DCPAPERS

Application Profile Driven Data Acquisition for Knowledge Graph and Linked Data Generation in Crowdsourced Data Journalism

Nishad Thalhath
Graduate School of Library, Information and Media Studies,
University of Tsukuba
Japan
nishad@slis.tsukuba.ac.jp

Mitsuharu Nagamori
Faculty of Library, Information and Media Studies
University of Tsukuba
Japan
nagamori@slis.tsukuba.ac.jp

Tetsuo Sakaguchi
Faculty of Library, Information and Media Studies
University of Tsukuba
Japan
saka@slis.tsukuba.ac.jp

Abstract

Application Profiles are a collection of vocabularies mixed and matched from different namespaces and customized for the local application. Application Profiles act as a constrainer as well as an explainer of the metadata for every dataset. Linking data and generating knowledge graphs are the general challenges that information processing communities are trying to address constructively. Application Profiles can act as a means of linking data and generating knowledge graphs. The authors propose application profile driven questionnaire creation for data linking in the perspective of crowdsourced data acquisition; especially where there are challenges in adapting a single vocabulary or a limited number of adaptable domain-specific vocabularies. This paper presents a proof of concept study based on the proposed approach by adapting existing standards and tools, with the notion that similar methods are applicable in related use-cases.

Keywords: Application Profile; Crowd sourcing; Data Journalism; Linked Data; Knowledge Graphs

1 Introduction

Application Profiles are the most suitable way to build consensus on Metadata, and well defined URIs help the data be interoperable and linkable. Application Profiles are a collection of vocabularies mixed and matched from different namespaces and customized for the local application. Application Profiles acts as a constrainer as well as an explainer of the metadata for every dataset. Well defined metadata application profiles will streamline knowledge graph generation for multi-source datasets. This paper examines the possibilities of defining and developing MAP for questionnaires and crowdsourcing data acquisition by introducing general-purpose ontologies and vocabularies, also attempts to utilize Wikidata as a broader vocabulary resource for promoting the use of linkable concepts. The authors evaluate this proposal in the context of data journalism. Data journalism projects are vibrant, with varying demands and purposes, making them less suitable for adopting any common vocabularies or ontologies. Developing and maintaining custom vocabularies are expensive processes for smaller projects in terms of resources and skill requirements.

The authors explain how a simplified MAP can be created for data journalism projects and demonstrates the possibilities of adopting it to acquire and disseminate data in some case studies. The paper is intended to show some practical insights on reliable and reasonable MAP and vocabulary development for data journalism projects, with a notion to improve the quality and quantity of linkable data obtained from them. Knowledge graphs and graph databases are getting accepted widely. Also, the data handlers are seeking an efficient mechanism to store, express, and distribute unstructured, scalable, diverse, and incomplete data (Ali et al., 2021). Resource Description Framework (RDF) (Cyganiak et al., 2014) is emerging as a universal data model and helps the Semanticweb express and consume linked data and knowledge graphs. Sharing the profile of

20th International Conference on Dublin Core and Metadata Applications 2022 (DC 2022)

https://doi.org/10.23106/dcmi.953151686

the data, especially in actionable formats, will help the data stakeholders ensure the structure and validate the content seamlessly (Thornton et al., 2019).

1.1 Data Journalism

Data journalism is a journalism specialty reflecting the increased role that numerical data is used in the production and distribution of information in the digital era. It reflects the increased interaction between content producers (journalist) and several other fields such as design, computer science, and statistics. Data journalism has been widely used to unite several concepts and link them to journalism. Some see these as levels or stages leading from the simpler to the more complex uses of new technologies in the journalistic process (Gray & Bounegru, 2019).

In the coming future, data journalism will be expected to consume more of non-numerical data and concepts from Open Data and Linked Data. Being a form of Social informatics, modern Journalism has broader perspectives in its growing popularity on the adaptation of Data Journalism (Howard, 2014).

Data Journalism is the promising element of forthcoming Journalism, in which the conventional news reporting will be expected to turn towards the quest of fact-finding across the vast available sets of Open Data, Linked Open Data (LOD) and semantic web technologies. Considering the sheer scale and range of digital information now available, applied metadata concepts and de-facto standards of Linked Data practices will make this adoption smoother, reliable, expandable and bidirectional. The current state of Data Journalism is mostly depended on numerical data and limited to either data-driven journalism or data-assisted journalism. Adoption of Linked Data and publishing the supporting data along with the news will help to identify or eliminate 'Fake news,' disinformation, hoaxes, and propaganda as well as evaluate and explore related and relevant information. Adopting such standards is a crucial step towards data-literacy and journalistic integrity in an information-driven global community. This research aims to understand the significance and possible adoption of Metadata and Linked Open Data in future trends of Data Journalism and social semantic journalism.

Data Journalism reflects the increased interaction between content producers (journalist) and several other fields such as design, computer science, and statistics. Data Journalism has been widely used to unite several concepts and link them to journalism. Some see these as levels or stages leading from the simpler to the more sophisticated uses of new technologies in the journalistic process (Gray et al., 2012). In the coming future, Data Journalism will be expected to consume more of non-numerical data and concepts from Open Data and Linked Data. Being a form of Social informatics, modern Journalism has broader perspectives in its growing popularity on the adaptation of Data Journalism (Howard, 2014).

1.2 Application Profiles

Application Profiles are data element schemas from various namespaces mixed and customized for a specific application (Heery & Patel, 2000). MAPs are the best mechanism to express consensus of any metadata instance by documenting the elements, policies, guidelines, and vocabularies for that particular implementation along with the schemas, and applicable constraints. Application profiles also provide the term usage specifications and support interoperability by representing domain consensus, alignment, and the structure of the data (Baca, 2016) (Hillmann, 2006).

1.2.1 Expressing Metadata Application Profiles

Metadata application profiles (MAP) provides a means to structure and customise the metadata instance by documenting the elements, policies, guidelines, and vocabularies for that particular implementation along with the schemas, and applicable constraints(Heery & Patel, 2000). MAP gives the specific syntax guidelines and data format, description set profiles (DSP), domain consensus and alignment(Baca, 2016) (Hillmann, 2006). DSP format is obtained from the DCMI guidelines, expressed through XML or RDF. With the emergence of Semantic Web concept, application profiles are expressed using JSON-LD and OWL with increasing use cases using validation tools like Shape Expressions (ShEx) [1] or Shapes Constraint Language (SHACL) (Kontokostas & Knublauch, 2017). Authoring formats like Yet Another Metadata Application Profiles (YAMA) (Thalhath et al., 2019) provide options to use an intermediary formats such as YAML to author application profiles and convert them to actionable formats. There are domain specific authoring tools for application profiles, such as

[1] http://shex.io/

https://doi.org/10.23106/dcmi.953151686

BIBRAME Profile Editor (Development & Office, 2021) provides interactive interfaces to edit specific type of application profiles.

1.2.2 Recents developments in Application Profiles

One of the recent developments in application profiles is the DCMI application profile interest group's (DCMI, 2021) ongoing efforts with DCMI Application Profile vocabulary[2], which defines the elements of an application profile. The profile is intended to define and constrain the property-value pairs in metadata instances. These pairs are statements about something that the metadata describes and grouped into shapes.

The group is also working on DC Tabular Application Profiles (DC TAP)[3], which provides a vocabulary and a format for creating table-based application profiles. DCTAP is defined in tabular format, with each row in a TAP table as a single metadata statement. These statements may be combined to form shapes.

Using application profiles to fulfill actionable use-cases are also getting accepted among communities. For example, in Wikidata, one of the prominent public knowledge graph, uses ShEx Compact Syntax (ShExC) to express the schemas of Wikidata entities. These Wikidata Schemas helps to test subsets of Wikidata to verify if they conform to a specific schema [4].

1.3 Application Profile development for Data Journalism Projects

Developing Application Profiles is an expensive process with a disproportionate incentive. The lack of expertise in developing application profiles is one of the primary reasons for its limited availability. Lack of metadata application profiles reduces interoperability and shrinks the options to generate linkable data. Most of the social informatics projects are covering either broader subjects of interdisciplinary concepts. Adopting a single vocabulary or ontology sometimes limits the scope as well as use cases.

1.3.1 Data Journalism and crowd sourcing data

One of the significant acquisition phases of data journalism projects is through questionnaires. Developing these questionnaires, along with metadata application profiles, will make the whole process of acquisition, access, and dissemination of data will be more LOD friendly. Another approach is to define metadata with URIs for tabular data as well. Standards like CSV on the Web (CSVW)(Tennison, 2016) and toolkits like OpenRefine[5] can use these URIs and minimal metadata to describe or generate meaningful linkable data.

1.3.2 Wikidata as a vocabulary source for data journalism Projects

Wikidata can be treated as a controlled vocabulary resource in places where we can use it to replace literals. Multi-lingual labels in Wikidata permits the use of any Wikidata entity to be mapped with a persistent URI irrespective of the language. In the context of Social Informatics projects, the stakeholders can always find the relevant resources from Wikipedia and use the corresponding Wikidata entity as a URI, with suitable labels. Concepts are always linked to the same URI resources irrespective of the language or domain of the project. A higher level of interoperability with precise concept mapping can be obtained from this approach. (Thalhath et al., 2021) Wikidata, so as Wikipedia is a domain-independent knowledge graph and the URIs servers valid machine-actionable as well as human-interpretable resources. The machine actionability gives another possibility of using these URIs to utilize the scope of knowledge extraction processes further.

[2] https://www.dublincore.org/groups/application_profiles_ig/dctap_primer/

[3] https://github.com/dcmi/dctap/blob/main/TAPprimer.md

[4] https://www.wikidata.org/wiki/Wikidata:WikiProject_Schemas

[5] https://openrefine.org/

https://doi.org/10.23106/dcmi.953151686

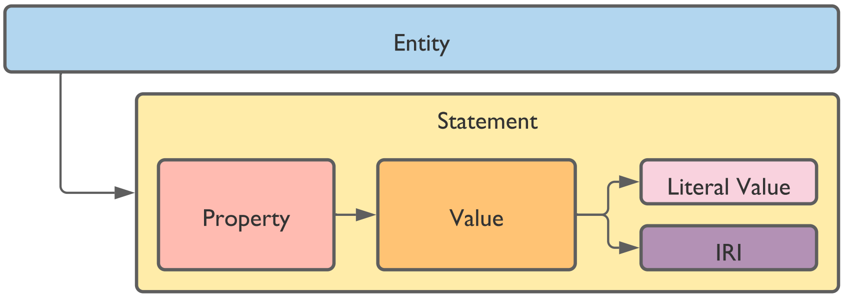

Figure 1: Application profile structure

1.4 Problems this approach attempted to address

1.4.1 Metadata consensus

Developing consensus of metadata is critical in terms of ensuring the reusability of data. This proposal is to promote the creation of more linkable and interoperable datasets. Such approaches can bring maximum acceptance and longevity for the data by making it adaptable in various use-cases. Application profiles can act as a lightweight ontology in expressing the local application of terms, which profiles the metadata in a comprehensible form for both human and machine consumers. However, application profiles are not intended to express the semantics of the profile, which makes them not to be a substitute for ontologies but more as a consensus builder for the local application of higher ontologies.

1.4.2 Interoperability of data

Interoperable data can help the stakeholders utilize simpler workflows and make the data acceptable in various other scenarios than the pre-defined use-cases. Interoperable metadata makes the data to be interoperable and reusable; thus, the data be more satisfiable for the FAIR (Findable, Accessible, Interoperable, and Reusable data) models of data publishing (Wilkinson et al., 2016).

1.4.3 Linked data generation

A metadata profile cannot address data linking as a drop-in solution, but it can act as a guideline to include linkable concepts in the data and model the data to be easily represented by URIs than literals. An application profile will provide the essential blueprint of the linking, and it can also be used in validating the produced LOD datasets.

2 Methods

Application Profiles are mainly derived in the context of Resource Description Format (RDF) concepts. A general overview of an application profile can be more similar to that of the RDF structure. Application profiles are a collection of entities, and entities can be further made of statements about the entities. Statements are composed of a property and a value. In general, structuring data to this universal data model of RDF makes it more convenient in generating RDF, based on the profile and acquired tabular data. Statement values can be either a literal, URL, or reference to another entity. A simple structure of a profile is explained in Figure 1.

Building an application profile is required to find property elements from existing vocabularies and mix and match them to make them suitable for the local application; in this paper, the application is the questionnaire for data acquisition. A group of questions can be an entity and questions can be properties with the answers mapped to values. A simplified example of this concept is expressed in Figure 2.

https://doi.org/10.23106/dcmi.953151686

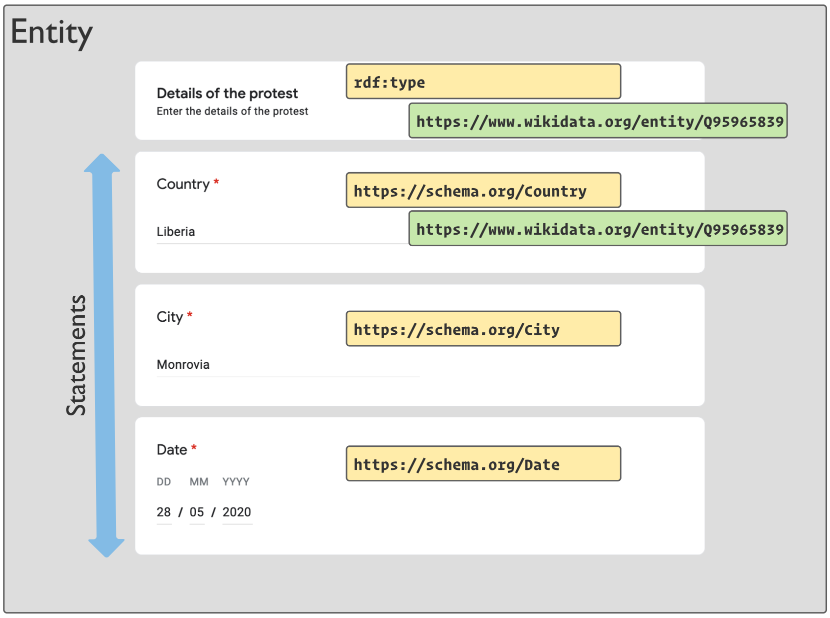

Figure 2: A sample form mapped to Application profile

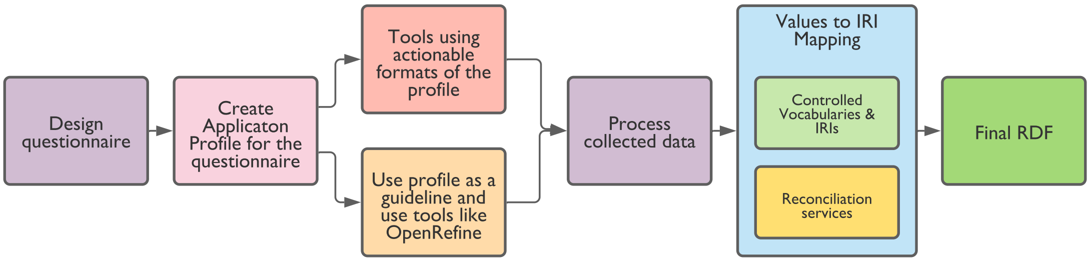
Design questionnaire
Create Applicaton Profile for the questionnaire
Tools using actionable formats of the profile
Use profile as a guideline and use tools like OpenRefine
Process collected data
Values to IRI Mapping
Controlled Vocabularies & IRIs
Reconciliation services
Final RDF

Figure 3: A simple workflow for generating linked data based on the application profile

2.1 Proof of concept workflow

A sample scenario of collecting a list of cities where any George Floyd protests were conducted is set up to demonstrate the workflow. A sample questionnaire is made for the crowdsourced data collection. For the proof of concept demonstration, the resulting dataset from the questionnaire is simulated from the map data available at Wikidata commons[6]. The sample questionnaire is transformed into an application profile in DCTAP format. The resulting dataset is reconciled with OpenRefine against Wikidata to link URIs of the cities instead of the literals. In the same way, countries were also reconciled to use the Wikidata URIs.

[6] https://commons.wikimedia.org/wiki/Data:George_Floyd_protests.map

Table 1: DCTAP Representation of the application profile

shapeID	propertyID	propertyLabel	mandatory	repeatable	valueNodeType	valueConstraint
:protest		Protest	y	y		
	rdf:type	Instance of	y	n	URI	wd:Q95965839
	sdo:Country	Country	y	n	URI	
	sdo:City	City	y	n	URI	
	sdo:Date	Date	n	y	literal	xsd:date

https://doi.org/10.23106/dcmi.953151686

An RDF schema as per the application pofile was defined in the OpenRefine RDF extension, and a LOD dataset is generated. The resulting RDF is validated with a validation schema in Shape Expressions (ShEx) generated from the DCTAP application profile.

```
BASE   <http://example.org/>
PREFIX sdo: <http://schema.org/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

<protest> {
  rdf:type IRI {1} ;
  sdo:Country IRI {1} ;
  sdo:City IRI {1} ;
  sdo:Date xsd:date ? ;
}
```

Example 1: ShExC representation of the Application Profile

```
{
  "@context": "http://www.w3.org/ns/shex.jsonld",
  "type": "Schema",
  "shapes": [
    {
      "type": "Shape",
      "id": "http://example.org/protest",
      "expression": {
        "type": "EachOf",
        "expressions": [
          {
            "type": "TripleConstraint",
            "predicate": "http://www.w3.org/1999/02/22-rdf-syntax-ns#type",
            "valueExpr": {
              "type": "NodeConstraint",
              "nodeKind": "iri"
            },
            "min": 1,
            "max": 1
          },
          {
            "type": "TripleConstraint",
            "predicate": "http://schema.org/Country",
            "valueExpr": {
              "type": "NodeConstraint",
              "nodeKind": "iri"
            },
            "min": 1,
            "max": 1
          },
          {
            "type": "TripleConstraint",
            "predicate": "http://schema.org/City",
            "valueExpr": {
              "type": "NodeConstraint",
              "nodeKind": "iri"
            },
            "min": 1,
            "max": 1
          },
          {
            "type": "TripleConstraint",
            "predicate": "http://schema.org/Date",
            "valueExpr": {
              "type": "NodeConstraint",
              "datatype": "http://www.w3.org/2001/XMLSchema#date"
            },
            "min": 0,
            "max": 1
          }
        ]
      }
    }
  ]
}
```

Example 2: ShExJ representation of the application profile

3 Results

Application profiles can be converted to actionable formats such as ShEx, or it can be used to convert the data to RDF using more straightforward tools like OpenRefine with RDF extension. The advantage of having a detailed profile as a framework for RDF conversion will provide a level of convenience in generating RDF. Application profiles can be used to create CSVW expression in JSON, and there are tools to convert CSV to RDF using CSVW as a mapping format.

The authors devised an opensourced static publishing mechanism to browse the resulting knowledge graph. This dataset can be explored using SPARQL within the web browser without any server-sided technologies. This is implemented with quadstore,[7] a javascript-based RDF store and YASGUI(Rietveld & Hoekstra, 2013) as the query and display interface. The resulting proof of concept dataset is accessible at https://nishad.github.io/gf-protest-kg.

4 Discussion

There are different kinds of tooling available around semantic web publishing. However, building a streamlined workflow around these tools and creating more use-cases for these tools will help eliminate any duplication of the efforts in tooling and will help to follow a systematic approach in producing knowledge graphs for smaller communities. Publishing profiles with datasets will allow communities to reuse the available data and promote interoperable profiling mechanisms. Attempts like Profiles vocabulary (Car, 2019) is supporting profile publishing with a semantic web approach.

Publishing findable and accessible datasets are a step ahead of conventional concepts of data sharing (Poline, 2019). Vocabularies like DCAT(Riccardo Albertoni et al., 2020), VoID (Keith Alexander et al., 2011) , and Schema.org[8] helps in describing datasets. These mechanisms will eventually help services like google dataset search to find and index datasets in a highly accessible and interoperable manner (Natasha Noy & Dan Brickley, 2017) (Natasha Noy, 2018).

4.1 Limitations of this propsal

This paper is merely presenting the concepts of Application profile driver questionnaires in data acquisition for data journalism projects. These methods won't be suitable for complex questionnaires or elaborate data collection workflows. Generating RDF of any other forms of Knowledge graphs requires additional layers of skills for the end-users. An application profile will only act as a guideline to the process and makes the whole data set RDF ready. The efficiency of using existing tools to create RDF from collected data using a profile is highly dependent on the use-cases and capabilities of the stakeholders.

[7] https://beautifulinteractions.github.io/node-quadstore/

[8] https://schema.org/Dataset

https://doi.org/10.23106/dcmi.953151686

4.2 Limitations of this proof of concept

Creating application profiles demands a minimal level of skills and may not be easier for stakeholders with limited knowledge in semantics web concepts. Even though DCTAP is a simplified intermediary format for application profiles, creating actionable expressions from application profiles requires dedicated tooling and skilled interaction. The workflow proposed in this paper may not be satisfactory for many practical scenarios.

The sample questionnaire demonstrated in this work, and the simulated data is barely minimal. In the real-world scenario, profiles require complex shapes, and more than one shape or entity is required to encapsulate the data into reusable and expressive knowledge graphs.

4.3 Future Work

The authors are working on an application profile framework for crowdsourced data acquisition along with a set of reusable tools to use application profiles and collected data to generate RDF. This framework and toolkits are usable in many data journalism as well as social informatics projects. The framework will be covering a set of use-cases and many sample questionnaires to demonstrate the possibilities.

5 Conclusion

Linking data from different domains will help to develop new possibilities of information-driven knowledge discovery. Developing consensuses of data through metadata application profiles and linking concepts through URIs will improve the quality and availability of linkable data. In terms of Social Informatics, especially in data journalism, will bring more insights and impacts for the social, political, cultural, and economic contexts. In the long run, it is anticipated to link data from governance and the public administration to data journalism projects and vice versa. Thus the impact of such projects can be improved positively. By bringing in the idea of Application Profiles, data journalism projects can mix and match terms from different vocabularies and generate linkable datasets as well as knowledge graphs.

References

Ali, W., Saleem, M., Bin, Y., Hogan, A., & Ngomo, A.-C. N. (2021). A survey of rdf stores and sparql engines for querying knowledge graphs. https://doi.org/10.20944/preprints202104.0199.v1

Baca, M. (2016). Introduction to Metadata. Retrieved April 10, 2019, from http://www.getty.edu/publications/intrometadata

Car, N. (2019). *The profiles vocabulary* (W3C Note) [https://www.w3.org/TR/2019/NOTE-dx-prof-20191218/]. W3C.

Cyganiak, R., Lanthaler, M., & Wood, D. (2014). *RDF 1.1 concepts and abstract syntax* (W3C Recommendation) [https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/]. W3C.

DCMI. (2021). Dcmi/dctap [original-date: 2020-12-01T13:17:07Z]. Retrieved April 20, 2021, from https://github.com/dcmi/dctap

Development, N., & Office, M. S. (2021). *Bibframe profile editor*. Network Development; MARC Standards Office. https://github.com/lcnetdev/profile-edit

Gray, J., Chambers, L., & Bounegru, L. (2012). *The data journalism handbook: How journalists can use data to improve the news*. O'Reilly Media. https://datajournalismhandbook.org/handbook/one

Gray, J., & Bounegru, L. (2019). *Data journalism handbook 2: Towards a critical data practice*. European Journalism Centre. https://datajournalism.com/read/handbook/two

Heery, R., & Patel, M. (2000). Application Profiles: Mixing and Matching Metadata Schemas. *Ariadne*, *25*. http://www.ariadne.ac.uk/issue/25/app-profiles/

Hillmann, D. (2006). Metadata standards and applications [publisher: Metadata Management Associates LLC]. Retrieved April 25, 2019, from http://managemetadata.com/

Howard, A. B. (2014). The Art and Science of Data-Driven Journalism.

Keith Alexander, Richard Cyganiak, Michael Hausenblas, & Jun Zhao. (2011). Describing Linked Datasets with the VoID Vocabulary. Retrieved April 20, 2020, from https://www.w3.org/TR/void/

Kontokostas, D., & Knublauch, H. (2017). *Shapes constraint language (SHACL)* (W3C Recommendation) [https://www.w3.org/TR/2017/REC-shacl-20170720/]. W3C.

https://doi.org/10.23106/dcmi.953151686

Natasha Noy. (2018). Making it easier to discover datasets [Library Catalog: www.blog.google]. Retrieved April 23, 2020, from https://blog.google/products/search/making-it-easier-discover-datasets/

Natasha Noy, & Dan Brickley. (2017). Facilitating the discovery of public datasets [Library Catalog: ai.googleblog.com]. Retrieved April 10, 2020, from http://ai.googleblog.com/2017/01/facilitating-discovery-of-public.html

Poline, J.-B. (2019). From data sharing to data publishing. *MNI open research*, *2*. https://doi.org/10.12688/mniopenres.12772.2

Riccardo Albertoni, David Browning, Simon Cox, Alejandra Gonzalez Beltran, Andrea Perego, & Peter Winstanley. (2020). Data Catalog Vocabulary (DCAT) - Version 2. Retrieved April 20, 2020, from https://www.w3.org/TR/vocab-dcat-2/

Rietveld, L., & Hoekstra, R. (2013). YASGUI: Not Just Another SPARQL Client. In P. Cimiano, M. Fernández, V. Lopez, S. Schlobach, & J. Völker (Eds.), *The Semantic Web: ESWC 2013 Satellite Events* (pp. 78–86). Springer. https://doi.org/10.1007/978-3-642-41242-4_7

Tennison, J. (2016). *CSV on the web: A primer* (W3C Note) [https://www.w3.org/TR/2016/NOTE-tabular-data-primer-20160225/]. W3C.

Thalhath, N., Nagamori, M., Sakaguchi, T., & Sugimoto, S. (2019). Yet Another Metadata Application Profile (YAMA): Authoring, Versioning and Publishing of Application Profiles. *International Conference on Dublin Core and Metadata Applications*, *0*, 114–125. Retrieved January 31, 2020, from https://dcpapers.dublincore.org/pubs/article/view/4055

Thalhath, N., Nagamori, M., Sakaguchi, T., & Sugimoto, S. (2021). Wikidata Centric Vocabularies and URIs for Linking Data in Semantic Web Driven Digital Curation. In E. Garoufallou & M.-A. Ovalle-Perandones (Eds.), *Metadata and Semantic Research* (pp. 336–344). Springer International Publishing. https://doi.org/10.1007/978-3-030-71903-6_31

Thornton, K., Solbrig, H., Stupp, G. S., Labra Gayo, J. E., Mietchen, D., Prud'hommeaux, E., & Waagmeester, A. (2019). Using shape expressions (shex) to share rdf data models and to guide curation with rigorous validation. In P. Hitzler, M. Fernández, K. Janowicz, A. Zaveri, A. J. Gray, V. Lopez, A. Haller, & K. Hammar (Eds.), *The semantic web* (pp. 606–620). Springer International Publishing. https://doi.org/10.1007/978-3-030-21348-0_39

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., . . . Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship [Number: 1 Publisher: Nature Publishing Group]. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

https://doi.org/10.23106/dcmi.953151686