Design of a Framework to Support Reuse of Open Data about Agriculture

Alec Gordon[1], Mohammad Sadnan Al Manir[1],
Brandon Smith[1], Amir Rezaie[1] and Christopher J.O. Baker[1]

Department of Computer Science
University of New Brunswick, Saint John, Canada
{alec.gordon}[at]live.com
{sadnan.almanir,brandon.smith,amir.rezaie,bakerc}[at]unb.ca

Abstract. Online Datasets in Open Data Portals typically have minimal metadata and users wishing to consider their reuse in extended analyses are poorly served. One approach is to find and re-annotate the metadata according to subject-specific, community adopted vocabularies. In support of this we explore a multi-tiered framework combining the capabilities of a crawler, a tagger and a recommendation engine, as well as tools for the provisioning of data as discoverable services. We provide details of prototype scale implementations of these components and a cursory evaluation of the tagger for subject-specific metadata enrichment using the Global Agricultural Concept Scheme (GACS).

1 Introduction

1.1 Open Data

Open Data is emerging as a best practice for provisioning citizens and stakeholders with data that can be explored and reused to support decision making. National governments, Non-government organizations (NGOs) and corporations are increasingly making datasets available online for public access. With the growing number of datasets and end users, the task of identifying the precise topics of the data is increasingly relevant. Open Data Portals (ODP) provide search capabilities to look for data sets based on custom collections of tags assigned by system managers. More than 2000 such portals[1] exist worldwide. Despite their proliferation, few portals have adopted systematic tagging procedures and the lack of high quality Open Data descriptions is one of the key factors hindering the utility of these portals as a go to archive for discovering online data [1]. In particular, some researchers have identified open challenges, namely low-rate tag adoption by users, the existence of multiple different tags for the same meaning, and redundancy of some tags in Open Governmental Data Portals [1]. Further to this, there is a need for more granular subject specific tags and more efficient ways of applying metadata to datasets in ODPs [2].

[1] https://opendatainception.io/

1.2 Open Data about Agriculture

A rudimentary search for Open Data tagged with the term *agriculture* identified datasets in a variety of country/geography specific portals [3] including USA[2], UK[3], France[4], Australia[5], Canada[6], Netherlands[7], and the continent of Africa[8]. These datasets are published in a range of data formats and the permitted modes of access can vary also. The following formats were found; CSV, XML, HTML, GML, FGDB/GDB, PDF, DOC, ArcGIS, KML, ODT, ZIP, API, ArcGIS Map Service, XLSX, JSON and RDF/OWL. Given that ODP's often use a limited number of tags, a more granular breakdown of the specific subtopics is necessary for the domain of agriculture. Recently, the Agriculture Open Data Package (AgPack[9]) has introduced 14 key data categories on agriculture policy and food security perspectives that can be applied to datasets, albeit such tags are not yet in common use in ODPs.

Earlier approaches to publishing structured Open Data have leveraged community adopted controlled vocabulary terms and dataset definitions expressed in Resource Description Framework (RDF) serialization formats, known as as Linked Open Data [4]. This approach affords users the option to query over linked data using the SPARQL[10] query language. One such deployment of this approach is the Agronomic Linked Data project (AgroLD) [5] which provides access to data resources about plants in the form of an RDF graph for domain experts, such as bioinformaticians. The extent to which target data is readily discoverable and queryable depends on the skills of the end users who need to be proficient with SPARQL and related tools.

In recent years, the Global Open Data for Agriculture and Nutrition (GODAN[11]) project has advocated for the publication of open data and the creation of ecosystems where agricultural data is Findable, Accessible, Interoperable, and Reusable (FAIR) [6].

1.3 Target Functionality and Design Challenges

The current state of the ODPs containing agricultural data provides a good motivation for the creation of a dedicated infrastructures that supports comprehensive Open Data exploration for potential reuse. Primarily, users want to i) search for and query across globally distributed agricultural datasets based on multiple keywords and defined relations, and ii) retrieve integrated data in

[2] `https://catalog.data.gov/dataset`
[3] `https://data.gov.uk/`
[4] `https://www.data.gouv.fr/en/datasets/`
[5] `https://data.gov.au/dataset`
[6] `https://open.canada.ca/data/en/dataset?portal_type=dataset`
[7] `https://data.overheid.nl/data/dataset`
[8] `https://africaopendata.org/dataset`
[9] `https://opendatacharter.net/agriculture-open-data-package/`
[10] `https://www.w3.org/TR/sparql11-overview/`
[11] `https://www.godan.info/`

a unified standard format so that they are compatible and readily usable with third party tools.

In order for an infrastructure to support these capabilities it needs to address the following tasks: (i) regular crawling of the Web for sites related to agriculture, (ii) screening of Open Data files and indexing them, (iii) downloading and scanning the files for key agriculture vocabulary terms, (iv) generating subject specific metadata for the data files, (v) recommending relevant datasets based on curated metadata, (vi) change management and revision of metadata, (vii) provision of data resources as discoverable Web services, and (viii) publishing data according to interoperability standards.

In this paper we propose a multi-tier framework, Section 2, for the harvesting of Open Data files, subject specific enrichment of metadata, and the provisioning of Open Data as services. Using the target use case of Open Data about agriculture and leverage of the Global Agricultural Concept Scheme (GACS) we provide details of prototype scale implementations, Section 3, and a cursory evaluation of the tagger in, Section 4. In Section 5 we briefly discuss the framework in the context of the target functionality and list future work. Section 6 contains concluding remarks.

2 Framework

The multi-tier framework presented in in Figure 1 provides a solution to support better discovery and reuse of Open Agricultural Data.

As shown in Figure 1, the *Data Sources* column displays two sources of data: i) Open Agricultural Datasets which are generated and collected based on typical agricultural activities, and ii) the Controlled Vocabulary of Agriculture and Nutrition such as the Global Agricultural Concept Scheme (GACS) consisting of standard vocabularies which are agricultural concepts mapped from three well known sources: the AGROVOC multilingual agricultural thesaurus by the Food and Agricultural Organization (FAO) of the United Nations, the CAB Thesaurus by the Centre for Agriculture and Biosciences International (CABI), and the NAL Thesaurus by the US National Agricultural Library [7].

In *Phase 1*, the country-specific ODPs hosting the Open Agricultural Datasets are crawled and indexed for further processing. The crawler uses seed URLs of the ODPs as inputs, fetches contents such as text, data, and hyperlinks from recursively-linked pages, parses and stores them as segments, from which an index is then created. Off-the-sheft crawlers[12] and indexers[13] can also be used for this purpose.

In *Phase 2*, the index is enriched and updated using a *tagger* Individual data files are downloaded and parsed, and relevant tags are added based on a custom scoring algorithm that ranks words matching to the controlled vocabulary.

In *Phase 3*, a Semantic Recommendation System is used to suggest relevant datasets to end users, which can then be further curated in preparation for

[12] `http://nutch.apache.org/`
[13] `http://lucene.apache.org/solr/`

integration with other datasets. Further enrichment of metadata using mapping to external ontologies can be incorporated also.

In *Phase 4*, access to data as services is provided using SADI Semantic Web services [8]. Services are generated by Valet SADI [9] over fully enriched semantic metadata descriptions mapped to data schemes. Services are deployed in a service-registry and can be discovered, invoked, orchestrated into workflows and executed automatically using a SADI specific semantic query client.

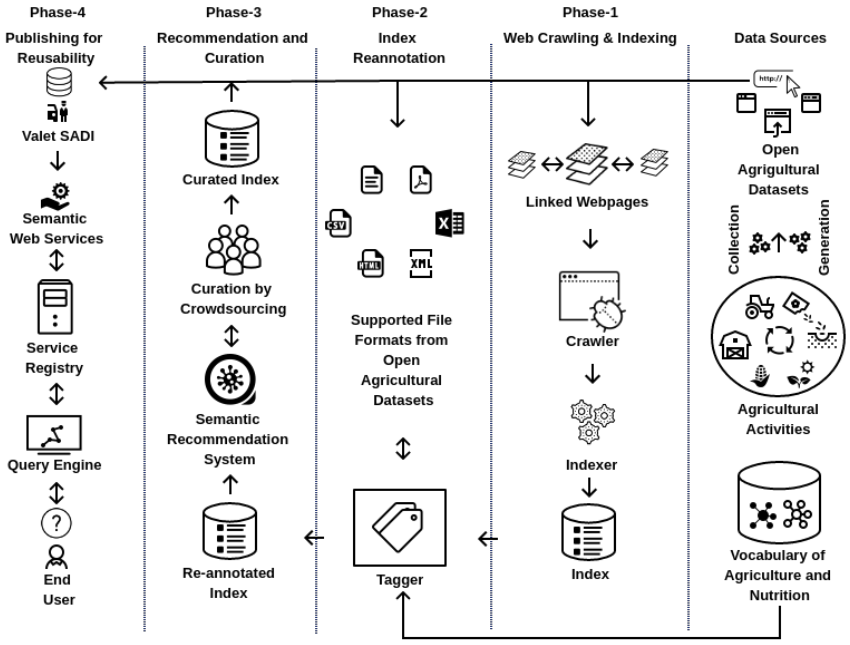

Fig. 1. A multi-tier framework for supporting reuse of Open Data about agriculture. In *Phase 1* webpages hosting datasets are crawled and indexed. In *Phase 2* the index is re-annotated using agriculture-specific controlled vocabularies. In *Phase 3* datasets are recommended to the end users against a search-term using a recommendation system. This phase also allows curation by crowd-sourcing. In *Phase 4*, access, interoperability and queryability across datasets are provisioned via SADI Semantic Web services.

3 Implementation

The development of the framework is ongoing and the implementation is at the preliminary stage, albeit a light-weight crawler, tagger, and recommendation engine have been developed and are undergoing testing. Here we provide an outline of these components with particular emphasis on the performance of the tagger, which plays an essential role for the subsequent phases to be successful.

3.1 Crawler

The crawler in *Phase 1* recursively scans through the ODP pages and sub-pages describing each dataset and their URLs. The crawler saves this information locally in segments which are parsed and structured into fields by an indexer. An index of the datasets containing descriptions and metadata is created. The file formats currently supported by the crawler are Zip (.zip), Microsoft Excel (.xls, .xlsx), Portable Document Format (.pdf), Comma-separated values (.csv) and Text (.txt). Similar functionalities are provided by the recently introduced *Dataset Search*[14] by Google™.

3.2 Tagger

The tagger in *Phase 2* is used to enrich the descriptions of the datasets by adding metadata from expert-authored controlled vocabularies. The core features of the tagger are the use of (i) an in memory vocabulary graph generated from a controlled vocabulary file and (ii) a custom scoring algorithm based on lexical matching of terms in data files to the terms in the vocabularies.

The current implementation of the tagger uses the vocabularies from GACS to create a graph where the nodes in the graph are terms or concepts. Before a node is created in the vocabulary graph, stemming is applied so that each term is reduced to its root form. The concept hierarchies of the vocabulary contain both broader concepts (as in superclass in ontologies) which identify parent nodes and narrower concepts (as in subclass in ontologies) which identify child nodes.

Scoring of Annotations The tagger reads each word from the input data file and applies stemming. It then searches for both an exact match and a stem match in the vocabulary graph. If a lexical match, with or without stemming, to a concept is detected, a score is added to the term and to each of its broader concept terms in the graph based on their depth in the hierarchy. The narrower concepts (more specific and lower down the hierarchy) are assigned lower scores to avoid the selection of concepts that don't provide significant information. Once scoring is complete, the upper 3rd percentile of concepts are selected as annotations for the document. This provides a barrier excluding tags that are unrelated to the content of a document but are still contained in it, such as terms from sources and references. The current deployment of the tagger excludes matches to geographical locations because of their widespread use and marginal relevance in the current study.

Augmented Tagging with Broader Concepts To illustrate how the scoring provides additional tagging to the datasets a simple example is shown for illustration and intelligibility purposes. The tagger was run on the dataset titled

[14] `https://toolbox.google.com/datasetsearch`

Wheat/Barley and their Products[15] hosted at the Open Government Portal[16] maintained by the Government of Canada. This file contains mentions of *Wheat* and *Barley* but not *Cereals*.

Table 1 shows tags annotated to the Open Data file with and without the introduction of the scoring technique. Without the implemented scoring technique (tagging of lexical and stem-based matches to GACS) the tagger can identify only terms directly mentioned in the files. Using the adopted scoring technique the term *Cereals*, the parent term for *Wheat* and *Barley* in GACS is retrieved.

Tags with lexical/stem matching	Tags with scoring
import, export, wheat, barley, permits	cereals

Table 1. Annotations for the dataset titled 'Wheat/Barley and their Products'

The GACS hierarchy[17], shown below, for the preferred term *wheat* illustrates how the broader concept *cereals* is related to the narrower. Moreover, the scoring can be extended to retrieve multiple parent terms in the hierarchy including cases where multiple inheritance may occur.

$$... > crops > fieldcrops > graincrops > cereals > wheat$$

Metadata and tags provided when the file was submitted to an ODP can be enriched in a systematic way by using the tagger, namely with lexically matched terms found in GACS. The scoring algroithm additionally provides subject specific tags that are broader in scope. In the subsequent phase of the framework only the enriched datasets, including the lexically matched tags and the broader augmented tags, are used by the recommendation engine to filter and categorize data according to users' interests, *Phase 3*.

3.3 Semantic Recommendation Engine

The recommendation engine in *Phase 3* currently uses a *content-based filtering* method, where extensive tagging of data files and custom scoring of matched tags is employed to determine the level of similarity between files. The engine uses the initial preferences of a user, which can be obtained from tagging an online publication specified by the user.

Upon request for recommendation, all datasets are scored according to their relevance to the tags within the user's profile. Scoring is done by multiplying the normalized weight of each tag by the normalized weight of a matching tag within the user's profile. The cumulative score for each document is then compared pairwise and the highest scoring documents are returned to the user as a

[15] `https://open.canada.ca/data/en/dataset/3a4e7f9b-64d2-432f-8394-15f6814aad62`
[16] `https://open.canada.ca/en`
[17] `http://browser.agrisemantics.org/gacs/en/page/C212`

recommendation. Additionally, a history of the suggested files is stored within the user's profile to avoid repeat runs offering the same recommendations. The engine was tested for both programmatic functionality and the quality of the recommended datasets. Preliminary test results show the greater the numbers of annotations, the better the relevance of the recommended datasets. Extension of the recommendation engine will include the use of additional community developed ontologies and inferencing based on subsumption, transitivity.

4 Preliminary Results of the Tagger

The tagger was run on a machine running Ubuntu 17.10 server with a 4-core 3 GHz processor and 8 GB memory. During the experiment, the tagger tried to match data from 212 CSV datasets hosted on FAOStat[18] and Data.gov[19] to the beta version of GACS controlled vocabulary. The outcome of the initial experiments showed that the scoring worked surprisingly well for most of the datasets. As is to be expected, the tagger worked best when data files contained meaningful agriculture related terms and performed worst when data files contained terms mostly as names, identifiers and numeric values.

Table 2 shows an analysis of results derived after running the tagger on 5 random datasets. The *Topics* column indicates what type of information the data files contain, the *Tags* column indicates if the value of a score crossed the threshold to select any tags or not, and the *Outcome* indicates whether the matching performance of the tagger is best case, moderate case, worst case or resulted in a false positive. For some data files the selected tags were found to be false-positive as well as false-negative. Due to space constraints a rigorous analysis of the tagger is beyond the scope of this paper. However, in testing it was found that Open Datasets are very broad in scope and their composition is complex as they often are published as spreadsheets, invoices, and statistical reports. Often, the rows and columns can only be explained by an expert in the subject area or by the data provider. It is also difficult to interpret when numerical values with units are present.

Thus, although automatic tagging may work for some datasets, for many other datasets it is prone to errors. Therefore, it is recommended that the tags added automatically be verified manually by experts before approving the files for use in the recommender system in the subsequent phase.

5 Discussion

We have outlined a framework designed to address the challenges described in Section 1.3. In addition, we have been able to corroborate the general feasibility of our approach in so far as harvesting, tagging and recommending files to users. At the current time the tools implemented in this framework

[18] `http://www.fao.org/faostat/en/#data`
[19] `https://www.data.gov/`

Title of the dataset	Topics	Tags	Outcome
Incidental catch at BC marine finfish aquaculture sites	Time, location, facility, common and scientific name of the fish	321 Words were tried, 266 matches, and 9 top scoring tags	Best case
Adult Salmon Health (Snorkel Surveys) Cape Breton Highlands	waterbody, species, age and quality	58 Words were tried, 51 matches, and 2 top scoring tags	Moderate case
USDA FSA Farm Payment Name/ Address File for 2008	Names and addresses	None matched	Worst case
USDA FSA Farm Payment File for 2010	Identifiers and numerical values	None matched	Worst case
Pineapple - Average retail price per pound and per cup equivalent, 2013	Packages and market price	fertilizers	False positive

Table 2. Performance summary of the Tagger on datasets

are yet to mature and more experiments are required to assess and improve their performance. The idea of harvesting files in ODPs likely motivated the development of *Dataset Search* by Google^{TM} where users are provided with an overview of the metadata assigned by the original publisher of the datasets. In our pilot studies, we were able to further enrich the metadata for individual files providing agriculture specific tags from GACS that extend beyond the metadata provided by the dataset publisher. Compared to the techniques described in related work [10,1], the tagging approach implemented in our framework finds tags by traversing each word of the data file and by applying lexical and semantic matching to an expert-curated, subject specific controlled vocabulary instead of reusing the existing tag libraries shared between ODPs. These portals tend to use tags that are generally broad in scope as opposed to subject specific. Our methodology additionally has the benefit of being domain agnostic and alternate vocabularies other than GACS could be supplemented e.g. for Open Data files about health topics.

With end users in mind, the recommendation system we implemented was designed to support users who are looking for recently published candidate data files and consider them for reuse. In addition, it can support users wishing to participate in crowdsourcing and provisioning of data as services. Indeed, the greater goal for the framework includes the provision of Open Data as services over which ad hoc queries can be run. This is possible if the data can be sufficiently well structured, annotated with metadata and could support meaningful queries across data sets. Given that our system is still in development and since

we have not processed large volumes of Open Data files we have yet to determine the extent to which Open Data files can be readily made available as services. We have proposed to leverage SADI Semantics Web services given that registries of SADI services, along with associated query tools, can support the target functionality where complex workflows of combined data retrieval and data analytics services can be run. Moreover we can point to recent work where researchers [11,12] report the use of the SADI Semantic Web services in agriculture for surveillance tasks in precision irrigation and precision dairy farming use cases. More recently we conducted pilot studies in the creation of services for a decision support system in agricultural operations management. SADI services were created to fetch target trait data for eggplant varieties and compute costs, revenue and profits for individual eggplant varieties. User provided values for market prices and estimated crop yields were required as inputs [13]. Whereas these services were build manually, more recent reports show the utility of Valet SADI for the automated generation of services in the domain of *malaria analytics* [14,15], where a registry of services specific to malaria insecticide resistance surveillance queries was built.

6 Conclusion

We have presented a prototype to annotate Open Data files with subject specific tags on agriculture. The target objective is to make Open Data in ODPs more discoverable and intelligible for potential data reuse purposes. We have proposed to do this using a multi-phase approach involving crawling and indexing of Open Datasets, a custom tagging approach leveraging lexical term matching and a scoring algorithm. Files enriched with tags in this way are then made available to a recommendation engine to support alerting of end users. Subsequent to this we proposed the provisioning of data as services with semantic descriptions to support ad hoc federation of data in response to complex user queries.

References

1. Alan Tygel, Sören Auer, Jeremy Debattista, Fabrizio Orlandi, and Maria Luiza Machado Campos. Towards cleaning-up open data portals: A metadata reconciliation approach. In *ICSC*, pages 71–78. IEEE Computer Society, 2016.
2. Wei Wei, Zhanglong Ji, Yupeng He, Kai Zhang, Yuanchi Ha, Qi Li, and Lucila Ohno-Machado. Finding relevant biomedical datasets: the UC San Diego solution for the bioCADDIE Retrieval Challenge. *Database*, 2018(1):bay017, 2018.
3. David Corsar and Peter Edwards. Challenges of open data quality: More than just license, format, and customer support. *J. Data and Information Quality*, 9(1):3:1–3:4, 2017.
4. Christian Bizer, Tom Heath, Kingsley Idehen, and Tim Berners-Lee. Linked data on the web (ldow2008). In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 1265–1266, New York, NY, USA, 2008. ACM.
5. Stella Zevio, Nordine El Hassouni, Manuel Ruiz, and Pierre Larmande. Agrold indexing tools with ontological annotations. In *Proceedings of the 9th International*

Conference Semantic Web Applications and Tools for Life Sciences, Amsterdam, The Netherlands, December 5-8, 2016., 2016.

6. Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3, 2016.
7. Thomas Baker, Caterina Caracciolo, Anton Doroszenko, and Osma Suominen. GACS core: Creation of a global agricultural concept scheme. In *Metadata and Semantics Research - 10th International Conference, MTSR 2016, Göttingen, Germany, November 22-25, 2016, Proceedings*, pages 311–316, 2016.
8. Mark Wilkinson, Benjamin Vandervalk, and Luke McCarthy. The Semantic Automated Discovery and Integration (SADI) Web service Design-Pattern, API and Reference Implementation. *Journal of Biomedical Semantics*, 2(1):8, 2011.
9. Mohammad Sadnan Al Manir, Alexandre Riazanov, Harold Boley, Artjom Klein, and Christopher J. O. Baker. Valet SADI: provisioning SADI web services for semantic querying of relational databases. In *IDEAS*, pages 248–255. ACM, 2016.
10. Alexandre Passant. LODr - A Linking Open Data Tagging System. In *Proceedings of the First Social Data on the Web Workshop (SDoW2008)*, Karlsruhe, Germany, October 27 2008.
11. Wilfried Wöber, Klemens Gregor Schulmeister, and Christian Aschauer et al. agriOpenLink: Adaptive Agricultural Processes via Open Interfaces and Linked Services. In M. Clasen, M. Hamer, S. Lehnert, B. Petersen, and B. Theuvsen, editors, *GIL Jahrestagung*, volume 226 of *LNI*, pages 157–160. GI, 2014.
12. Slobodanka Dana Kathrin Tomic, Wilfried Wöber, and Sandra Hörmann et al. Enabling Semantic Web for Precision Agriculture: a showcase of agriOpenLink Project. In A. Filipowska, R. Verborgh, and A. Polleres, editors, *SEMANTiCS (Posters Demos)*, volume 1481 of *CEUR Workshop Proceedings*, pages 26–29. CEUR-WS.org, 2015.
13. Mohammad Sadnan Al Manir, Bruce Spencer, and Christopher J. O. Baker. Decision Support for Agricultural Consultants With Semantic Data Federation. *IJAEIS*, 9(3):87–99, 2018.
14. Jon Haël Brenas, Mohammad Sadnan Al Manir, Christopher J. O. Baker, and Arash Shaban-Nejad. A malaria analytics framework to support evolution and interoperability of global health surveillance systems. *IEEE Access*, 5:21605–21619, 2017.
15. Jon Haël Brenas, Mohammad Sadnan Al Manir, Kate Zinszer, Christopher J. O. Baker, and Arash Shaban-Nejad. Exploring semantic data federation to enable malaria surveillance queries. In *Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth - Proceedings of MIE 2018, Medical Informatics Europe, Gothenburg, Sweden, April 24-26, 2018*, pages 6–10, 2018.