Advances in Social Science, Education and Humanities Research, volume 281
4th International Conference on Social Science and Contemporary Humanity Development (SSCHD 2018)

Studies on Publication Related Ontologies

Yun-Liang ZHANG[1,2], Fang YUAN[1,*]
[1]Institute of Scientific & Technical Information of China, Beijing 100038
[2]Key Laboratory of Rich-media Knowledge Organization and Service of Digital Publishing Content
zhangyl@istic.ac.cn, yuanfang@istic.ac.cn
*Corresponding author

Keywords: Ontologies, Publication, Comparison, Digital Publishing, Knowledge Organization Systems, Application.

Abstract. In the paper the new trend of use ontologies in the publication related fields in knowledge organization and service are analyzed. Some typical ontologies in both China and abroad are discussed and these ontologies are compared from different dimensions of construction and details. At last some suggestion about how to construct and use publication related ontologies are offered.

Introduction

The publishing industry has made great contributions to the individuals and the whole society for a long time. Through the publication and popularization of books, the monopoly of information and knowledge has been broken, so that almost everyone can access, understand and learn the information and knowledge needed. It is just because of the vigorous development of the publishing industry that the difficulty of obtaining knowledge is reduced, and more people can continuously increase their knowledge and supplement their own knowledge system through reading. Although the publishing industry has been challenged more as the types of media become more abundant, the publishing industry itself has its irreplaceable advantages. Most publishing companies have strict selection criteria and publishing process, and the professional division of the editing and processing team is clear and meticulous, which guarantees the authority and value of the publication to a certain extent. However, in recent years, due to the influence of the Internet, especially the mobile Internet, people's reading habits have changed. People are more likely to accept relatively short, graphical, multimedia content, and it is difficult to read formal publications. In order to cope with this situation and play the content advantage better, the digital publishing industry is developing rapidly [1], and it is increasingly recognized by the publication industry.

Digital publishing, represented by semantic publishing, offers a greater possibility for knowledge services. But semantics can be understood at least in two ways. The first is based on traditional linguistics and its branch names computational linguistics, and the second is based on the Semantic Web. There is a certain correlation between the two, but there is a big difference in terms of connotation and strategy. From the perspective of natural language understanding, semantics can be completely analyzed from content and its context, so it does not need more rules for the use of natural language writing, and emphasizes the use of late technology and knowledge base to restore the author's semantics.

The semantics in the Semantic Web emphasizes the elimination of the term ambiguity, defines a term uniquely through several related terms, and forms a public cognition and the computer processes the standard to achieve unambiguous expression. Therefore, the artificially defined properties of semantics in the Semantic Web are stronger, and the related techniques of the Semantic Web from the understanding and processing emphasize the addition of an explicit labeling layer embedding in the natural language, thus achieving the elimination of ambiguity and extensive association. The two methods can be combined with each other. Relatively speaking, most complicated understanding and annotation work is done manually, but the natural language understanding technology automatically preprocesses some relatively simple semantic expressions to save labor. So as one of the most

Copyright © 2019, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

important knowledge organization systems [2], ontology has been applied in a lot of fields [3] and can be also play an important role in publication.

Publication Related Ontologies out of China

Publishing related ontologies, includes concepts and metadata content for the description, association, and processing of different types of publications such as books, journals, papers, and data etc.

VIVO-ISF Ontology

VIVO is an open source software developed and promoted by the Library of Cornell University. It is used by hundreds of organizations in 25 countries. It provides an open ontology of the VIVO-ISF (namely the VIVO Integrated Semantic Framework) that reflects the basic ontology of academic activities. The latest version of the ontology framework is 1.6, covering most academic publishing resources, including books, journals, conferences, degrees and other related papers, patents, as well as databases, including various audio and video media, web pages, lectures, manuscripts, etc.

The ontology [4] contains 414 classes, and mainly includes *Addressing, Calendar, Geographical, Identification*, and so on. The total number of properties is 415, 198 are object properties include *provided by, contains*, and 217 are the data properties such as *abbreviation* and *has Code*.

Different application units can build instances based on their own resource situation and focus. Take the VIVO-based system built by Cornell University as an example [5]. The system example contains about 26,000 people, 4,400 events, 29,000 events, 12,000 institutions, 151,000 studies, and 2000 topics.

The Bibliographic Ontology

The Bibliographic Ontology, abbreviated as BIBO [6], is created by Frédércick Giasson and Bruce D'Arcus. It is mainly used to represent bibliographic information and citation relations. The official also gives detailed examples, which can be used to describe books, conference papers, articles, Legal cases, letters, manuscripts, etc.

The ontology [7, 8] contains 69 concept classes, of which the main categories include collection, document, organization, person, agent, event, etc. At the level of the large class, a large number of classes including FOAF and Dublin Core Metadata DC are borrowed. There are a total of 108 ontology properties, including 53 object properties such as *cited_by, produced_in*, etc. There are also a large number of these properties from FOAF and DC, such as *dcterms:title, dcterms:language, foaf:dipicition, foaf:homepage*, etc., the data properties are 55, include abstract, DOI, etc.

The PAV Ontology

Digital publishing must involve a large number of digital information resources, and there are also a large number of digital information resources and electronic digital archives on the Internet. In the process of publishing or releasing, the sources of digital resources are often involved, which is important for ensuring the accuracy of scientific content. In response to such demands, W3C organize a source information working group and proposed the source ontology The PROV Ontology. This ontology provides the basis for the expression, exchange, and integration of such digital resource information generated by different systems and different contexts. This ontology combines the information access and retrieval service (PROV-AQ) with the data model (PROV-DM) to enable information exchange in specific domain-based information applications such as via networks.

The PAV Ontology [9] is an ontology formed on the basis of The PROV Ontology, including proof, authoring, and versioning. The ontology contains 31 concept classes, of which the main categories include *Activity, Agent, Entity, Influence, Instantaneouse Event, Location, Role* and so on. There are a total of 50 properties, including 44 object properties including *at Location* and *had_activity*, and 6 data properties including *at Time* and *value*.

The ontology is designed for a relatively broad source of information. The researcher proposes an operational role that clearly distinguishes operational digital resources, such as author, contributor, and curator, based on the need to track network resource creation and version information, forming a PAV ontology [10] that complements the PROV Ontology.

Semantic Web Technology Evaluation Ontology

Semantic Web Technology Evaluation Ontology, also abbreviated as SWETO, is a related ontology of computer science publishing data constructed by the University of Georgia in the United States [11]. The ontology has relatively simple structure but a lot of examples. The source of the construction is the well-known English literature database DBLP developed and maintained by the Trier University, Germany. The ontology contains 114 concept classes, of which the main categories include *Person, Place, and Thing*, among which the *Thing* subclass has more subclasses. The total number of properties is 69, of which the object properties include 13 listed *author in, classified by*, etc., and the data properties include 56 such as *publication keywords and paper numbers*. In August 2007, the instances described by RDF have 2.4 million resources (including Examples of classes, properties, and classes), 3.06 million literals (which can be used as values for data properties), and 3.74 million Resource-to-Resource Triples, 7.27million Resource-to-Literal Triples.

Entity instances include *Person, Articles in Proceedings, Journal Articles, Webpages of persons, Proceedings, Book Chapters, Books*, etc. Properties include *publication-has-author, contained in proceedings, cites publication*, etc.

Publication Related Ontologies of China

PUBlication Ontology

PUBlication Ontology (PUBO), is a multi-level publication content resource model proposed by the standard "*CY/T 102.1-2014 Digital Content Object Storage, Reuse and Exchange Specification Part 1: Object Model*" in the publishing industry of China. The model belongs to the ontology model, so only the architecture discussed here.

The ontology [12] is expected to be able to uniformly model the digital content resources of publications such as magazines, periodicals, general books (including books, multi-volume books), and multimedia e-books. PUBO classifies digital resources into three broad categories: collection classes (including document containers and document collections and their subclasses), document resource classes (including document classes and their subclasses), and proxy classes (agent classes and its subclasses), specifically defined a total of 60 classes. In order to standardize the values of certain properties, PUBO defines eight enumeration classes in the form of controlled terms, namely metadata type, document status type, auxiliary type, cover type, publication packaging type, single page type, Reference type, product type. The ontology defines 48 data properties, *including productManifest, readingOrder, contains, associatedMedia, embeddedAudio, embeddedVideo, embeddedFont, isContentSection, isComponentOf, generates*. The ontology defines 107 data properties, and there are subclass relationships among some data properties. From the perspective of large classes, data properties are roughly divided into *date properties, file format properties, identifier properties, position attribute, product format properties, and title. Properties, agent information properties, other properties*, and more. The architecture of the PUBO ontology itself is quite large. It contains most of the classes, object properties and data properties of digital publishing. It can cover the ontology construction requirements in most cases. Even if it is not perfect, it only needs a small amount of expansion. What is more valuable is that the design of the ontology model is not out of thin air, but fully considers the reality of the digital publishing industry itself. At the same time, the ontology makes extensive reference to the existing domestic and international mature standard systems and built-up ontology, such as EPUB, BIBO, CEBX, Schema, OAI-ORE, CNMARC, CNONIX, MARC21, DC, METS, FOAF, SKOS, etc.

Education Publication Ontology

Anhui Education Network Publishing Co., Ltd. under the support of the National Science and Technology Support Program project “Dynamic Digital Publishing Service Model Support Technology R&D” project constructed the ontology of education field (EPO) [13]. The class of EPO is derived from the education science, education, education at all levels, schools, education management, teaching, curriculum, teaching materials, school facilities, teaching aids, educators, Students have a total of 12 facets. At the same time, further determine the relationship between some types of instances, that is, special concepts. The conceptual relationship of the ontology is divided into two categories, one is education-related relationship, and the other is digital publishing-related relationship. The education-related relationship setting is mainly based on the "CELTS-42 Basic Education Resource Metadata Application Specification", which was published by the Ministry of Education Information Technology Standards Committee in 2012. The "Code" first describes the core metadata elements and makes unambiguous definitions. On the basis of the above, several examples are constructed. The examples include both the concept of education ontology and the knowledge association between concepts. The ontology contains at least 4 major categories and 18 relationship types. Some of them about the human and organization include *Educate_to* (teachers impart knowledge to students), *Belong_to* (affiliation relationships such as students and grades) etc., and in about educational resources there are *HasTitle* (titles for educational resources), *Creator_is* (create educational resources) etc. There are at least five types of relationships in digital publishing, such as PubKey_of (keywords for digital publications), PubCover_to (target audience for digital publications), PubAccess_to (access to digital publications), PubID_of (identification for digital publications), PubCopyright_of (copyright of digital publications), etc.

Knowledge Element Ontology

Publishing House of Electronics Industry has constructed a Knowledge Element Ontology (KEO)seven-tuple knowledge element representation model containing *names, labels, subject terms, content, content categories, morphological categories, and sources* as a model of knowledge elements [14]. This model is equivalent to the data properties described by the ontology instance, and the establishment includes at least *is-a, include, have instances, have definitions, have sources, have laws, have icons*, and other object properties to describe the association between different entities. The agency uses the ontology to describe the physical related knowledge, from the specific the technical route shows that the Publishing House adopts the ontology, but fully considers the existing foundation, and fully associates the entity with the existing thesaurus in the subject of the seven-tuple, and the other semantic information is reflected in the rest. Since the annotation techniques with thesaurus are very mature, the work has great feasibility.

Digital Publishing Domain Ontology

Wuhan University has built a Digital Publishing Domain Ontology [15]. The ontology construction is mainly divided into two stages, namely, the topic vocabulary construction stage and the ontology transformation stage. The thesaurus construction stage first extracts candidate topic words from the corpus, conducts effective concept screening, normalizes the synonymous concept group names, then classifies the concepts, and adjusts the classification system according to the actual situation. Complete the compilation and revision of the thesaurus. The ontology transformation stage first determines the semantic relationship between concepts and further describes it as the ontology model. Then the constructive software Protégé is used to complete the formal description of the ontology, and its plug-in is used to realize the visual display of ontology content. The constructed ontology includes about 680 concepts, including a 4-5 level classification system with 11 top categories as *00 theory, 01 policy regulations, 02 standard specifications, 03 technology, 04 tools, 05 processes, 06 cases, 07 industry, 08 digital publishing products, 09 related concepts, 10 institutions*.. The basic semantic relations including equivalence relation, hierarchical relationship and related relationship are further constructed, and the thesaurus is formed, and the relationship subdivision is

further based on the thesaurus. Specifically, it includes *Equals/Is synonym of, Has part/Is part of, Has type/Is type of ,Has instance/Is instance of, Has tool/Is tool of, Offer/Offered by, Develop/Developed by, Has standard/Is standard of9In Relation to10Manage/Manage by* et al. After automatic construction and manual verification, a total of more than 3,500 pairs were established. After the ontology modeling is completed, WebProtégé (an online version of Protégé) is used to adopt a multi-person online collaboration method to add relationships to concepts, and to check the hierarchical structure, class names and property values of the ontology, and complete the construction of the ontology.

Publication Related Ontologies Analysis

This paper studies 9 publication related ontologies of China and abroad, the details are shown in table 1, which including 5 foreign ontology, namely VIVO-ISF ontology, BIBO, PAV ontology, SWETO, comic book ontology CBO, and 4 domestic ontologies, that is PUBO, EPO, KEO, and DPDO.

Table 1, The details of the 9 publication related ontologies

Ontology Name	**Construction organizations**	**Organization Categories**	**Countries**
VIVO-ISF Ontology	VIVO Community (Led by Cornell University)	Community (University)	International (U.S.)
BIBO	BIBO Community	Community	International
PAV Ontology	Massachusetts General Hospital, Harvard University	Hospital & University	U.S.
SWETO	University of Georgia	University	U.S.
CBO	Kent State University	University	U.S.
PUBO	Led by State Administration of Press, Publication, Radio, Film and Television of The People's Republic of China	Government	China
EPO	Anhui Education Network Publishing Co., Ltd.	Publication Intitution	China
KEO	Publishing House of Electronics Industry	Publication Intitution	China
DPDO	Wuhan University	University	China

Among these ontology, there are six ontology mainly about publishing, including PAV ontology, SWETO, PUBO for publishing content resources, EPO, KEO, DPDO, related to bibliographic and document services. There are three mainly about bibliography and literature services, namely VIVO-ISF ontology, BIBO, and CBO.

The details of the construction information are shown in Table2. VIVO-ISF ontology, BIBO, and PUBO for publishing content resources are advocated by non-profit organizations or the governments, and they are widely used. The EPO and KEO are funded by government projects. The rest ontologies are experimental and developed by universities and the research institutes. DPDO has no instances and it is close to the structure of the thesaurus. VIVO-ISF ontology, BIBO, PAV ontology, PUBO themselves are models, and examples can be added in.

Table 2, The construction information of the 9 publication related ontologies

Ontology Name	**Construction Platform**	**Construction Method**	**Ontology References or Data sources**	**Instances**	**Instance magnitude**
VIVO-ISF Ontology	Unknown	Unknown	DC, FOAF, SKOS	It's schema, can add instances	Not applicable (Cornell University 10^5)
BIBO	Unknown	Unknown	FOAF, DC	It's schema, can add instances	Not applicable
PAV Ontology	Unknown	Based on PROV-O	DC, PROV-O, OPM, Provenance Vocabulary	It's schema, can add instances	Not applicable
SWETO	Semagix Freedom	Automatic Extraction	DC , FOAF	Have instances	10^7
CBO	Protégé	Unknown	DC, Schema.org	Have instances	10^3
PUBO	Unknown	Unknown	EPUB, BIBO, CEBX, Schema.org, OAI-ORE, CNMARC, CNONIX, MARC21, DC, METS, FOAF, SKOS	Have instances	Not applicable
EPO	Unknown	Unknown	CELTS-42	It's schema, can add instances	Unknown
KEO	Unknown	Unknown	Unknown	Have instances	Unknown
DPDO	WebProtégé	Based on Thesaurus	9 volumes of the series of Digital Publishing Theory, Technology and Practice	No instances	Not applicable

Summary

From the publication related ontologies constructed by different institutions, we know that in different scenarios of digital publishing and subsequent knowledge service, we should construct different ontologies to support. When construct publication related ontology, there are some exist ones we can borrow from. Usually there are some universal sources we can include in as references, in which DC and FOAF are most popular. And of course the ontologies would be used more widely if more instances are constructed and linked.

Acknowledgement

This research was financially supported by CKCEST Project Program (Grant No. CKCEST-2018-1-26), ISTIC Key Project Program (Grant No. ZD2018-07) and National Digital Composite Publishing System Project (Grant No. XWCB-ZDGC-FHCB/28). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

[1] XU Lifang, LIU Hongjin, CONG Ting. Conspectus on Digital Publication, Publishing House of Electronics Industry, Beijing, 2013.

[2] ZENG wen, LI ying, HAN hongqi, et al. Study on the Method of Data Organization and Management, Technology Intelligence Engineering, 2, 1 (2016) 109-113.

[3] ZHANG Dezheng, XIE Yonghong, LI Man, et al. Construction of Knowledge Graph of Traditional Chinese Medicine Based on the Ontology, Technology Intelligence Engineering, 3,1(2017)35-42.

[4] VIVO Core Ontology on http://vivoweb.org/sites/vivoweb.org/files/vivo-isf-public-1.6.owl#.

[5] VIVO Cornell University on http://vivo.cornell.edu/.

[6] Bibliographic Ontology Use case Examples on http://bibliographic-ontology.org/examples.

[7] The Bibliographic Ontology on https://raw.githubusercontent.com/structureddynamics/Bibliographic-Ontology-BIBO/master/bibo.owl.

[8] Bibliographic Ontology Specification on http://bibliographic-ontology.org/specification.

[9] The PROV Ontology on https://www.w3.org/TR/prov-o/.

[10] Ciccarese P, Soiland-Reyes S, Belhajjame K, et al. PAV ontology: provenance, authoring and versioning, Journal of biomedical semantics, 4, 37 (2013) 1-22.

[11] Aleman-Meza B, Hakimpour F, Arpinar I B, et al. SwetoDblp ontology of Computer Science publications [J]. Web Semantics: Science, Services and Agents on the World Wide Web, 5, 3 (2007)151-155.

[12] Wang Wenqing, Liu Chuntong, Zhang Yuexiang, et al. PUBO: Construction of Publication Ontology of Digital Resource, Journal of Academic Libraries, 33, 3 (2015) 88-95.

[13] RUAN Huai-wei. Research on Ontology Model of Learning Resources for Digital Publishing, Computer and Information Technology, 21, 5, (2013) 52-54.

[14] Li Hong. The Re-organization of Digital Content Resource Based on knowledge element and Ontology, Digital communication World, 10 (2015) 53-54.

[15] Si Li, Chen Yuxue, Zhuang Xiaozhe. The Construction of a Digital Publishing Domain Ontology Based on Thesaurus, Publishing Journal, 23, 6 (2015) 80-84.