DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications

Poster—Research on the construction and application of the semantic-based ontology model for biographical data

Li Si
School of Information Management, Wuhan University, China
lsiwhu@163.com

Yao Liu
School of Information Management, Wuhan University, China
2016302480135@whu.edu.cn

Abstract

Biography, a historical genre dating back to antiquity, is a scholarly resource in history and prosopography(Shimizu et al., 2020). It records the lives of individuals and serves as a witness to social customs and historical scenes. However, most biographies follow the traditional form of non-structured text and cannot reveal the semantic features of biographers(Hyvönen et al., 2019), which has hindered information mining and retrieval. Semantic Web technologies provide a solution through deep semantic descriptions. However, few studies explore the semantic modeling of biography. Given this, we constructed a biographical ontology based on existing ontology and terms extracted from an autobiography. On this basis, semantic applications, such as semantic retrieval, annotation, and reasoning, can be explored.

Keywords: ontology model; semantic model; biography

1. Data and Proposed Methodology

1.1. Data

To examine the fundamental elements for the description of biography, an autobiography of a renowned Chinese library scientist and educator Peng Feizhang was used as the primary data source. The autobiography has hundreds of rare photos and manuscripts, 318 pages and 300,000 words. It is rich in characters and events with complex relations, which ontology can demonstrate. Besides, external data such as historical studies and wiki are used as supplements.

1.2. Proposed Methodology

An ontology defines the common concepts used to represent an area of knowledge(Gruber, 1993). Concepts, their properties, and relationships are the main components of an ontology. After integrating mature ontology construction methods, the steps followed in developing the biographical ontology are described below.

(1) Requirement Analysis

The ontology is a framework for sequencing and revealing individuals' family background, academic achievement, professions, social status, kinship, social relationships, and academic networks. It aims to provide a reusable, scalable, and shareable model to organize, link, and present biographical data.

(2) Reusing existing ontologies

An intensive literature search was conducted to find related ontologies, and the conceptual framework of the China Biographical Database (CBDB) is one among them. It was a cooperative project between Harvard University and Peking University, establishing multiple forms of relationships among complex physical objects. Besides, common ontologies are used as supplements, such as FOAF, Time, GeoNames, BIBO, CERIF, shlgen, and CIDOC-CRM.

https://doi.org/10.23106/dcmi.953126255

(3) Knowledge acquisition

The Natural Language Processing and Information Retrieval Sharing Platform (NLPIR), a popular text processing tool for Chinese, was used, including sentence segmentation, part-of-speech tagging, and named entity recognition. The final result was 322 terms related to people, 178 terms related to time, 131 terms related to events, 95 terms related to organizations, and 45 terms related to places.

(4) Definition of the classes and the class hierarchy

Based on CBDB and knowledge acquisition, the core classes of the ontology are identified as Agent, Event, Time, Place, and Resource. For the expansion and supplementation of the core classes, the study set 19 subordinate classes regarding FOAF, shlgen, Time, BIBO, and CERIF.

(5) Definition of the properties and relations

Biographical data have many properties and relations that will be the basis for data mining and discovery. Based on some attributes of generic ontologies, the biographical ontology was proposed, containing 24 classes, 60 properties, and 51 relationships. For brevity, the figure below shows the main classes and relationships.

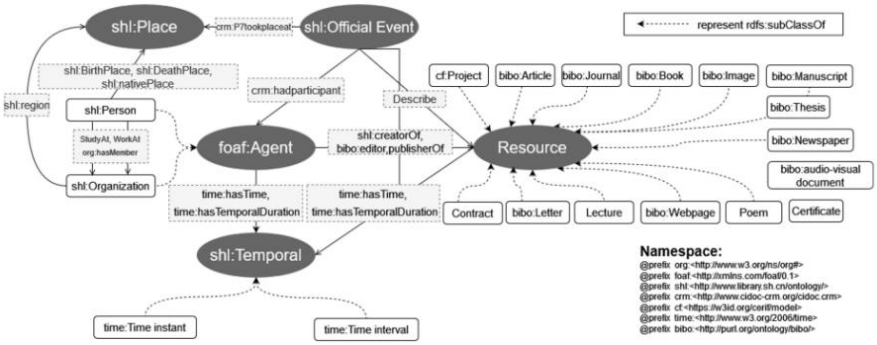

FIG.1.Semantic-based ontology model for biographical data (partial)

(6) Ontology formalization

Protégé was used as an editor for the ontology using the Web ontology language combined with description logic axioms. The ontology was filled with data collected from Mr.Peng's autobiography. TABLE 1 shows the metrics of the developed ontology.

TABLE 1: Metrics of the developed ontology model for biographic data

Parameters	Counts
Axiom	13519
Logical axioms count	12360
Declaration axioms count	1159
Class count	24
Object property count	51
Data property count	60
Individual count	995

2. Expected Benefits of the Ontology Model

2.1. Semantic retrieval

In terms of the central composition aspect of ontology, semantic retrieval can be carried out from the classes, properties, and relations. Supposing the information to be queried is all students who have been Mr.Peng's students. With the DL Query module and the relation called "Study At," we can get information about 111 alumni and then use the OntoGraf plugin to get the alumni network, as shown in the figure below.

https://doi.org/10.23106/dcmi.953126255

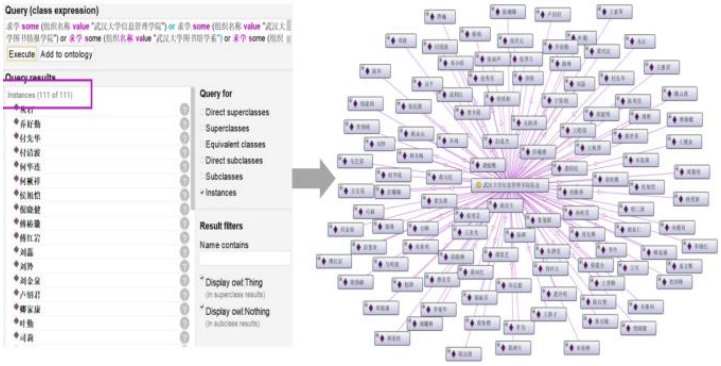

FIG.2.Semantic retrieval and visualization result of alumni of School of Information Management Wuhan University

2.2. Semantic Annotation

We conducted semantic annotation of biography based on the constructed ontology, transforming unstructured text into structured data. The study took the biography of Zhao Shiliang, a librarian, as an example, and FIG.3. shows the result. Firstly, the properties include the date and place of birth and death. Secondly, the related organizations, groups, academic resources, time, and place for the content semantics were extracted and annotated.

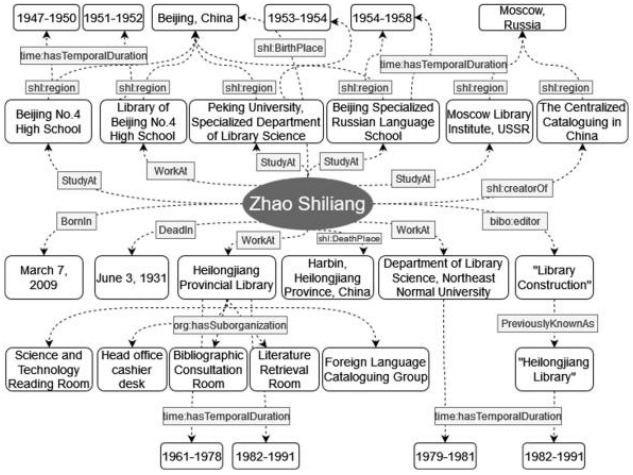

FIG.3.Semantic annotation result of Zhao Shiliang's biography

2.3. Semantic Reasoning

The ontology-based DL query can obtain explicit knowledge but not implicit information. The ontology-based HermiT reasoner can be used to achieve the implicit information using semantic data of the autobiography. The query of predecessor institutions of the School of Information Management at Wuhan University was used as an example. The six predecessor institutions of the school can be obtained simultaneously by running the inference rule since the adjusted institutions are related by "predecessor institutions" and "successor institutions."

References

Gruber, T. R. (1993). A translation approach to portable ontology specifications. Knowledge Acquisition, 5(2), 199-220.

Hyvönen, E., Leskinen, P., Tamper, M., Rantala, H., Ikkala, E., Tuominen, J., et al. (2019). BiographySampo – Publishing and enriching biographies on the semantic web for digital humanities research, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 11503 LNCS, pp. 574-589).

Shimizu, C., Hitzler, P., Hirt, Q., Rehberger, D., Estrecha, S. G., Foley, C., et al. (2020). The enslaved ontology: Peoples of the historic slave trade. Journal of Web Semantics, 63.

https://doi.org/10.23106/dcmi.953126255