Research on Metadata Specification of Service Data Resources Under the Background of Data Transaction

Yaqing Si\textsuperscript{1,*}, Yinghua Chen\textsuperscript{1}, Jing Su\textsuperscript{1}, Jin Zheng\textsuperscript{2}, and Siyao Qin\textsuperscript{1}

\textsuperscript{1}School of Economics and Management, BUPT, 100876, Haidian District, Beijing, China
\textsuperscript{2}Electric Power Research Institute, CSG, 510663, Huangpu District, Guangzhou, China

Abstract. This paper designs the service data resource metadata standard system and catalog vocabulary by investigating the status of service data resource descriptions of mainstream data trading platforms at home and abroad, and referring to W3C’s official recommendation standard DCAT and Dublin core metadata. The service data resource metadata interaction model is designed by referring to the RDF/XML framework. The service data resource metadata standard applying RDF/XML grammar specification will effectively support the registration and authentication of data products, transaction process management, data demanders’ understanding and use of purchased data, and market supervision.

1 Introduction

With the innovation of the modern service industry development model, the data generated by consumers in the process of service increases exponentially. In this paper, a data set with a certain scale that records the service process, collected by the service provider, is defined as "service data resource". The current service data resources are facing some outstanding problems in the transaction process: poor quality of service data resources; description of service data products Lack of unified standards; difficulty in pricing transactions; difficulty in proof of illegal use of data, etc.

The above-mentioned problems have led to the dilemma of fewer types of service data goods, less transaction volume, and lower transaction efficiency. How to solve the existing problems on the basis of protecting the rights of the data subject, promote the circulation of the data market, and facilitate the mining of the meaning and value behind the service data resources are urgent problems that need to be solved in data transactions. The key entry point to solve the above problems lies in the research on the standard specification of product description and interaction of service data resources.

In the Internet-based data trading market, the following five product forms exist for service data resources: data terminals, data files, data source services (API), information authentication services, and data processing services. In this article, data file products will be used as examples to propose a metadata standard solution for data file products.
2 Design principles and functions of metadata specification

Compared with other data resources, service data resources are multi-dimensional, multi-layered, dynamic, and complex. The formulation of service data resource metadata specifications requires three aspects: data resource demand side, provider side, and transaction management side. For consideration and analysis, the principles to be followed for the design of the metadata specification are as follows: the use of metadata needs to facilitate the commercialization of service data resources and the management of the transaction process, and the description and management costs are small [1].

Analysis of the previous general metadata standards, combined with the characteristics of service data resources and transactions, the proposed metadata standards should support the realization of the following six functions.

—Description: The multi-dimensional and multi-level description of service data products is the most basic and important function of metadata.

—Retrieval: Use metadata to better organize service data resources and establish correlations between resources.

—Choice: The demand side can have a basic understanding of the data products without directly browsing the entire contents of the service data products, can decide whether to purchase related data products, and can personally configure the data products to be purchased.

—Management: The use of metadata makes it easier for transaction managers to increase the efficiency of service data resource identification, product display, dynamic pricing, and transaction process records.

—Interaction: In order to make it easier for the demand side to retrieve data products, the metadata needs to be in a machine-readable form to support the complete human-computer interaction process.

—Pricing: After the data demander personally configures (cuts) the data goods to be purchased according to his own needs, the system automatically determines the price of the new data goods through the pricing model automatically with the support of metadata.

3 Metadata specification framework design

Based on literature research, the internationally accepted metadata specification framework is divided into the following three categories: ISO-based models, Dublin core metadata-based models and RDF-based models [2].

![Metadata specification framework](image)

Figure 1. Summary of metadata specification framework model.
Because W3C's DCAT vocabulary and Dublin core metadata can meet some of the design principles and functions of the service data resource metadata specification, this article will focus on reusing part of the DCAT vocabulary and Dublin core metadata. Proposed Service Data Catalog Vocabulary (SDCAT).

4 Data file product metadata catalog

The data file product refers to the data set file generated in advance. After the data file is confirmed and the unique identification is obtained, the data file product can be displayed on the web page for the data demander to browse the data product, personalize the order, purchase and download the data set file. Data file product metadata (see Table 1 for details) mainly contains eight types of data information, which are summarized below.

Table 1. Glossary of product file metadata catalog.

| The serial number | Element type | Metadata Element | Standard reuse and supplement | If necessary |
|-------------------|--------------|------------------|-------------------------------|--------------|
| 1                 | Data digital object unique identifier | sdcat: dataDOI | necessary | necessary |
| 2                 | Dataset title | dct: title | necessary | necessary |
| 3                 | Confirmation mark | sdcat: confirmationMark | necessary | necessary |
| 4                 | Confirmation mark issuing organization code | sdcat: CMIOC | necessary | necessary |
| 5                 | Brand name | sdcat: brandName | optional | optional |
| 6                 | Brand image | sdcat: brandImage | optional | optional |
| 7                 | Source data hash value | sdcat: sourceHash | necessary | necessary |
| 8                 | Product form | sdcat: productForm | necessary | necessary |
| 9                 | Product type | sdcat: productType | necessary | necessary |
| 10                | Sales status | sdcat: salesStatus | necessary | necessary |
| 11                | Keyword | dct: keyword | necessary | necessary |
| 12                | Content description | dct: description | necessary | necessary |
| 13                | Industry | sdcat: industry | necessary | necessary |
| 14                | Theme taxonomy | dct: themeTaxonomy | necessary | necessary |
| 15                | Field description | sdcat: fieldDescription | necessary | necessary |
| 16                | Format | dct: format | optional | optional |
| 17                | Byte size | dct: byteSize | necessary | necessary |
| 18                | Dataset quality | sdcat: quality | optional | optional |
| 19                | Data sources | sdcat: dataSources | necessary | necessary |
| 20                | Sampling method | sdcat: samplingMethod | necessary | necessary |
| 21                | Sampling frequency | sdcat: samplingFrequency | necessary | necessary |
| 22                | Issued | dct: issued | necessary | necessary |
| 23                | Time period | dct: temporal | necessary | necessary |
| 24                | Sample data link | sdcat: sampleDataLink | necessary | necessary |
4.1 Data identification information

The important element in the data identification information is the "Data Digital Object Unique Identifier". In order to effectively realize the storage, transaction, application and service of service data resources, each service data product needs to be uniquely identified; the "Confirmation Mark" element is a metadata file generated by the data provider through product characteristics, apply for the confirmation token according to the confirmation agreement of the trading alliance, which determines the ownership of the data provider in the data transaction and the scope of transfer of ownership. All data products must obtain the "Confirmation Mark" and "Data Digital Object Unique Identifier" in order to trade in the market. At the same time, in order to ensure the reliability of the authorization label, it is necessary to record the "Confirmation Mark Issuing Organization Code" in the metadata for later verification and traceability.

The "Source Data Hash Value" element refers to the hash value obtained by hashing the original data set through a hash function, and the purpose of recording this element is to facilitate management and the definition and verification of ownership before and after data transactions [3], using it as a one-to-one association between metadata and the source data set, and cannot be tampered with.

The "Product Form" element represents the product attributes of the service data resource in the trading platform. The "Product Type" element specifically distinguishes data products as original products, authorized products or resale products.

4.2 Summary of data content

Overall describe the content characteristics of data products as a whole, and more importantly describe the specific conditions of each part of the data from the field, focusing on reflecting the multi-dimensionality and polymorphism of the data products, so that the
demand side of the content of the data products from overall to partial. Can have a more specific and clear understanding. The data content summary is composed of five elements. The "Field Description" element describes the multi-dimensionality and polymorphism of each field in the data product in a three-level nesting manner.

The three-level nested metadata description draws on the ideas of the XBRL Dimensions technical specification to define the polymorphic information in the data document. The key to polymorphic description of data is to describe the association between the dimensions related to polymorphism and construct a hierarchical structure. The dimensions constructed in the hierarchy include: main item, hypercube, dimension, field, and field value. The main item is a specific data item, which can be associated with one or more hypercubes. The hypercube binds a series of dimensions related to the main item. It is a collection of multiple possible dimensions. The dimension at this level is an abstract concept, which is divided into display dimensions and hidden dimensions. The display dimension refers to the dimension in which all the domain values can be listed through enumeration, and the implicit dimension represents the dimension in which the domain value range is determined by means of constraints. A domain is a set of all domain values used to describe a dimension. A domain value is a member of a domain.

The first-level metadata elements "Field Description" can be refined with five second-level metadata elements: "Field name", "Field type", "Field value factor", "Whether the field is selectable" and "Whether the field is configurable" description. Second-level metadata elements Under "Field type", there are three third-level metadata elements: "Field format", "Field value - start value" and "Field value - termination value".

![Diagram](image.png)

Figure 2. "Field Description" Dimensions module association diagram.

"Field value factor" refers to the value weight index of each field, and limits the value coefficient of all fields to add up to 1. The meaning of setting the "field value coefficient" element is that the data demander can personally select the relevant fields and ranges in the service data product. The pre-defined value coefficient can facilitate the data provider to combine other pricing requirements and automatically generate service data products dynamically and price.

The "Whether the field is selectable" element describes whether the data field is a demand-side selectable field, and the "Whether the field is configurable" element describes
whether the data field can be personally selected by the user within a certain value range. The above two elements determine whether the data product can be tailored and the scope of tailoring to meet the personalized configuration (trimming) needs of the data consumer.

"Field format" has made a further step in the storage form of the field. Refine, describe and constrain from a specific format. When the field is a configurable field, it means that the user can choose between the "start value" and "termination value" of the field.

### 4.3 Data format

The data format category mainly describes the storage format, scale and quality of service data products, and mainly includes the following three metadata elements: "Format", "Byte Size" and "Dataset Quality".

The "Dataset Quality" element field is included in the core metadata element in order to enable the demand side to more objectively understand the quality of the data from the third-party evaluation results. Since the data resources used for data transactions need to determine the quality of the resources, buyers and sellers need to determine the quality of the data through professional third-party institutions or data service parties [4].

### 4.4 Data sources

The data source category is mainly used to reflect the source and method of data collection. The metadata elements mainly include: "Data Sources", "Sampling Method", and "Sampling Frequency".

The product type of the data file product can be an original product, an authorized product, or a resale product. When the data is an authorized product or a resale product, in order to clearly understand the process from the generation of the data to the first sale to the re-sale, the metadata sets the "Data Sources" element, and design the secondary element to expand the description range. The data acquisition channel is expanded from three aspects: first, the specific source of the original data collection, such as a specific service of a website platform; second, the organization and relevant person in charge of collecting the original data, which is convenient for reviewing the certified data products; Third, when the product type is resale data, it is necessary to determine whether the provider of the data product signed a purchase contract at the time of purchase to avoid legal disputes over data ownership and other rights.

The metadata element of the data source class provides a basis for tracing the source of service data resources.

### 4.5 Data release update information

The data release update information is mainly used to reflect the timeliness of the data in the data file, record the earliest release time of the service data product, and the coverage of the data set in the data file in the time dimension.

### 4.6 Data sample

The data sample is to display the data form and related content in a small amount of data, so that the demander can more intuitively understand the content, form, and type of the data. Mainly include: "Sample Data link", "Sample Data Format" and "Sample Record Number". It is not easy to save specific sample data in the metadata, so the data controller can upload the sample data to the web page or apply for DOI for storing the sample data, and only need
to record the URL link or DOI in the metadata. Data demanders can view and download sample data on the platform, or use the DOI parser to view sample data.

4.7 Data provider information

Metadata includes data provider information from the perspective of platform transactions and security compliance, to help data demanders understand the situation of data product providers, so that when data products have problems in property rights or other aspects, they can be traced to the relevant responsible person.

4.8 Others

With the increasing variety of service data resources, the ontology characteristics of data resources will also become more diversified, and the expansion of the core metadata of service data resources will also increase. Therefore, the metadata elements also include descriptions of "Language", "Failure Time", "Associated Data" [5], "Customer Group", etc.

5 Summary and prospect

At present, the standardization and standardization of metadata of data resources in China is not high. This article has made a preliminary attempt in the reuse and expansion of international general metadata, which basically solves the problems of service data resources in data description, transaction and management Problems, but with the rapid development of the modern service industry, the types of data generated during the service process are also more diverse, such as 3D data, VR data, multi-modal data, etc. The description in other aspects will be further discussed and studied. It makes the application and description of service data resources based on metadata have wider application prospects.

This work was supported by National Key R&D Program of China (2017YFB1401100).

References

1. J. Zhang, F. Li, M. Tang. Design and application of contract document description metadata specification. Books and intelligence work, 61(08), 106-111, (2017).
2. M. M. Morsy, A. M. Castronova, et al. Design of a metadata framework for environmental models with an example hydrologic application in HydroShare, Environmental Modelling & Software, 93(C), 3-28, (2017).
3. A.M. Fu, J.Y. Song, M. Su, S. Li. Client safety Destressing scheme for ciphertext data in cloud storage, Acta Electronica Sinica, 45(12), 2863-2872, (2017).
4. Metadata quality dashboard. European Data Portal, http://www.europeandataportal.eu/mqa-service/en/dashboard.html, (2016).
5. W.B. Huang. Research on collection resource aggregation based on metadata association, Information theory and practice, 38(04), 74-79, (2015).