A Metadata Model for Data Goods

Markus Spiekermann[1], Daniel Tebernum[1], Sven Wenzel[1], and Boris Otto[1,2]

[1]Fraunhofer ISST, Digitization in Service Industries, Dortmund, Germany
{firstname.surname}@isst.fraunhofer.de
[2]TU Dortmund, Chair of Supply Net Order Management, Dortmund, Germany
{firstname.surname}@tu-dortmund.de

Abstract. Digital transformation implies the development of data-driven business models and thus the management of data goods. While marketplaces for data are being established and platforms for the exchange of data are being created, companies have to adapt their data management to the increasing requirements. One central question can be deduced: How can data goods be described in a standardized way? This paper describes the development of the metadata model for data goods M4DG. The M4DG, based on an analysis of existing data marketplaces and metadata models of related topics, makes it possible to describe data sources with defined properties. This creates a unified understanding of the properties of data goods to facilitate selection and trading. We are convinced that the M4DG will contribute to the practical design of data management.

Keywords: data management, metadata model, data goods, data economy

1 Introduction

The role of data has changed from an entrepreneurial resource, used in the individual processes of a company, to an independent product. For example, platforms such as OpenWeatherMap[1] not only use the data from connected weather stations as their business resource. In addition to free access to weather information, they also offer more detailed and finely granular data in exchange for money. Among other things, this development can be attributed to the emergence of platforms and marketplaces on which data of different categories, different characteristics and under different conditions are offered. On the other hand, a general change in companies' business models can be seen. As a result of the expansion of digital transformation in companies, there is a shift from standardized and material-based business models to highly individual, hybrid or purely digital products [1]. In contrast to material goods, however, data has special characteristics that prevent the simple transfer of established processes and rules of material trade [2, 3]. This has an impact on the supply chain and data exchange between companies.

[1] https://openweathermap.org, accessed 10.11.2017

Multikonferenz Wirtschaftsinformatik 2018,
March 06-09, 2018, Lüneburg, Germany

For the emerging trade of data goods and their cross-company exchange, a uniform understanding of the properties of the data offered or purchased is necessary. A standardized description of data sources can improve interoperability and processes and rules for data management can be defined. This would pave the way for a data supply chain and a sovereign exchange of data, which is aimed at by many current initiatives [4, 5]. This raises the research question of whether and how data sources can be modeled in the sense of a classical economic good, a so-called data good. This paper proposes a model that allows data goods to be managed like corporate assets.

This paper describes the qualitative approach to the creation of a model for the standardized mapping of data goods. It shows how a suitable model was developed from existing approaches of creating metadata models in cooperation with companies and how a first version of a possible metadata model for data goods was developed and tested. The remainder of the paper is structured as follows: First, the relevant approaches, related topics and technologies within the thematic area are examined. The following is an explanation of the scientific methodology used. Section 4 describes the corresponding procedure for implementing the model design by explaining the qualitative data collection and the resulting decisions for the model development and an evaluation. In the following, the use of the model in the data backend of a software component is described in a practical application case in section 5. Section 6 summarizes the results of the report and gives an outlook on further research.

2 Background

2.1 Data Goods

The importance of data for business success has been growing steadily since the introduction of electronic data processing and the automation of production processes in companies. Although research and industry have been discussing the role of data as an asset since the 1980s, digitization has for the first time led to the use of the term “data-driven business” [6]. Digitization is not possible without data. The individual development stages in the role of data are not disjoint, but can be found in parallel in companies. This results in the so-called “data paradox”. On the one hand, data is a consequence of the above-mentioned developments such as digitization and industry 4.0 (because machines, smart services, etc. produce an increasing amount of “big data”). On the other hand, it is a resource for the production of services or even the product itself. We consider data as an independent product that can be traded between companies. For this reason, we call data in this context “data goods”, analogous to economic goods. When considering data as a product, it is essential to consider the special properties of the product. This is evident, for example, in the economic analysis of a price determination, a qualitative evaluation or the definition of user authorizations. The challenges of considering data as an economic asset are not a new topic in scientific literature [6–10]. Moody and Walsh describe the seven “laws” of information by highlighting the differences between data and material goods [3]. Despite the fact that the topic has been dealt with, there is still no practical approach to the topic. While

other intangible assets such as patents, trademarks or human capital are shown in the balance sheet, this is currently not possible for data. As part of the process of digital transformation, however, the scientific community is once again increasingly concerned with the economic consideration of the resource for digital business processes and its value added. Data trading can also represent an independent business model [11].

2.2 Data Marketplaces and Platforms

We observed several independent data marketplaces where sellers can offer their data (see Table 1). The marketplaces allow the providers to make their data available to interested parties [12]. Various price models have been developed in which data can be made freely available or monetized via models such as costs per call-off, costs per volume or flat rates [11, 13]. Data marketplaces can therefore be defined as a platform that allows you to offer data for sale and to specify the rules and prices for the data [14]. Data marketplaces, as a platform for trading data, can have different architectures. One possibility is the creation of a central instance (cloud computing approach), where the data is offered by different vendors via a central point. The advantages of this approach are the high traceability through a central search for products, a uniform interface for the provision of data goods and the possible specification of formats and distribution channels.

Another platform architecture is the decentralized approach [15]. In this case, the data items are offered decentrally, i.e. independently by the sales staff. This gives the seller control over his data and enables him to regulate the use of the data. This means that data to be sold is not offered to the public but only in certain partnerships, such as a supply chain. A major disadvantage of this architecture is the large number of different interfaces, formats, price models and the difficulty of finding interesting data.

This variety of possible descriptions and interfaces is a well-known problem of data management, which is becoming increasingly important due to the increasing importance of cross-company data exchange. This illustrates the development of platforms for managing data and secure trade in digital goods between companies. One of these platforms is the internationally established Industrial Internet Consortium (IIC), which has been developing the architecture of a standardized platform for connecting human beings and machines since 2014 [5]. In Germany, a network of various institutions and associations is developing a reference model specifically for the context of Industry 4.0; the so-called Reference Architecture Model for Industry 4.0 (RAMI 4.0) [16]. The Industrial Data Space, a combination of research project and industry consortium, which describes a reference architecture and strives for a prototypical implementation, chooses a holistic approach with the focus on preserving data sovereignty when trading data goods [15, 17].

2.3 Metadata Models

A well-known metadata model is the Dublin Core Metadata Element Set (DCE) [18]. The DCE does not focus on the description of a specific application area, but on the

general description of resources. However, DCE is limited to the basic properties such as *author*, *format*, *language*, *ID number* or *title*. In February 2009, the Dublin Core Metadata Initiative registered the DCE as ISO Standard 15836 [19]. For other models, DCE often serves as a basis for the extension of certain application contexts [20–23] due to its simplicity, comprehensibility and flexibility.

Based on the DCE, Maali et al. developed the Data Catalog Vocabulary (DCAT) [20]. DCAT is designed to increase interoperability between different data catalogs and simplify the search for data sources. For this purpose, properties of different metadata models from the field of data cataloguing have been compared and commonalities identified, which were then combined into a new model. Another metadata model based on the DCAT is used in the Comprehensive Knowledge Archive Network (CKAN). CKAN provides an open source software tool that allows data to be published on a platform as a self-hosted service. This is mainly used by public authorities or non-profit organizations [24]. Within the scope of Open Data initiatives, CKAN and its metadata model have become a de-facto standard. CKAN is also limited to the essential properties of the data sources to describe them [25].

Another metadata model in the context of describing digital assets is the Asset Description Metadata Schema (ADMS). It is a standard accepted by W3C and also an extension of existing models. ADMS is based on DCAT and defines itself as a special variant [26]. ADMS allows the description of assets within data catalogs. An asset can be a data source, vocabulary, or other metadata models. However, none of the metadata models found provides economic properties for a data source.

2.4 Metadata Models for Data Goods

Indicators supporting the decision to develop a new data model for data goods were identified in the previous background notes. Data is increasingly becoming the product itself. There are already metadata models for immaterial goods [20, 26, 27]. However, according to our research, there is no metadata model that reflects the requirements concerning data goods (see Chapter 2.1). The lack of a standard makes it difficult, for example, to develop federal data marketplaces that can easily exchange their offers among each other. Therefore, it is necessary to create a uniform understanding of the properties of data goods.

3 Methodological Approach

The creation of the desired metadata model for data goods followed the methodological approach of Design Research (DR) [28]. In particular, we chose the method of Action Design Research (ADR), which focusses on generalizing the results through an iterative approach to the creation of artifacts and application in several use cases. This prevents the results from being influenced too much by a single individual application. The iterative process of ADR goes through the phases “Problem Formulation”, “Building, Intervention, and Evaluation”, “Reflection and Learning” and “Formalization of Learning”. The procedure implies a close exchange between science and practice, in

which both sides contribute their experience, knowledge and ideas to develop a solution to the problem [29]. The development of the model described in this paper was supported by practitioners in three workshops in the first half of 2017. The first partner is from the pharmaceutical industry. Employees from the area of data purchasing and data modeling were involved. The second partner is from the consulting sector. The Head of Business Design and the Head of IT supported us here. As a third partner, experts from the field of data modelling supported us. The goal of the partners is to provide a unified description of data sources, so that communication and exchange via automated interfaces can be simplified. From a scientific point of view, the main aim is to contribute to the gain of knowledge in the field of data management and to create a basis for further research.

In a first step, the scientists reviewed the current literature and conducted an analysis of existing data marketplaces (Table 1). Findings from related research projects that deal with data management have also been incorporated here. Afterwards, a first workshop with the members of the pharmaceutical company was held. They collected possible characteristics of data sources in a brainstorming session. The focus was on properties that support the idea of a company-wide inventory based on a standardized description and establishment of a transparent mapping of governance processes. Afterwards, a first metadata model for data goods was developed based on the research and workshop results. In a second workshop, the model was presented and discussed to the consulting company. This has yielded new findings that are incorporated into the next iteration of the development of the scientific artifact within the ADR process. The same procedure as in the second workshop was repeated in a third one with the data modelling experts.

4 Metadata Model for Data Goods

4.1 Initial Research

In order to determine the necessary properties for a description of data goods, we have carried out a literature analysis to include relevant work within the subject area. The literature analysis was conducted after Webster and Watson [30]. Springerlink, DBLP Computer Science Bibliography, Elsevier and ScienceDirect were used for the research. The keywords “Data Catalog”, “Data Source Description”, “Data Source Metadata”, “Data Asset”, “Data Inventory”, “Data Marketplace”, “Data Governance”, “Data Economy” and “Data Management” were used for the initial search. Further keywords were identified by screening the found literature. Only those works were considered which show in the abstract that they are positioned in the field of data and data source description or consider economic aspects of data. Literature that was not available in English or German was ignored. We found two existing vocabularies, Data Catalog Vocabulary (DCAT) and Asset Description Metadata Schema (ADMS), which had a strong relation to the desired result. While DCAT allows a rough description of data sources, the ADMS has been developed as a more specialized metamodel to describe vocabularies. In addition to literature research, we have looked at existing web-based

data marketplaces. Due to the focus on data as a product, the analysis investigated those data marketplaces where data is offered for sale in order to monetize them. The data marketplaces considered represent internet platforms that allow a data exchange between a seller and a buyer. The data marketplaces considered in this report have been selected by means of a manual internet search for corresponding platforms. The list does not claim to be exhaustive, as new marketplaces are constantly being created, especially in this area, or access is only granted against payment. Table 1 shows which properties are represented by which data marketplace. By comparing the found properties of data sources from literature and the found properties of data marketplaces, we found a large consensus. We chose ADMS as a starting point for the further development of our model, as it has been identified during our initial research as the metamodel with the closest relation, the representation of vocabulary as a special class of data goods. Due to the aspired goal of a high level of standardization and generalization, the design decision was therefore made to adapt the ADMS [26], taking into account the different focus and specialization on data sources.

Table 1. Overview of the properties represented in data marketplaces

	dmi.io[2]	*Azure*[3]	*datastreamx*[4]	*Mashape*[5]	*bids4bytes*[6]
Title	X	X	X	X	X
Type				X	
Category	X	X	X	X	X
Dateadded	X	X			X
Provider	X	X	X	X	X
Description	X	X	X	X	X
Pricinginformation	X	X	X	X	X
Example	X	X	X	X	
Licensing information	X	X	X	X	X
User rating	X			X	
Structure	X		X	X	
URL (API)	X		X	X	
Numberofdatarecords		X			
Quality information		X			
Deliveryperiod		X			
Format		X	X	X	X
Deliverymethod		X	X		
Update cycle		X			X
Alternatives			X		

[2] https://dmi.io/, accessed 09.11.2017
[3] https://datamarket.azure.com/browse/data/, accessed 09.11.2017
[4] http://www.datastreamx.com/, accessed 10.11.2017
[5] https://market.mashape.com/, accessed 10.11.2017
[6] https://www.bids4bytes.de/, accessed 10.11.2017

4.2 Development of the M4DG

The following describes the development of the M4DG from a global perspective, so that a holistic picture is in the focus, and not the individual workshops. In order to determine the necessary variations and to map the special requirements in a new metamodel, we carried out a four-stage process. In the first stage, we formed so-called "building blocks" in order to enable a rough classification of the determined properties and to visualize them. The characteristics can be divided into the areas of *organization*, *economy*, *technology* and *compliance*. The building blocks allow a general understanding of the categories of the individual properties and their structuring.

In the second stage, we examined the primary concepts of ADMS. The ADMS divides the properties into 3 concepts, the *repository*, the *asset* and the *distribution*. This classification is also suitable for describing a data source:

- *Repository:* A data source has a storage location. This storage location is marked as a repository. It describes the data catalog of a company or business area.
- *Asset:* The concept of the asset describes the actual data source and its main properties. The characteristics of an asset are primarily organizational and economic properties that enable the management and retrieval of a data source.
- *Distribution:* The distribution of a data source describes the characteristics of data provision. This concept mainly describes the technical characteristics of the data source and is an important source of information for interactions such as data access.

During the third stage, we assigned the identified characteristics to the individual concepts. We merged thematically dependent or related properties into classes that were assigned to the concepts. At this point, even though the models build upon each other, the differences between the models and their specializations become clear. Georgiev et al. show a similar approach with the RAGE metamodel [27].

In the fourth stage, the development ends with the typing of the various attributes in order to achieve the highest possible degree of standardization.

For the description of data goods, we added the missing classes and properties identified from analyses of marketplaces and scientific literature to the ADMS model. Table 2 lists these properties added to the metamodel for data goods with a description.

Table 2. Properties added to the first concept level

Property	*Description*
Curator	In addition to the property of publisher, we propose another for curator information. The curator is the responsible person for the data source.
Subscriber	Applications, processes or individuals that use the data source.
Accessibility	Define access level (e.g. 'public', 'club', 'private') in competitive markets where public availability means less competitiveness.
Owner	For data sources, the publisher must not always be the owner, so we added the property to solve this issue.

Property	*Description*
Rating	Compared to material goods a recommendation play more and more an important role. Because of this, we added a rating property for the product data as well.
Alternatives	By cataloguing *company's* data sources, the alternative property allows to tag alternative sources for information retrieval.
Origin	For data goods, it is necessary to know the origin and add territorial notices.
DataSetCount	The dataset count describes how many datasets are included in the data source.
UpdateCycle	Specifies the frequency at which the data source is usually updated.
Encryption	Encryption property is used to describe the encryption type of the data source and the used algorithm.
AuthorizationType	Describes whether and if so, which logon method is used.
CompressionType	Describes whether and if so, how the data is compressed.
AccessType	Describes how the data can be retrieved technically.
Sample	Here an example data record can be stored, which allows a first insight into the data source.
Volume	Describes if the volume of the data source is static or dynamic.

In addition to adding properties to existing classes, new classes have also been added to the model, if they serve to describe data as a commodity (see Table 3). Especially economic attributes were added to the M4DG, as they have not received much attention in previous metadata models for data sources.

Table 3. Added classes to represent economic properties

Class	*Property*	*Example*
BillingModel	metric	Pricing model for a data source
BillingModel	price	Pricing information for the data source
BillingModel	discount	Information on discounts that play a role in the purchase of data goods
	payment	The method of paying for the data
LicenceDocument	SLA	Service Level Agreements between the publisher and the user of the data source. Especially in the B2B environment also an interesting factor for data.
LicenceDocument	restriction	Restrictions on the use of data
LicenceDocument	rights	Rights of use for the data
RightsStatement	rightsHolder	Who owns rights to the data
RightsStatement	actions	What can be done with the data
RightsStatement	accessRights	Who can access the data

Like DCAT and ADMS, we also use existing vocabularies, like *dcterms* or *foaf*, to typify the properties of the new classes [31, 32]. This allows a common way of typification and common understanding when transferring the model into an information system for practitioners.

4.3 Resulting Metadata Model

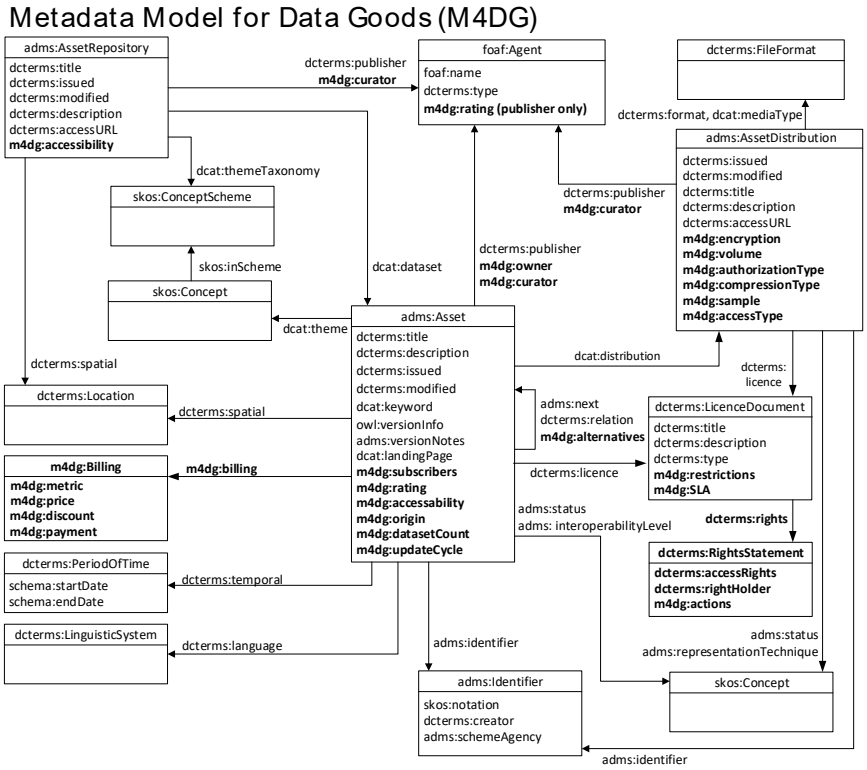

Figure 1. The metadata model for data goods (M4DG)

The result is a metadata model derived from the ADMS, which contains the necessary properties for managing data. The metadata model shown in Figure 1was developed by looking at existing marketplaces, discussions with companies and the analysis of existing literature. We added necessary properties during the design phase and we removed properties that were not related to the context under consideration. Wherever possible, we used existing vocabularies to define and unambiguously typify the properties of the model. By using the 3-levels *repository*, *asset* and *distribution* and the removal and collection of individual properties in classes and themes the data model remains extensible for future versions of the M4DG.

5 Evaluation of the Metadata Model

The goal of the metadata model is to describe data sources, and thus data goods, in a uniform way and consider them as a future resource in companies. For this reason, if data is considered as product, it should be inventoried like any other product. In collaboration with an worldwide active partner from the pharmaceutical industry, who purchases a lot of data from third parties (e.g. patient data or drug studies), we developed an inventory software that enables internal and external data goods to be managed and which supports the mapping of data governance processes, such as assigning responsibilities or evaluating data goods. We developed our software prototype as a web application that enables users to collect existing data sources, search for already registered data goods or display statistics and indicators on inventoried data sources.

The backend of this application uses the M4DG metadata model presented in this paper. Therefore, we implemented M4DG in this project as a mongoose[7] scheme. It enables storing and validating application data in the document-based database MongoDB[8].

Thus we can collect the data sources used by a company in a central repository and catalogue them according to a uniform model. The uniform description and management of the data sources enabled some potentials to be made visible, such as the avoidance of duplicates through the internal search function, the simplified determination of alternative data sources based on entities and simplified control options, including user evaluation. Furthermore, the inventory can be extended with external data sources such as from data marketplaces and industrial platforms by means of a uniform interface.

6 Summary and Outlook

Companies face new challenges in handling and managing data. This report introduces the Metadata Model for Data Goods (M4DG), a metadata model for describing data sources with a focus on using information in an economic context. For this purpose, we examined and compared descriptions of data goods on existing marketplaces on the internet. Subsequently, we identified related metadata models that could serve as the basis for a standardized design. Based on the Asset Description Metadata Schema (ADMS), we developed a model by adding properties and attributes and structuring them using the concept of classes. We examined the model critically from a practical point of view and adapted it to the knowledge gained from practical experience through workshops with companies. Finally, we were able to verify the practicability of the M4DG by using it as a database schema for the inventory of data sources for an international company which purchases a lot of data sources and has to manage all these data sources.

[7] http://mongoosejs.com/, accessed 01.09.2017
[8] https://www.mongodb.com/, accessed 01.09.2017

In the context of our future research, we want to use M4DG as the basis for a standardized description of data sources. We are looking forward to further evaluate the model in cooperation with our current partners and new companies from other industry sectors in order to improve the design and the general validity of the model through further iterations in the sense of the DSR. A first step in this direction could be the extension of the data basis by looking at data marketplaces that have been released in the meantime. Data marketplaces previously ignored due to a lack of availability of information could also be considered again. We are aware that the selection of marketplaces is a weak point and we would like to improve it in the following works. With this contribution we would like to encourage the community to carry out further research in this field.

7 Acknowledgements

This work was partly supported by Fraunhofer Center of Excellence for Logistics and IT.

References

1. Otto, B., Bärenfänger, R., Steinbuß, S.: Digital Business Engineering. Methodological Foundations and First Experiences from the Field (2015)
2. Choi, S.-Y., Stahl, D.O., Whinston, A.B.: The economics of electronic commerce. Macmillan Technical Publishing Indianapolis, IN (1997)
3. Moody, D.L., Walsh, P. (eds.): Measuring the Value Of Information-An Asset Valuation Approach (1999)
4. Otto, B., Jürjens, J., Schon, J., Auer, S., Menz, N., Wenzel, S., Cirullies, J.: Industrial Data Space. Digitale Souveränität über Daten. Whitepaper (2016)
5. Object Management Group, Inc.: Industrial Internet Consortium, http://www.iiconsortium.org/
6. Otto, B.: Quality and value of the data resource in large enterprises. Information Systems Management 32, 234–251 (2015)
7. Hirshleifer, J.: Where are we in the theory of information? The American Economic Review 63, 31–39 (1973)
8. Hilton, R.W.: The determinants of information value. Synthesizing some general results. Management Science 27, 57–64 (1981)
9. Stigler, G.J.: The economics of information. Journal of political economy 69, 213–225 (1961)
10. Repo, A.J.: The value of information. Approaches in economics, accounting, and management science. Journal of the American Society for Information Science (1986-1998) 40, 68 (1989)
11. Muschalle, A., Stahl, F., Löser, A., Vossen, G. (eds.): Pricing approaches for data markets. Springer (2012)
12. Schomm, F., Stahl, F., Vossen, G.: Marketplaces for data. An initial survey. ACM SIGMOD Record 42, 15–26 (2013)

13. Heckman, J.R., Boehmer, E.L., Peters, E.H., Davaloo, M., Kurup, N.G.: A Pricing Model for Data Markets. iConference 2015 Proceedings (2015)
14. Armstrong, A.A., Durfee, E.H. (eds.): Mixing and memory. Emergent cooperation in an information marketplace. IEEE (1998)
15. Otto, B., Auer, S., Cirullies, J., Jürjens, J., Menz, N., Schon, J., Wenzel, S.: Industrial data space. Digital souvereignity over data (2016)
16. Hankel, M., Rexroth, B.: The reference architectural model industrie 4.0 (rami 4.0). ZVEI (2015)
17. Otto, B., Lohmann, S., Auer, S., Brost, G., Cirullies, J., Eitel, A., Ernst, T., Haas, C., Huber, M., Jung, C.: Reference architecture model for the Industrial Data Space. Fraunhofer-Gesellschaft, Munich (2017)
18. Weibel, S., Kunze, J., Lagoze, C., Wolf, M.: Dublin core metadata for resource discovery (1998)
19. International Organization for Standardization: ISO 15836:2009 - Information and documentation -- The Dublin Core metadata element set, https://www.iso.org/standard/52142.html
20. Maali, F., Cyganiak, R., Peristeras, V. (eds.): Enabling Interoperability of Government Data Catalogues. Springer (2010)
21. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J. (eds.): Describing Linked Datasets (2009)
22. Tambouris, E., Manouselis, N., Costopoulou, C.: Metadata for digital collections of e-government resources. The Electronic Library 25, 176–192 (2007)
23. Alasem, A.: An overview of e-government metadata standards and initiatives based on Dublin Core. Electronic Journal of e-Government 7, 1–10 (2009)
24. CKAN Association: CKAN instances around the world, https://ckan.org/about/instances/
25. Marienfeld, F., Schieferdecker, I., Lapi, E., Tcholtchev, N. (eds.): Metadata aggregation at govdata.de. An experience report. ACM (2013)
26. Shukair, G., Loutas, N., Peristeras, V., Sklarß, S.: Towards semantically interoperable metadata repositories. The asset description metadata schema. Computers in Industry 64, 10–18 (2013)
27. Georgiev, A., Grigorov, A., Bontchev, B., Boytchev, P., Stefanov, K., Bahreini, K., Nyamsuren, E., van der Vegt, W., Westera, W., Prada, R. (eds.): The RAGE software asset model and metadata model. Springer (2016)
28. Alan, R.H. von, March, S.T., Park, J., Ram, S.: Design science in information systems research. MIS quarterly 28, 75–105 (2004)
29. Sein, M.K., Henfridsson, O., Purao, S., Rossi, M., Lindgren, R.: Action design research. MIS quarterly, 37–56 (2011)
30. Webster, J., Watson, R.T.: Analyzing the past to prepare for the future. Writing a literature review. MIS quarterly, xiii–xxiii (2002)
31. The Dublin Core Metadata Initiative: DCMI Metadata Terms, http://dublincore.org/documents/dcmi-terms/
32. Brickley, D. and Miller, L.: FOAF vocabulary specification 0.91. Technical report, ILRT (2007)