Semantic Web Oriented Approaches for Smaller Communities in Publishing Findable Datasets

Nishad Thalhath\(^1\), Mitsuharu Nagamori\(^2\), Tetsuo Sakaguchi\(^2\), Deepa Kasaragod\(^3\), and Shigeo Sugimoto\(^2\)

\(^1\) Graduate School of Library, Information and Media Studies, University of Tsukuba, Tsukuba, Japan
nishad@slis.tsukuba.ac.jp

\(^2\) Faculty of Library, Information and Media Studies, University of Tsukuba, Tsukuba, Japan
\{nagamori,saka,sugimoto\}@slis.tsukuba.ac.jp

\(^3\) Department of Neurobiology, Graduate School of Biomedical and Health Sciences, Hiroshima University, Hiroshima, Japan
deepa@hiroshima-u.ac.jp

Abstract. Publishing findable datasets is a crucial step in data interoperability and reusability. Initiatives like Google data search and semantic web standards like Data Catalog Vocabulary (DCAT) and Schema.org provide mechanisms to expose datasets on the web and make them findable. Apart from these standards, it is also essential to optionally explain the datasets, both its structure and applications. Metadata application profiles are a suitable mechanism to ensure interoperability and improve use-cases for datasets. Standards and attempts, including the Profiles (PROF) and VoID vocabularies, as well as frameworks like Dublin core application profiles (DCAP), provide a better understanding of developing and publishing metadata application profiles. The major challenge for domain experts, especially smaller communities intending to publish findable data on the web, is the complexities in understanding and conforming to such standards. Mostly, these features are provided by complex data repository systems, which is not always a sustainable choice for various small groups and communities looking for self-publishing their datasets. This paper attempts to utilize these standards in self-publishing findable datasets through customizing minimal static web publishing tools and demonstrating the possibilities to encourage smaller communities to adopt cost-effective and simple dataset publishing. The authors express this idea though this work-in-progress paper with the notion that such simple tools will help small communities to publish findable datasets and thus, gain more reach and acceptance for their data. From the perspective of the semantic web, such tools will improve the number of linkable datasets as well as promote the fundamental concepts of the decentralized web.

Keywords: Metadata · Data publishing · Datasets · Findable datasets · Static publishing · Semantic web

© Springer Nature Switzerland AG 2021
E. Garoufallou and M.-A. Ovalle-Perandones (Eds.): MTSR 2020, CCIS 1355, pp. 234–242, 2021.
https://doi.org/10.1007/978-3-030-71903-6_23
1 Introduction

1.1 Significance of Publishing Findable Datasets

Good data management practices are a prerequisite in any research or scientific quest for finding new knowledge or reusing existing knowledge. This is of significant importance to various data stakeholders including academia, industry, government agencies, funding bodies, journal publishers. Interoperability and reusability of data allows for transparency, reproducibility and extensibility of the data and hence the scientific knowledge.

One of the underlying guidelines for data management that ensures data findability is the FAIR principle [17]: —Findability, Accessibility, Interoperability, and Reusability formulated by FORCE11 – The Future of Research Communications and e-Scholarship. FAIR principles are not protocols or standards but are underlying guidelines that rely on data/metadata standards to allow for data to be reused by machines and humans [8].

The data citation principles from Data Citation Synthesis Group: Joint Declaration of Data Citation Principles by FORCE11 highlights the necessity of creating citation practices that are both human understandable and machine-actionable [5]. A recent trend of clear emerging consensus to support the best data archiving and citation policies with different stakeholders interest in view and comprising the privacy, legal and ethical implications has been seen [4]. The life sciences community are also warming up to idea of the need for better metadata practices in terms of annotating the experiments; for example, the information accompanying the nucleic acid sequencing data [13].

1.2 Standards in Defining Datasets

Dublin Core, Data Catalog Vocabulary(DCAT) [2], VOID -Vocabulary for Interlinked Data [9], schema.org\(^1\), etc. are some of the standards and vocabularies for defining data/metadata. These are based on FAIR principles and ensure data longevity and stewardship. W3C’s DCAT is an resource description file (RDF) vocabulary for data catalogs. It is based on six main classes and is designed to facilitate interoperability between data catalogs published on the Web. Dublin Core\(^\text{TM}\) Metadata Element (Dublin Core) and additional vocabularies referred to as “DCMI metadata terms” from Dublin Core\(^\text{TM}\) Metadata Initiative (DCMI) is described for use in RDF vocabularies for making data linkable\(^2\). The heterogeneity of metadata standards required across different dataset maintenance combining the Dublin Core vocabularies and SKOS vocabularies is addressed by DCAT. VOID is complementary to the DCAT model, uses RDF based schema for describing metadata. Schema.org provides framework to provide supporting information descriptions of data which makes data discovery easier through services like Google Dataset Search engine [15].

\(^1\) https://schema.org/.
\(^2\) https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.
1.3 Explaining the Data

Metadata application profiles (MAP) provide a means to express the schema customising the metadata instance by documenting the elements, policies, guidelines, and vocabularies for that particular implementation along with the schemas, and applicable constraints [6] including specific syntax guidelines and data format, domain consensus and alignment [1,7].

Data profiles are profiles that express the structure and organization data sources like CSV and JSON. Frictionless data DataPackage [16] is another example of data profiles which provides JSON schemas mostly on tabular data. Although CSVW is used in expressing data profiles of tabular data, it is more suitable in explaining linkable data from tabular data resources [14].

Profiles vocabulary [3] is an RDF vocabulary to semantically link datasets with the related profiles. Profile resources may be human-readable documents (PDFs, textual documents) or machine-actionable formats (RDF, OWL, JSON, JSON-LD), constraint language resources used by specific validation tools (SHACL, ShEx), or any other files that define specific roles within profiles.

1.4 Challenges for Smaller Communities

The major challenge for domain experts in smaller communities say researchers in small labs or data journalists, intending to publish findable data on the web, is the complexities in understanding and conforming data publishing standards. Mostly, these features are provided by complex data repository systems, which is not always a sustainable choice for various small groups and communities looking for self-publishing their datasets [12].

This is a work in progress paper which aims to incorporate FAIR based concepts into a minimal and sustainable digital publishing platform for data from smaller communities like small research labs, independent data journalists as well as organization or individuals who prefer not to depend on big data sharing platforms for political and technical reasons or for the sake of platform ownership.

2 Methods

2.1 Addressing the Problems in Self-publishing Datasets

Dataset publishing systems need to cater to more specific requirements than general digital repositories or web publishing platforms. They need to address emerging specifications and use cases on finding and curating datasets. One of the significant challenges for smaller stakeholders in adapting such a sophisticated system is its resource requirements. Most of the regular data publishing systems demand a bigger scale of resources, such as server-side processing and database dependency. Maintaining such infrastructure for a long time poses monetary overheads.
Along with its resource requirement, ensuring the security and updating the software stacks demands further resources. In the long run, any such unmaintained systems may expose security vulnerabilities and often causes dependency conflicts on its hosted infrastructure. Due to various challenges in maintaining such resource-intensive data publishing platforms, stakeholders with minimal resources will face challenges in ensuring the longevity of their datasets. For post-project datasets such as data collected from data journalism projects or crowdsourcing projects which prefer to self publish their data, than depending on a centralized data repositories may also find it challenging to maintain the platform for an uncertain duration.

2.2 Guidelines for Findable Dataset Publishing

There are different dataset indexing services built on various standards, principles, and interests. Data search services like Google Dataset Search\(^3\) and Socrata Open Data API\(^4\) are some of the emerging search systems designed for dataset searching. This attempt at this stage is primarily focusing on Google Dataset Search recommendations\(^5\) as a guideline and expand the scope to cover more standards and specifications from the Semantic Web perspective. Google Dataset Search helps to find datasets irrespective of the publishing platform. The guidelines are intended for the providers to describe their data in a machine-actionable way that can better help the search engines and indexers understand the content of the resource. Some of the salient requirements are the provenance, rights, and basic descriptive metadata of the datasets \([10]\). Some of the key challenges addressed by these recommendations are defining the constitutes of datasets, identifying the datasets, relating datasets to each other, linking metadata between related datasets, and finally describing the content of datasets \([11]\).

As a minimal prototype for the proposal, and Google Dataset Search as the initial focus of finding services, the authors included schema.org Dataset\(^6\) markup as main JSON-LD expression and DCAT with DublinCore as RDFa expression within the rendered HTML to explain the dataset. Since Schema.org is evolving as a well-accepted and unified vocabulary in describing resources for search engines, this approach has added advantage of better search engine visibility. The alternate DCAT and DublinCore RDF/RDFa expressions will also increase the reachability in other types of data finding mechanisms.

2.3 Extending a Static Publishing System

Hugo web publishing framework has emerged as one of the fastest framework for building websites\(^7\). Hugo, written in Go language, is open source based static

\(^3\) https://datasetsearch.research.google.com/.
\(^4\) https://dev.socrata.com/.
\(^5\) https://developers.google.com/search/docs/data-types/dataset.
\(^6\) https://schema.org/Dataset.
\(^7\) https://gohugo.io/.
site generator which is fast and flexible and supports different content types and
taxonomies which are extensible into multiple content output formats. Dynamic
JSON format based i18n provide translation module for multilingual publishing.

Under static publishing, files are updated on the file system thus making it
easily scalable and less resource intensive. Static publishing also allows for easier
management of content backups without having to deal with security issues or
server side maintenance involving CMS based approaches.

The advantage of extending existing frameworks for static publishing allows
communities to 1). Integrate data publishing to their existing websites, 2). Extend an existing open source projects, thereby reducing the risks of longevity
of the proposal, 3). Remove the redundant dependency on the efforts, thus sup-
porting the growth of the depending project 4). Allow minimal resource require-
ments for development and maintenance.

2.4 Template Driven Output Formats

Template driven output formatting are easier to customise, share and reuse. They
can be adapted to any similar system. Hugo’s Go-based templating\(^8\) supports
logic to build templates for any level of complex design. Go language’s html and
text templates are capable of generating virtually any text based formats\(^9,\)\(^10\). Different output formats include 1). HTML and embedded JSON-LD, 2). RDF
files for content negotiation.

2.5 Semantically Linking Published Profiles and Catalogues

The proposed solution uses Profiles vocabulary [3] to semantically link datasets,
and data catalogs to corresponding profiles. Schema.org based JSON-LD expres-
sion is used in linking datasets to data catalogs. DCAT and Dublin Core based
RDF expression is used to provide an alternative RDF expression of the pub-
lished item. This multi-level semantic linking will help the published instance to
be findable in different ways. Semantically linking datasets to data catalogues
and collections also links datasets to its corresponding profiles (Fig. 1). Gener-
ating RDF and JSON-LD helps to support finding profiles of the data.

2.6 Similar Attempts

Some of the similar approaches are enlisted below with the advantages and dis-
advantages: 1). Quire\(^11\) from J. Paul Getty Trust is a Hugo based multiformat
publishing format which is customised for book publishing and not datasets. 2).
DKAN\(^12\) is a Drupal 8-based, microservice-architected, schema-centered, API-
first, front-end decoupled, open data platform. It requires a web-server with PHP

\(^8\) https://gohugo.io/templates/.
\(^9\) https://golang.org/pkg/text/template/.
\(^10\) https://golang.org/pkg/html/template/.
\(^11\) https://gettypubs.github.io/quire.
\(^12\) https://getdkan.org/.
Fig. 1. Semantic relationship of dataset and application profile

support and a database. 3). JKAN\(^\text{13}\) is a Jekyll based small footprint backend-free open data portal, which work with GitHub pages as well. However, it doesn’t have semantic web features. 4). CKAN\(^\text{14}\) is designed for bigger data portals and is less semantic web friendly. It requires enormous resources to maintain the portal. 5). Invenio\(^\text{15}\) which is Python based and developed for larger digital repositories like zenodo. However, it is an overkill for smaller communities.

3 Resulting Project

3.1 HuDot: A Minimalist Publishing System

This is a work in progress project that proposes to integrate these concepts into a reusable implementation based on the Hugo web publishing framework. The work in progress project is titled HuDot and shall be made available under MIT license. Content within HuDot is organized on items and collections. Page types are initially dataset, collection, profiles, and a text type for regular pages. A general overview and the relationship of these pages are illustrated in Fig. 2.

The proposed minimalist publishing system is an open-source project under MIT license, which is a permissible license. The project’s longevity is maintained as long the Hugo dependencies are maintained. The authors intend to keep the templates interoperable with any Hugo implementation so that the data publication can be seamlessly integrated with any Hugo based websites. The emerging

\(^{13}\) https://jkan.io/.
\(^{14}\) https://ckan.org/.
\(^{15}\) https://invenio-software.org/.
The popularity of Hugo as a static publishing solution for various communities can make easy adoption of these templates in independently publishing datasets. The whole HuDot systems is designed by limiting any external dependencies other than Hugo. This may help the adapters in keeping their sites irrespective of the longevity of the HuDot project. As the project progresses, detailed documentation explaining the semantic web approaches in developing these templates and customizing them will also be added.

4 Discussion

4.1 Novelty and Impact

This approach brings in the newly proposed Profiles vocabulary for semantically linking and expressing profiles of datasets and encourages stakeholders in publishing metadata application profiles and data profiles along with their datasets. The most crucial gap this attempt tries to fill in is proposing an easier option to deal with the complex requirements in resource and knowledge for publishing findable datasets. Limited resource communities and independent data publishers find it challenging to create essential semantic web oriented resources in expressing a findable dataset. Most of the time, such datasets end up in cloud file storage or linked to web pages, which may not express the content and nature of the dataset. Eventually, this makes it difficult to be exposed as a data resource. Creating adaptable and less resource-intensive tools which support semantic web technologies by design will increase the adaptability of such technologies. This may also reduce the cost and efforts for individuals and resource-limited communities in integrating semantic web concepts.

4.2 Limitations of the Approach and Future Work

Our approach is not a replacement for any full-fledged data archival and data curation systems like Dkan or hosted systems like zenodo. To use a content management framework like Hugo and customize the templates requires certain
level of technical expertise and it can be a challenge for many users. Many of
the communities may find it difficult to customize templates for non HTML
outputs. Extending the utility with improved HuDot based tooling to generate
linkable and linked (open) data from datasets using associated profiles is to be
considered. This will also involve providing extended documentation and theme
support for easier adaptation.

5 Conclusion

The authors have described a work-in-progress project on developing a simple
static site publishing tool based on Hugo to help smaller data stakeholders like
data journalists, individual research labs to publish findable data. This would
help towards promoting more reach and acceptance of dataset in terms of self-
publishing. From the perspective of the semantic web, such tools will improve
the number of linkable datasets as well as promote the fundamental concepts of
the decentralized web.

Acknowledgements. This work was supported by JSPS KAKENHI Grant Numbers
JP18K11984 (to Nagamori) and 19K206740A (to Kasaragod).

References

1. Baca, M.: Introduction to metadata (2016). http://www.getty.edu/publications/
intrometadata
2. Browning, D., Perego, A., Albertoni, R., Cox, S., Beltran, A.G., Winstanley, P.: Data
catalog vocabulary (DCAT) - version 2. W3C Recommendation, W3C (2020).
https://www.w3.org/TR/2020/REC-vocab-dcat-2-20200204/
3. Car, N.: The profiles vocabulary. W3C Note, W3C (2019). https://www.w3.org/
TR/2019/NOTE-dx-prof-20191218/
4. Cousijn, H., et al.: A data citation roadmap for scientific publishers. Sci. Data 5(1),
180259 (2018). https://doi.org/10.1038/sdata.2018.259
5. Crosas, M.: Joint declaration of data citation principles - FINAL (2013). https://
www.force11.org/datacitationprinciples
6. Heery, R., Patel, M.: Application profiles: mixing and matching metadata schemas.
Ariadne 25 (2000). http://www.ariadne.ac.uk/issue/25/app-profiles/
7. Hillmann, D.: Metadata standards and applications. Metadata Management Asso-
ciates LLC (2006). http://managemetadata.com/
8. Jacobsen, A., et al.: FAIR principles: interpretations and implementation con-
siderations. Data Intelligence 2(1–2), 10–29 (2019). https://doi.org/10.1162/
dint_r_00024
9. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing linked datasets
with the VoID vocabulary. https://www.w3.org/TR/void/
10. Noy, N.: Making it easier to discover datasets (2018). https://blog.google/
products/search/making-it-easier-discover-datasets/, library Catalog: www.blog.
google
11. Noy, N., Brickley, D.: Facilitating the discovery of public datasets (2017). http://
ai.googleblog.com/2017/01/facilitating-discovery-of-public.html, library Catalog:
ai.googleblog.com
12. Sansone, S.A., et al.: FAIRsharing as a community approach to standards, repositories and policies. Nature Biotechnol. 37(4), 358–367 (2019). https://doi.org/10.1038/s41587-019-0080-8

13. Stevens, I., Mukarram, A.K., Hörtenhuber, M., Meehan, T.F., Rung, J., Daub, C.O.: Ten simple rules for annotating sequencing experiments. PLoS Comput. Biol. 16(10), e1008260 (2020). https://doi.org/10.1371/journal.pcbi.1008260

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

15. Thalhath, N., Nagamori, M., Sakaguchi, T.: MetaProfiles - a mechanism to express metadata schema, privacy, rights and provenance for data interoperability. In: Ishita, E., Pang, N.L.S., Zhou, L. (eds.) ICADL 2020. LNCS, vol. 12504, pp. 364–370. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-64452-9_34

16. Walsh, P., Pollock, R.: Data package. https://specs.frictionlessdata.io/data-package

17. Wilkinson, M.D., et al.: The FAIR guiding principles for scientific data management and stewardship. Sci. Data 3(1), 160018 (2016). https://doi.org/10.1038/sdata.2016.18. https://www.nature.com/articles/sdata201618