US Open Government Data Metadata Standards and Implications: Postprint

Authors: Si Li, Zhao Jie

Date: 2023-08-26T00:00:00+00:00

Abstract

[Purpose/Significance] This paper analyzes the metadata system and specific standards of the U.S. open government data portal Data.gov as a case study, aiming to provide references for the development of China's open government data metadata standards. [Method/Process] Through case analysis, it summarizes the architectural framework of U.S. open government data metadata standards. [Result/Conclusion] The U.S. open government data metadata standards are divided into metadata standards for dataset content description and dataset format description, employing different standards for raw datasets and geospatial datasets; it also points out that China may use the metadata standard system of Data.gov as a reference when constructing its own open government data metadata standards.

Full Text

Preamble

Volume 62, Issue 3, February 2018
ChinaXiv Cooperative Journal

U.S. Open Government Data Metadata Standards and Their Implications

Si Li[1], Zhao Jie[2]
[1]Center for Studies of Information Resources, Wuhan University, Wuhan 430072
[2]School of Information Management, Wuhan University, Wuhan 430072

Abstract

[**Purpose/Significance**] This paper examines the metadata standards of Data.gov, the U.S. open government data portal, analyzing its metadata system and specific standards to provide references for constructing metadata standards for China's open government data. [**Method/Process**] Using case analysis methods, this study summarizes the structural framework of U.S. open government data metadata standards. [**Result/Conclusion**] U.S. open government data metadata standards are divided into dataset content description metadata standards and dataset format description metadata standards, employing different standards for raw datasets and geospatial datasets. The paper suggests that China can draw lessons from Data.gov's metadata standard system when building its own open government data metadata standards.

Keywords: metadata, open government data, metadata standards, Data.gov
Classification Number: G254.31
DOI: 10.13266/j.issn.0252-3116.2018.03.011

Open government data emerged from the data openness movement and government information disclosure initiatives. Government data refers to data and information produced by or commissioned by government agencies, and "open" means it can be freely used, reused, and redistributed by anyone. Since open government data can increase government transparency, enhance the social and commercial value of data, and improve citizen participation in government activities, countries worldwide—including the United Kingdom, United States, Australia, Canada, and China—have actively engaged in open government data initiatives and built their own open government data platforms to provide unified access and utilization of open government data.

Open government data platforms typically integrate multi-source datasets in the form of data catalogs, and metadata standards serve as an important means for catalog management. They have been listed as one of the quality evaluation indicators for open government data in countries such as the United Kingdom and China. China's national open government data platform is currently under construction, with existing platforms primarily at the local government level, such as those in Beijing and Shanghai. However, these local platforms suffer from issues including inconsistent standards, insufficient metadata information, incomplete dataset descriptions, lack of machine-readable formats, and low interoperability. Given the importance of metadata standards for dataset integration in open government data platforms and the current state of metadata standardization in China, research on constructing open government data metadata standards is urgently needed.

Domestic scholars have conducted relevant research on open government data metadata standards, primarily focusing on mature foreign platforms. Studies have examined platforms in the United Kingdom, Australia, Canada, New Zealand, and other countries. For instance, Zhao, Liang, and Duan studied the UK's Data.gov.uk platform, analyzing its CKAN format records (including

CSV and JSON) and GEMINI geospatial metadata standard from perspectives of file structure, element composition, and rules. Huang and Li investigated Australia's Data.gov.au platform, examining its three metadata standards—AGLS metadata standard, ANZLIC geospatial metadata standard, and DCAT data catalog vocabulary—from the aspects of element composition, data format, and syntactic structure. Wang and Tang compared Australia's AGLS and New Zealand's NZGLS metadata standards, focusing on standard establishment, element definitions, and relevant qualifiers. Yu, Zhai, and Lin introduced W3C's recommended standard DCAT and New York State's metadata scheme, analyzing and summarizing the achievements and characteristics of U.S., EU, and Irish government open data metadata construction. Wu and Huang provided a detailed review of metadata policies and standards in the U.S., U.K., Canada, and EU, comparing metadata formats, frameworks, elements, data catalog vocabularies, and controlled vocabularies across these countries. Huang and Lin surveyed metadata description specifications for government data portals in the U.K., U.S., Canada, New Zealand, and EU. However, none of these reviews provide a detailed study specifically focused on U.S. open government data metadata standards.

1. U.S. Open Government Data Metadata Standards

The United States was among the early adopters of open government data policies. Its open government data platform, Data.gov, was launched in May 2009. As of November 15, 2017, 197,990 datasets from 198 organizations had been published on the platform, with datasets continuously being updated.

Data.gov organizes and manages datasets from various agencies in the form of a data catalog. The platform does not physically store data resources directly but instead harvests dataset catalog information through metadata harvesting, presenting summary information and providing access and download links to the actual datasets. This metadata harvesting is implemented based on unified metadata standard descriptions. Data.gov organizes its open data resources into three categories: raw datasets, geospatial datasets, and data tools, employing different metadata standards to describe these three types of resources with distinct characteristics. Data tools refer to various APIs developed based on raw and geospatial datasets. This paper focuses on analyzing metadata standards for describing raw datasets and geospatial datasets.

Based on differences in dataset description approaches, the metadata standards involved can be divided into two categories: dataset content description metadata standards and dataset format description metadata standards, as shown in Table .

1.1 Dataset Content Description Metadata Standards

These standards describe the information content of datasets themselves. Different types of datasets require different content description aspects and thus

employ different metadata standards. According to resource types in Data.gov, dataset content description metadata standards are categorized into raw dataset content description metadata standards and geospatial dataset content description metadata standards.

1.1.1 Raw Dataset Content Description Metadata Standard Raw datasets are provided by federal government agencies or their affiliated institutions. Data.gov uses a unified open government data metadata standard—Project Open Data Metadata Schema (hereinafter referred to as POD v1.1)—to describe raw dataset information. This standard, proposed by the Project Open Data initiative, defines metadata that datasets should follow. It is a hierarchical vocabulary specifically designed for dataset description, built upon the Data Catalog Vocabulary (DCAT). By February 2015, it had been updated to version 1.1. The following analysis examines this standard from three perspectives: description content, field categories, and metadata elements.

(1) Description Content. The description content primarily involves two aspects of dataset content information: external content information—covering people, organizations, time, space, licenses, and rights involved throughout the dataset creation and publication process; and internal content information—such as title, description, and tags. Metadata standard description information refers to metadata version details.

(2) Field Categories. According to the objects described by metadata fields, the standard's fields are divided into three types: catalog fields, dataset fields, and dataset distribution fields. Metadata fields can be categorized as required fields and non-required fields. Required fields are further divided into immutable required fields (always required) and conditional required fields (required if certain conditions are met). Immutable required fields must be present in every dataset, while conditional required fields must be included only when datasets meet specific conditions. For example, the "bureauCode" field is required only when the dataset belongs to a U.S. federal government agency, and the "distribution" field is required only when the dataset has an "accessURL" or "downloadURL." Expanded fields are non-required fields.

(3) Metadata Elements. The POD v1.1 metadata standard contains 45 elements covering three aspects: data catalog, dataset, and dataset distribution, as shown in Table .

Table POD v1.1 Metadata Standard Field Categories and Elements

Field Category	Field Type	Elements
Catalog Fields	Immutable Required Fields	Metadata standard version, dataset

Field Category	Field Type	Elements
	Expanded Fields	Metadata context, metadata catalog identifier, metadata type, data dictionary
Dataset Fields	Immutable Required Fields	Title, description, tags, last update date, publisher, contact name and email, unique identifier, public access level, bureau code, program code
	Conditional Required Fields	License, rights, spatial information, temporal information, distribution
	Expanded Fields	Metadata type, dataset update frequency, data standard, data quality, data dictionary, data dictionary type, dataset plan, release date, dataset language, dataset landing page, unique IT investment identifier, related documents, system of records, theme category
Dataset Distribution Fields	Conditional Required Fields	Access URL, download URL, media type

Field Category	Field Type	Elements
	Expanded Fields	Metadata type, data standard, data dictionary, data dictionary type, human-readable description, format, title

Catalog Fields: Describe the entire public data list catalog file, which is a collection of multiple dataset descriptions. Catalog fields contain six elements: two immutable required fields and four expanded fields. The immutable required fields specify the datasets included in the catalog and the metadata standard version they follow, while expanded fields describe the metadata standards used for datasets.

Dataset Fields: Describe the content characteristics of individual dataset objects, comprising 29 elements: 10 immutable required fields, 5 conditional required fields, and 14 expanded fields. Analysis of these fields reveals that immutable required fields primarily describe basic information related to dataset creation, themes, identification, public access, and attribution; conditional required fields mainly cover publication information, temporal and spatial information, and licensing rights; expanded fields describe metadata information, data quality, and update status.

Dataset Distribution Fields: Describe information related to dataset access, containing 10 elements: no immutable required fields, 3 conditional required fields, and 7 expanded fields. Although there are no immutable required fields, each set of dataset distribution fields must include either an access URL or download URL. Conditional required fields primarily provide specific access addresses, including access and download URLs. Expanded fields describe metadata information for the dataset.

The analysis shows that catalog fields, dataset fields, and dataset distribution fields have a nested relationship. The “dataset” sub-element in catalog fields can be subdivided into all elements in dataset fields, and the “distribution” sub-element in dataset fields can be subdivided into all elements in dataset distribution fields.

When designing the metadata standard, Data.gov aligned it with the platform’s dataset presentation approach. It first presents a summary list of all datasets in catalog form, including dataset titles, owning agencies, and content abstracts, along with dataset acquisition formats. It then provides detailed content descriptions for individual datasets and describes distribution information for datasets with access and download methods, thereby helping users find needed datasets.

Meanwhile, catalog fields, dataset fields, and dataset distribution fields are all designed under the framework of required fields (including immutable and conditional) and non-required fields (expanded fields). Immutable required fields form the core of each field type, describing essential attributes; conditional required fields apply to datasets containing specific information; expanded fields primarily declare metadata information, used data dictionaries and formats, dataset presentation, and other relevant information.

1.1.2 Geospatial Dataset Content Description Metadata Standard
Geospatial metadata describes map, GIS file, image, and other location-based data resources, typically included in Data.gov's GeoPlatform.gov sub-platform. As of November 15, 2017, Data.gov contained 197,990 geospatial datasets—approximately 1.8 times the number of non-geospatial datasets. In addition to following the POD v1.1 metadata standard, geospatial datasets must also comply with specialized geospatial metadata standards. Geospatial information in these datasets includes three main types: (1) spatial values and attribute features of geographic entities, such as latitude/longitude and scale; (2) information about carriers bearing spatial data, such as referenced GIS systems and map display frames; and (3) elements related to resource objects or users, such as provenance and usage constraints.

Data.gov's geospatial data primarily follows two geospatial metadata standards: ISO 19115-2 Imagery and Gridded Data Metadata Standard and the Content Standard for Digital Geospatial Metadata (CSDGM).

(1) ISO 19115-2 Imagery and Gridded Data Metadata Standard. ISO 19115 is a standard for geographic information content and description. Data.gov uses Part 2 of this standard—Imagery and Gridded Data (ISO 19115-2:2009) for description, as shown in Table .

Table Main Components of ISO 19115-2 Imagery and Gridded Data Metadata Standard

Component	Description
Metadata Root Information	Root element containing metadata information
Spatial Representation Information	Geospatial representation information
Reference System Information	Spatial and temporal reference system information
Metadata Extension Information	User-specific extension information describing resources
Data Quality Information	Information uniquely identifying resources
Portrayal Catalog Information	Physical parameters and other attribute information contained in resources

Component	Description
Metadata Constraint Information	Information about resource publishers and how to access resources
Application Schema Information	Information about resource quality, processing steps, and sources
Metadata Maintenance Information	Information identifying portrayal catalogs used by resources
Data Acquisition Information	Usage constraints for metadata and resources
	Application schema information used to build datasets
	Maintenance information for metadata and described resources
	Tools, platforms, operations, and other relevant information for data acquisition

Table lists 13 main components of this metadata standard and explains their content. The ISO 19115-2 metadata standard covers all aspects from initial content description through quality assessment and data publication to later data usage and maintenance, demonstrating good completeness. Based on description objects, these 13 components can be divided into descriptions of datasets and metadata standards themselves. The specific geographic features of datasets are primarily described through fields in spatial representation information and reference system information. For example, for the geographic feature "Spatial Extent," the standard describes the absolute position of a geographic entity through four boundary longitudes or latitudes (east, west, north, south).

(2) Content Standard for Digital Geospatial Metadata (CSDGM). CSDGM (latest version 1998) is a metadata standard developed by the U.S. Federal Geographic Data Committee (FGDC) for describing digital geospatial datasets. It defines metadata content related to positioning, acquisition, use, and publication of digital geospatial datasets, organized through a hierarchical structure of data elements and compound elements that records information content, definitions, and domain values. A data element refers to a logically simple data item, while a compound element is a group of data elements and/or other compound elements—that is, compound elements have subordinate sub-elements.

This metadata standard divides geographic data information into 11 modules for description, with specific fields under each module to detail data or datasets, as shown in Table . Each module begins with the compound element's name and definition, followed by production rules defining the compound element's composition. Modules for describing geographic features are Spatial Data Organization Information and Spatial Reference Information. Spatial Data Organization Information contains specific fields describing digital geospatial data, divided

into four parts: indirect spatial reference, direct spatial reference method, point and vector object information, and grid object information. These describe geographic data's position, size, distance, and other features through how datasets reference geographic locations and through point, vector, and grid data. Spatial Reference Information refers to the dataset's reference framework, encoding method, and coordinate description, serving as the reference system for specific fields in spatial data organization.

Table Specific Elements of CSDGM Metadata Standard

Module	Description	Key Elements
Metadata Root Information	Basic dataset information	Citation, description, time period of content, data status, spatial domain, keywords, access constraints, use constraints, contact point, browse graphic, dataset credit, security information, native dataset environment, cross-reference
Data Quality Information	General dataset quality assessment	Attribute accuracy, logical consistency report, completeness report, positional accuracy, cloud cover
Spatial Data Organization Information	Dataset spatial representation mechanism	Indirect spatial reference, direct spatial reference method, point and vector object information, grid object information
Spatial Reference Information	Dataset reference framework, encoding, coordinate description	Longitude coordinate system definition, latitude coordinate system definition
Entity and Attribute Information	Dataset content details including entity types and attributes	Detailed description (entity type, attributes), overview description
Distribution Information	Publisher and selection information for acquiring datasets	Distributor, resource description, distribution liability, standard order process, custom order process, technical prerequisites, available time period

Module	Description	Key Elements
Metadata Reference Information	Metadata accuracy and responsible party	Metadata date, metadata review date, future metadata review date, metadata contact, metadata standard name, metadata standard version, metadata time convention, metadata access constraints, metadata use constraints, metadata security information, metadata extensions
Citation Information	Dataset citation methods	Originator, publication date, publication time, title, edition, geospatial data presentation form, series information, publication information, other citation details, online linkage, larger work citation
Time Period Information	Event dates and times	Single date/time, multiple date/time, date/time range
Contact Information	Identities and communication methods of individuals/organizations related to datasets	Primary contact, primary contact organization, contact position, contact address, contact telephone, TDD/TTY telephone, facsimile telephone, email address, hours of service, contact instructions

Both ISO 19115-2 and CSDGM are used to describe geospatial datasets in Data.gov. The former is an international standard, while the latter is a U.S. national standard. Both share similarities in metadata framework construction and module settings, including metadata root information, identification information, data quality information, spatial representation information, spatial reference information, and distribution information. Both effectively describe geographic features and attributes related to dataset content, quality, creation, and acquisition.

1.1.3 Dataset Content Description Metadata Standard Mapping To facilitate the creation of POD v1.1 metadata records for datasets by various agencies, FGDC member organizations have established mappings between CSDGM and ISO 19115 geospatial metadata standards and POD v1.1. ISO 19115

shares more common elements with POD v1.1 than CSDGM does. Common elements among ISO 19115, CSDGM, and POD v1.1 are primarily concentrated in POD v1.1's dataset fields and distribution fields.

In dataset fields, common elements among the three standards include title, description, keywords, revision, publisher name, contact name, contact email, identifier, access level, bureau code, program code, spatial information, temporal information, and theme. In distribution fields, common elements are download URL and media type. For identical elements, the three standards are largely similar in description, though CSDGM and ISO 19115 have more detailed field settings than POD v1.1. Values for the same fields may differ slightly. For example, the "revision" field in POD v1.1 is defined as "Last Update," while the corresponding item in CSDGM is "publication date," and ISO 19115 includes resource maintenance frequency, data citation revision date, and first data citation date. Despite these differences, the three standards are fundamentally mappable.

1.2 Dataset Format Description Metadata Standards

Dataset format description metadata standards refer to the standards used to present dataset content, ensuring datasets are both human-readable and machine-readable. Data.gov primarily uses two dataset format description metadata standards: JSON and ISO 19139 Geographic Information—Metadata—XML Schema Implementation.

1.2.1 JSON Open data policies require dataset metadata to be described in JSON format to enable unified metadata harvesting by data catalogs. JSON is a lightweight, text-based data interchange format that is easy to read, parse, and generate, optimizing data exchange. The format is based on two structural components: (1) name-value pairs, typically implemented through objects, records, structures, dictionaries, hash tables, keyed lists, or associative arrays; and (2) ordered lists of values, typically implemented through arrays, vectors, lists, or sequences. A JSON structure example is as follows:

```
{
  "conformsTo": "https://project-open-data.cio.gov/v1.1/schema",
  "dataset": [
    {
      "accessLevel": "public",
      "bureauCode": ["018:10"],
      "contactPoint": {
        "fn": "Jane Doe",
        "hasEmail": "mailto:jane.doe@agency.gov"
      },
      "description": "This dataset provides national statistics on the production of widgets
    }
```

```
  ]
}
```

This snippet describes the metadata standard the dataset follows, access level, bureau code, contact information, and description, clearly showing the machine representation of dataset metadata as "attribute": "value" pairs.

1.2.2 ISO 19139 Geographic Information—Metadata—XML Schema Implementation Standard ISO 19139 is the XML record format and validation specification for ISO 19115 Geographic Information—Metadata, serving as the XML encoding of ISO 19115, published in 2007. In Data.gov, raw datasets use ISO 19115-2 for content description and ISO 19139 XML record format for format description. An example dataset XML structure is as follows:

```
<gmi:MI_{Metadata} xmlns:gco="http://www.isotc211.org/2005/gco"
  xmlns:gmd="http://www.isotc211.org/2005/gmd"
  xmlns:gmi="http://www.isotc211.org/2005/gmi"
  xmlns:gmx="http://www.isotc211.org/2005/gmx"
  xmlns:gsr="http://www.isotc211.org/2005/gsr"
  xmlns:gss="http://www.isotc211.org/2005/gss"
  xmlns:gts="http://www.isotc211.org/2005/gts"
  xmlns:gml="http://www.opengis.net/gml/3.2"
  xmlns:xlink="http://www.w3.org/1999/xlink"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.isotc211.org/2005/gmi
  http://www.ngdc.noaa.gov/metadata/published/xsd/schema.xsd">
  <gmd:fileIdentifier>...</gmd:fileIdentifier>
  <gmd:language>...</gmd:language>
  <gmd:characterSet>...</gmd:characterSet>
  <gmd:hierarchyLevel>...</gmd:hierarchyLevel>
  <gmd:contact>...</gmd:contact>
  <gmd:dateStamp>...</gmd:dateStamp>
  <gmd:metadataStandardName>...</gmd:metadataStandardName>
  <gmd:metadataStandardVersion>...</gmd:metadataStandardVersion>
  <gmd:identificationInfo>...</gmd:identificationInfo>
  <gmd:contentInfo>...</gmd:contentInfo>
  <gmd:distributionInfo>...</gmd:distributionInfo>
  <gmd:metadataMaintenance>...</gmd:metadataMaintenance>
</gmi:MI_{Metadata}>
```

This structure shows the first-level hierarchy in XML representation of a dataset. ISO 19139 first declares namespaces used in the description, then sequentially describes dataset metadata information: file identifier, language, character set, hierarchy level, contact information, date, metadata standard name, metadata standard version, identification information, content information, distribution information, and maintenance information. Each first-level element contains second-level, third-level, and deeper descriptive information, with the descrip-

tion hierarchy determined by ISO 19115-2.

1.3 Characteristics of U.S. Open Government Data Metadata Standards

Based on the above analysis, U.S. open government data metadata standards exhibit three key characteristics. First, metadata descriptions are detailed and highly targeted. The platform's metadata descriptions cover both dataset content and format aspects, using different standards tailored to dataset types (raw datasets vs. geospatial datasets) to highlight their respective features. Second, the standards are adapted from existing international standards with localization adjustments. POD v1.1, for example, was built on DCAT with additions such as "bureauCode" specific to U.S. federal agencies. Third, the metadata standards share the same hierarchical structure as the open government data platform, organized clearly according to the data catalog-dataset-dataset distribution hierarchy.

Metadata standards in the U.S., U.K., and Australia all cover three aspects: data catalog, raw datasets, and geospatial datasets. The differences lie in the specific standards chosen to describe these resources. For data catalogs, all three use DCAT or DCAT-based standards: the U.S. uses catalog fields in POD v1.1, the U.K. uses CKAN record formats, and Australia follows DCAT. For raw datasets, the U.S. uses POD v1.1 dataset and distribution fields with JSON for content and format description; the U.K. uses CSV and JSON formats; and Australia uses the AGLS metadata standard. For geospatial datasets, the U.S. uses ISO 19115-2, CSDGM, and ISO 19139; the U.K. uses GEMINI; and Australia uses the ANZLIC geospatial metadata standard.

2. Implications for Constructing Metadata Standards for China's Open Government Data Platforms

China is currently constructing open government data platforms at national and regional levels, with plans to establish a unified government data open platform by the end of 2018 to enable reasonable and moderate opening of public data resources. Metadata standard construction is a crucial component of this process. According to China's first "Local Government Data Open Platform Report" in 2017, the number and items of metadata entries vary across local platforms, resulting in differences in dataset description detail and presentation methods that hinder integration across platforms. Huang and Wang found that among 13 existing local government data open platforms, only those in Beijing, Shanghai, Wuxi, and Qingdao provide relatively detailed metadata, while others are quite simple. Qing and Zhao discovered that most datasets on Beijing's government data open website only offer CSV format downloads with simple information, lacking detailed content. Chen et al. noted that existing local government open data portals generally lack standardized metadata systems, making it difficult for users to understand and utilize dataset information, including provenance.

Zhao and Mo introduced DC, VoID, DCAT, and other metadata standards for catalog description, dataset description, relationship description, and access description to support effective data sharing, discovery, and management after opening. Therefore, establishing unified metadata standards is essential for standardized dataset description and for harvesting and timely updating datasets stored across various agencies and departments through metadata harvesting.

Based on analysis of U.S. open government data metadata standards, China should consider the following key points when constructing its own metadata standards:

2.1 Metadata Standard Selection Should Be Broadly Applicable and Universal

Constructing a unified government data open platform requires unified metadata standards to ensure datasets follow consistent specifications throughout the entire process from collection to publication. There are two approaches to selecting unified metadata standards: (1) adopting existing internationally accepted open government data metadata standards, such as POD v1.1, ISO 19115-2, CSDGM, and JSON used by the U.S., which facilitates subsequent compatibility with datasets from other countries; or (2) developing metadata standards tailored to China's national conditions based on existing international standards, considering differences in resource types and data providers between U.S. and Chinese government datasets. The latter is more effective, as China's local government data open platforms have already established their own dataset description methods. Unified metadata standards should be compatible with these existing local standards. During construction, basic metadata elements and extended metadata elements can be extracted from existing international standards and local Chinese standards, serving as required and non-required fields respectively to accommodate datasets with varying levels of detail.

2.2 Metadata Standards Should Distinguish Dataset Types

The U.S. open government data platform categorizes datasets into raw datasets and geospatial datasets, using different metadata standards for each, particularly selecting standards that can characterize geographic features for geospatial datasets. When constructing its own metadata standards, China should also select and develop targeted metadata standards based on different dataset types to distinguish them, fully demonstrate dataset characteristics, and enable more rational and effective dataset utilization.

2.3 Metadata Standards Should Be Interoperable

Metadata standard interoperability has two aspects: (1) interoperability among metadata standards within the open government data platform, such as the mapping between ISO 19115, CSDGM, and POD v1.1 used by the U.S. platform;

and (2) interoperability between constructed metadata standards and their reference standards, such as the mapping between POD v1.1 and its foundation DCAT, as well as with Schema.org. Therefore, China's unified open government data platform metadata standards should establish interoperability with both local municipal metadata standards and their reference standards.

2.4 Metadata Standards Should Include Both Content and Format Descriptions

Datasets must be standardized in both content description and format description to achieve both human-readability and machine-readability. For content description, China can combine foreign dataset description standards with its own dataset characteristics to build suitable metadata standards. For format description, it can directly adopt foreign standards such as JSON.

Based on these key points, China's open government data platform metadata standard construction should follow this approach: First, identify dataset types in the platform and provide targeted metadata descriptions for different types, selecting or building different metadata standards for three main resource types—data catalog, raw datasets, and geospatial datasets—following the examples of the U.S., U.K., and Australia. Second, determine basic description aspects such as content and format, selecting appropriate format description metadata standards based on content description standards. Third, determine element attributes and granularity for Chinese datasets, and based on existing universal standards like DCAT and DC, localize them according to Chinese dataset characteristics while refining description granularity to include all types of people, time, organizations, and content and format features involved from dataset origin to publication, laying the foundation for diverse user retrieval and secondary data development.

Conclusion

This paper analyzed the dataset composition and metadata standards for different dataset types on the U.S. open government data platform Data.gov. The platform categorizes datasets into raw datasets, geospatial datasets, and data tools. Raw datasets are described using the POD v1.1 metadata standard, while geospatial datasets use ISO 19115-2 and CSDGM metadata standards. These three standards all describe dataset content information and share some common fields. In addition to content description metadata standards, there are format description metadata standards: JSON and ISO 19139. Based on this analysis of U.S. open government data metadata standards, we argue that China should select broadly applicable metadata standards, adopt different standards for different dataset types, ensure interoperability among standards, and include both content and format descriptions when constructing its own metadata standards.

References

[1] Open government data [EB/OL]. [2017-03-09]. https://opengovernmentdata.org/. [2] Zheng L, Gao F. Research on China's open government data platforms: Framework, status, and recommendations [J]. E-Government, 2015(7): 8-16. [3] Open Data Institute. Open data certificate [EB/OL]. [2017-11-17]. https://certificates.theodi.org/en. [4] Sun L, Li G. Research on constructing an analytical model for government open data applications [J]. Library and Information Service, 2017, 61(3): 97-108. [5] Yu M, Zhai J, Lin Y. Research on core metadata for local government open data in China [J]. Journal of Intelligence, 2016, 35(12): 98-104. [6] Zhai J, Yu M, Lin Y. Comparison of major global government open data metadata schemes [J]. Information Studies: Theory & Application, 2016, 39(19): 31-39. [7] Zhao R, Liang Z, Duan P. Metadata standards for UK government data open sharing: Investigation and implications from Data.gov.uk [J]. Library and Information Service, 2016, 60(19): 31-39. [8] Huang R, Li N. Metadata standards for Australian open government data: Investigation and implications from Data.gov.au [J]. Library Journal, 2017(5): 87-94. [9] Wu L, Huang Y. Research progress on metadata standards for open government data platforms [J]. Library Science Research, 2017(6): 14-21. [10] Wang L, Tang Y. Metadata standards for Australia and New Zealand government website construction [J]. Library and Information Service, 2004(S1): 410-413. [11] Huang R, Lin Y. Investigation and analysis of foreign open government data description specifications [J]. Library and Information, 2017(4): 113-121. [12] Project Open Data. Project Open Data metadata schema v1.1 [EB/OL]. [2017-02-20]. https://project-open-data.cio.gov/v1.1/schema/. [13] W3C. Data catalog vocabulary (DCAT) [EB/OL]. [2017-02-20]. https://www.w3.org/TR/vocab-dcat/. [14] Federal Geographic Data Committee. Geospatial metadata [EB/OL]. [2017-03-08]. https://www.fgdc.gov/metadata. [15] Federal Geographic Data Committee. Organization of the standard [EB/OL]. [2017-03-09]. https://www.fgdc.gov/metadata/csdgm/organization.html. [16] Project Open Data. Metadata resources for schema v1.1 [EB/OL]. [2017-05-09]. https://project-open-data.cio.gov/v1.1/metadata-resources/#field-mapping. [17] Introducing JSON [EB/OL]. [2017-05-08]. http://www.json.org/. [18] Catalog sample [EB/OL]. [2017-05-11]. https://project-open-data.cio.gov/v1.1/examples/catalog-sample.json. [19] Federal Geographic Data Committee. ISO 191** suite of geospatial metadata standards [EB/OL]. [2017-03-08]. https://www.fgdc.gov/metadata/iso-suite-of-geospatial-metadata-standards. [20] XML file [EB/OL]. [2017-05-11]. https://catalog.data.gov/harvest/object/cd79c269-d065-4a46-8a5c-0aacd2654bdb. [21] Zhao R. Investigation and research on government-linked open datasets [J]. Library and Information, 2016(4): 102-112. [22] Huang R, Wang C. Investigation and analysis of China's government data open platforms [J]. Information Studies: Theory & Application, 2016, 39(7): 50-55. [23] Qing Q, Zhao R. Research on the status of Beijing municipal government data opening [J]. Journal of Intelligence, 2016, 35(4): 177-182. [24] Chen H, Zhai J, Yuan C, et al. Research and application of provenance metadata for open

government data [J]. Journal of Intelligence, 2017, 36(6): 148-155. [25] Zhao L, Mo L, Chen M. Resource description methods for government data opening [J]. Library and Information Service, 2017, 61(6): 115-121.

Author Contributions

Si Li: Guided paper writing and revised research ideas.
Zhao Jie: Identified research topic, conducted website investigations and literature review, wrote and revised the paper.

Investigation and Enlightenment of Metadata Standards of American Open Government Data

Si Li[1], Zhao Jie[2]
[1]Center for Studies of Information Resources, Wuhan University, Wuhan 430072
[2]School of Information Management, Wuhan University, Wuhan 430072

Abstract: [Purpose/significance] This paper takes the metadata standards of Data.gov, an open government data website, as an example, and analyzes its metadata system and specific standards so as to provide references for the construction of our country's open data metadata standards. [Method/process] Using the method of case analysis, the paper summarized the system structure of the metadata standard of American open government data. [Result/conclusion] The metadata standards of American open government data can be divided into two categories, which are dataset content metadata standards and dataset format metadata standards. Data.gov uses different metadata standards to describe original datasets and geospatial datasets respectively. We can reference the metadata standards system in Data.gov in the construction of metadata standards of Chinese open government data.

Keywords: metadata; open government data; metadata standard; Data.gov

Note: Figure translations are in progress. See original paper for figures.