A Survey of Metadata Elements for Provenance Provision in China Open Government Data Portals

Chunqiu Li
School of Government,
Beijing Normal University,
China
lichunqiu@bnu.edu.cn

Yuhan Zhou
School of Government,
Beijing Normal University,
China
18811350608@163.com

Kun Huang
School of Government,
Beijing Normal University,
China
huangkun@bnu.edu.cn

Abstract

The open government movements facilitate the transparency and sharing of government data. Provenance of open government data (OGD) describes source information related to who, how, where, when and other information over the lifecycle of OGD. Provenance of OGD should be tracked for high-quality and trustworthiness of OGD. Currently, OGD portals provide provenance through general metadata elements, such as creator, provider, creation date, publication date, and issued time. In China, local OGD portals define their own metadata profiles. However, these metadata elements in different OGD portals vary and there is no specific and well-defined provenance description scheme for OGD in China. Therefore, this paper is purposed to survey the current provision situation of provenance metadata elements in 42 China OGD portals and conduct the unification of provenance elements based on the survey results. This research is meaningful to facilitate formal description of provenance information in China OGD portals.

Keywords: metadata; provenance; government dataset; open government data

1. Provenance Necessity for Open Government Data

Provenance is quite widely used for trustworthiness, integrity and reproducibility for data-centric sciences. Provenance is especially critical in the web environment for data that are aggregated from distributed data sources and for data that are derived from other data sources.

In the open data context, it is required to provide high-value OGD with good provenance referring to the details about the origins of OGD, such as who created the OGD, how the OGD was created. Over the OGD lifecycle, OGD might be changed from local data to regional data, national data, and international data. OGD might be modified and become to new data. Provenance of OGD also includes the information about how the OGD was modified or manipulated. Provenance is becoming more and more important for data quality in the open context. We need to know how provenance of OGD is provided by OGD portals in China with the development and efforts of open government movements.

Thus, this paper focuses on the survey of current situation of provenance provision in China OGD portals. We also unified the metadata elements for provenance information provided by 42 local OGD portals in China. Both the survey and unification process were conducted from dataset level and distribution level based on Data Catalog Vocabulary (DCAT) published in 2014 for government data interoperability. The research work presented in this paper is part of our research project, which aims to propose a formal description scheme for provenance metadata based on the survey results.

2. Data Catalog Vocabulary for Interoperability of Government Data

DCAT is an RDF vocabulary to describe government data catalogs on the web. DCAT defines a set of classes and properties to describe information of government data. DCAT published in

2014 defines three main classes, i.e., dcat:Catalog, dcat:Dataset, dcat:Distribution. According to the definition of DCAT, the three classes respectively mean as follows:

"dcat:Catalog represents the catalog.

dcat:Dataset represents a dataset in a catalog.

dcat:Distribution represents an accessible form of a dataset as for example a downloadable file, an RSS feed or a web service that provides the data ."

DCAT is used to facilitate interoperability between data catalogs published on the web. Metadata can be easily discovered and consumed with the use of DCAT terms. In this paper, we identified and unified provenance elements of OGD based on the dataset level and distribution level defined in DCAT. This research selected DCAT for the interoperability of metadata among China OGD portals.

3. Provenance-centric Survey of OGD Portals in China

3.1. Selection of China Local OGD Portals

The selection principles are designed as follows: (1) the portal is accessible; (2) the portal is approved by governments with "gov.cn" in the domain name; (3) the portal provides accessible and structural datasets through download URI or API; (4) the local portal represents the city above prefecture level. Finally, we selected 42 local OGD portals including 8 provincial level, 7 sub-provincial level, and 27 prefecture level portals.

3.2. Provenance Provision Current Situation of China Local OGD Portals

We surveyed the 42 China local OGD portals and recorded the metadata elements representing source information about OGD, such as maintainer, created time, update time, and other provenance elements. The provenance elements are listed for each local government city as shown in Table 1. We collected provenance related metadata elements from the dataset level and distribution level following DCAT, respectively.

From the total number of elements both on dataset level and distribution level, the metadata element numbers for provenance information vary from 1 to 12 among 42 portals. It is noted that only two portals provide detailed information to describe history data, for example, version information, update time information, download link, etc.

TABLE 1: Metadata elements for provenance provision in the 42 China local OGD portals.

Local Government City	**Metadata Elements for Provenance Provision on Dataset Level**	**Metadata Elements for Provenance Provision on Distribution Level**	**Total Number of Elements**
Guiyang	release date, provider department, provider address, data maintainer, source system, latest update date, open pattern/open property, update frequency, online resource link	upload time, download link, interface address	12
Ningbo	institution, update period, update date, expiration date, open type, resource release time	download link, history data (version number, data update time, version invalidated time, download link)	12
Beijing	resource publication date, resource ownership department, update time	download link, upload time, interface address, history data（update time, download link）	9
Zhejiang	release department, update date, information resource provider, information resource release date, online resource link, update frequency	upload time, download link, interface address	9
Zhongshan	update time, department name, dataset ending date, update frequency, open level	download link, history data, interface address	8

Shenzhen	release date, update date, data provider, update frequency, open property	interface address, created time, download link	8
Ningxia/Guangzhou/ Harbin/Foshan/Huizhou /Jiangmen	open status, source department, release time, last update, update frequency	upload time, download link, API interface	8
Shandong and other 16 cities	open status, source department, release time, last update, update frequency	upload time, download link, API interface	8
Guangdong/Zaoqing	update frequency, created time, last modified time, open level, data provider department	upload time, download link	7
Guizhou	data provider, latest update time, security level/open property, update frequency	upload time, download link, interface address	7
Shanghai	first release date, data provider, release date, open property, update frequency	download link, API address	7
Zhanjiang	update time, update frequency, release date, data provider address, open property, data provider department	download link	7
Wuhan	update time, data source, data release date, source link	download link, history record, API interface	7
Dongguan	source department, update period, latest update	update time, download link, API interface	6
Jiangxi	update date, data source/provider, release date, update period	download link	5
Suzhou	affiliation, release time, open property	upload time, download link	5
Meizhou	provider, update date	update time, download link	4
Yangzhou	provider, release date, sharing level	download link	4
Jingmen	release date, data provider department	download link	3
Wuxi	/	download link	1

3.3. Unification of Collected Metadata Elements for Provenance Description

The use of unified metadata elements assists in the exchange of OGD among OGD portals. However, we found that the sampled 42 OGD portals use distinct element names for the same meaning. Therefore, we unified the collected metadata elements for provenance description in China local OGD portals from the dataset level and distribution level as shown in Table 2. We used brief labels with clear meaning for unification as listed in the "Unified Element" column.

TABLE 2: Unification of metadata elements on Dataset and Distribution level

Dataset Level		**Distribution Level**	
Unified Element	**The Element Unifies the Following Elements**	**Unified Element**	**The Element Unifies the Following Elements**
release time	release date, information resource release date, resource release time, resource publication date, release time, first release time, data release time, release date	release time	release date, release time
update time	latest update date, update date, update time, last update, last modification time, last update time, latest update	history data	history data, history record
update frequency	update frequency, update period	API interface	Interface address, API interface, API address
online resource link	online resource link, online resource link address, source link		
provider	provider institution, information resource provider, data provider department, data provider		
provider information	provider address, data provider address		
resource ownership department	resource ownership institution, affiliation department		
access limitation	open pattern, open property, open level, open status, security level, open property, sharing level		

3.4. Analysis of Coverage Fraction of Collected Provenance Elements

We classified the elements from both dataset and distribution level, respectively. From the counted number of portals that provide the same element, we calculated the coverage fraction of each collected provenance element as shown in Table 3. The elements with high coverage fraction indicate that they are widely recognized and used in China OGD portals. In this research, we defined the elements with coverage fraction above 70% as mandatory elements, and the elements with coverage fraction below 70% as recommended elements.

TABLE 3: Coverage fraction of collected provenance element in the 42 China local OGD portals.

Class	Metadata Element	Number of Portals Providing the Element	Coverage Fraction	Obligation Level
dcat:Dataset	provider	38	90.48%	Mandatory
	update time	38	90.48%	Mandatory
	release time	37	88.10%	Mandatory
	update frequency	35	83.33%	Mandatory
	access limitation	33	78.57%	Mandatory
	online resource link	3	7.14%	Recommended
	dataset ending date	2	4.76%	Recommended
	provider information	2	4.76%	Recommended
	resource ownership department	2	4.76%	Recommended
	data maintainer	1	2.38%	Recommended
	release institution	1	2.38%	Recommended
	source system	1	2.38%	Recommended
	created time	1	2.38%	Recommended
dcat:Distribution	download link	41	97.62%	Mandatory
	API interface	32	76.19%	Mandatory
	release time	30	71.43%	Mandatory
	history data	4	9.52%	Recommended
	update time	2	4.76%	Recommended
	upload time	1	2.38%	Recommended
	created time	1	2.38%	Recommended

Based on the analysis of the coverage fraction, we could get the following conclusions: (1) On the dataset level, the provenance elements named "provider", "update time", "release time", "update frequency" and "access limitation" are widely used by China local OGD portals; (2) On the distribution level, the provenance elements named "download link", "API interface" and "release time" are widely used by China local OGD portals; (3) The other elements with occurrence within 2 to 4, such as "history data", "data maintainer" and "release institution" are not widely used. However, the information about the derivation history and responsible agents of OGD are important provenance, which shall be recommended for provision in China OGD portals.

This work reports the current outcome of our on-going project. We are trying to reuse classes and properties of DCAT and PROV for provenance description in China OGD portals. Due to the space limitation, we will not introduce our final research achievement in detail. Here, we only give a simple example of the RDF data following our proposed provenance description scheme. The below example describes the download URL, access URL, issued date and other information about an instance of dcat:Distribution related to Beijing Hospital List.

```
:BeijingHospitalList.csv
    a dcat:Distribution, prov:Entity;
    dcat:downloadURL <http://www.bjdata.gov.cn/cms/web/download/attachmentID=13>;
    dcat:accessURL
<http://www.bjdata.gov.cn:8/cms/web/APIInterface/userApply.jsp?id=136&key=PID>;
    dct:issued "2018-11-12"^^xsd:date;
    owl:versionInfo "2.0";
    prov:wasMemberOf :ChinaHostpitalList;
    prov:alternateOf :BeijingHospticalList.xls.
```

4. Discussion

4.1. Analysis of the Provenance Provision Survey of China OGD Portals

Based on the above survey findings, we identified and summarized provenance provision situation of China OGD portals: the lack of a unified provenance description schema, and incomplete provenance provision of OGD. The provenance information provided by China OGD portals is mostly embedded in webpages without the use of Semantic technologies and machine-readable representation in RDF syntax. However, machine-readable provenance documentation is recommended and required for high-quality OGD.

Open Data Certificate proposed by the Open Data Institute is classified to four levels: Bronze Level, Silver Level, Gold Level, and Platinum Level. Platinum is the highest level which requires the following points: (1) data uses a machine-readable format; (2) data uses open standard machine-readable formats; (3) machine-readable provenance documentation; (4) machine-readable metadata (documentation), etc. Therefore, recording provenance of OGD as structured description is required for interoperable data exchange of government data. In the future, China OGD portals shall pay more attention to and make more efforts of machine-readable provenance description for interoperability of government data.

4.2. Comparison of Provenance Elements between Foreign and Domestic OGD Portals

It is useful to figure out the similarities and differences of provenance elements between foreign and domestic OGD portals through comparison. Therefore, we extracted and recorded the metadata elements for provenance information provided by four typical OGD portals as shown in Table 4.

TABLE 4: Provenance Elements Provided by Four Foreign OGD Portals

Portal	Provenance Elements
European Data Portal	Harvested from, Source, Last Updated, Created, Contact Point, Issued, Temporal, Publisher, Modified, Conforms To, Provenance, Page, access url, download url, License, rights, identifier
Data.gov	modified, publisher, contactPoint, programCode, describedBy, accessLevel, bureauCode, distribution, identifier, rights, license, references
Data.gov.uk	Published by, Last updated, Licence, Data links (Link to the data, File added), Additional information (Added to data.gov.uk, Access constraints, dataset reference date, Frequency of update, Responsible party, Source), mandate, lineage, Contact (Enquiries, Freedom of Information (FOI) requests), Licence information
Data.gov.au	creator, date, availability, publisher, audience, contributor, coverage, mandate, relation, rights, source

On one hand, same with China cases, the four OGD portals in EU, US, UK and AU also mostly provide responsible agents, date information and modification information as important provenance of OGD. The responsible agents usually refer to publisher, creator, and contributor. The date information usually refers to created date and updated date. Modification information usually refers to update frequency.

On the other hand, we identified the following differences: (1) The foreign OGD portals emphasize license information while China OGD portals rarely provide such provenance elements; (2) The foreign OGD portals define the detailed constraints of provenance elements in their metadata schemas. The China OGD portals usually do not define clarified metadata schemas.

5. Conclusion and Future Work

This paper shows and analyzes the survey results of provenance provision in 42 sampled China local OGD portals. We reported the widely used provenance elements based on the survey results and unified provenance elements of the sampled portals. We also briefly compared provenance metadata elements in China OGD portals with typical OGD portals in other countries.

For the future work, we plan to propose a framework for provenance description of China OGD in RDF syntax. The proposed framework designed with a basis on DCAT and PROV defines the following contents: meanings of provenance elements, constraints of provenance elements, usage of provenance elements, etc. The design and verification of the proposed framework will be introduced in another journal paper.

References

Attard, Judie, Fabrizio Orlandi, Simon Scerri, and Sören Auer. (2015). A Systematic Review of Open Government Data Initiatives. Government Information Quarterly. Volume 32, Issue 4, pp. 399-418.

Data.gov. Retrieved, March 24, 2019, from http://www.data.gov/

Data.gov.uk. Retrieved, March 24, 2019, from http://data.gov.uk/

Data.australia.gov.au. Retrieved, March 24, 2019, from https://www.australia.gov.au/

Data Catalog Vocabulary. Retrieved, March 24, 2019, from https://www.w3.org/TR/vocab-dcat/

EUROPEAN DATA PORTAL. Retrieved, March 24, 2019, from https://www.europeandataportal.eu/

Maali, Fadi, Richard Cyganiak, and Vassilios Peristeras. Enabling Interoperability of Government Data Catalogues. EGOV 2010, LNCS 6228, pp. 339-350, 2010.

Open Data Certificate. Retrieved, March 24, 2019, from http://certificates.theodi.org/en/about/badgelevels

Yu, Liyang. A Developer's Guide to the Semantic Web. Chapter 12. Other Recent Applications: data.gov and Wikidata. Springer, pp. 561, 2014.

Zhai, Jun, Hongyu Chen, and Changfeng Yuan. (2017). Provenance Metadata of Open Government Data based on PROV-JSON. http://dx.doi.org/10.1145/3085228.3085229