set of workflows to publish, share, search and manage datasets. CKAN is the portal powering web sites like Datahub, the Europe's Public Data portal or the U.S Government's open data portal[11].

CKAN is a complete catalog system with an integrated data storage and powerful RESTful JSON API. It offers a rich set of visualization tools (e.g. maps, tables, charts) as well as an administration dashboard to monitor datasets usage and statistics. CKAN allows publishing datasets either via an import feature or through a web interface. Relevant metadata describing the dataset and its resources as well as organization related information can be added. A Solr[12] index is built on top of this metadata to enable search and filtering.

[9] `http://opendatasupport.eu`
[10] `http://ckan.org`
[11] `http://data.gov`
[12] `http://lucene.apache.org/solr/`

The CKAN data model[13] contains information to describe a set of entities (dataset, resource, group, tag and vocabulary). CKAN keeps the core metadata restricted as a JSON file, but allows for additional information to be added via "extra" arbitrary key/value fields. CKAN supports Linked Data and RDF as it provides a complete and functional mapping of its model to Linked Data formats.

2.6 DKAN

DKAN[14] is a Drupal-based DMS with a full suite of cataloging, publishing and visualization features. Built over Drupal, DKAN can be easily customized and extended. The actual data sets in DKAN can be stored either within DKAN or on external sites. DKAN users are able to explore, search and describe datasets through the web interface or a RESTful API.

The DKAN data model[15] is very similar to the CKAN one, containing information to describe datasets, resources, groups and tags.

2.7 Socrata

Socrata[16] is a commercial platform to streamline data publishing, management, analysis and reusing. It empowers users to review, compare, visualize and analyze data in real time. Datasets hosted in Socrata can be accessed using RESTful API that facilitates search and data filtering.

Socrata allows flexible data management by implementing various data governance models and ensuring compliance with metadata schema standards. It also enables administrators to track data usage and consumption through dashboards with real-time reporting. Socrata is very flexible when it comes to customizations. It has a consumer-friendly experience giving users the opportunity to tell their story with data. Socrata's data model is designed to represent tabular data: it covers a basic set of metadata properties and has good support for geospatial data.

2.8 Schema.org

Schema.org[17] is a collection of schemas used to markup HTML pages with structured data. This structured data allows many applications, such as search engines, to understand the information contained in Web pages, thus improving the display of search results and making it easier for people to find relevant data.

Schema.org covers many domains. We are specifically interested in the `Dataset` schema. However, there are many classes and properties that can be used to describe organizations, authors, etc.

[13] http://docs.ckan.org/en/ckan-1.8/domain-model.html
[14] `http://nucivic.com/dkan/`
[15] `http://docs.getdkan.com/dkan-documentation/dkan-developers/dataset-technical-field-reference/`
[16] `http://socrata.com`
[17] `http://schema.org`

2.9 Project Open Data

Project Open Data (POD)[18] is an online collection of best practices and case studies to help data publishers. It is a collaborative project that aims to evolve as a community resource to facilitate adoption of open data practices and facilitate collaboration and partnership between both private and public data publishers.

The POD metadata model[19] is based on DCAT. Similarly to DCAT-AP, POD defines three types of metadata elements: Required, Required-if(conditionally required) and Expanded (optional). The metadata model is presented in the JSON format and encourages publishers to extend their metadata descriptions using elements from the "Expanded Fields" list, or from any well-known vocabulary.

3 Metadata Classification

A dataset metadata model should contain sufficient information so that consumers can easily understand and process the data that is described. After analyzing the models described in the section 2, we find out that a dataset can contain four main sections:

- **Resources**: The actual raw data that can be downloaded or accessed directly via queryable endpoints. Resources can come in various formats such as JSON, XML or RDF.
- **Tags**: Descriptive knowledge about the dataset content and structure. This can range from simple textual representation to semantically rich controlled terms. Tags are the basis for datasets search and discovery.
- **Groups**: Groups act as organizational units that share common semantics. They can be seen as a cluster or a curation of datasets based on shared categories or themes.
- **Organizations**: Organizations are another way to arrange datasets. However, they differ from groups as they are not constructed by shared semantics or properties, but solely on the dataset's association to a specific administration party.

Upon closed examination of the various data models, we group the metadata information into eight main types. Each section discussed above should contain one or more of these types. For example, resources have general, access, ownership and provenance information while tags have general and provenance information only. The eight information types are:

- **General information**: The core information about the dataset (e.g., title, description, ID). The most common vocabulary used to describe this information is Dublin Core[20].

[18] `http://project-open-data.cio.gov/`
[19] `https://project-open-data.cio.gov/v1.1/schema/`
[20] `http://dublincore.org/documents/dcmi-terms/`

- **Access information**: Information about dataset access and usage (e.g., URL, license title and license URL). In addition to the properties in the models discussed above, there are several vocabularies designed specially to describe data access right e.g. Linked Data Rights[21], the Open Digital Rights Language (ODRL)[22].
- **Ownership information**: Authoritative information about the dataset (e.g. author, maintainer and organization). The common vocabularies used to expose ownership information are Friend-of-Friend (FOAF)[23] for people and relationships, vCard [120] for people and organizations and the Organization ontology [35] designed specifically to describe organizational structures.
- **Provenance information**: Temporal and historical information about the dataset creation and update records, in addition to versioning information (e.g. creation data, metadata update data, latest version). Provenance information coverage varies across the modeled surveyed. However, its great importance lead to the development of various special vocabularies like the Open Provenance Model[24] and PROV-O [130]. DataID [98] is an effort to provide semantically rich metadata with focus on providing detailed provenance, license and access information.
- **Geospatial information**: Information reflecting the geographical coverage of the dataset represented with coordinates or geometry polygons. There are several additional models and extensions specifically designed to express geographical information. The Infrastructure for Spatial Information in the European Community (INSPIRE) directive[25] aims at establishing an infrastructure for spatial information. Mappings have been made between DCAT-AP and the INSPIRE metadata. CKAN provides as well a spatial extension[26] to add geospatial capabilities. It allows importing geospatial metadata from other resources and supports various standards (e.g. ISO 19139) and formats (e.g. GeoJSON).
- **Temporal information**: Information reflecting the temporal coverage of the dataset (e.g. from date to date). There has been some notable work on extending CKAN to include temporal information. `govdata.de` is an Open Data portal in Germany that extends the CKAN data model to include information like `temporal_granularity`, `temporal_coverage_to` and `temporal_granularity_from`.
- **Statistical information**: Statistical information about the data types and patterns in datasets (e.g. properties distribution, number of entities and RDF triples). This information is particularly useful to explore a dataset as it gives detailed insights about the raw data when provided properly. VoID is the only model that provides statistical information about a dataset. VoID defines properties to express different statistical characteristics of datasets like

[21] `http://oeg-dev.dia.fi.upm.es/licensius/static/ldr/`
[22] `http://www.w3.org/ns/odrl/2/`
[23] `http://xmlns.com/foaf/spec/`
[24] `http://open-biomed.sourceforge.net/opmv/`
[25] `http://inspire.ec.europa.eu/`
[26] `https://github.com/ckan/ckanext-spatial`

the total number of triples, total number of entities, total number of distinct classes, etc. However, there are other vocabularies such as SCOVO [102] that can model and publish statistical data about datasets.
– **Quality information**: Information that indicates the quality of the dataset on the metadata and instance levels. In addition to that, a dataset should include an openness score that measures its alignment with the Linked Data publishing standards [129]. Quality information is only expressed in the POD metadata. However, `govdata.de` extends the CKAN model also to include a `ratings_average` field. Moreover, there are various other vocabularies like daQ [36] that can be used to express datasets quality. The RDF Review Vocabulary[27] can also be used to express reviews and ratings about the dataset or its resources.

4 Towards A Harmonized Model

Since establishing a common vocabulary or model is the key to communication, we identified the need for an harmonized dataset metadata model containing sufficient information so that consumers can easily understand and process datasets. To create the mappings between the different models, we performed various steps:

– Examine the model or vocabulary specification and documentation.
– Examine existing datasets using these models and vocabularies. `http://dataportals.org` provides a comprehensive list of Open Data Portals from around the world. It was our entry point to find out portals using CKAN or DKAN as their underlying DMS. We also investigated portals known to be using specific DMS. Socrata, for example, maintains a list of Open Data portals using their software on their homepage such as `http://pencolorado.org` and `http://data.maryland.gov`.
– Examine the source code of some portals. This was specifically the case for Socrata as their API returns the raw data serialized as JSON rather than the dataset's metadata. As a consequence, we had to investigate the Socrata Open Data API (SODA) source code[28] and check the different classes and interfaces.

CKAN	DKAN	POD	DCAT	VoID	Schema.org	Socrata
resources	resources	distribution	dcat:Distribution	void:Dataset→ void:dataDump	CreativeWork:keywords	attachments
tags	tags	keyword	dcat:Dataset→ :keyword	void:Dataset→ :keyword	Dataset:distribution	tags
groups	groups	theme	dcat:Dataset→ :theme	-	CreativeWork:about	category
organization	organization	publisher	dcat:Dataset→ :publisher	void:Dataset→ :publisher	-	-

Table 1. Data models sections mapping

The first task is to map the four main information sections across those models. Table 1 shows our proposed mappings. For the ontologies (DCAT, VoID),

[27] `http://vocab.org/review/`
[28] https://github.com/socrata/soda-java/tree/master/src/main/java/com/socrata/model

the first part represents the class and the part after → represents the property. For Schema.org, the first part refers to the schema and the second part after : refers to the property.

Table 2 presents the full mappings between the models across the information groups. Entries in the CKAN marked with ∗ are properties from CKAN extensions and not included in the original data model. Similar to the sections mappings, for the ontologies (DCAT, VoID), the first part represents the class and the part after → represents the property. However, sometimes the part after → refers to another resource. For example, to describe the dataset's maintainer email in DCAT, the information should be presented in the `dcat:Dataset` class using the `dcat:contactPoint` property. However, the range of this property is a resource of type `vcard` which has the property `hasEmail`.

For Schema.org, similar to the sections mapping, the first part refers to the schema and the second part after : refers to the property. However, if the property is inherited from another schema we denote that by using a → as well. For example, the size of a dataset is a property for a `Dataset` schema specified in its `distribution` property. However, the type of `distribution` is `dataDownload` which is inherited from the `MediaObject` schema. The size for `MediaObject` is defined in its `contentSize` property which makes the mapping string `Dataset:distribution`→ `DataDownload`→ `MediaObject:contentSize`.

Examining the different models, we noticed a lack of a complete model that covers all the information types. There is an abundance of extensions and application profiles that try to fill in those gaps, but they are usually domain specific addressing specific issues like geographic or temporal information. To the best of our knowledge, there is still no complete model that encompasses all the described information types.

HDL aims at filling this gap by taking the best from these models. HDL is currently modeled in JSON[29] but converting it to a standalone OWL ontology is part of our future work.

The CKAN model controls the values to be used in describing some dataset properties. For example, the `resource_type` property can have the values: file: direct accessible bitstream, file.upload: file uploaded to the CKAN FileStore[30], api, visualization, code: the actual source code or a reference to a code repository and documentation. However, using the Roomba tool [3], we managed to generate portal-wide reports about the representation of various fields in CKAN portals. The goal behind these reports is to find what are the frequent fields data publishers are adding as `extras` fields.

We created two "key:object meta-field values" reports using Roomba. The first one aims to collect the list of `extras` values using the query string `extras>value:extras>name` and the second one is to list the file types specified for resources using the query string `resources>resource_type:resources>name`. We run the report generation process on two prominent data portals: the Linked

[29] `https://github.com/ahmadassaf/opendata-checker/blob/master/model/hdl.json`

[30] `http://docs.ckan.org/en/ckan-1.8/filestore.html`

Open Data (LOD) cloud hosted on the Datahub containing 259 datasets and the Africa's largest open data portal, OpenAfrica[31] that contains 1653 datasets.

After examining the results, we noticed that for OpenAfrica, 53% of the datasets have contain additional information about the geographical coverage of the dataset (e.g. spatial-reference-system, spatial_harvester, bbox-east-long, bbox-north-long, bbox-south-long, bbox-west-long). In addition, 16% of the datasets have additional provenance and ownership information (e.g frequency-of-update, dataset-reference-date). For the LOD cloud, the main information embedded in the extras fields are about the structure and statistical distribution of the dataset (e.g. namespace, number of triples and links). The OpenAfrica resources did not specify any extra resource types. However, in the LOD cloud, we observe that multiple resources define additional types (e.g. example, api/sparql, publication, example).

Roomba easily enables to perform such tests and to gather a detailed view about the kind of missing information data publishers require in the core model. We further plan to run Roomba on various portals to collect more information about such missing data to include it in HDL.

5 Conclusion and Future Work

In this paper, we surveyed the landscape of various models and vocabularies that described datasets on the web. Since establishing a common vocabulary or model is the key to communication, we identified the need for an harmonized dataset metadata model containing sufficient information so that consumers can easily understand and process datasets. We have identified four main sections that should be included in the model: resources, groups, tags and organizations. Furthermore, we have classified the information to be included into eight types. Our main contribution is a set of mappings between each properties of those models. This has lead to the design of HDL, an harmonized dataset model, that takes the best out of these models and extends them to ensure complete metadata coverage to enable data discovery, exploration and reuse.

At the moment, HDL is available as a hierarchical JSON file. As part of our future work, we plan to refine HDL and present it as a fully fledged OWL ontology. At the moment, HDL contains some values that were frequently defined in CKAN extras fields. However, we plan to broaden our analysis of these values by running Roomba on additional portals and present the top results as enumerations, ensuring a fine-grained representation of a dataset. We further plan to create mappings between HDL and all the various models to ensure full compatibility. These mappings, for example, can be used to extend Roomba allowing it to perform metadata profiling on other portals like DKAN. Finally, we plan to create a set of supporting tools that allow validation of generation of HDL profiles.

[31] `http://africaopendata.org/`

Table 2: Harmonized Dataset Models Mappings

Data Model	CKAN	DKAN	POD	DCAT	VoID	Schema.org	Socrata
General Information	id	id	identifier	dcat:Dataset→ dct:identifier			id/externalId
	private	private	accessLevel				privateMetadata
	state	state					publicationStage
	type	type				Thing:additionalType	
	name	name				Thing:name	name
	isopen						
	notes	notes	description	dcat:Dataset→ dct:description	void:Dataset→ dct:description	Thing:description	description
	title	title	title	dcat:Dataset→ dct:title	void:Dataset→ dc:title	Thing:name	name
	num_resources				void:Dataset→ void:documents		
	num_tags						
			conformsTo	dcat:Dataset→ dct:conformsTo	void:Dataset→ dct:conformsTo		
			language	dcat:Dataset→ dct:language	void:Dataset→ dct:language	CreativeWork:inLanguage	
			accuralPeriodicity	dcat:Dataset→ dct:accuralPeriodicity	void:Dataset→ dct:accuralPeriodicity		
access information	license_title	license_title	license	dcat:Distribution→ dct:license	void:Dataset→ dct:license		license→ name
	license_id						licenseId
	license_url					CreativeWork:license	license → termsLink
	url	url	landingPage	dcat:Dataset→ dcat:landingPage		Thing:url	
			rights	dcat:Distribution→ dct:rights	void:Dataset→ dct:rights		
	attribution_text∗						attribution
							attributionLink
provenance	version					CreativeWork:version	
	revision_id						
	metadata_created	metadata_created		dcat:Distribution→ dct:created	void:Dataset→ dct:created	CreativeWork:dateCreated	
	metadata_modified	metadata_modified	modified	dcat:Distribution→ dct:modified	void:Dataset→ dct:modified	CreativeWork:dateModified	
	revision_timestamp	revision_timestamp					
			issued	dcat:Distribution→ dct:issued	void:Dataset→ dct:issued	CreativeWork:datePublished	
			temporal	dcat:Dataset→ dct:temporal	void:Dataset→ dct:temporal	Dataset:temporal	
ownership	maintainer	maintainer	contactPoint→ fn	dcat:Dataset→ dcat:contactPoint→ vcard:fn		CreativeWork:producer→ Thing:name	owner→ displayName / owner→ ScreenName
	maintainer_email	maintainer_email	contactPoint→ hasEmail	dcat:Dataset→ dcat:contactPoint→ vcard:hasEmail		CreativeWork:producer→ Person:email	
	owner_org					CreativeWork:sourceOrganization:LegalName	
	author			dcat:Dataset→ dct:creator→ foaf:Person:givenName	void:Dataset→ dct:creator→ foaf:Person:givenName	CreativeWork:author→ Thing:name	
	author_email	author_email		dcat:Dataset→ dct:creator→ foaf:Person:mbox	void:Dataset→ dct:creator→ foaf:Person:mbox	CreativeWork:author→ Person:email	
			bureauCode				
			programCode				
	description					CreativeWork:sourceOrganization→ Thing:description	
			isPartOf			CreativeWork:isPartOf	
						CreativeWork:hasPart	
			systemOfRecords				
			describedBy				
			describedByType				
GeoSpatial	spatial-text∗		spatial	dcat:Dataset→ dct:spatial	void:Dataset→ dct:spatial	Dataset:spatial	
	geographical_granularity∗						
							bbox
							layers
							bboxCrs
							namespace
Temporal			temporal	dcat:Dataset→ dct:temporal	void:Dataset→ dct:temporal	Dataset:temporal	
	temporal_granularity∗						
	temporal_coverage_to∗						
	temporal_coverage_from∗						
Quality	ratings_average∗		dataQuality			CreativeWork:aggregateRating	
Organization							

Continued on next page

Table 2 Harmonized Dataset Models Mappings

Data Model	CKAN	DKAN	POD	DCAT	VoID	Schema.org	Socrata
General Information	title		name	dcat:Dataset→ dct:creator→ foaf:Organization:givenName	void:Dataset→ dct:creator→ foaf:Organization:givenName	CreativeWork:sourceOrganization:LegalName	
	description					CreativeWork:sourceOrganization→ Thing:description	
	id						
	type					CreativeWork:sourceOrganization→ Thing:additionalType	
	name					CreativeWork:sourceOrganization→ Thing:name	
	image_url						
	state						
	is_organization						
	approval_status						
provenance	revision_timestamp		subOrganizationOf			CreativeWork:sourceOrganization:subOrganization	
	revision_id						
Resources							
general	resource_group_id	resource_group_id					
	id	id					blobId
	size	size		dcat:Distribution→ dcat:byteSize		Dataset:distribution→ DataDownload→ MediaObject:contentSize	
	state	state					
	hash						
	description	description	description	dcat:Distribution→ dct:description		Dataset:distribution→ DataDownload→ Thing:description	
	format	format	format	dcat:Distribution→ dct:format	void:Dataset→ dct:format	Dataset:distribution→ DataDownload→ MediaObject:encodingFormat	
	mimetype	mimetype	mediaType	dcat:Distribution→ dcat:mediaType			
	mimetype_inner						
	name	name	title	dcat:Distribution→ dct:title		Dataset:distribution→ DataDownload→ Thing:name	filename / name
	position						
	resource_type					Dataset:distribution→ DataDownload→Thing:additionalType	
			describedBy				
			describedByType				
			conformsTo				
access information	cache_url						
	url-type						
	url	url	downloadURL	dcat:Distribution→ dcat:downloadURL	void:Dataset→ void:dataDump	Dataset:distribution→ DataDownload→ Thing:url	
			accessURL	dcat:Distribution→ dcat:accessURL		Dataset:distribution→ DataDownload→ MediaObject:contentUrl	accessPoints
	webstore_url						
provenance	cache_last_updated						
	revision_timestamp	revision_timestamp					
	webstore_last_updated						
	created	created				Dataset:distribution→ DataDownload→ CreativeWork:dataCreated	created_at
	last_modified	last_modified				Dataset:distribution→ DataDownload→ CreativeWork:dataModified	updated_at
	revision_id	revision_id					
Groups							
General	display_name	display_name					
	description	description					
	title	title					
	image_display_url	image_display_url					
	id	id					
	name	name					
	subgroups*						
Tags							
General	vocabulary_id	vocabulary_id		dcat:Dataset→ dcat:theme→ skos:ConceptScheme			
	display_name			dcat:Dataset→ dcat:keyword			
	name	name		dcat:Dataset→ dcat:theme→ skos:Concept			
	state						
	id	id					
Provenance	revision_timestamp						

Acknowledgments

This research has been partially funded by the European Union's 7th Framework Programme via the project Apps4EU (GA No. 325090).

References

1. Z. Abedjan, T. Gruetze, A. Jentzsch, and F. Naumann. Profiling and mining RDF data with ProLOD++. In *30th IEEE International Conference on Data Engineering (ICDE)*, pages 1198–1201, 2014.
2. M. Acosta, A. Zaveri, E. Simperl, and D. Kontokostas. Crowdsourcing Linked Data quality assessment. In *12th International Semantic Web Conference (ISWC)*, 2013.
3. A. Ahmad, S. Aline, and T. Raphaël. Roomba: Automatic Validation, Correction and Generation of Dataset Metadata. In *24th World Wide Web Conference (WWW), Demos Track*, Florence, Italy, 2015.
4. H. Aidan, H. Andreas, and D. Stefan. ReConRank: A Scalable Ranking Method for Semantic Web Data with Context. In *2nd Workshop on Scalable Semantic Web Knowledge Base Systems*, 2006.
5. K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. Describing Linked Datasets. In *2nd International Workshop on Linked Data on the Web (LDOW)*, 2009.
6. M. Alistair and B. Sean. SKOS Simple Knowledge Organization System Reference. W3C Recommendation, 2009. `http://www.w3.org/TR/skos-reference/`.
7. J. Anja, C. Richard, and B. Chrstian. State of the lod cloud. `http://lod-cloud.net/state/`.
8. F. Annika. Quality Characteristics of Linked Data Publishing Datasources. Master's thesis, Humboldt-Universität zu Berlin, 2010.
9. I. Antoine and S. Ed. SKOS Simple Knowledge Organization System Primer. W3C Working Group Note, 2009.
10. A. Assaf, E. Louw, A. Senart, C. Follenfant, R. Troncy, and D. Trastour. RUBIX: a framework for improving data integration with linked data. In *International Workshop on Open Data (WOD'12)*, pages 13–21, 2012.
11. A. Assaf and A. Senart. Data Quality Principles in the Semantic Web. In *6th International Conference on Semantic Computing ICSC '12*, 2012.
12. A. Assaf, A. Senart, and R. Troncy. SNARC - An Approach for Aggregating and Recommending Contextualized Social Content. In *The Semantic Web: ESWC 2013 Satellite Events, Revised Selected Papers*, pages 319–326, 2013.
13. S. Auer, J. Demter, M. Martin, and J. Lehmann. LODStats - an Extensible Framework for High-performance Dataset Analytics. In *18th International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, pages 353–362, Galway, Ireland, 2012.
14. C. Avitha, G. S. Sadasivam, and S. N. Shenoy. Ontology Based Semantic Integration of Heterogeneous Databases. *European Journal of Scientific Research*, page 115, 2011.
15. S. Ben. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *IEEE Symposium on Visual Languages*, pages 336–343, 1996.
16. H. Bernhard and P. Niko. DSNotify: Detecting and Fixing Broken Links in Linked Data Sets. In *8th International Workshop on Web Semantics*, 2009.
17. S. Besiki, G. Les, T. M. B., and S. L. C. A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, 2007.
18. C. Bizer and R. Cyganiak. Quality-driven information filtering using the wiqa policy framework. *Jorunal of Web Semantics*, 7(1), 2009.
19. C. Böhm, G. Kasneci, and F. Naumann. Latent Topics in Graph-structured Data. In *21st ACM International Conference on Information and Knowledge Management (CIKM)*, pages 2663–2666, Maui, Hawaii, USA, 2012.
20. C. Bohm, F. Naumann, Z. Abedjan, D. Fenz, T. Grutze, D. Hefenbrock, M. Pohl, and D. Sonnabend. Profiling linked open data with ProLOD. In *26th International Conference on Data Engineering Workshops (ICDEW)*, 2010.
21. K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *ACM International Conference on Management of Data (SIGMOD)*, 2008.

22. D. Boyd and K. Crawford. Six provocations for big data. *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, 2011.
23. S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *7th International Conference on World Wide Web (WWW'98)*, 1998.
24. C. Buil-Aranda and A. Hogan. SPARQL Web-Querying Infrastructure: Ready for Action? In *12th International Semantic Web Conference (ISWC)*, 2013.
25. B. Christian. Evolving the Web into a Global Data Space. In *28th British National Conference on Advances in Databases*, 2011.
26. B. Christian, L. Jens, K. Georgi, A. Sören, B. Christian, C. Richard, and H. Sebastian. DBpedia - A Crystallization Point for the Web of Data. *Journal of Web Semantics*, 7(3), 2009.
27. B. Christian, H. T, and B.-L. T. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2009.
28. M. Christian, H. Bernhard, and I. Antoine. Finding quality issues in SKOS vocabularies. *Theory and Practice of Digital Libraries*, 2012.
29. B. Christoph, L. Johannes, and N. Felix. Creating voiD Descriptions for Web-scale Data. *Journal of Web Semantics*, 9(3):339–345, 2011.
30. M. Cornolti, P. Ferragina, and M. Ciaramita. A Framework for Benchmarking Entity-annotation Systems. In *22nd World Wide Web Conference (WWW)*, 2013.
31. R. Cyganiak, H. Stenzhorn, R. Delbru, S. Decker, and G. Tummarello. Semantic Sitemaps: Efficient and Flexible Access to Datasets on the Semantic Web. In *5th European Semantic Web Conference (ESWC)*, pages 690–704, Tenerife, Spain, 2008.
32. R. Cyganiak, J. Zhao, M. Hausenblas, and K. Alexander. Describing Linked Datasets with the VoID Vocabulary. W3C Note, 2011. `http://www.w3.org/TR/void/`.
33. A. S. da Silva, D. Barbosa, J. M. B. Cavalcanti, and M. A. S. Sevalho. Labeling Data Extracted from the Web. In *On The Move Confederated International Conferences*, pages 1099–1116, 2007.
34. M. d'Aquin and E. Motta. Watson, More Than a Semantic Web Search Engine. *Semantic Web Journal*, 2011.
35. R. Dave. The Organization Ontology. W3C Recommendation, 2014. `http://www.w3.org/TR/vocab-org`.
36. J. Debattista, C. Lange, and S. Auer. daQ, an Ontology for Dataset Quality Information. In *Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference (WWW 2014)*, 2014.
37. R. Delbru, N. Toupikov, and M. Catasta. Hierarchical link analysis for ranking web data. In *7th European Semantic Web Conference (ESWC)*, 2010.
38. T. Di Noia, R. Mirizzi, V. C. Ostuni, D. Romito, and M. Zanker. Linked Open Data to Support Content-based Recommender Systems. In *8th International Conference on Semantic Systems - I-SEMANTICS '12*, 2012.
39. C. Didier, U. Ricardo, B. Andreas, and L. Jens. CROCUS: Cluster-based ontology data cleansing. In *2nd International Workshop on Semantic Web Enterprise Adoption and Best Practice*, 2014.
40. B. Diego, F. Sergio, and F. Iv"an. Cooking HTTP content negotiation with Vapour. In *4th Workshop on Scripting for the Semantic Web (SFSW'08)*, 2008.
41. K. Dimitris, Z. Amrapali, A. Sören, and L. J. TripleCheckMate: A Tool for Crowdsourcing the Quality Assessment of Linked Data. *4th Conference on Knowledge Engineering and Semantic Web*, 2013.

42. L. Ding, T. Finin, A. Joshi, R. Pan, and R. Cost. Swoogle: A semantic web search and metadata engine. In *13st ACM International Conference on Information and Knowledge Management (CIKM)*, 2004.
43. N. Douglas. Developing Spatial Data Infrastructures: The SDI Cookbook, 2004. `http://www.gsdi.org/docs2004/Cookbook/cookbookV2.0.pdf`.
44. P. Emmanuel, B. Christian, K. David, and L. Ryan. Fresnel: A Browser-Independent Presentation Vocabulary for RDF. In *5th International Semantic Web Conference (ISWC'06)*, pages 158–171, 2006.
45. D.-A. Ernesto, D. Lucas, S.-T. Lars, and N. Wolfgang. Real-time top-n recommendation in social streams. In *6th ACM conference on Recommender systems - RecSys*, 2012.
46. S. Evren, S. Michael, and W. Evan. Opening, Closing Worlds - On Integrity Constraints. In *5th OWLED Workshop on OWL: Experiences and Directions*, 2008.
47. B. Eytan, R. Itamar, M. Cameron, and A. Lada. The role of social networks in information diffusion. In *21th International Conference on World Wide Web (WWW'12)*, 2012.
48. M. Fadi and E. John. Data Catalog Vocabulary (DCAT). W3C Recommendation, 2014. `http://www.w3.org/TR/vocab-dcat/`.
49. T. Fawcett. An Introduction to ROC Analysis. *Pattern Recogn. Lett.*, 2006.
50. B. Fetahu, S. Dietze, B. Pereira Nunes, M. Antonio Casanova, D. Taibi, and W. Nejdl. A Scalable Approach for Efficiently Generating Structured Dataset Topic Profiles. In *11th European Semantic Web Conference (ESWC)*, 2014.
51. T. Finin, Z. Syed, J. Mayfield, P. Mcnamee, and C. Piatko. Using Wikitology for Cross-Document Entity Coreference Resolution. In *AAAI Spring Symposium on Learning*, 2009.
52. G. Flouris, Y. Roussakis, and M. Poveda-Villalón. Using provenance for quality assessment and repair in linked open data. In *2nd Joint Workshop on Knowledge Evolution and Ontology Dynamics (EvoDyn'12)*, 2012.
53. B. Forchhammer, A. Jentzsch, and F. Naumann. LODOP - Multi-Query Optimization for Linked Data Profiling Queries. In *International Workshop on Dataset PROFIling and fEderated Search for Linked Data (PROFILES)*, Heraklion, Greece, 2014.
54. P. Frischmuth, S. Auer, S. Tramp, J. Unbehauen, K. Holzweißig, and C. Marquardt. Towards Linked Data based Enterprise Information Integration. In *Workshop on Semantic Web Enterprise Adoption and Best Practice Co-located with 12th International Semantic Web Conference (ISWC'13)*, 2013.
55. P. Frischmuth, J. Klímek, S. Auer, S. Tramp, J. Unbehauen, K. Holzweißig, and C.-M. Marquardt. Linked Data in Enterprise Information Integration. 2012.
56. M. Frosterus, E. Hyvönen, and J. Laitio. Creating and Publishing Semantic Metadata about Linked and Open Datasets. In *Linking Government Data*. 2011.
57. M. Frosterus, E. Hyvönen, and J. Laitio. DataFinland - A Semantic Portal for Open and Linked Datasets. In *8th Extended Semantic Web Conference (ESWC)*, pages 243–254, 2011.
58. C. Fürber and M. Hepp. SWIQA - A Semantic Web information quality assessment framework. 2011.
59. G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The Vocabulary Problem in Human-System Communication. *Communications of The ACM*, 30(11):964–971, 1987.
60. T. Giovanni, C. Richard, C. Michele, D. Szymon, D. Renaud, and D. Stefan. Sig.ma: Live views on the Web of data. *Journal of Web Semantics*, 8(4), 2010.

61. G. Gouriten and P. Senellart. API Blender: A Uniform Interface to Social Platform APIs. In *21th International Conference on World Wide Web (WWW'12)*, 2012.
62. C. Guéret, P. T. Groth, C. Stadler, and J. Lehmann. Assessing Linked Data Mappings Using Network Measures. In *9th European Semantic Web Conference (ESWC)*, 2012.
63. A. Harth, K. Hose, M. Karnstedt, A. Polleres, K.-U. Sattler, and J. Umbrich. Data Summaries for On-demand Queries over Linked Data. In *19th World Wide Web Conference (WWW)*, 2010.
64. A. Harth, S. Kinsella, and S. Decker. Using naming authority to rank data and ontologies for web search. In *8th International Semantic Web Conference (ISWC)*, 2009.
65. O. Hassanzadeh, S. Duan, A. Fokoue, A. Kementsietsidis, K. Srinivas, and M. J. Ward. Helix: Online Enterprise Data Analytics. In *20th International Conference Companion on World Wide Web (WWW'11)*, pages 225–228, 2011.
66. A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres. Weaving the pedantic web. 2010.
67. A. Hogan, J. Umbrich, A. Harth, R. Cyganiak, A. Polleres, and S. Decker. An empirical survey of Linked Data conformance. *Journal of Web Semantics*, 2012.
68. K. Houda, A. Ghislain, R. Giuseppe, and R. Troncy. Aggregating Social Media for Enhancing Conference Experiences. In *1st Internationl Workshop on Real-Time Analysis and Mining of Social Streams*, 2012.
69. R. Isele, J. Umbrich, C. Bizer, and A. Harth. LDspider: An Open-source Crawling Framework for the Web of Linked Data. In *9th International Semantic Web Conference (ISWC), Posters & Demos Track*, 2010.
70. C. Iván and B. Alejandro. Semantic contextualisation of social tag-based profiles and item recommendations. In *12th Internationl Conference on E-Commerce and Web Technolgoies*, 2011.
71. P. Jain, P. Hitzler, K. Janowicz, and C. Venkatramani. There's No Money in Linked Data, 2013. `http://knoesis.wright.edu/faculty/pascal/pub/nomoneylod.pdf`.
72. M. James and D. E. Almasi. Open data: Unlocking innovation and performance with liquid information. Technical report, McKinsey Business Technology Office, 2001.
73. L. Jens and S. Soeren. DL-Learner: Learning Concepts in Description Logics. *Journal of Machine Learning Research*, 2009.
74. A. Jentzsch. Profiling the Web of Data. In *13th International Semantic Web Conference (ISWC), Doctoral Consortium*, Trentino, Italy, 2014.
75. D. Jeremy, L. Santiago, L. Christoph, and A. Sören. LUZZU - A framework for linked data quality assessment. *CoRR*, abs/1412.3750, 2014.
76. J. M. Juran and A. B. Godfrey. *Juran's quality handbook*. McGraw Hill, 1999.
77. K. B. K., S. D. M., and W. R. Y. Information quality benchmarks: product and service performance. *Communications of the ACM*, 2002.
78. T. Käfer, A. Abdelrahman, J. Umbrich, P. O'Byrne, and A. Hogan. Observing Linked Data Dynamics. In *10th European Semantic Web Conference (ESWC)*, 2013.
79. H. Kang and B. Shneiderman. MediaFinder: an interface for dynamic personal media management with semantic regions. In *Conference on Human Factors in Computing Systems (CHI)*, pages 764–765. ACM, 2003.

80. C. Keet, M. del Carmen Suárez-Figueroa, and M. Poveda-Villalón. The Current Landscape of Pitfalls in Ontologies. In *International Conference on Knowledge Engineering and Ontology Development (KEOD)*, 2013.
81. S. Khatchadourian and M. P. Consens. ExpLOD: Summary-based Exploration of Interlinking and RDF Usage in the Linked Open Data Cloud. In *7th Extended Semantic Web Conference (ESWC)*, pages 272–287, Heraklion, Greece, 2010.
82. J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *ACM Journal*, 1999.
83. M. Konrath, T. Gottron, S. Staab, and A. Scherp. SchemEX - Efficient Construction of a Data Catalogue by Stream-based Indexing of Linked Data. *Journal of Web Semantics*, 16, 2012.
84. D. Kontokostas, P. Westphal, S. Auer, S. Hellmann, J. Lehmann, R. Cornelissen, and A. Zaveri. Test-driven Evaluation of Linked Data Quality. In *23rd International Conference on World Wide Web (WWW'14)*, 2014.
85. Kovács-Láng. Global Terrestrial Observing System. Technical report, GTOS Central and Eastern European Terrestrial Data Management and Accessibility Workshop, 2000.
86. C. J. Kowalski. On the Effects of Non-Normality on the Distribution of the Sample Product-Moment Correlation Coefficient. *Journal of the Royal Statistical Society*, 1972.
87. S. Lalithsena, P. Hitzler, A. Sheth, and P. Jain. Automatic Domain Identification for Linked Open Data. In *IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, pages 205–212, 2013.
88. A. Langegger and W. Woss. RDFStats - An Extensible RDF Statistics Generator and Library. In *20th International Workshop on Database and Expert Systems Application (DEXA)*, pages 79–83, 2009.
89. S. LaValle, E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz. Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review*, 2011.
90. P. Lawrence, B. Sergey, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, 1998.
91. M. Lenzerini. Data Integration: A Theoretical Perspective. In *21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 233–246, 2002.
92. J. Leskovec and C. Faloutsos. Sampling from Large Graphs. In *12thth ACM International Conference on Knowledge Discovery and Data Mining (KDD'12)*, 2006.
93. H. Li. Data Profiling for Semantic Web Data. In *International Conference on Web Information Systems and Mining (WISM)*, pages 472–479, 2012.
94. G. Limaye, S. Sarawagi, and S. Chakrabarti. Annotating and Searching Web Tables Using Entities, Types and Relationships. *VLDB Endowment*, pages 1338–1347, 2010.
95. E. Mäkelä. Aether - Generating and Viewing Extended VoID Statistical Descriptions of RDF Datasets. In *11th European Semantic Web Conference (ESWC), Demo Track*, Heraklion, Greece, 2014.
96. P. Marco and G. Siva. Investigating topic models for social media user recommendation. In *20th International Conference on World Wide Web (WWW'11)*, 2011.

97. N. Marie, F. Gandon, M. Ribière, and F. Rodio. Discovery Hub: On-the-fly Linked Data Exploratory Search. In *The 9th International Conference on Semantic Systems*, 2013.
98. B. Martin, B. Ciro, E. Ivan, F. Markus, K. Dimitris, and H. Sebastian. DataID: Towards Semantically Rich Metadata for Complex Datasets. In *10th International Conference on Semantic Systems*, 2014.
99. V. Mateja. LODGrefine - LOD-enabled Google Refine in Action. In *8th International Conference on Semantic Systems - I-SEMANTICS '12*, 2012.
100. S. Max, B. Christian, and P. Heiko. Adoption of the Linked Data Best Practices in Different Topical Domains. In *13th International Semantic Web Conference (ISWC)*, 2014.
101. P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia Spotlight: Shedding Light on the Web of Documents. In *7th International Conference on Semantic Systems*, 2011.
102. H. Michael, H. Wolfgang, R. Yves, F. Lee, and A. Danny. SCOVO: Using Statistics on the Web of Data. In *ESWC*, 2009.
103. N. Mihindukulasooriya, R. Garcia-Castro, and M. E. Gutiérrez. Linked Data Platform as a novel approach for Enterprise Application Integration. In *4th International Workshop on Consuming Linked Data (COLD'13)*, 2013.
104. B. Mike. Deconstructing the Google Knowledge Graph. `http://www.mkbergman.com/1009/deconstructing-the-google-knowledge-graph`.
105. R. J. Miller and P. Andritsos. Schema Discovery. *IEEE Data Engineering Bulletin*, 26:40–45, 2003.
106. T. Nickolai, U. J, and D. Renaud. DING! Dataset ranking using formal descriptions. In *2nd International Workshop on Linked Data on the Web (LDOW)*, 2009.
107. A. Nikolov, M. d'Aquin, and E. Motta. What Should I Link to? Identifying Relevant Sources and Classes for Data Linking. In *Joint International Semantic Technology Conference (JIST)*, 2011.
108. H. Olaf and Z. Jun. Using web data provenance for quality assessment. In *8th International Semantic Web Conference (ISWC)*, 2009.
109. S. Osma and M. Christian. Assessing and Improving the Quality of SKOS Vocabularies. *Journal on Data Semantics*, 2013.
110. S. Osma and H. Eero. Improving the quality of SKOS vocabularies with skosify. In *The 18th International Conference on Knowledge Engineering and Knowledge Management*, 2012.
111. H. Patricia. *Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works*. Getty Research Institute, 2010.
112. E. Peukert, J. Eberius, and E. Rahm. AMC - A framework for modelling and comparing matching systems as matching processes. In *IEEE 27th International Conference on Data Engineering (ICDE'11)*, 2011.
113. E. Peukert, J. Eberius, and E. Rahm. A Self-Configuring Schema Matching System. In *IEEE 28th International Conference on Data Engineering (ICDE'12)*, 2012.
114. A. Phil and S. Gofran. Asset Description Metadata Schema (ADMS). W3C Working Group Note, 2013. `http://www.w3.org/TR/vocab-adms`.
115. M. PN, M. Hannes, and B. Christian. Sieve: linked data quality assessment and fusion. 2012.
116. M. Poveda-Villalón, M. Suárez-Figueroa, and A. Gómez-Pérez. Validating Ontologies with OOPS! In *18th International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, 2012.

117. D. Preotiuc-Pietro, S. Samangooei, T. Cohn, N. Gibbins, and M. Niranjan. Trendminer: An architecture for real time analysis of social media text. In *6th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2012.
118. N. Press. Understanding Metadata. Technical report, National Information Standards Organization, 2004.
119. D. C. Reis, P. B. Golgher, A. S. Silva, and A. F. Laender. Automatic Web News Extraction Using Tree Edit Distance. In *13th International World Wide Web Conference (WWW'04)*, pages 502–601, 2004.
120. I. Renato and M. James. vCard Ontology - for describing People and Organizations. W3C Interest Group Note, 2014. `http://www.w3.org/TR/vcard-rdf`.
121. E. Ruckhaus, O. Baldizan, and M.-E. Vidal. Analyzing Linked Data Quality with LiQuate. In *11th European Semantic Web Conference (ESWC)*, 2014.
122. A. Rula and A. Zaveri. Methodology for Assessment of Linked Data Quality. In *1st Workshop on Linked Data Quality (LDQ)*, 2014.
123. D. Soergel. Thesauri and ontologies in digital libraries. In *2nd ACM/IEEE-CS Joint Conference on Digital Libraries*, 2002.
124. C. Soumen, D. B. E., S. K. Ravi, R. Prabhakar, R. Sridhar, T. Andrew, G. David, and K. Jon. Mining the web's link structure. *Computer*, 1999.
125. T. Steiner and S. Mirea. SEKI@home or Crowdsourcing an Open Knowledge Graph. In *1st International Workshop on Knowledge Extraction & Consolidation from Social Media (KECSM'12)*, Boston, USA, 2012.
126. F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A Core of Semantic Knowledge. In *16th International World Wide Web Conference (WWW)*, 2007.
127. Z. Syed, T. Finin, V. Mulwad, and A. Joshi. Exploiting a Web of Semantic Data for Interpreting Tables. In *2nd Web Science Conference*, 2010.
128. J. Tao, L. Ding, and D. L. McGuinness. Instance Data Evaluation for Semantic Web-Based Knowledge Management Systems. In *42nd Hawaii International Conference on System Sciences, HICSS'09*, pages 1–10, 2009.
129. B.-L. Tim. Linked Data - Design Issues. W3C Personal Notes, 2006. `http://www.w3.org/DesignIssues/LinkedData`.
130. L. Timothy, S. Satya, and M. Deborah. PROV-O: The PROV Ontology. W3C Recommendation, 2013. `http://www.w3.org/TR/prov-o`.
131. A. Tonon, M. Catasta, G. Demartini, P. Cudré-Mauroux, and K. Aberer. TRank: Ranking Entity Types Using the Web of Data. In *12th International Semantic Web Conference (ISWC)*, 2013.
132. G. Tummarello, R. Delbru, and E. Oren. Sindice.com: Weaving the open linked data. In *6th International Semantic Web Conference (ISWC)*, 2007.
133. S. Umberto and T. Raphaël. oMAP: Combining Classifiers for Aligning Automatically OWL Ontologies. In *6th International Conference on Web Information Systems Engineering*, 2005.
134. R. Usbeck, M. Röder, A.-C. Ngonga-Ngomo, C. Baron, A. Both, M. Brümmer, D. Ceccarelli, M. Cornolti, D. Cherix, B. Eickmann, P. Ferragina, C. Lemke, A. Moro, R. Navigli, F. Piccinno, G. Rizzo, H. Sack, R. Speck, R. Troncy, J. Waitelonis, and L. Wesemann. GERBIL - General Entity Annotation Benchmark Framework. In *24th World Wide Web Conference (WWW)*, 2015.
135. Z. Valentina and C. L. Social ranking: uncovering relevant content using tag-based recommender systems. In *2nd ACM conference on Recommender systems - RecSys*, 2008.
136. G. Vickery. Review of Recent Studies on PSI-use and Related Market Developments. Technical report, EC DG Information Society, 2011.

137. H. Wache, T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Hübner. Ontology-Based Integration of Information - A Survey of Existing Approaches. In *IJCAI Workshop: Ontologies and Information*, pages 108–117, 2001.
138. J. Wang and F. H. Lochovsky. Data Extraction and Label Assignment for Web Databases. In *12th International World Wide Web Conference (WWW'03)*, pages 187–196, 2003.
139. W. R. Y. and S. D. M. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 1996.
140. A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. Quality Assessment Methodologies for Linked Open Data. *Semantic Web Journal*, 2012.