Construction of the STKOS Terminology Publishing and Sharing Service Platform

Authors: FU Honghu, Zhang Zhixiong, Liu Jianhua, Qian Li, Wang Ying, Fu Honghu

Date: 2016-01-25T00:00:00+00:00

Abstract

[Objective] To design and implement the STKOS Term Publishing and Sharing Service System. [Background] As a mega-thesaurus, the Scientific and Technological Knowledge Organization System (STKOS) needs to be provided to users to advance knowledge services and promote knowledge sharing, thus requiring its publication through a sharing service platform. [Methods] Based on investigation of relevant international projects and systems, combined with the characteristics and application requirements of STKOS, the system' s functional framework is designed, and key issues in system implementation including application scenarios, data exchange formats, data structures, visualization, and multi-version management are analyzed to complete the overall system construction. [Results] Under the scenario of tens of millions of data entries, the STKOS Term Publishing and Sharing Service System platform is implemented. [Conclusion] This system supports the management and publishing of STKOS data, facilitates the disclosure of knowledge system content, and provides users with browsing, retrieval, and personalized customized downloading of the knowledge organization system.

Full Text

Construction of STKOS Term Publishing and Sharing Service Platform

Fu Honghu[1], Zhang Zhixiong[1], Liu Jianhua[1,2], Qian Li[1,2], Wang Ying[1]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract

[**Objective**] To design and implement the STKOS term publishing and sharing service system. [**Context**] As a metathesaurus, the Science and Technology Knowledge Organization System (STKOS) needs to be made available to users to advance knowledge services and promote knowledge sharing, which requires its publication through a sharing service platform. [**Methods**] Based on an investigation of relevant international projects and systems, and combined with the characteristics and application requirements of STKOS, the functional framework of the system was designed. Key issues in system implementation, including application scenarios, data exchange formats, data structures, visualization, and multi-version management, were analyzed to complete the overall system construction. [**Results**] Under the scenario of ten-million-level data volume, the STKOS term publishing and sharing service platform was implemented. [**Conclusions**] This system supports the management and publication of STKOS data, facilitates the revelation of knowledge system contents, and provides users with browsing, retrieval, and personalized customized download capabilities for the knowledge organization system.

Keywords: Term service; STKOS; Knowledge service
Classification: G353.1

1 Introduction

To further strengthen the construction of China' s scientific and technical information resources and services, consolidate and develop the national scientific and technical literature information assurance system, and more effectively utilize massive amounts of scientific and technical literature information, the National Science and Technology Documentation Center led the organization and implementation of the National Science and Technology Support Plan project "Knowledge Organization System Construction and Demonstration Application for Foreign Language Scientific and Technical Literature Information" [1]. Based on the integration of hundreds of relevant thesauri, standards, and ontologies in foreign scientific and technological fields, the project constructed the Science and Technology Knowledge Organization System (STKOS), forming a metathesaurus. The STKOS publishing service system is the user-facing service system of this project, providing scientific and technical term publishing and sharing services, and is a terminology service system oriented toward end users.

Many foreign institutions have already conducted relevant research and construction of terminology service systems, such as OCLC Terminology Services [2], STAR (Semantic Technologies for Archaeological Resources) [3], JISC-HILT [4], UMLS Terminology Services [5], and VocBench [6]. Existing papers have provided detailed introductions to foreign terminology registries and services [7], and have comprehensively summarized the functions and services of terminology service systems, which will not be elaborated here.

The primary objective of this paper is to design and implement a publishing

and service platform for the STKOS knowledge organization system for the purpose of sharing and reusing STKOS research and construction achievements. Through in-depth investigation of relevant international projects and systems, and based on a summary of their experiences, functional, and technical characteristics, combined with the features and application requirements of STKOS, the system platform design and development were completed. The system design references successful foreign terminology service systems such as UMLS Terminology Services and VocBench. Built upon the established STKOS, the system provides external services including retrieval and query, browsing and navigation, and detailed information display, enabling users to conveniently acquire, access, and utilize the science and technology knowledge organization system. The system also provides STKOS download functionality, allowing the output of STKOS fragments in specific disciplinary domains according to standard formats to promote the sharing and application of STKOS, and offers multiple approaches for browsing and utilizing STKOS and its subsets to meet the needs of different institutions and users for domain knowledge ontologies. The construction of the STKOS term publishing and sharing service platform enables the sharing and publication of cross-disciplinary, ten-million-level scientific and technical metathesaurus, realizes multi-view content revelation of the thesaurus (multi-view result revelation, multi-dimensional visual analysis), provides users with customization and download capabilities, promotes thesaurus content sharing, and achieves the goal of maximizing the sharing and reuse of science and technology knowledge organization system achievements.

2 System Function Framework

The overall construction objective of the STKOS term publishing and sharing service platform is to build STKOS term publishing and sharing services based on the established Science and Technology Knowledge Organization System (STKOS) and the STKOS-based knowledge ontology engine. The platform provides external services including retrieval and query, browsing and navigation, and detailed information display, enabling users to conveniently acquire, access, and utilize the science and technology knowledge organization system. Simultaneously, to promote broader application of STKOS, the platform provides STKOS customization and integration capabilities, including input, output, download, and integration functions for third-party knowledge organization systems. The backend management system of the STKOS term publishing and sharing service platform primarily manages various versions of STKOS, controls user operation permissions, and provides maintenance tools for adding and editing STKOS content.

According to the system construction objectives, the overall framework of the STKOS term publishing and sharing service platform comprises three modules: the frontend service system, the customization and integration system, and the backend management system. The overall system framework is shown in Figure 1 [Figure 1: see original paper].

Figure 1 STKOS Term Publishing and Sharing Service System Framework

(1) The frontend service system primarily interacts with users, providing various explicit knowledge services, including retrieval and query, browsing and navigation, and detailed information display, enabling users to conveniently acquire, access, and utilize the science and technology knowledge organization system. Through the frontend service system, users can obtain required data information.

(2) The customization and integration system supports user-defined knowledge organization system functionality. Users can utilize this module to independently select and customize required knowledge organization systems to achieve personalized services, defining their own required knowledge organization fragments (with knowledge sources from multiple thesauri and multiple categories), and enabling multi-format downloads of customized knowledge organization fragments. Potential application scenarios include: users needing to exclude unwanted thesauri or those unusable in local applications, users requiring a customized subset with multiple data output options and filters, etc.

(3) The backend management system is primarily responsible for user permission management, user operation behavior management, publication and organization management of various versions of knowledge organization systems in the platform, and management of revised content within the knowledge organization systems.

The STKOS term publishing and sharing service system is built upon the underlying STKOS and various source thesauri and knowledge organization engines. The entire service platform will employ account login for permission control when providing external services, distinguishing different services according to user roles. The system data flow is shown in Figure 2 [Figure 2: see original paper].

3 Technical Implementation Details

3.1 Application Scenario Analysis

The STKOS term publishing and sharing service platform provides terminology services based on the scientific and technical metathesaurus. In addition to machine user access interfaces, the platform focuses on the query and access requirements of human users via web browsing. From the perspective of human user application scenarios, users are categorized into two types: general users and information science domain experts. Differentiated service interfaces are provided according to the differences in these two user groups' knowledge backgrounds and usage objectives.

3.2 Data Exchange Format

The term publishing and sharing service platform provides terminology publishing services based on STKOS but does not itself process or produce STKOS data. It needs to receive specific data content from the STKOS compilation working group. The format and quality of the received data directly affect STKOS services. STKOS publishing metadata defines the data reception specification and the specific format (Schema) for STKOS data content received from the STKOS compilation working group. The data reception module parses and imports the received data into the term publishing and sharing service platform according to the metadata defined by the data reception Schema.

STKOS publishing metadata defines six major categories of basic data models [8], including source terms, source thesauri, scientific and technical terms, STKOS standard concepts, category classes, and category tables, and provides detailed descriptions of the relationships among these basic data types, as shown in Figure 3 [Figure 3: see original paper].

Figure 3 Basic Structure of Publishing Metadata Schema [8]

3.3 Basic Information Data Structure

STKOS data integrates hundreds of source thesauri in the scientific and technological fields, containing both data information from source thesauri and integrated/processed data information. It is characterized by diverse data sources, large data volume, presence of upper-level categories, and multiple types of data relationships.

The design of the data structure must consider the inherent characteristics of STKOS data, the requirements of term publishing services, and must comply with relevant standards and specifications. Therefore, the design primarily references the ISO 25964-1 data model, STKOS publishing metadata, and the UMLS data model. The ISO 25964-1 data model defines thesaurus-related standards and specifications and provides data model recommendations. STKOS publishing metadata provides the STKOS data model, whose design primarily references ISO 25964-1 and UMLS data models. UMLS integrates over a hundred thesauri in the medical domain, and its data shares many similarities with STKOS in content and form, making it highly referential.

The design approach for the data model is to reference the ISO 25964-1 data model, follow STKOS publishing metadata specifications, draw on relevant design concepts from UMLS, and allow for appropriate data redundancy. The STKOS publishing metadata specification defines three levels of data types: source terms, basic terms, and concepts. The relationships between terms and concepts are primarily built upon source terms, reflecting relationships among source terms, while term-concept relationships are mainly established between concepts and source terms.

STKOS publishing metadata defines six major categories of basic data models,

including source terms, source thesauri, scientific and technical terms, STKOS standard concepts, category classes, and category tables. On this basis, combined with the practical application requirements of the publishing service system, the design focuses on a three-layer data structure of standard concepts–scientific and technical terms–source terms, as well as various concept and term relationship data.

3.4 Data Visualization

Multiple types of semantic relationships exist in STKOS, such as term–term relationships, term–concept relationships, term–category relationships, concept–concept relationships, concept–category relationships, and category–category relationships. In STKOS, concepts and terms constitute the main body of knowledge, while relationships between concepts are primarily manifested through relationships between terms. Visualization revelation in knowledge organization systems is an important means of displaying internal relationships within knowledge organizations.

STKOS comprises multiple thesaurus systems with high-level category specifications, featuring characteristics such as inter-thesaurus concept mapping, synonyms, preferred terms, and various relationship types. Therefore, effectively revealing the various relationships between a concept and other related concepts is a complex problem.

Based on knowledge organization system visualization and combined with STKOS characteristics, a multi-dimensional semantic visualization revelation model for STKOS is constructed to deeply reveal data relationships from multiple dimensions, including term-intrinsic relationships, associative relationships, hierarchical relationships, semantic relationships, and evolutionary relationships. This multi-dimensional STKOS concept relationship visualization method can effectively reveal the relationships between a concept and other concepts in STKOS, thereby revealing the rich logical associations inherent in the entire knowledge organization system. It enables users to deeply understand the position of a particular concept within the entire STKOS through various visualized relationships, helping users comprehend and utilize the entire knowledge organization system. Cytoscape Web [9] and D3.js [10] were selected as the foundational visualization tools and were further developed through secondary development.

3.5 Data Version Management

The publication of multi-version data forms the foundation for data version evolution analysis and version update notifications in STKOS. Version management performs operations such as import, browsing, archiving, and export on various different types of STKOS versions, providing support for users to apply STKOS data across versions. Through STKOS version management, it is possible to trace the evolution of STKOS across different periods and versions,

and to promptly reflect modifications and changes within the current STKOS version, providing cross-version data interfaces for various functional modules of the STKOS publishing service system.

The design primarily references the UMLS version history change record file format and the latest thesaurus change table in the VocBench agricultural thesaurus management platform of the Food and Agriculture Organization (FAO) of the United Nations. The UMLS version history change record file reflects major changes in the official UMLS version compared to the previous version, while VocBench records thesaurus revisions within the same version through the latest thesaurus change table. The combination of these two approaches can precisely satisfy the requirements of the STKOS backend management system to both preserve officially released multi-version data and promptly reflect the latest revision information in the current version.

The design requires complete archival recording of information for each STKOS historical version as long-term preserved original records that cannot be revised. The current service version is based on the most recent historical version, with timely additions of the latest revision information from the processing platform for the current period. The change record table should be able to comprehensively reflect various information changes between historical versions and within the currently active service version.

3.6 Multi-core Index Construction

STKOS publishing services are characterized by large data volume, complex relationships, and multi-dimensional retrieval. To meet the system' s rapid response requirements, the indexing system requires optimization. The system index is built on Solr.

Based on analysis of data query relationships, two categories comprising five types of indexes were designed to support multi-angle retrieval and querying of concepts: basic data indexes (including concept indexes and term indexes) and relationship indexes (including associative relationship indexes, hierarchical relationship indexes, and co-occurrence relationship indexes). Basic data indexes include categories, concepts, terms, atoms, attributes, and other basic information.

4 System Features

4.1 Retrieval and Results Display

Retrieval is the core module of the frontend service system. The primary retrieval targets are concept terms. Users can select specific retrieval methods as needed and set retrieval conditions to query required information. The system provides associative retrieval, i.e., retrieving concept terms that have specific relationships with a designated term or concept. It also provides enhanced or expanded retrieval, offering prompts for other concepts that have semantic

associations with the search term through term relationships and related information.

Users can limit retrieval conditions from multiple aspects, and the system provides retrieval suggestion functionality. Search terms are highlighted in retrieval results, and update prompts are displayed for recently updated entries. For retrieval result sets, three display forms are provided: list view, category faceting, and source thesaurus faceting. For detailed retrieval result information, multiple views are provided, including text view, RDF view, and graphical view. The system retrieval result detail display interface is shown in Figure 4 [Figure 4: see original paper].

Figure 4 Retrieval Result Details

4.2 Multi-dimensional Visualization of Retrieval Results

Multi-dimensional visualization of retrieval results enables users to comprehensively understand the retrieval target and its position within the entire knowledge system from multiple dimensions. Visualization results are implemented across dimensions including term-intrinsic relationships (i.e., important attribute relationships of the terms themselves), associative relationships, hierarchical relationships, semantic relationships, and evolutionary/merging relationships. Figure 5 [Figure 5: see original paper] shows the multi-dimensional visualization result presentation interface. Users can switch between different dimensional views by selecting different tabs.

Figure 5 Multi-dimensional Visualization Results

4.3 Browsing and Personalized Customization

The system provides browsing by category system. Categories reflect the disciplinary domain structure of the entire knowledge organization system. Through category browsing and display, users can quickly locate their areas of interest and understand the overall knowledge structure and summary information of knowledge content in those domains. Furthermore, to facilitate user convenience, the system provides subset customization based on category browsing. After customization, a subset of concepts, terms, and related information within the selected scope is generated and made available for user download. Figure 6 [Figure 6: see original paper] shows the user interface for category browsing and subset customization. Users can select content of interest for subset customization through category navigation and initial letter navigation.

Figure 6 Category Browsing and Customization

5 Discussion and Future Considerations

The STKOS term publishing and sharing service platform is an important platform for realizing the sharing of STKOS construction achievements and holds sig-

nificant importance for knowledge organization and knowledge services. Based on specific issues encountered during the system construction process, several key aspects should be noted when building terminology services:

(1) **Data Exchange Format.** Data exchange format includes two aspects: the format for data received by the terminology service system and the format for data publication. The design of the received data format must be based on the thesaurus processing system data format, while taking into account both the requirements of the publishing service system for data content revelation and efficiency demands, achieving both standardization and flexibility. The published data format must consider the needs of users and third-party systems, as well as the mapping relationships between data content and other knowledge organization systems, and should follow data exchange standards.

(2) **Underlying Data Model.** Since STKOS integrates hundreds of thesauri in the scientific and technological fields with diverse data sources, and has constructed an upper-level category system and performed term concept merging processes, the underlying data model must be able to completely preserve all types of information content to ensure comprehensive revelation of various knowledge contents.

(3) **Version Management.** As knowledge domains develop, relevant terms and concepts also evolve. Terminology service systems built upon knowledge organization systems must fully consider future compatibility. Multi-version management of knowledge organization system content provides the data foundation for data version evolution analysis and version update notifications.

(4) **Visualization.** Visualization is an intuitive form that helps users understand and recognize various data objects and their relationships within knowledge organization systems, and is an important means of displaying internal relationships within knowledge organizations. Visualization forms should conform to human cognitive habits, be easy to understand, and thereby effectively reveal various relationships between a concept and other related concepts.

Furthermore, the current terminology service primarily uses the standardized STKOS metathesaurus as its main content. The STKOS metathesaurus is a controlled vocabulary; however, in actual services, uncontrolled vocabularies are of great significance to users. How to establish associations between controlled and uncontrolled vocabularies to realize corresponding services is also an important issue that needs to be considered in future system construction.

References

[1] Construction of STKOS Sharing Service Platform [EB/OL]. [2015-03-03]. http://sharestkos.las.ac.cn/. (in Chinese)

[2] Terminology Services [EB/OL]. [2015-03-03]. http://www.oclc.org/research/projects/termservices/.

[3] Semantic Technologies for Archaeological Resources [EB/OL]. [2015-03-03]. http://hypermedia.research.southwales.ac.uk/kos/star/.

[4] JISC-HILT [EB/OL]. [2015-03-03]. http://pure.strath.ac.uk/portal/files/384452/strathprints014046.pdf.

[5] Unified Medical Language System [EB/OL]. [2015-03-03]. http://www.nlm.nih.gov/pubs/factsheets/umls.ht

[6] VocBench [EB/OL]. [2015-03-03]. http://aims.fao.org/vest-registry/tools/vocbench-2.

[7] Ou Shiyan. A Review of Foreign Terminology Registries and Terminology Services [J]. Journal of Library Science in China, 2014, 40(5): 110-126. (in Chinese)

[8] Song Wen. Design of STKOS Publishing Metadata [EB/OL]. [2015-03-03]. http://168.160.16.186/conference/dome_ch/2012/downloads/pdf 发言稿/宋文 _STKOS 发布元数据设计.pdf. (in Chinese)

[9] Cytoscape Web [EB/OL]. [2015-03-03]. http://www.cytoscape.org/index.html.

[10] D3.js [OL]. [2015-03-03]. http://d3js.org/.

Author Contributions

Fu Honghu: Detailed system design, data analysis, partial system development, paper writing;
Zhang Zhixiong: System framework design;
Liu Jianhua: Preliminary research, system framework design;
Qian Li: Detailed system design, system development;
Wang Ying: Data analysis, data exchange format design.

Dates

Received: March 3, 2015
Revised: March 31, 2015

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