Creating Core Ontology for Social Sciences Empirical Data Integration

Dmitry Kudryavtsev[a], Tatiana Gavrilova[b] and Alena Begler[c]
Graduate School of Management St. Petersburg University, Saint-Petersburg, Russia

Keywords: Ontology Reuse, Empirical Data Integration, Empirical Research Datasets, Knowledge Engineering.

Abstract: There exist several dozens of metadata standards for empirical research data, making it difficult for users to choose and apply such standards. Consequently, the integration of datasets from similar empirical studies for further knowledge acquisition is highly constrained. To resolve this problem, an ontology for social science research data integration (Empirion-core) has been developed. The ontology reuses existing data integration schemas: DDI-RDF Discovery Vocabulary, Generic Statistical Information Model, Core Ontology for Scientific Research Activities, Data Catalog Vocabulary, and DCMI Metadata Terms. It consists of five subontologies that provide concepts for empirical datasets description: Information resource ontology, Research activity ontology, Research coverage ontology, Measurement ontology, and Sampling ontology.

1 INTRODUCTION

The volume and number of datasets generated by empirical research are increasingly becoming a challenge for professionals, who use them to extract knowledge and carry out comparative studies. More specifically, the amount of such data is increasing, while their accessibility and reuse potential are declining (Vines et al., 2014). Over two thousand repositories for open research data exist (according to the re3data.org web portal). Each of these repositories can store up to several hundred thousands open datasets. From this perspective, therefore, research data represents one of major open data categories alongside the government data (Vetrò et al., 2016; Mouromtsev, Jens et al., 2015).

Unfortunately, majority of these datasets cannot be reused due to lack of data integration opportunities. One of the issues is lump of metadata schemas that vary among repositories and research groups. The inconsistency in data description leads to data fragmentation and inhibits knowledge acquisition, as it implies working with separate datasets, whereas certain tasks can be handled more efficiently using multiple integrated datasets. Increasing amount and heterogeneity of data results in the growing need for new data integration solutions, as well as general principles of organization, storing and distribution (Atkinson, Gesing, Montagnat, & Taylor, 2017; Wilkinson, 2016). This problem is particularly important for social science data as it is relatively cheap and therefore less hardly managed.

The advancement of semantic technologies, including the ones designed for data integration, can be regarded as a response to this challenge (Kudryavtsev & Gavrilova, 2020; Lenzerini, 2011; Li et al., 2013). Ontologies have already been used for research data description. Existing application can be divided into three types:

- metadata schemas associated with a specific type of data or repository – for instance, DCAT, a recommendation of the World Wide Web Consortium for data catalogue integration (Archer, 2014), or DataCite Metadata Schema (Starr & Gastl, 2011) built for citing data through the DataCite web portal;
- domain-specific ontologies;
- ontologies of research activity such as ontologies of scientific experiments EXPO (Soldatova & King, 2006), ontology of research activity (Zagorulko & Zagorulko, 2015), The SWRC

[a] https://orcid.org/0000-0002-1798-5809
[b] https://orcid.org/0000-0003-1466-8100
[c] https://orcid.org/0000-0003-4375-1106

Kudryavtsev, D., Gavrilova, T. and Begler, A.
Creating Core Ontology for Social Sciences Empirical Data Integration.
DOI: 10.5220/0010146502670274
In *Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2020) - Volume 2: KEOD*, pages 267-274
ISBN: 978-989-758-474-9

(Semantic Web Research Community) Ontology (Sure, Bloehdorn, Haase, Hartmann, & Oberle, 2005), (KA)2 ontology (Benjamins & Fensel, 1998), a schema of the Integrated Scientific Information Space of the Russian Academy of Sciences (Bezdushny et al., 2000).

Such ontologies (especially metadata schemas) do create an infrastructure for datasets storing and description, however, they are not fully support data reuse as they are lacking variable-level refinement.

The approach under discussion is aimed to overcome this limitation with a creation of an ontology for empirical datasets description (Empirion-core). The ontology is defined as “formal, explicit specification of a shared conceptualization” (Gruber, 1993; Studer, Benjamins, & Fensel, 1998) that can serve as a knowledge base schema (Villazon-Terrazas et al., 2017). The proposed ontology is a core ontology in a sense that it includes a set of concepts that are both necessary for social science empirical data description and are not domain-specific (Ruy et al., 2017). The development of ontology is based on the scenario of knowledge reuse (Suarez-Figueroa, Gómez-Pérez, & Fernández-López, 2012) that includes the following stages: 1) specification of ontology requirements; 2) analysis of ontology reuse resources; 3) conceptualization; 4) formalization; 5) software implementation. This paper presents the first three stages (one by section).

2 REQUIREMENTS TO ONTOLOGY FOR EMPIRICAL RESEARCH DATA INTEGRATION

While existing schemas describe dataset as a whole, some dataset-related tasks require understanding of the variables inside the dataset. For example, researcher or analyst might be interested in datasets, where (examples for a managerial research provided in brackets):

- concept or phenomenon or construct X is examined *(e.g. a company’s human capital)*;
- the relationship between concept X and concept Y is examined *(e.g. the relationship between a company’s human capital and its performance)*;
- variable A is used to evaluate concept X *(e.g. variable ‘the proportion of staff members with higher education’ used to evaluate the concept ‘company’s human capital’)*;
- a specific data collection method was used *(e.g. a survey)*;
- data was collected in a specific region or a region with specific characteristics *(e.g. in a developing country / emerging market)*;
- samples included members of the population with specific characteristics *(e.g. with disabilities)*;
- variable Z was considered as an influencing factor *(e.g. company size)*;
- the study was aimed at resolving a problem X *(e.g. investment decision-making)*;
- data was collected in the last 5 years;
- specific equipment or technologies were used during data collection (*e.g. eye movement tracking*).

To answer such questions ontology should contain terms that allow to address exact variables in a dataset. This is important because different datasets might contain data for similar variables that can be integrated. For example, different managerial research of the concept “customer loyalty” may use different variables (or metrics) such as Net Promoter Score (NPS) or Repurchase Ratio. The principle of variable-based integration is shown at Fig. 1.

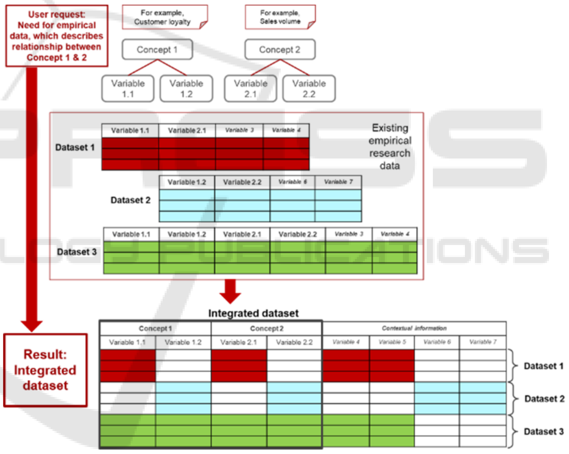

Figure 1: A principle of datasets integration.

The backbone of the ontology development was three main assumptions. The first, the ontology should be, above all, suitable for the description of research data in social sciences (see subject areas in All Science Journal Classification Codes, ASJC), including cognitive research, management, and economics. The second, it should be extensible as any research group can create a conceptual model for specific variables. The third, it should itself extend existing research metadata infrastructure as the field of research publishing (including data publishing) is covered by the variety of schemes and vocabularies.

3 ANALYSIS OF RESOURCES FOR REUSE

To answer discussed questions, an ontology should include the terms:

- dataset (data array);
- concept (phenomenon, construct);
- variable;
- data collection method;
- sample;
- member of the population;
- place of the study / data collection;
- time of the study / data collection;
- data collection equipment / technologies.

Thus, the ontology should incorporate a data description format for empirical research data, empirical research ontology, ontology of research activity, etc. Aside from this, it should also be able to integrate domain-specific extensions.

The given terms are partially covered by existing ontology resources, suitable for reuse. To construct the requisite ontology, the following ontologies were used as a basis:

- DDI-RDF Discovery Vocabulary (disco) (Bosch, Gregory, Cyganiak, & Wackerow, 2013);
- Generic Statistical Information Model (GSIM) (UNECE, 2013);
- Core Ontology for Scientific Research Activities (COSRA) (Campos, Reginato, & Almeida, 2019);
- Data Catalog Vocabulary (DCAT) (Archer, 2014);
- DCMI Metadata Terms ("DCMI Metadata Terms," n.d.).

DDI-RDF Discovery Vocabulary (disco) (Bosch, Gregory, Cyganiak, & Wackerow, 2013), for describing statistics data and metadata, focused primarily on questionnaires. Data Documentation Initiative (DDI) were developed for the social, behavioral, and economic sciences data management (Bosch, Gregory, Cyganiak, & Wackerow, 2013). This standard deals with social science data, data covering human activity, and other data based on observational methods measuring real-life phenomena. DDI formally describes the main concepts and common practices in this domain and puts stress on both microdata and aggregated data. It concentrates on microdata – data about the attributes and properties of population units. DDI offers the reuse of metadata of existing studies (e.g. questions, variables) for designing other studies, an important ability for repeated surveys and for comparison purposes (Vardigan et al, 2008). The DDI-RDF Discovery Vocabulary represents a research dataset organization as a set of variables. It also provides connection of the dataset with related research entities such as instruments (for example, questionnaire) and concepts under investigation.

Generic Statistical Information Model (GSIM) (UNECE, 2013), is dedicated to statistics data description. It proposes a general framework of internationally agreed definitions, attributes and relationships that illustrate the fragments information that are used in official statistics or other open information objects. GSIM may be supposed as a common language to describe information as a background of the common statistical production procedures from the identification of user requirements to the dissemination of the statistical results and outcomes (GSIM Brochure; UNECE, 2013). We reused some of the elements from Concept Group, which defines the meaning of information to provide understanding of what the data are measuring (Clickable GSIM v1.2) and from Structure Group, which describes the information structure within the statistical process (Clickable GSIM v1.2).

The main value of the GSIM for the research datasets representation is a three-level approach to the variable understanding: (1) as a something that can be measured; (2) as its representation as a measurement of a particular kind; and (3) as a concrete measure. The third can be understood as data values in a dataset, the second as variables with the information needed to interpret them, and the first as all the possible variable that may occur in datasets.

Core Ontology for Scientific Research Activities (COSRA) (Campos, Reginato, & Almeida, 2019), a domain-independent ontology for describing research processes related to data collection. The ontology provides classes for the detailed description of the activities necessary to interpret a dataset. Its scope is similar to the proposed ontology with two important differences. At first, the proposed ontology focuses on social science datasets and consider corresponding standards (DDI, GSIM), while COSRA is more generic and is deductively created from upper ontology (Unified Foundational Ontology, UFO). A strong dependence on UFO is not great for Social Science, where a majority of domain ontologies are Basic Formal Ontology (BFO) compliant. At second, the proposed ontology is dedicated to description of variables in the empirical research datasets and goes in more detail at a dataset level while in less detail at a context level. However, the context related COSRA classes are relevant to the social science datasets description. Additionally, the proposed ontology reuses some elements of COSRA's modular structure.

Data Catalog Vocabulary (DCAT) (Archer, 2014) – a recommendation of the World Wide Web Consortium aimed at increasing the dataset interoperability. The majority of research related metadata vocabularies (including the mentioned DDI-RDF Discovery Vocabulary) is aligned with DCAT. It describes the dataset as a part of a catalog and provides description of dataset's physical properties as well as time and space coverage.

DCMI Metadata Terms ("DCMI Metadata Terms," n.d.) – Dublin Core, a metadata collection that provides a minimal necessary information for digital asset description and widely used by academic literature publishers and research data repositories. It contains such concepts as author and creation date and thus necessary for reuse in the proposed model.

Despite all the mentioned models describe datasets in detail they provide only a dataset-level description. Namely, majority of them allow to say something about the dataset, but not about variables it contains. The proposed ontology aimed to extend the mentioned vocabularies to allow variable-level integration. Thus, it extends the mentioned vocabularies with the additional level of details.

4 CONCEPTUAL ONTOLOGY SCHEMA FOR EMPIRICAL RESEARCH DATA (EMPIRION-CORE)

The architecture of the Empirion-core ontology resembles the Core Ontology for Scientific Research Activities (COSRA), which includes Research activity ontology, Sampling ontology, Preparation ontology and Measurement ontology. But since the Empirion-core ontology is targeted at Social Science data we mostly reused concepts from DDI-RDF Discovery Vocabulary (DISCO) and Generic Statistical Information Model (GSIM). The Empirion-core ontology consists of 5 subontologies:

1. *Information resource ontology* that considers and describes empirical research dataset as a kind of information resource. This ontology references *disco:Dataset* and *dcat:Resource* classes and provides their connections with different types of *Metadata* using corresponding classes. This is important as research dataset is often accompanied be the metadata in the separate files.
2. *Research activity ontology* with a focus on types of research activities, agents (or actors) of research activity and methods of data collection. This ontology references classes cosra:ResearchActivity and disco:Instrument to reflect how dataset was collected. These classes relate to the *Research coverage ontology* classes.
3. *Research coverage ontology* provides a context of the research activities and thus locates dataset in space (*Period of Time* class) and time (*Location* class) and connects it with the information about *Object* and *Subject of research*.
4. *Measurement ontology* is especially important in the context of variable-based data integration. The central concept of this ontology is *gsim:Variable*, it also includes the associated concepts such as measurement unit, value domain etc. The goal of this ontology is to provide meaningful variables description. For example, the same measurement may be presented using different measurement units. To allow the data integration for such cases the ontology references classes *gsim:MeasurementUnit* that connects variable with its unit of measure and *gsim:ValueDomain* that allows to describe range of values.
5. *Sampling ontology* describes units of research such as sample and target population. This ontology extends *Research coverage ontology* with the *Target population* and *Sample* classes. The latter characterizes the former and relates to *Measurement ontology*.

The more detailed reuse and mapping to existing ontologies is represented in Table 1. All the subontologies are connected through relations between their classes (see Figure 3). The key class is *Data Set* that *has representation* in particular files and is *described* by metadata. This physical representation of dataset (described by *Information resource ontology*) relates to *Research coverage ontology* concepts as dataset *has geographical* and *temporal coverage* as well as *subject* and *object coverage.* The research object should *include* target population that is *characterized* by the research sample (the concepts of *Sampling ontology*). The dataset is a result of a research activity that is described by the corresponding ontology. The research activity *is performed* by an agent and *uses* some methods of data collection. Finally, to interpret the information in the dataset it should be connected with *Measurement ontology*: the dataset *contains* variables that *measures* concepts.

5 CONCLUSIONS

The paper proposes an ontology for integration of the empirical datasets obtained in various research studies. Similar problems are encountered in a number of areas of the social sciences, and the approach discussed in the paper will support the integration and merging of such datasets in order to extract new knowledge from existing data. The proposed approach is based on the ontology engineering paradigm and principles.

The proposed Empirion-core ontology allows to describe datasets obtained from empirical research. Empirion-core combines and merges existing ontological and non-ontological resources, and supplements them with the new necessary concepts (terms) allowing to display metadata schemes used in existing data sets of empirical research.

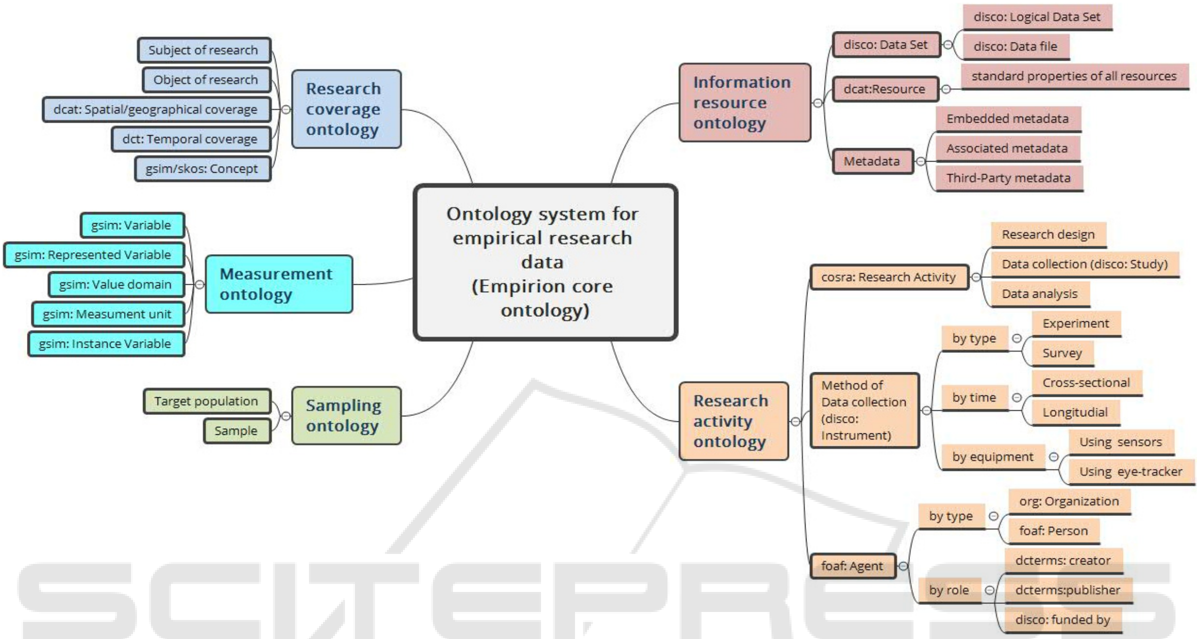

Figure 2: Composition of an ontology system for empirical research data (Empirion-core), upper level.

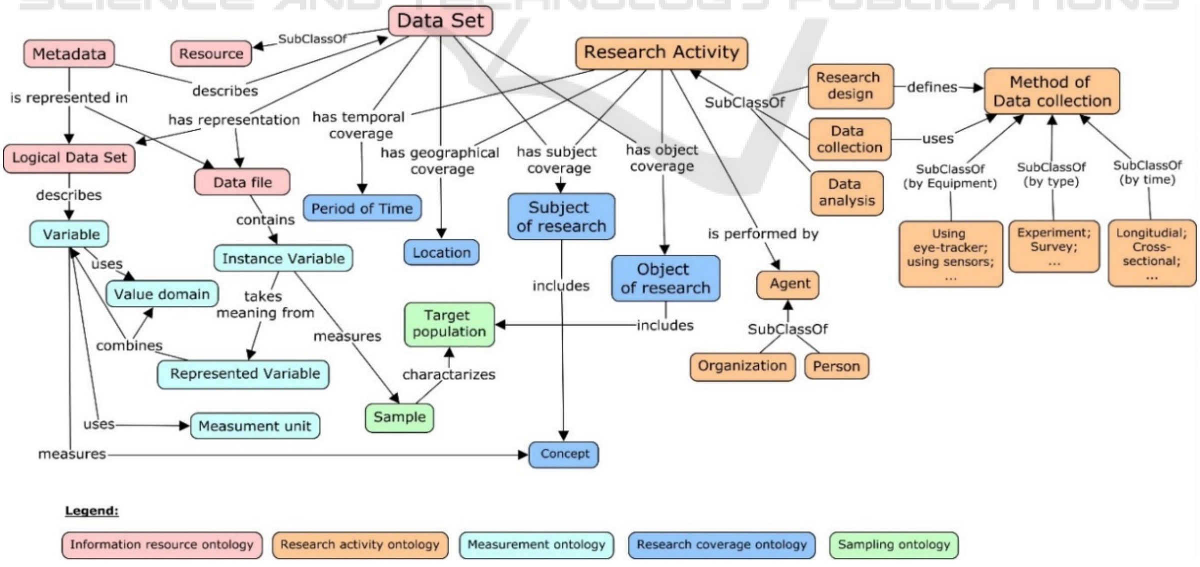

Figure 3: Relationships in Empirion-core ontology.
Reused concepts specified in the Table 1 for the sake of readability.

Table 1: Reuse and mapping to existing ontologies and non-ontological resources.

Empirion-core elements	DDI-RDF Discovery Vocabulary (DISCO) (Bosch, Gregory, Cyganiak, & Wackerow, 2013)	Generic Statistical Information Model (GSIM) (UNECE, 2013)	Core Ontology for Scientific Research Activities (COSRA) (Campos, Reginato, & Almeida, 2019)	Other ontologies and non-ontological resources
Information resource ontology				
Data Set	Data Set			
Logical Data Set	Logical Data Set			
Data file	Data file			
Metadata				
Embedded metadata				(Duval, Hodgins, Sutton, & Weibel, 2002)
Associated metadata				
Third-Party metadata				
Research activity ontology				
Research activity			Research Activity	
Research design				
Data collection	Study		Measurement	
Data analysis				
Method of Data collection	Instrument			
Agent				foaf: Agent
Research coverage ontology				
has subject coverage				dc: Subject dct: has subject coverage
has object of research			Researchable Entity	
Concept	Universe	Concept		skos:Concept
has geographical coverage			Geographic point	dc: Spatial Coverage dcat: Spatial/geographical coverage
has temporal coverage				dc: has temporal coverage dcat: temporal coverage
Measurement ontology				
Variable	Variable	Variable	Measure Scale	
Represented variable		Represented variable		
Value domain		Value domain	Scale Value	
Measurement unit		Measurement unit	Measure Unit	
Instance variable		Instance variable	Measured Value	
Sampling ontology				
Target population	Universe	Universe	Sampled Entity	
Sample		Population	Sample	

ACKNOWLEDGEMENTS

The reported study was funded by RFBR, project number 20-07-00854.

REFERENCES

Archer, P. (2014). Data Catalog Vocabulary (DCAT). Retrieved July 20, 2019, from https://www.w3.org/TR/vocab-dcat/

Atkinson, M., Gesing, S., Montagnat, J., & Taylor, I. (2017). Scientific workflows: Past, present and future. *Future Generation Computer Systems*, *75*, 216–227. https://doi.org/10.1016/j.future.2017.05.041

Benjamins, V. R., & Fensel, D. (1998). Community is Knowledge! in (KA)2. In *Proceedings of the 11th Workshop on Knowledge Acquisition, Modeling, and Management (KAW'98)* (pp. 1–18).

Bezdushnyi, A. N., Zhizhchenko, A. B., Kulagin, M. V., & Serebryakov, V. A. (2000). Integrated information resource system of the Russian Academy of Sciences and a technology for developing digital libraries. *Programming and Computer Software,* 26(4), 177-185.

Bosch, T., Gregory, A., Cyganiak, R., & Wackerow, J. (2013). DDI-RDF discovery vocabulary: A metadata vocabulary for documenting research and survey data. In *Proceedings of the WWW2013 Workshop on Linked Data on the Web (LDOW2013)* (Vol. 996). CEUR Workshop Proceedings.

Campos, P. M. C., Reginato, C. C., & Almeida, J. P. A. (2019). Towards a Core Ontology for Scientific Research Activities. *Lecture Notes in Computer Science*, *11787*, 3–12. https://doi.org/10.1007/978-3-030-34146-6_1

Clickable GSIM v1.2 URL: https://statswiki.unece.org/ display/clickablegsimDCMI Metadata Terms. (n.d.). Retrieved from https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

Duval, E., Hodgins, W., Sutton, S., & Weibel, S. L. (2002). Metadata Principles and Practicalities. D-Lib Magazine, 8(4)

Generic Statistical Information Model (GSIM). (n.d.). Retrieved from https://statswiki.unece.org/display/gsim

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199–220. https://doi.org/10.1006/knac.1993.1008

ISO 19156:2011: Geographic information – Observations and measurements. *International Standard* (2011)

Kudryavtsev D., Gavrilova T. (2020) An Overview of Practical Ontology Implementation in Decision Support Systems. In: Arseniev D., Overmeyer L.

Lenzerini, M. (2011, October). Ontology-based data management. *In Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 5-6).

Li, Y. F., Kennedy, G., Ngoran, F., Wu, P., & Hunter, J. (2013). An ontology-centric architecture for extensible scientific data management systems. *Future Generation Computer Systems*, 29(2), 641-653.

Mouromtsev D., Jens, L., Semerhanov I., & Navrotskiy M. (2015). Study of current approaches for Web publishing of open scientific data. *Journal Scientific and Technical Of Information Technologies, Mechanics and Optics,* 100(6), 1081-1087.

Ruy, F. B., Guizzardi, G., Falbo, R. A., Reginato, C. C., & Santos, V. A. (2017). From reference ontologies to ontology patterns and back. *Data and Knowledge Engineering,* 109(March), 41–69. https://doi.org/10.1016/j.datak.2017.03.004

Soldatova, L. N., & King, R. D. (2006). An ontology of scientific experiments. *Journal of The Royal Society Interface*, *3*(11), 795–803. https://doi.org/10.1098/rsif.2006.0134

Starr, J., & Gastl, A. (2011). *IsCitedBy: A metadata scheme for DatacCite*. https://doi.org/10.1045/january2011-starr

Studer, R., Benjamins, V. R. R., & Fensel, D. (1998). Knowledge Engineering: Principles and methods. Data & Knowledge Engineering, 25(1), 161–198. https://doi.org/10.1016/S0169-023X(97)00056-6

Suarez-Figueroa, M. C., Gómez-Pérez, A., & Fernández-López, M. (2012). The NeOn Methodology for Ontology Engineering. In *Ontology Engineering in a Networked World* (pp. 9–34). https://doi.org/10.1007/978-3-642-24794-1

Sure, Y., Bloehdorn, S., Haase, P., Hartmann, J., & Oberle, D. (2005). The SWRC ontology – Semantic Web for research communities. *Proceedings of the 12th Portuguese Conference on Artificial Intelligence – Progress in Artificial Intelligence (EPIA 2005)*, *3803*, 218–231. https://doi.org/10.1007/11595014_22

The Generic Statistical Information Model (GSIM) Brochure. URL: https://statswiki.unece.org/display/gsim/Brochures

Vardigan, M., Heus, P., and Thomas, W. Data Documentation Initiative: Toward a Standard for the Social Sciences. International Journal of Digital Curation 3, 1 (2008), 107–113.Villazon-Terrazas, B., Garcia-Santa, N., Ren, Y., Srinivas, K., Rodriguez-Muro, M., Alexopoulos, P., & Pan, J. Z. (2017). Construction of Enterprise Knowledge Graphs (I). In *Exploiting Linked Data and Knowledge Graphs in Large Organisations* (pp. 87–116). Springer. https://doi.org/10.1007/978-3-319-45654-6.

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

Villazon-Terrazas, B., Garcia-Santa, N., Ren, Y., Srinivas, K., Rodriguez-Muro, M., Alexopoulos, P., & Pan, J. Z. (2017). Construction of Enterprise Knowledge Graphs (I). *In Exploiting Linked Data and Knowledge Graphs in Large Organisations* (pp. 87–116). Springer. https://doi.org/10.1007/978-3-319-45654-6

Vines, T. H., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., … Rennison, D. J. (2014). The availability of research data declines rapidly with article age. *Current Biology*, *24*(1), 94–97. https://doi.org/10.1016/j.cub.2013.11.014

United Nations Economic Commission for Europe (UNECE) Generic Statistical Information Model (GSIM): Communication Paper for a General Statistical Audience (Version 1.1, December 2013) URL: https://statswiki.unece.org/display/gsim/GSIM+Communication+Paper.

Wilkinson, M. D. (2016). Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

Zagorulko, Y., & Zagorulko, G. (2015, September). Ontology-based technology for development of intelligent scientific internet resources. In International Conference on Intelligent Software Methodologies, Tools, and Techniques (pp. 227-241). Springer, Cham.