Leveraging Guides to Empower Open Data Research

Christina Christodoulakis, Moshe Gabel, and Angela Demke Brown

Department of Computer Science, University of Toronto, Canada
{christina, mgabel, demke}@cs.toronto.edu

Abstract. Data packages in Open Data repositories often contain data guides: supplementary materials with information supporting interpretation and consumption of contents of files containing tabular data. This short paper describes the design of a system that discovers, unifies and links metadata from guide files to Open tabular data. Enriching tabular data will facilitate tasks like table search, interpretation, and integration for Open Data users such as scientists and journalists.

Keywords: Open Data · Metadata Discovery · Tabular Data.

1 Introduction

Governments and industry are embracing Open Data as they recognize the impact on scientific, economic, social, and environmental development of communities [2]. While this data is freely available, discoverability, accessibility and reusability remain significant barriers from a stakeholder perspective [5,6].

Automatically annotated header lines of tabular data found in Open Documents, such as in [3], provide some information about the contents of the files, usually encoded in attribute names, that can be used for search and integration [4]. Packages in Open Data repositories often contain *data guides*, which are supplementary materials often in tabular format with information supporting interpretation and consumption of contents of files containing tabular data and are associated with a specific set of tables in data files. Metadata described in guides may include extended semantics and contextual information for tables, attributes, and attribute values, as well as other metadata such as languages used, data types, formatting, units, scales, etc.

Users wanting to efficiently search, interpret, and combine tabular data from Open Documents cannot easily benefit from them as is currently no automated way of leveraging guides. Some of the main challenges to solving this include diverse contents, structures, and naming conventions. Figure 1 shows an example adapted from a set of real Open Data files. It shows a data table extracted from the file `electricity.csv`, and guide annotations extracted from a supplementary file `dictionary.csv`.[1]

[1] CSV is a popular Open Data format widely used in a variety of domains for its simplicity and effectiveness in storing and disseminating data, and is frequently used to describe data guides.

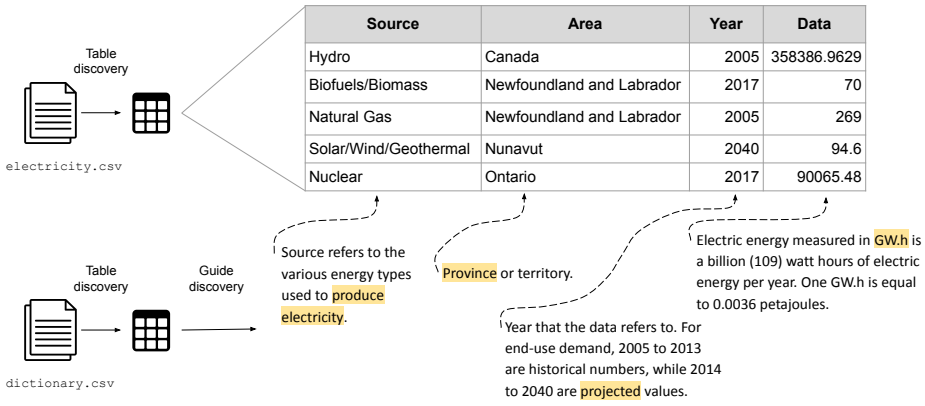

Fig. 1. An example Open Data table extracted from file `electricity.csv` is matched to attribute guides extracted from guide file `dictionary.csv`.

We use the example in Figure 1 to demonstrate several benefits of automatic guide discovery, unification, and linking:

- **Table Search:** Given only the data table in Figure 1 and the query `"projected electricity generation per Canadian province"` an information retrieval engine is unlikely to return this table in top ranked results. However, if `dictionary.csv` is identified as a guide file, and the metadata it describes is properly extracted and combined with the data table annotations, the user can now discover this table successfully, as the description of the attribute `Year` (identified and extracted from a guide file) informs us that there is a range of years for which the data is historical and another range for which it is **projected**, and the description of the attribute `Area` informs us that the values are **provinces** or territories.
- **Interpretation:** A user consuming the data in Figure 1 must know that for this particular table, data recorded up to 2013 is historical, while the rest is projected. Value guides can indicate aggregation rows (e.g., `Area` = `Canada`), and explain value semantics in greater detail.
- **Integration:** Consider a user that wants to build a data set of electricity production across North America, and discovers a data set for US electricity production across states, with electricity production measured in KWh (kilowatt-hour), however, their seed data set as seen in Figure 1 records data in GWh (gigawatt-hour). Knowing the units and scale of the data will add a much needed step of data conversion before integration.

We focus on discovering, unifying, and linking guides from CSV files to tables annotated in CSV files such that approaches supporting tasks like search and integration may benefit from previously unused rich metadata. This requires designing an end to end system that addresses table discovery, guide discovery and linking of the two. Such a system could be a great asset in connecting open data sleuths such as scientists and journalists with open data tables.

2 System Design and Implementation

We are designing a system which will scan CSV files crawled from an Open Data repository to discover and annotate tables. The system will process the annotated tables to discover guides, which it will extract and unify to a common schema. Following that, the system will discover the links between unified guides and annotated tables. Finally, the system will present users with an interface for table and guide annotation, browsing, and review.

Guide discovery and extraction: While some Open Data repositories support and encourage annotating published resources as guides, more often than not such files are not explicitly annotated, and do not follow a single naming convention. Furthermore, formatting of guide files is not uniform, making automatic extraction of guide elements challenging. We studied a random sample of 100 Open Data packages with CSV guide resources crawled from Open Data repositories and observed guide files with guide information related to tables, table attributes, and attribute values. Ideally their respective guides are each well structured tables, with each row corresponding to guides of a table, table attribute, or attribute value respectively. In practice, we frequently observe element guide information presented as merged tables (e.g., attribute and attribute value guides combined in a single guide table, guides for attributes of multiple tables combined in a single guide table, etc.), rotated, or even nested.

Data Model: Following our observations on Open CSV guides, we design a unifying data model for guides as well as software that extracts and maps information from existing guides to the unified model. For the unifying model, we identified a set of guide fields for table attributes (see Table 1) and attribute values. In a preliminary evaluation on a second random sample of 50 guide files, our model captures the information in these files successfully.

Table 1. Guide fields identified for table attributes. OPT indicates optional fields, MLT indicates multilingual.

Field	OPT	MLT	Description
Header	-	-	Short text for identifying the attribute.
Title	-	✓	A version of the header in natural language text.
Description	✓	✓	A detailed definition of the attribute in natural language text.
Note	✓	✓	Natural language text with context on the attribute data.
Unit	✓	-	Attribute value units of measurement (e.g., L, mpH, $, %, etc.).
Scale	✓	-	Scale of the reported attribute values (e.g., billions, 10^{-2}, etc.).
Domain	✓	-	Legal values for an attribute. Defined by a range or dictionary.
Datatype	✓	-	E.g., text, integer, Boolean, date, etc.

Open Document data tables may be published with partially or fully encoded data values. For example, encoded values may be used to represent missing or

redacted information (e.g., `*`, `n/d`, `x`, `NA`, etc.), or an encoding scheme may be used for all data values (e.g., replacing province names with province codes given a value dictionary). Furthermore, guides may contain rich descriptions for non-encoded values. We identify a value guide fields as a subset of fields used for attribute guides, namely, a title, description, and notes.

Prototype implementation: Our prototype has several components; a Web frontend in JavaScript, a Flask-based backend supported by a PostgreSQL relational database implementing the model we name GuideDB, and a Lucene [1] backend for indexing and customized search. We provide an API for writing and reading JSON annotations to and from the database.

Tables in Open CSV (comma separated value) files can pose a significant challenge to identify due to significant variety in structure and formatting. For table discovery in CSV files we use Pytheas, a weighted rule-based table discovery system for CSV files [3].[2] We are currently designing and implementing a hybrid rule-based/learning-based approach to automatically identify guide tables, classify table structure into a set of known designs, and unify guide information into a common schema. We are also designing a customized ranking algorithm based on Lucene and table and attribute similarity distance functions that take into account annotated guides to support table search and integration.

Via the Web interface, users can manually add or generate automated table annotations on CSV files, save automated annotations, or edit or generate their own. Users can annotate multilingual column headers, scales, and units, identify guide tables, extract and normalize guide fields, and match guide fields to a data table, a table attribute, or an attribute value. The interfaces support editing of annotations persisted in GuideDB. We use this interface to annotate the ground truth against which we will evaluate our automatic guide discovery and unification methods.

References

1. Apache Lucene, https://lucene.apache.org/
2. Capgemini Consulting: Creating Value through Open Data: Study on the Impact of Re-use of Public Data Resources (2015), accessed: 2019-09-23
3. Christodoulakis, C., Munson, E., Gabel, M., Brown, A.D., Miller, R.J.: Pytheas: Pattern-based table discovery in CSV files. PVLDB **13**(11), 2075–2089 (2020)
4. Liu, Y., Bai, K., Mitra, P., Giles, C.L.: Tableseer: Automatic table metadata extraction and searching in digital libraries. In: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries. p. 91–100. JCDL '07, Association for Computing Machinery, New York, NY, USA (2007)
5. Miller, R.J., Nargesian, F., Zhu, E., Christodoulakis, C., Pu, K.Q., Andritsos, P.: Making open data transparent: Data discovery on open data. IEEE Data Eng. Bull. **41**(2), 59–70 (2018)
6. Máchová, R., Hub, M., Lněnička, M.: Usability evaluation of open data portals: Evaluating data discoverability, accessibility, and reusability from a stakeholders' perspective. Aslib Journal of Information Management **70** (05 2018)

[2] https://github.com/cchristodoulaki/Pytheas