Design and Application of a Fine-Grained Semantic Description Model for Modern Newspaper Resources: A Case Study of Shengjing Times (Postprint)

Authors: Sun Shaodan, Deng Jun, Chang Yanyu, Zhang Zishu, Shen Yong

Date: 2023-04-01T15:51:24+00:00

Abstract

[Purpose/Significance] This study designs a scientific and standardized fine-grained semantic description model for modern newspaper resources to deeply reveal their characteristics and interrelationships, providing a reference for the effective management, organization, knowledge discovery, and knowledge services of these resources. [Method/Process] By analyzing the logical structure, physical layout, and content information of modern newspaper resources, this research approaches from both domain ontology and metadata description perspectives, reusing the CIDOC-CRM ontology conceptual model and EAD, DC, "Metadata Specification for Ancient Books". Taking "Shengjing Times" as a case study, a fine-grained semantic description model for modern newspaper resources is designed, and the semantic description model is implemented in RDF/XML language using the Oxygen XML tool to achieve element interoperability and model application. [Results/Conclusion] The study provides an operable fine-grained semantic description model for modern newspaper resource organization, offering fundamental support for the construction of modern newspaper resource databases, standardized newspaper management, and application system development, thereby promoting the development, utilization, and sharing of modern newspaper resources.

Full Text

Preamble

Design and Application of a Fine-Grained Semantic Description Model for Modern Newspaper Resources: A Case Study of *Shengjing Times*

Sun Shaodan[1], **Deng Jun**[1], **Chang Yanyu**[1], **Zhang Zishu**[1], **Shen Yong**[2]
[1] School of Business and Management, Jilin University, Changchun 130012 [2] School of Public Health, Jilin University, Changchun 130022

Abstract: [Purpose/Significance] This paper designs a scientific and standardized fine-grained semantic description model for modern newspaper resources to deeply reveal the characteristics and relationships of these resources, providing references for effective management, organization, knowledge discovery, and knowledge services. [Method/Process] By analyzing the logical structure, physical layout, and content information of modern newspaper resources, this study approaches the problem from two perspectives: domain ontology and metadata description. Reusing the CIDOC-CRM ontology conceptual model and EAD, DC, and *Ancient Books Metadata Specification*, the paper designs a fine-grained semantic description model for modern newspaper resources using *Shengjing Times* as a case study, and employs Oxygen XML tools to describe the semantic model in RDF/XML language to achieve element interoperability and model application. [Result/Conclusion] The study provides an operable fine-grained semantic description model for modern newspaper resource organization, offering foundational support for database construction, standardized newspaper management, and application system development, thereby promoting the development, utilization, and sharing of modern newspaper resources.

Keywords: modern newspaper resources; semantic description model; *Shengjing Times*; RDF/XML **Classification Number:** G254 **DOI:** 10.13266/j.issn.0252-3116.2022.07.004

Modern newspaper literature comprehensively documents the tremendous social transformations of modern times, carrying the imprints of their era and serving as an important information source for research in social, political, economic, cultural, and journalism history. Modern newspapers originated during the embryonic stage of Western capitalism. For instance, German Johannes Gutenberg invented movable metal type printing in the second half of the 15th century, leading to the emergence of newspaper prototypes called "newsbooks." By the 17th century, as capitalism flourished in Europe, representative newspapers emerged in various countries. Compared with the West, modern Chinese newspapers appeared relatively late. The first batch was founded by foreign missionaries in China, such as *Chashisu Monthly Statistical Record*, *Eastern and Western Monthly Examination*, *Shenbao*, and *Shengjing Times*. These foreign-founded newspapers carried subjective political inclinations, primarily disseminating Western culture. The earliest daily newspaper independently run by Chinese nationals was *Zhaowen Xinbao*, founded by Ai Xiaomei in Hankou in 1873 [1]. Subsequently, the Hundred Days' Reform, Xinhai Revolution, and New Culture Movement brought reformists, revolutionaries, and emerging national bourgeoisie onto the historical stage, with newspapers such as *Shiwu Bao*, *Zhixin Bao*, and *Guowen Bao* emerging as platforms for public opinion propaganda. The May Fourth Movement, marking the beginning of modern Chinese history, successfully introduced Marxism to China, leading to the founding

of revolutionary newspapers including *Xiangdao*, *New Youth*, *Communist*, and *Chinese Youth*. Thus, modern Chinese newspapers primarily aimed at national salvation, focusing on "enlightenment" and "revolution" [2]. During this period (1840-1949), newspapers featured diverse columns and all-encompassing content, ranging from major domestic and international political news to everyday life of ordinary citizens, with meticulous reporting that vividly captures the panorama of modern Chinese history. These publications played important roles in social transformation and historical turmoil, possessing precious historical, academic, and corrective value for historical facts [3].

In recent years, a new round of technological revolution has flourished, with digital technologies continuously updating. Document digitization has become an inevitable historical trend, bringing new development opportunities to modern newspaper resources. However, due to their age, these newspapers suffer from aging, acidification, fragility, and severe damage, facing critical preservation crises. To rescue and protect modern newspaper resources and promote excellent traditional culture, academic organizations and commercial institutions have employed modern information technology to achieve long-term digital preservation and utilization. For example, cultural institutions led by the National Library of China launched the "Revolutionary Documents and Republican Period Documents Protection Plan," encompassing large quantities of modern newspaper documents. This project received high-level government attention, being included in the 13th Five-Year Plan for National Economic and Social Development of the People's Republic of China in 2016 and written into the Ministry of Culture's "13th Five-Year Plan for Cultural Development and Reform" and the "National Cultural Development and Reform Plan for the 13th Five-Year Period" in 2017. Other national libraries have also undertaken historical newspaper digitization activities through compilation of reference books, construction of specialized newspaper databases, photocopying, microfilm reproduction, and digital processing, providing users with browsing and retrieval functions. Projects such as the Library of Congress Digital Newspaper Program, European Newspaper Digital Project, Finnish Library Digital Newspaper Project, and Australian National Library Digital Newspaper Project have advanced the digital construction process and laid foundations for efficient development and utilization.

However, through investigation of domestic modern newspaper resource databases and index libraries, this study finds that these databases lack unified and standardized newspaper semantic description models as support. They often simply pile up and display text and image resources, with single retrieval methods across multiple databases. They lack deep revelation of newspaper content and multidimensional semantic relationship mining and organization, remaining at the level of brief descriptions of physical carriers. This severely limits users' possibilities for obtaining fine-grained information, prevents rapid location of target resources, and affects service quality. Therefore, it is necessary to design a comprehensive, standardized, and interoperable fine-grained semantic description model for modern newspaper resources that

extracts structured knowledge to meet users' complex information acquisition and retrieval needs, thereby improving knowledge service efficiency. Against this background, this paper designs a fine-grained semantic description model for modern newspaper resources based on their logical structure, physical layout, and content characteristics, approaching from both domain ontology and metadata description perspectives. Using *Shengjing Times* as a case study, the paper aims to provide foundational support for modern newspaper resource database construction, standardized newspaper management, and application system development, promoting the development, utilization, and sharing of modern newspaper resources.

2 Related Research

Literature review reveals that scholars primarily focus on four aspects: newspaper digitization project construction, newspaper rescue and long-term preservation, data quality inspection during digitization, and newspaper resource knowledge organization.

2.1 Newspaper Digitization Project Construction

P. Tonijala et al. comprehensively introduced the U.S. National Digital Newspaper Program and proposed embedding newspaper resources into education and teaching processes [4]. R. Atanassova et al. analyzed the website construction history and functional modules of the European Library digital newspaper project [5-6] to meet digital humanities researchers' knowledge needs. Domestic scholars primarily used modern newspapers collected by the National Library, Peking University Library, Capital Library, Shanghai Library, or regional libraries as examples to analyze current status from perspectives of digitized newspaper varieties, quantities, types, time ranges, content selection, retrieval and reading functions, and service methods [7], proposing corresponding solutions and recommendations [8].

2.2 Newspaper Rescue and Long-Term Preservation

A. Krahmer used the collaborative project between the University of North Texas and Stanford University, The Texas Digital Newspaper Program (TDNP), as an example to elaborate on digital preservation strategies [9]. M. Georgieva discussed newspaper rescue and long-term preservation strategies from a project management perspective, using the Nevada Digital Newspaper Project as a case study [10]. Domestic scholars explored digitization technologies and tools for local modern newspapers, analyzed the necessity and advantages of digitization, and proposed relevant recommendations [11-12].

2.3 Newspaper Data Quality Inspection

J. Jarlbrink et al. analyzed digital noise issues in the Swedish National Library's historical newspaper digitization process [13], such as inconsistent Optical Char-

acter Recognition (OCR) quality, loss of value during carrier format conversion, and quality control risks in digital outsourcing. Digital noise is a focal issue in newspaper digitization, directly affecting resource development and utilization. Domestic scholars discussed quality inspection issues in Republican newspaper digitization practice, including newspaper format recognition, OCR text recognition, record identification numbers, newspaper titles, publication dates, editions, and columns [14].

2.4 Newspaper Knowledge Organization

Scholars primarily discuss metadata description standards for newspaper resources. In digitization practice projects, bibliographic metadata standards are generally used to roughly define element characteristics. For example, the Library of Congress Digital Newspaper Program uses the Metadata Object Description Schema (MODS) in METS documents [15]; the Finnish National Library's historical newspaper digitization project primarily references DC standards to describe elements such as newspaper title, publisher, and publication date [16]; the National Library of China uses MARC format for Republican digital newspaper description, mainly recording content features, carrier forms, and source information [17]. These approaches primarily reference mature metadata standards and roughly reveal newspaper resource features in practice, lacking deep mining and indexing of content, with single metadata descriptions across databases that fail to comprehensively describe semantic features and relationships.

In theoretical research on newspaper resource metadata description, scholars discuss metadata description analysis, metadata-assisted user interactive retrieval [18], identification of user retrieval patterns, and localized metadata standards [19]. For example, J.H. Rho conducted detailed metadata element design and application for the Korean colonial newspaper *Chosun Ilbo*, delving into newspaper knowledge content units, analyzing newspaper attributes from article and advertisement metadata, and designing metadata standards [20] to achieve localization and promote long-term preservation. P. Fafalios used *The New York Times* (1987-2007) as a data source to construct Resource Description Framework (RDF) graphs using archival description metadata and semantic information, attempting to solve semantic information retrieval issues for newspaper archival resources [21]. T. Bogaard et al. used log analysis to explore the utility of Dutch National Library historical newspaper metadata in user search behavior, identifying search patterns [22]. Domestic scholars primarily discuss historical newspaper digitization, preservation strategies, and database construction, lacking deep semantic description and organization, with scarce literature results. Representative works include Ding Xiaolei et al., who roughly designed newspaper metadata at the edition and article levels referencing DC standards [23]; and Wang Jing et al., who focused on metadata cataloging rules for three resource types (main text, advertisements, images) and analyzed the functional construction of the *Shibao* database [24]. These two studies briefly revealed

formal features but lacked semantic depth.

In summary, domestic research on modern newspapers is limited, especially scarce in library and information science regarding newspaper knowledge organization, with insufficient depth. Under the impact of new liberal arts construction and digital humanities waves, modern newspaper resource knowledge organization deserves academic attention. Library and information science should leverage its strengths to conduct comprehensive semantic description and revelation, fully realizing newspapers' documentary and historical value. To fill this research gap, this paper deeply considers newspaper resource characteristics and constructs a comprehensive fine-grained semantic description model from both ontology and metadata perspectives, using *Shengjing Times* as a case study and employing RDF/XML for resource interoperability and practical application, promoting efficient organization and utilization of modern newspaper resources and enhancing knowledge organization capabilities and service levels.

3 Semantic Description Methods

Current academic semantic description methods for knowledge mainly include two types: metadata standards and domain ontology models. Mature metadata standards and domain ontologies can provide references for constructing modern newspaper resource semantic description models. By reviewing metadata standards and domain ontologies with similar properties to modern newspaper resources, appropriate components can be extracted and reused to build the model.

3.1 Metadata Standards

Metadata, known as "data about data," abstracts basic data to higher dimensions and levels, consisting of elements, qualifiers, and attributes. It can describe content attributes and features of digital information resources, forming a standardized data description system for effective management, organization, and retrieval. This study reviews metadata standards applicable to modern newspaper resource description: MARC, DC, EAD, MODS, CADAL, and *Ancient Books Metadata Specification* (see Table 1).

Although these metadata standards differ in elements and qualifiers, most describe resources from content attributes and external structure. DC and MODS have broad applicability to various network information resources; EAD primarily describes archives and manuscripts, including textual documents, electronic documents, visual materials, and audio recordings [25-27], with high-level elements comprising EAD header, archival description, and front matter. MARC is mainly used for library bibliographic data description. CADAL formulated newspaper metadata cataloging specifications based on DC, reusing 15 DC elements and adding 2 custom elements (edition information and MARC records), resulting in relatively coarse granularity. CDWA, primarily for artworks and collections, contains 540 elements including classification, name, creator, time,

location, and related works, providing very comprehensive description. The *Ancient Books Metadata Specification* formulated by the State Administration of Cultural Heritage references CDWA and customizes some elements, including 23 elements with relatively detailed description.

3.2 Ontology Models

Ontology originated in philosophy as an abstract summary of things in the objective world. The knowledge engineering field borrowed this concept, redefining it as a formal and explicit specification of a shared conceptual model for describing concepts, attributes, and relationships. R. Studer et al. [34] define ontology as "a formal, explicit specification of a shared conceptual model." Both metadata and ontology are structured description methods for information resources. Metadata primarily explains physical characteristics and forms of information resources for effective management and retrieval, while ontology focuses on knowledge description, revealing content information such as entities (people, events, places, times, objects) and implicit relationships between concepts. Metadata has a resource-centered radial structure, whereas ontology has a decentralized three-dimensional network structure, with metadata elements serving as properties of concepts in ontology [35]. Common ontology models include FRBR, BIBFRAME, CIDOC_{CRM}, and FOAF. FRBR is a functional requirements framework for bibliographic records based on an "entity-relationship" model. BIBFRAME simplifies the FRBR model, summarizing three entity groups to achieve linked bibliographic data. FOAF is an RDF vocabulary following W3C standards for describing social networks between people. CIDOC_{CRM} uses object-oriented methods to define entities (concepts) and properties (relationships) in the cultural heritage domain, becoming an international standard (ISO 21127:2014) in 2014, with its May 2021 version including 81 classes and 160 properties. CRM has rich entity types and, besides describing cultural heritage resources, applies to other information resource types related to cultural relics. Therefore, the CRM ontology model is also suitable for modern newspaper resource description modeling.

4 Fine-Grained Semantic Description Model for Modern Newspaper Resources

4.1 Feature Analysis of Modern Newspaper Resources

Designing a semantic description model for modern newspaper resources requires comprehensive consideration of logical structure, physical layout, and content information. This paper uses the renowned modern newspaper *Shengjing Times* as an example to analyze relevant features and provide reference for model design.

Shengjing Times was a Chinese-language newspaper founded by Japanese national Nakajima Masao in Shenyang on October 18, 1906, distributed throughout Northeast China, North China, some southern cities, and even Southeast

Asian Chinese-speaking countries [36], ceasing publication in 1944. The newspaper focused on domestic current affairs and commentary, gathering extensive information on Northeast China's finance, commerce, transportation, education, and literature. It is an invaluable historical resource for studying Northeast Chinese military resistance against Japanese aggression, Beiyang warlord history, and modern Chinese history.

Figure 1 [Figure 1: see original paper] shows the content of *Shengjing Times* from October 25, 1906. The overall feature information includes newspaper title, format, volume number, issue number, page layout, and columns. Figures 1(a) and 1(d) primarily contain "advertisements," such as for Yokohama Specie Bank, Mitsui & Co., and Yanshou Pharmacy, with rich and diverse content. Figures 1(b) and 1(c) focus on "main text," with columns including editorials, Beijing news, Northeast China news, international news, special telegrams, official documents, 市井杂俎 (miscellaneous notes), and vernacular sections, all describing content involving entities such as "people, events, places, times, institutions, and officials," with diverse types including social news, political news, and literary fiction.

To ensure model applicability, besides investigating *Shengjing Times*, the authors examined hundreds of newspapers in the National Library of China's "Chinese Historical Documents Database: Modern Newspapers Database," summarizing overall formal features (see Table 2 element column) and concluding that modern newspapers primarily consist of "main text" and "advertisements" in logical structure. Therefore, model design must consider both physical carrier features and logical structure information to comprehensively deconstruct modern newspaper resource features and extract structured information.

4.2 Model Design Approach

The modern newspaper resource semantic description model must both identify and associate content-related entities (people, times, places, events, institutions, officials recorded in newspapers) and conduct fine-grained mining of descriptive information at the physical level according to logical structure, thereby fully describing overall resource features. This paper uses *Shengjing Times* as a case study to clearly demonstrate knowledge elements and intrinsic semantic relationships.

First, determine model entities and relationships. Through extensive investigation, relevant entity types are extracted, with entities as nodes and predicates as connections to build associations between modern newspaper resource entities and draw model diagrams.

Second, define model descriptive information, including descriptive metadata or administrative metadata. Both global description of modern newspaper resources and local description of content information according to logical architecture are required. Based on extracted attribute information, mature metadata standards are reused with some custom properties.

Finally, form a complete, high-quality, interoperable, and highly specific modern newspaper resource fine-grained semantic description model to achieve resource sharing and promote description, location, retrieval, and organization.

4.3 Model Design Process

4.3.1 Determining Model Entities and Relationships Entity relationships are determined based on entity types to establish clear and accurate associations. Using the event "Supervising the Yongding River" recorded in *Shengjing Times* as an example, this section constructs an entity-relationship diagram (see Figure 2 [Figure 2: see original paper]). In CRM, E5 represents the event entity, with participant E21 "Meng Xianyi," occurrence time E52 "1916," location E53 "Yongqing County," participating institution "Shunzhi Relief Bureau," and participating official "Yongding River Supervisor." Additionally, for person entity E21 "Meng Xianyi," there are subclass entities E67 "Meng Xianyi's birth year" and E69 "Meng Xianyi's death year," associated with entities E52 (time) and E53 (place). The person's official position was "prefect" at institution "Changchun Prefecture," though these changed across historical periods. Thus, entities of person, event, place, time, institution, and official in *Shengjing Times* form a network structure, with semantic relationships revealed to achieve knowledge organization and association analysis.

4.3.2 Defining Model Descriptive Information The entity-relationship model targets semantic content levels, while this section defines descriptive information from physical feature perspectives to build a complete fine-grained semantic description model. Through reviewing *Shengjing Times*, the authors summarized its overall feature information (see Table 2).

(1) Reused Metadata Standards Explanation. Reusing metadata requires full consideration of modern newspaper resource features to select appropriate standards. Modern newspapers differ from contemporary newspapers, especially in physical form and published content. Physical carriers primarily use oil, lead, and stone printing, while contemporary newspapers mostly appear as digital text. Published content focuses on modern China's semi-colonial and semi-feudal history and heroic resistance, reflecting historical changes and possessing significant archival and cultural relic value. Therefore, metadata description must consider these attributes.

This paper discusses modern newspaper resource description specifications from three dimensions: general resource metadata standards, archival resource metadata standards, and cultural relic resource metadata standards. DC, as an international general standard, has simplicity, flexibility, and compatibility, applicable to broad network information resources and modern newspapers, though with poor specificity, serving as a supplementary framework. EAD, as an archival metadata standard, describes archives and manuscripts and is suitable for modern newspapers as historical archives, with some elements reused for physical form description (provenance, identification numbers). The *An-*

cient Books Metadata Specification is a research outcome of the national science and technology support plan project "Research and Demonstration on Standard System and Key Standards for Digital Protection of Cultural Relics," releasing 62 standards for cultural relics, oracle bones, maps, murals, rubbings, and ancient books, applicable to ancient book and cultural relic resource description. Modern newspapers possess cultural relic attributes, with consistent description elements for creation, publication, material dimensions, digital objects, and holding institutions, making it suitable for reuse.

(2) Defining Model Global Descriptive Information. The description model includes 21 elements: 8 from *Ancient Books Metadata Specification* (denoted as sach), 8 from DC (denoted as dc), 2 from EAD (denoted as ead), and 3 custom elements (newspaper, denoted as np). See Table 2 for details.

Using *Shengjing Times* as an example, Figure 3 [Figure 3: see original paper] illustrates the global descriptive information.

(3) Defining Model Local Content Descriptive Information. Based on Section 3.1, modern newspaper resources primarily consist of "main text" and "advertisements." Therefore, content description must address both aspects, reusing relevant metadata standards and customizing some elements. See Table 3 for details.

Using *Shengjing Times* as an example, Figure 4 [Figure 4: see original paper] provides a schematic diagram of main text and advertisement information based on Table 3.

In summary, the modern newspaper resource semantic description model deconstructs resource features from a fine-grained knowledge element perspective, from internal semantic associations to external logical associations. It extracts six entity types (person, time, place, event, institution, official), builds relationships between them using CIDOC-CRM, extracts global and local descriptive information, analyzes content features from "main text" and "advertisement" dimensions, reuses *Ancient Books Metadata Specification*, EAD, and DC standards, and ultimately forms a clear and complete fine-grained semantic description model.

5 Application of the Fine-Grained Semantic Description Model for Modern Newspaper Resources

5.1 XML-Based Model Application

To apply the model to specific resource management and storage, achieving interoperability, data exchange, and resource sharing, this paper uses XML (Extensible Markup Language) for description. XML, recommended by W3C, is an extensible editing language with simplicity, high extensibility, strong interoperability, and convenient network transmission [37], providing flexible markup extension mechanisms. This paper places the semantic description model in XML,

using Oxygen XML editing tools to create XSD format documents. Oxygen XML Editor integrates XML viewing and editing functions, providing tools for XML creation and development with automatic code validation, tag detection, syntax highlighting, and support for all XML standards with high extensibility.

This paper uses Oxygen XML software to edit the description model, building Element and ComplexType, introducing DC, EAD, sach standards, customizing np standard, and creating namespaces. Figure 5 [Figure 5: see original paper] shows the operation interface.

5.2 RDF/XML-Based Model Application

Building on Section 4.1, this paper further adopts Resource Description Framework to encapsulate the description model. RDF, recommended by W3C based on XML, is a standard for describing network resources, encoding structured metadata for data exchange and reuse, providing an operable carrier and container for metadata [38]. RDF uses XML as a common syntax structure, providing clear semantic expression methods for adding structural constraints to XML [39]. RDF treats resources as objects identified by Uniform Resource Identifiers (URI) and uses triples (subject, predicate, object) to represent resource objects. Combining XML and RDF leverages their respective advantages for semantic description. RDF can reference different metadata schemes, encapsulating multiple metadata in a unified description model to achieve interoperability. RDF's normative syntax represents each statement with an rdf:Description element, using rdf:about attribute values to declare subject URI references.

This paper uses the W3C RDF Validator [40] to verify RDF documents, with results shown in Table 4 . The validator automatically generates RDF triples (subject, predicate, object) for *Shengjing Times*. For example, the subject is *Shengjing Times*, the predicate is resource description property, and the object is property value. RDF encapsulates different metadata using flexible triples <resource, property, property value> to describe newspaper resources, with all resources identified by unique URIs for structured presentation. This section presents only the resource property triples; future work will enrich the ontology description model using deep learning algorithms to identify entities (person, event, place, institution, official) in *Shengjing Times*, store them in relational databases, set entity relationships through foreign keys, use D2RQ tools to convert RDB relational data to RDF format, store them in Virtuoso database, and employ SPARQL for retrieval, achieving interconnected sharing and extended markup language for RDF models.

Conclusion

Modern newspaper resources contain rich historical information urgently requiring knowledge organization techniques for development. This paper analyzes physical layout, logical structure, and content information of modern newspaper resources, using *Shengjing Times* as a case study to construct a fine-grained

semantic description model from ontology and metadata dimensions, and applies the model using RDF/XML language. In ontology semantic description, CRM and other ontology conceptual models express entities and relationships in *Shengjing Times*, using the "Supervising the Yongding River" event as an example to construct semantic relationships among person, time, place, event, institution, and official entities. In metadata description, combining modern newspaper resource features, the study reuses *Ancient Books Metadata Specification*, EAD, and DC standards to construct semantic association diagrams for global and local descriptive information of newspaper resources.

The constructed fine-grained semantic description model for modern newspaper resources has theoretical and practical value. Theoretically, it broadens the application fields of ontology and metadata, applying knowledge organization theory to modern newspaper research, using metadata to reveal physical carrier features, locate resources, and achieve rapid navigation, discovery, and multidimensional semantic retrieval, while using ontology to parse hidden entities and construct interconnected semantic networks. Practically, the model applies to other modern newspaper resources and provides reference paths for addressing deficiencies in current modern newspaper databases regarding resource description, retrieval, and semantic services, promoting standardized management and refined services, improving utilization efficiency, fully realizing historical and documentary value, and inheriting social memory to develop excellent traditional Chinese culture. Future work will employ deep learning models to identify different entity types in newspapers to enrich instance content.

References

[1] Zhang Ke. Historical trajectory of newspaper origins, journalist birth, and evolution [J]. Shaanxi Archives, 2018(3): 24-26. [2] Yang Wen, Peng Junling. Accumulation and development of modern Chinese newspapers from the perspective of publishing cultural heritage protection [J]. Library Journal, 2015, 34(10): 78-84. [3] Kong Zhengyi. Analysis of the historical document value of modern Chinese newspapers [J]. Journal of Anhui University of Science and Technology (Social Science Edition), 2009, 11(4): 105-108. [4] Voices and viewpoints in chronicling America: uses of historical newspapers for education and outreach [EB/OL]. [2021-10-01]. http://library.ifla.org/id/eprint/1038/ [5] Neudecker C, Antonacopoulou A. Making Europe's historical newspapers searchable [EB/OL]. [2021-10-01]. http://library.ifla.org/id/eprint/1271/1/080-penn-en.pdf. [6] Improving the discovery of European historical newspapers [EB/OL]. [2021-10-01]. http://library.ifla.org/id/eprint/170-atnassova-en.pdf. [7] Ding Xiaolei. Exploration and practice of digitizing Republican local literature newspapers: a case study of Capital Library's Republican newspaper digitization [J]. Henan Library Science, 2016, 36(12): 98-100. [8] Duan Xiaolin. Research on the development status of Republican document databases [J]. Library Science Research, 2016(20): 42-45. [9] Krahmer A. Digital newspaper preservation through collaboration [J]. Digital library perspectives, 2016,

32(2): 73-87. [10] Georgieva M. Successful management of an outsourced large-scale digitization newspaper project: tips for effective collaboration, increased productivity, and outstanding deliverables [J]. Journal of archival organization, 2019, 16(1): 52-74. [11] Jin Caihong. Digitization of Republican newspaper microfilm in Sichuan: a case study of Sichuan University Library [J]. Journal of Sichuan Library Science, 2016(3): 17-19. [12] Chen Guixiang. Discussion on Republican newspaper digitization: a case study of Chongqing Library [J]. Sci-Tech Information Development & Economy, 2013, 23(4): 27-29. [13] Jarlbrink J, Snickars P. Cultural heritage as digital noise: nineteenth-century newspapers in the digital archive [J]. Journal of documentation, 2017, 73(6): 1228-1243. [14] Xiao Hong, Huai Yan. Analysis of quality inspection issues in Republican newspaper digitization practice [J]. Library Science Research, 2017(7): 61-78, 87. [15] Murray R L. Toward a metadata standard for digitized historical newspapers [C]//Proceedings of the 5th ACM/IEEE-CS joint conference on digital libraries. Denver: ACM Press, 2005: 330-331. [16] Ahonen E, Hyvonen E. Publishing historical texts on the semantic web: a case study [C]//2009 IEEE international conference on semantic computing. Berkeley: IEEE, 2009: 167-173. [17] Zhang Wei. Research on common problems in Republican newspaper digital acceptance: a case study of the National Library [J]. Library and Information Studies, 2019, 12(3): 72-79. [18] Allen R B, Schalow J. Metadata and data structures for the historical newspaper digital library [C]//Proceedings of the eighth international conference on information and knowledge management. Kansas: Association for Computing Machinery, 1999: 147-153. [19] Seifi L, Ahmadzadeh N, Pordel F. Digital preservation of old Persian periodicals in Iran with special reference to Iranian newspapers: strategies and challenges [C]//2015 4th international symposium on emerging trends and technologies in libraries and information services. Noida: IEEE, 2015: 81-85. [20] Rho J H. Metadata elements design and application for Japanese colonial newspaper 'Chosun Ilbo' issued in colonial Korea [J]. Journal of Korean Library and Information Science Society, 2019, 50(4): 137-158. [21] Fafalios P, Kasturia V, Nejdl W. Ranking archived documents for structured queries on semantic layers [C]//Proceedings of the 18th ACM/IEEE on joint conference on digital libraries. Fort Worth: Association for Computing Machinery, 2018: 155-164. [22] Bogaard T, Hollink L, Wielmaker J, et al. Metadata categorization for identifying search patterns in a digital library [J]. Journal of documentation, 2018, 75(2): 270-286. [23] Ding Xiaolei. Exploration and practice of digitizing Republican local literature newspapers: a case study of Capital Library's Republican newspaper digitization [J]. Henan Library Science, 2016, 36(12): 98-100. [24] Wang Jing, Shen Lili. Research on construction of Chinese historical newspaper databases: a case study of *Shibao* database [J]. Henan Library Science, 2018, 38(10): 106-108. [25] Feng Xiangyun, Xiao Long, Liao Sansan, et al. Comparative study of commonly used foreign metadata standards [J]. Journal of Academic Libraries, 2001(4): 15-21, 91. [26] Zhao Yi. Metadata design for network archival information retrieval [J]. Shanxi Archives, 2020(1): 54-61. [27] Song Xin, Lu Guoxuan. Research on metadata in digital construction of

palm-leaf archives [J]. Zhejiang Archives, 2021(3): 27-30. [28] MARC [EB/OL]. [2021-10-01]. http://www.loc.gov/marc/. [29] DC [EB/OL]. [2021-10-01]. http://dublincore.org/documents/dcmi-terms/. [30] CDWA [EB/OL]. [2021-10-01]. http://www.getty.edu/research/publications/electronic_{publications}/cdwa/. [31] EAD [EB/OL]. [2021-10-01]. http://www.loc.gov/ead/. [32] MODS [EB/OL]. [2021-10-01]. https://www.loc.gov/standards/mods/. [33] CADAL [EB/OL]. [2021-10-32]. http://cadal.edu.cn/index/home#page1. [34] Studer R, Benjamin V R, Fensel D. Knowledge engineering: principles and methods [J]. Data and knowledge engineering, 1998, 25(1/2): 161-197. [35] Xia Cuijuan. Knowledge integration of cultural memory resources: from heterogeneous resource metadata application profile to integrated ontology design [J]. Library and Information Knowledge, 2021(1): 53-65. [36] Guo Guangting. Building characteristic collection databases using TPI system: a case study of *Shengjing Times* database [J]. Henan Library Science, 2009, 29(6): 74-75. [37] Ren Ruijuan, Pu Demin, Miao Junmin, et al. DC metadata description technology based on XML/RDF [J]. Journal of Intelligence, 2002(9): 25-26. [38] RDF [EB/OL]. [2021-10-01]. https://www.w3.org/RDF/. [39] Lu Kui. Research on description and application of digital library information resources based on XML/RDF [D]. Hefei: Hefei University of Technology, 2003. [40] RDF Validator [EB/OL]. [2021-10-20]. https://www.w3.org/RDF/Validator/.

Author Contributions

Sun Shaodan: paper writing and revision; Deng Jun: research concept, paper writing and revision; Chang Yanyu: data collection; Zhang Zishu: assisted with data collection; Shen Yong: paper proofreading.

Design and Application of the Fine-Grained Semantic Description Model of Modern Newspaper Resources: Taking *Shengjing Times* as an Example

Sun Shaodan[1], Deng Jun[1], Chang Yanyu[1], Zhang Zishu[1], Shen Yong[2] [1] School of Business and Management, Jilin University, Changchun 130012 [2] School of Public Health, Jilin University, Changchun 130022

Abstract: [Purpose/Significance] This paper designs a scientific and standardized fine-grained semantic description model for modern newspaper resources, revealing their characteristics and relationships to provide references for effective management, organization, knowledge discovery, and knowledge services. [Method/Process] By analyzing logical structure, physical layout, and content, and reusing CIDOC-CRM, EAD, DC, and *Ancient Books Metadata Specification*, the model is designed using *Shengjing Times* as a case study and implemented in RDF/XML using Oxygen XML for interoperability. [Result/Conclusion] The model provides an operable fine-grained semantic description framework for modern newspaper organization, offering foundational sup-

port for database construction, standardized management, and application development, promoting resource development, utilization, and sharing.

Keywords: modern newspaper resources; semantic description model; *Shengjing Times*; RDF/XML

Note: Figure translations are in progress. See original paper for figures.