Metadata assessment for efficient open data retrieval

Chiao-Ling Kuo
Center for GIS, Research Center for Humanities and Social Sciences, Academia Sinica
128 Academia Road, Section 2
Nangang, 115 Taipei City, Taiwan
kuo@chiaoling.com

Han-Chuan Chou
Center for GIS, Research Center for Humanities and Social Sciences, Academia Sinica
128 Academia Road, Section 2
Nangang, 115 Taipei City, Taiwan
ossoumano@gmail.com

Abstract

After making data available to the public, the open data movement currently focuses on retrieving data efficiently and precisely for further application. Thus, this study proposes a metadata-based assessment mechanism toward metadata quality, spatial similarity, and temporal similarity analyses between user request and open data. By assessing the metadata of each dataset, metadata scoring in the mechanism contributes two scores, mandatory and optional scores, which are calculated using mandatory fields, and combination of recommended and optional metadata fields. Spatial similarity analysis is conducted by extracting keywords from user input or by addressing location on a map. Further, temporal similarity analysis is conducted by comparing a query and the open data pool. We obtain 36,000 open data from a main portal, the *advocate sharing platform*, in Taiwan for our experiments. A developed platform intuitively presents related data with score, similarity, and rank that help users obtain expected data. Our method effectively retrieves open data, can be a valuable demo site for open data promotion, and can be used as reference by other agencies worldwide.
Keywords: Open data, metadata, assessment, data quality, spatial similarity, temporal similarity.

1 Introduction

Open data is data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike (Open Knowledge International, 2019). Making data available to the public has many advantages, such as facilitating data reuse and public participation, sparking public creativity, increasing government transparency and accountability, and decreasing data production cost. Releasing open data online is a global trend, and many countries have enacted open data policies, and engaged in national or cross-national projects/programs to promote open data (make data open) (European Union, 2019; OpenDataSoft, 2019; UK Government, 2019; U.S. Government, 2019). In Taiwan, an open data portal, the Government Open Data Platform (National Development Council, 2019a) was established by the National Development Council in 2013. At present, over 36,000[1] open datasets are available on the Government Open Data Platform.

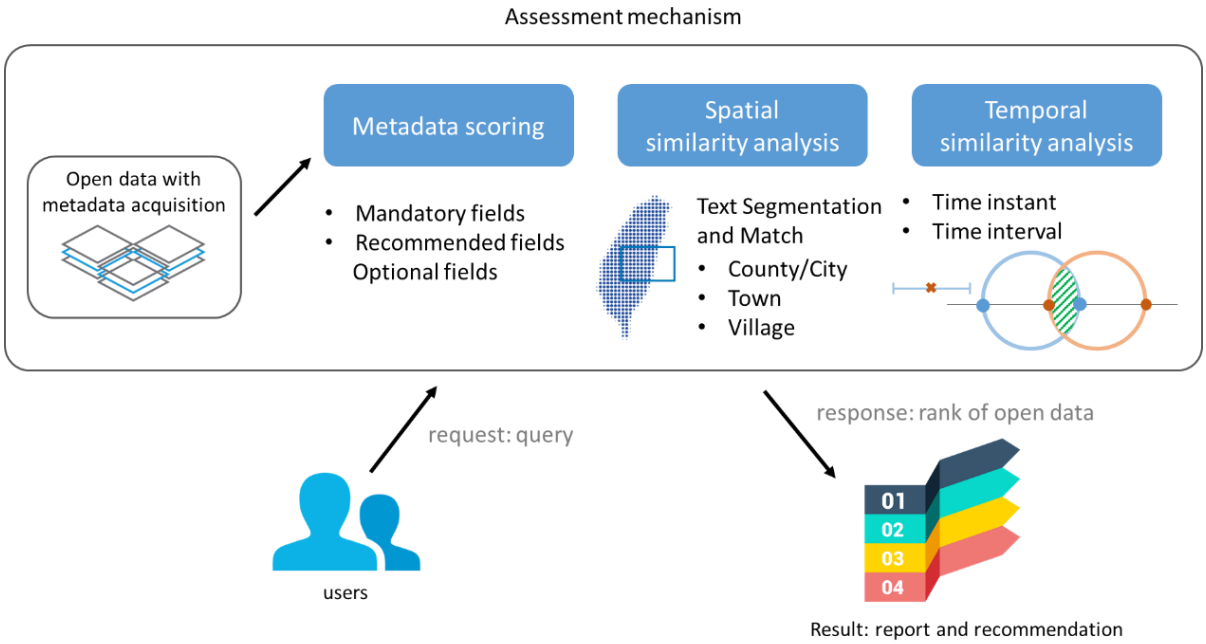
After making data available to the public, the open data movement is currently raising data and metadata accessibility worldwide. The movement is also improving data quality to support research, collaboration, transparency, and sustainability. Data quality assessment toward six dimensions, namely, retrievability, usage, completeness, accuracy, openness, and contactability, has been discussed and applied to CKAN-based portals by an Open Data Portal Watch framework(Umbrich, Neumaier & Polleres, 2015). Furthermore, Neumaier, Umbrich & Polleres (2016) proposed a metadata quality assessment metrics composed of existence, conformance, retrievability, accuracy, and openness based on the World Wide Web Consortium (W3C) Data Catalog Vocabulary (DCAT) metadata. (Vetrò et al, 2016) also suggested a measurement framework that contained traceability, currentness, expiration, completeness, compliance, understandability, and accuracy for open data assessment. To semantically retrieve open data, Degbelo et al (2016) proposed the semantic application programming interface (API). Our discussions indicate that data quality assessment and data retrieval effectiveness are still the major streams in the open data movement. Thus, we propose an efficient and simple mechanism for open data retrieval in this study. This mechanism includes metadata quality assessment, spatial similarity, and temporal similarity analyses between user requests and open datasets. The rest of this paper is arranged as follows. Sections 2 and 3 presents our method and provides the implementation results, respectively. Section 4 presents our conclusion and future work.

2 Method

To increase open data transparency and improve data retrieval confidence, a metadata-based assessment mechanism toward data quality and temporal and spatial similarities for supporting data accessibility is proposed in this research. Figure 1 illustrates the structure of the assessment mechanism, which has three components: metadata scoring, spatial similarity analysis, and temporal similarity analysis. Metadata

[1] Date retrieved: October 22, 2018

Figure 1: Open Data assessment mechanism.



plays a crucial role in examining data quality and retrieving data. Thus, metadata scoring is conducted by assessing the quality and validity of 35 fields from the metadata. While fields are provided, field values are eximined by cheching if they comply with specific formatting guidelines, logical consistency, conformance, etc. Spatial and temporal similarity analyses indicate the extent of similarity between query requests from users and open data in a portal from the spatial and temporal perspectives, respectively. The assessment mechanism can provide a detailed report of data quality, spatial similarity, and temporal similarity for open data recommendation. Additionals details are described in the following subsections.

2.1 Metadata scoring

Metadata contents are assessed using the proposed scoring method on the basis of the 35 fields including three categories, catalog, dataset, and distribution from the dataset metadata standard specifications (DMSS) (National Development Council, 2019b) for acquiring good-quality open data. Metadata fields have three types that are distinguished by necessity: mandatory, recommended, and optional fields. To clarify the completeness of mandatory and nonmandatory fields, mandatory (only mandatory fields are included) and optional (recommended and optional fields are included) scores are calculated. Each score is determined by examining the standard rules. For example, the value of the identifier field must be a 16-digit string composed of a 10-digit unit ID or agency ID and a 6-digit serial number combined by the character "-" (e.g., A41000000G-000001). Another example is the case if the change frequency of data matches the requirements/rules. If a written value of a field matches the rule of that field, then the field is marked. A mandatory score is obtained by a percentage number converted from the number of the marked fields of the total mandatory field. For example, 20/23*100 ≒ 87, which indicates that open data have a mandatory score of 87 (100 is full marks). The optional score is also calculated (23 mandatory, 10 recommended, and 2 optional fields). The higher the score, the better the quality of the metadata contents for the dataset.

2.2 Spatial similarity analysis

Spatial similarity analysis helps identify datasets that satisfy the conditions set by users from the spatial perspective. The proposed mechanism provides two modes for this process: the keyword search mode and the select-by-location mode. The keyword search mode parses user input. This mode matches the retrieved names of an administrative range or point of interest, such as the names of a city, town, village, place, or landmark, to several spatially relevant columns of the metadata, including title, description, and spatial range. The select-by-location mode allows users to query open data by drawing a polygon based on a location from a map in different administrative levels, such as county/city, town, or village. To fulfil the requirements of returning single or multiple data, users can type in a keyword or multiple keywords separated by commas in the keyword search mode, and select an option with either a single or multiple values in the select-by-location mode. Both options identify only open data with the highest value (single value) and all values (multiple values) with a descending order of intersection (unit: percentage).

2.3 Temporal similarity analysis

Temporal similarity analysis aims to retrieve datasets that fulfil user requirements by time. Metadata has four fields with regard to time: temporalCoverageFrom, temporalCoverageTo, issued (data published time), and modified field. In accordance with the property of temporal fields, time instant and time interval are two possible cases combined by temporal

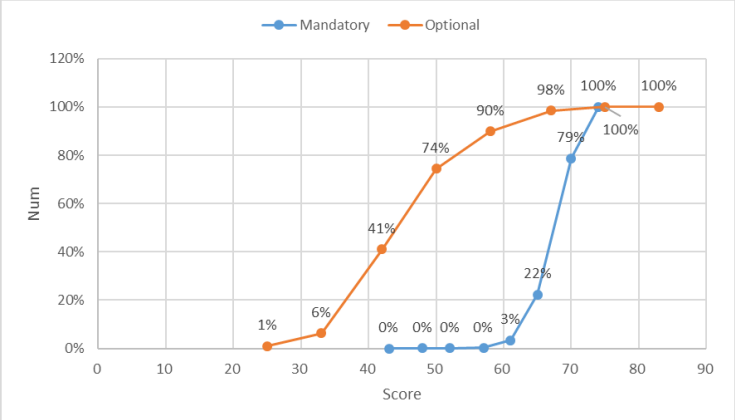
fields. That is, the issued and modified fields belong to time instant, whereas the temporalCoverageFrom and temporalCoverageTo fields form and belong to time interval. Because users may query data by time instant or time interval, four possible query combinations between user query and queried data are: point to point, point to line, line to point, and line to line. In such a line to line comparison, such as the time interval case from the queried data and the time interval query mode, can generate a similarity calculated by an intersection represented as a percentage. By contrast, the others such as time instant fields queried by the time interval mode or data with time interval fields queried by the time instant mode, uncertain comparisons of time while data without starting time or end time, or point to point comparison, can be represented by topological relationships, such as contains, containedby, disjoint, equal, touch, or overlap (Egenhofer & Franzosa, 1991). Using the percentage and topological relationships for the representation of the temporal similarity between user query and queried data is a good way of realizing the quality and the dimension of comparison towards time.

3 Results

Open data with metadata are collected through REST-based API by sending a URL request with an identifier from the government's open data platform (National Development Council, 2019a). The platform provides over 36,000 datasets composed of 18 categories, namely, Health During Pregnancy (493), Birth and Adoption (56), Schooling and Education (631), Military Service (210), Seeking Employment and Employment (486), Opening New Companies (544), Marriage (7), Investment and Financial Management (1747), Travel and Leisure (841), Transportation and Communications (1738), Medicine (915), Home Ownership and Moving (690), Elections and Voting (78), Living Safety and Quality (2386), Retirement (21), Elderly Care (231), Post-death Affairs (74), and Public Information (28,206)[2], from central and local government units. In this study, we assume that all datasets have metadata information according to the open data policy (National Development Council, 2013) addressing that a dataset must have metadata information while making data to public.

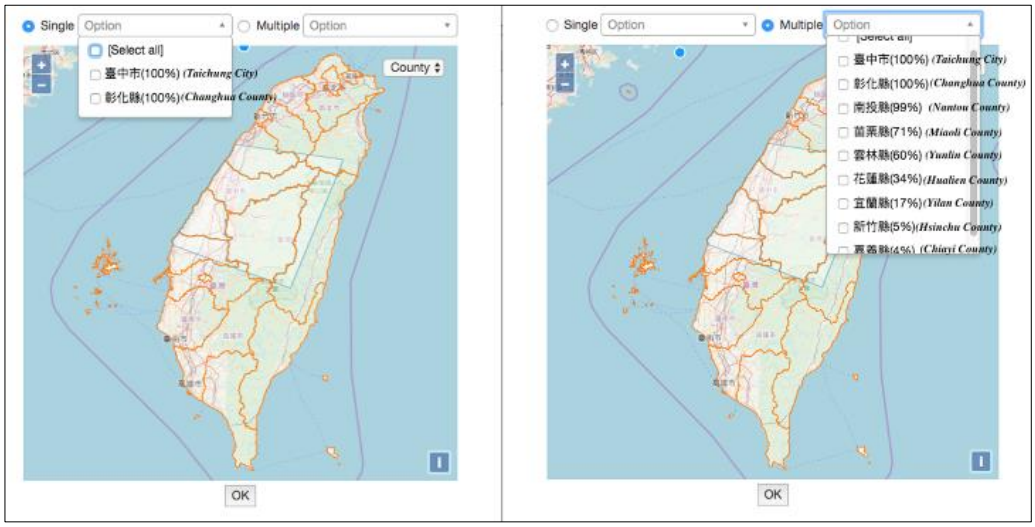
Mandatory and optional scores toward metadata scoring for initial metadata quality assessment are obtained by implementing the proposed approach for efficient open data retrieval. Figure 2 shows that approximately 80% of data obtained 70 and 20% obtained 74 for the mandatory score. By contrast, approximately 90% of data obtained 58 for the optional score among over 36,000 datasets (100 is full marks). The line chart indicates that metadata quality can still be improved considerably. Furthermore, we develop a platform (Fig. 3) in which users can query, understand, and retrieve ideal data by using the represented metadata score and the spatial and temporal similarities through spatial and temporal requests, respectively. The platform provides a keyword search function that can deal with non-spatial keywords, such as topics or themes, and spatial keywords, such as county names, among the 35 metadata fields. To provide a user-friendly platform interface, users can make a spatial query via a map by drawing a polygon, as shown in Fig. 3(a). By selecting a single or multiple radio option button, users can acquire datasets from one to multiple regions located at the county, town, or village levels. For example, on the left part of Fig. 3(a), users draw a polygon and select the single option at the county level. They can then obtain open datasets of Taichung City and Changhua County because the polygon has the largest intersection area in the aforementioned city and county. Conversely, on the right side of Fig. 3(a), users select the multiple option at the county level. Accordingly, they obtain open datasets of all intersecting counties/cities, Taichung City, Changhua County, and Nantou County, with an intersection percentage. Users can also directly evaluate the similarity of a spatial request. With regard to temporal similarity analysis, Fig. 3(b) shows a time instant query with a relationship similarity presentation, whereas Fig. 3(c) displays a time interval query with a percentage similarity presentation. The design of the open data assessment can simultaneously present data quality, spatial similarity, and temporal similarity. Such a design can also provide rank information for an intuitive and efficient access.

Figure 2: Metadata scoring result. Mandatory and optional scores.

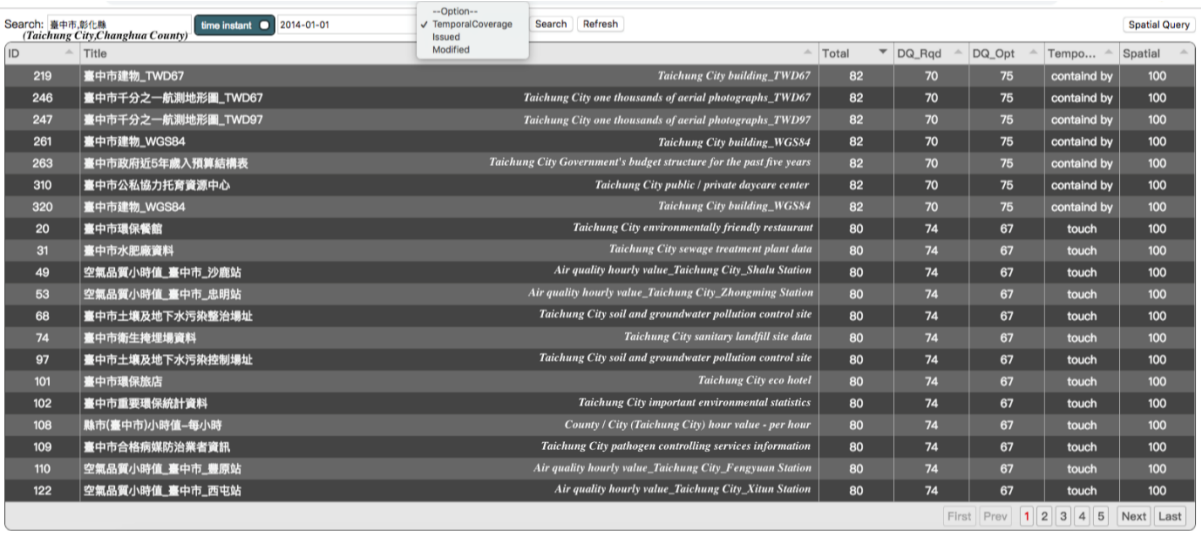


[2] The number in the bracket indicates the amount of open dataset in that category. The accessed date is February 1, 2019.

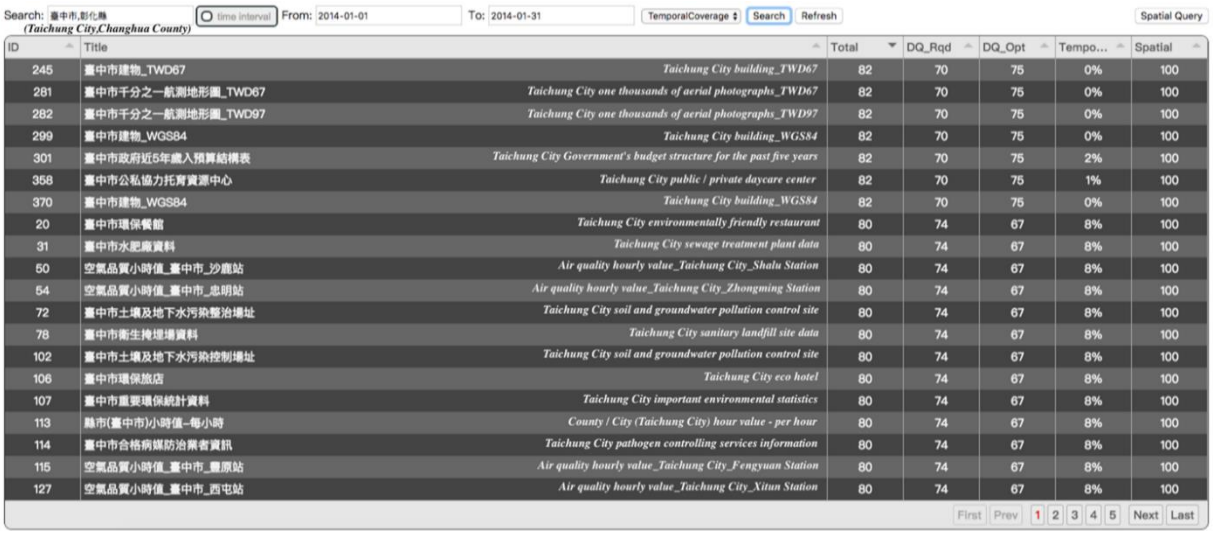
Figure 3: Implementation result. (a) With spatial similarity analysis result, (b) with temporal similarity analysis (time instant), and (c) with temporal similarity analysis (time interval).



(a)



(b)



(c)

4 Conclusion

A massive amount of open data is published online, and the open data movement currently focuses on efficient data retrieval for a wide range of uses and applications. Therefore, this study proposes a metadata-based mechanism that assesses data quality and analyses spatial and temporal similarities between user requests and open datasets. A metadata scoring composed of a mandatory score calculated by mandatory fields and an optional score calculated by recommended and optional fields is helpful in assessing data quality. Moreover, the design of spatial and temporal similarities presents matched open data, thereby satisfying user requirements. Our implementation, which involves more than 36,000 open data with a self-developed platform, demonstrates that the proposed method is feasible and can be adopted by other agencies or countries to establish an effective open data portal. In future studies, we will utilize data content to achieve precise spatial similarity. Furthermore, semantic similarity analysis will be included and conducted on metadata and data contents.

Acknowledgments

This work was supported by the Ministry of Science and Technology (MOST), Taiwan (R.O.C.), under grant MOST 107-2119-M-001-009-MY3.

References

Degbelo, A., Trilles Oliver, S., Kray, C., Bhattachaya, D., Schiestel, N., Wissing, J. & Granell Canut, C. (2016) Designing Semantic Application Programming Interfaces for Open Government Data.

Egenhofer, M. J. & Franzosa, R. D. (1991) Point-set topological spatial relations. *International Journal of Geographical Information System*, 5(2), 161-174.

European Union (2019). European Union Open Data Portal. Available online: https://data.europa.eu/euodp/home (accessed on 20 February 2019)

National Development Council (2013). Government Open Data Platform Use Specification. Available online: https://data.gov.tw/license/legacy (accessed on 22 February 2019)

National Development Council (2019a). data.gov.tw. Available online: https://data.gov.tw/en (accessed on)

National Development Council (2019b). Dataset metadata standard specifications. Available online: https://data.gov.tw/node/18252 (accessed on 20 February 2019)

Neumaier, S., Umbrich, J. & Polleres, A. (2016) Automated quality assessment of metadata across open data portals. *Journal of Data and Information Quality (JDIQ)*, 8(1), 2.

Open Knowledge International (2019). The Open Data Handbook. Available online: http://opendatahandbook.org/guide/en/what-is-open-data/ (accessed on 6 April 2019)

OpenDataSoft (2019). Open Data Inception - 2600+ Open Data Portals Around the World. Available online: https://opendatainception.io/ (accessed on 20 February 2019)

U.S. Government (2019). DATA.GOV. Available online: http://data.gov/ (accessed on)

UK Government (2019). data.gov.uk. Available online: http://data.gov.uk/ (accessed on 20 February 2019)

Umbrich, J., Neumaier, S., ; & Polleres, A. (2015) Quality assessment and evolution of open data portals 2015. IEEE.

Vetrò, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R. & Morando, F. (2016) Open data quality measurement framework: Definition and application to Open Government Data. *Government Information Quarterly*, 33(2), 325-337.