NSTL' s Practice and Exploration of Integrating and Utilizing Third-Party Source Metadata: Postprint

Authors: Yu Qianqian, Zhang Jianyong

Date: 2017-10-11T00:00:00+00:00

Abstract

【目的】将 WOS、Scopus 等第三方来源元数据应用到 NSTL 加工系统中。【应用背景】根据 NSTL 发展规划，需要从单纯自加工扩展到加工以及协商获取、购买第三方元数据等多渠道建设元数据方式。【方法】以 NSTL 加工规范为基础，实现与 WOS、Scopus 元数据的映射，分析第三方元数据特点对 NSTL 加工规范进行局部修订并映射，根据映射结果，将第三方元数据以 NSTL 加工规范格式输出并集成到 NSTL 加工系统中。【结果】实现第三方来源元数据快速、高效、低成本地集成整合到 NSTL 加工系统。【结论】WOS 元数据在 NSTL 加工系统中的应用，可以提高 NSTL 文献数据加工速度。有针对性地对现有元数据加工规范进行修订，为后续增加其他第三方资源构建了拓展框架。

Full Text

Preamble

ChinaXiv Cooperative Journal
Total Issue No. 266, 2016, Issue 1
Practices of NSTL Integrating and Utilizing Third-Party Metadata
Yu Qianqian, Zhang Jianyong
(National Science Library, Chinese Academy of Sciences, Beijing 100190)

Abstract

[**Objective**] To apply third-party source metadata such as Web of Science (WOS) and Scopus to the NSTL processing system. [**Context**] According to the NSTL Development Plan, it is necessary to expand from self-processing alone to multi-channel metadata construction methods, including negotiation, acquisition, exchange, and purchase of third-party metadata. [**Methods**] Based on the NSTL processing specifications, we implemented metadata mapping with

WOS and Scopus, analyzed the characteristics of third-party metadata to partially revise the NSTL processing specifications, and according to the mapping results, exported third-party metadata in NSTL specification format and integrated it into the NSTL processing system. [**Results**] Achieved rapid, efficient, and low-cost integration of third-party source metadata into the NSTL processing system. [**Conclusion**] The application of WOS metadata in the NSTL processing system can improve the speed of NSTL literature data processing. Targeted revisions to existing metadata processing specifications have built an expandable framework for subsequently adding other third-party resources.

Keywords: Web of Science; Scopus; NSTL; Metadata Mapping
Classification Number: G250.7

Introduction

The National Science and Technology Library (NSTL) "13th Five-Year" Development Plan proposes optimizing the national scientific literature resource guarantee system and expanding metadata resource acquisition methods. To this end, it is necessary to integrate, consolidate, and utilize third-party source metadata, expanding from sole self-processing to multi-channel metadata resource construction methods including negotiation, acquisition, exchange, donation, deposit, and purchase with domestic and foreign publishers and relevant information institutions. Therefore, based on the literature resource metadata processing specifications adopted by NSTL [1], we must thoroughly analyze the type characteristics and construction requirements of other source metadata, and establish and improve NSTL metadata specifications to more effectively integrate and utilize third-party source metadata.

Currently, different literature databases exhibit variations in metadata content and description methods, which creates obstacles to integrating and utilizing third-party resources. The conflict between metadata format diversity and the single-interface requirement of NSTL processing specifications makes interoperability between third-party source metadata and NSTL literature resource metadata inevitable [2-4]. Clarifying the content and organization methods of externally sourced metadata, establishing relevant rules to achieve mapping between third-party source metadata and NSTL literature resource metadata, and outputting external source metadata resources in NSTL metadata format represents one operational approach for NSTL to integrate and utilize third-party database data.

The Web of Science (WOS) database [5] and Scopus database [6] are internationally renowned databases that share similarities with NSTL in providing literature information services. Based on analysis of WOS metadata specifications, Scopus metadata specifications, and NSTL' s adopted literature resource processing specifications, and combined with relevant practices, this paper uses journal articles as an example to compare the metadata mapping content, mapping effectiveness, and metadata description methods among the three systems,

and proposes issues requiring attention during the mapping and utilization of third-party metadata, aiming to provide reference for metadata construction in relevant literature information systems and the utilization of existing third-party source metadata resources.

2. Journal Article Metadata Structure

Based on the modular design principles of DC metadata [7] and combined with analysis of metadata content from the three literature databases (WOS, Scopus, and NSTL), journal article metadata can be categorized into paper metadata, author metadata, author institution metadata, journal metadata, conference metadata, funding metadata, reference metadata, and citing literature metadata. Using entity analysis methods, the relationships among journal article entities are shown in Figure 1 [Figure 1: see original paper]. A journal article may be written by one or more authors; an author may belong to one or more institutions; a paper is published in a journal; it may originate from a conference; it may be affiliated with a funding source; it may have one or more references; and it may be cited by one or more papers.

The three databases contain different types of journal article metadata. As shown in Table 1 , WOS and Scopus describe all eight metadata categories, while NSTL lacks descriptions for conference, funding, and citing literature metadata. The main reasons are: WOS and Scopus use a single metadata schema to describe multiple literature types such as journal articles, conference papers, books, and patents. Therefore, if a journal article involves conference or funding information, relevant descriptions will appear. NSTL divides metadata schemas based on literature type, with conference metadata contained within the conference paper schema; the NSTL processing specifications do not include descriptions for funding data or citing literature data.

3. Metadata Mapping and Comparison

Based on NSTL journal article metadata (some fields are Required, denoted by R), this section compares WOS and Scopus descriptions of the same fields in paper metadata, author/institution metadata, journal metadata, and reference metadata, and analyzes the characteristics of metadata descriptions in different literature databases to learn from each other' s strengths and weaknesses, improve the completeness and compatibility of NSTL literature resource processing specifications, and better adapt to and support the integration of various third-party source metadata.

3.1 Paper Metadata Mapping Comparison

NSTL paper descriptive information constitutes the main body of the journal article descriptive metadata specification, including paper title, keywords, abstract, and classification information. Equivalent fields in WOS and Scopus originate from different metadata modules. For example, in WOS, title and

document type information come from paper metadata, while start page, end page, and total page count come from journal metadata; in Scopus, title, abstract, and document type information come from paper metadata, while start page, end page, and total page count come from journal metadata, and total reference count comes from reference metadata. The mapping of paper metadata among WOS, Scopus, and NSTL is shown in Table 2 .

As shown in Table 2, among the 22 NSTL paper metadata fields, WOS achieves mapping for 12 fields and Scopus for 16 fields, indicating that different third-party source metadata have varying mapping coverage with NSTL metadata. Among the unmapped fields are required fields paper_id and local_doi, which must be processed (e.g., output as empty tags) before mapped external data can be exported in NSTL Schema format.

For successfully mapped fields, differences in value types and field repeatability across databases also affect the output of external data in NSTL Schema format. For example, NSTL' s type, WOS' s doctype, and Scopus' s citation-type all describe document type but have different enumeration values, requiring specification of mapping methods from WOS and Scopus enumerations to NSTL literature types.

If an NSTL field is repeatable while the external data source field is not, values can be directly mapped. If an NSTL field is non-repeatable while the external data source field is repeatable, parsing rules must be specified to select one value from multiple values in the external data source as the unique value for the NSTL field. For example, mapping Scopus' s repeatable field citation-language xml:lang= "" to NSTL' s non-repeatable field language can be configured to take the first citation-language value.

Table 2 also shows that NSTL uses element-based descriptions, while WOS and Scopus primarily use attribute-based descriptions. For instance, title, page numbers, and reference counts all employ attribute-qualified elements, enabling better consolidation of descriptive content. Additionally, both WOS and Scopus have unique identifiers for journal articles: WOS uses the uid element to uniquely identify papers, while Scopus uses eid, pui, pii, and other identifiers.

Based on analysis of WOS and Scopus metadata description characteristics, we partially revised the NSTL Schema by adding external data source paper unique identifier fields using attribute-qualified elements to map with corresponding fields in external data sources. For example, we added extend_ids extend_id type= "" value= "", where the type attribute maps to the external data source' s unique identifier and value contains the identifier value. This approach both uniquely identifies papers from external data sources (distinguishing them from self-processed data) and provides an expandable framework for subsequently adding unique identifiers from other data sources.

3.2 Author/Institution Metadata Mapping Comparison

In NSTL, authors refer to writers of journal articles. In WOS and Scopus, paper authors share sub-elements with publishers, chart creators, translators, etc. Therefore, specifying role types or parent elements is necessary to achieve accurate mapping, as shown in Table 3 . Beyond mapping elements, WOS and Scopus both include descriptions for author first name, last name, corresponding author, institutional address, affiliated country and city, as well as author unique identifiers such as ResearcherID, ORCID, and AuthorID. Author unique identifiers play a crucial role in uniquely identifying authors, and the approach used for paper unique identifiers can be referenced to provide an expandable framework for adding author unique identifiers from external data sources to NSTL.

As shown in Table 3, among the six NSTL author/institution metadata fields, both WOS and Scopus can map five fields, achieving relatively high mapping coverage. There are also cases where the same field has different value types across databases. For example, the author sequence field has value type byte[8] in NSTL' s author_sequence, but positive Integer in WOS' s seq_no, requiring coordination for actual data retrieval.

Additionally, NSTL describes author and institution information sequentially, Scopus groups authors by institution, and WOS establishes one-to-one correspondence between authors and institutions through the addr_no attribute. If the addr_no attribute in the author name element matches the addr_no attribute in the address_spec element, that institution is identified as the author' s affiliation. This approach conveniently enables correspondence regardless of how many institutions an author has, avoiding duplicate records.

3.3 Journal Metadata Mapping Comparison

Journals are the carriers of journal articles. In NSTL, journal metadata includes journal description elements (the first 14 fields in Table 4) and volume/issue description elements (the last 3 fields in Table 4). In WOS and Scopus, volume/issue description elements are contained within journal description elements. Beyond the mapped fields in Table 4, WOS includes more detailed journal name abbreviations, volume/issue publication dates, and publisher address information, while Scopus also describes journal unique identifiers (srcid), document source URLs, and journal editor information. Unique identifiers for journals and volumes/issues are important for uniquely identifying publications and can similarly reference the approach used for paper unique identifiers to provide an expandable framework for adding external data source journal and volume/issue unique identifiers to NSTL.

Among the 17 NSTL journal metadata fields, WOS achieves mapping for 9 fields and Scopus for 10 fields. Unmapped required fields are handled the same way as in paper metadata. A single element in NSTL may correspond to multiple elements or multiple attributes within the same element in WOS and Scopus.

For example, NSTL has only an issue field without separate fields for supplements, special issues, or parts, but specifies cataloging rules for these within the issue field (e.g., if there is an issue number but the issue is divided into parts, record the part prefix as is; for supplements and special issues, fill after the issue number; if there is no issue number, fill supplement information directly) [9]. These cataloging rules can be used to extract and merge corresponding data from WOS and Scopus.

3.4 Reference Metadata Mapping Comparison

In NSTL, reference metadata includes citation author, title, source, volume/issue, and access path information. Reference information allows users to find related literature from the perspective of author research 脉络 [10]. WOS includes author, title, journal name, volume, and page information in references but lacks original reference information fields. Scopus includes both original information fields and split fields for author, title, etc. The mapping of reference metadata among the three systems is shown in Table 5 , with unmapped required fields handled as before.

4. Advantages and Disadvantages of Metadata Mapping Approaches

Through mapping metadata among WOS, Scopus, and NSTL, we can see that mapping is achievable for most metadata fields, and these fields are relatively important. Overall, metadata mapping enables accurate and efficient conversion of external data source data to NSTL data. Therefore, metadata mapping is a feasible and effective approach for NSTL to integrate and utilize third-party source metadata. The greater the number of mapped metadata fields, the more fully external data source data can be utilized.

The metadata mapping approach also provides insight into integrating other data sources such as WOS data into the NSTL joint data processing system and builds an expandable framework for subsequently adding other third-party resources.

Although metadata mapping resolves some differences in information organization and content revelation methods among the three databases, limitations remain. For example, it cannot avoid target information loss caused by incomplete field mapping, which affects the comprehensiveness and completeness of NSTL processed data. Additionally, source information loss caused by differences in metadata description granularity—WOS and Scopus contain many more detailed descriptive fields for authors, institutions, and journals that are not reflected in NSTL—these fields provide more granular revelation of literature resources, and output through metadata mapping results in loss of external data source data.

By comparing the description methods of metadata fields across these databases

with NSTL' s own specifications, we can learn from each other' s strengths to improve the completeness and compatibility of our metadata. Through analysis of WOS and Scopus metadata and their mapping with NSTL metadata, we can target revisions to the existing NSTL metadata Schema. For example, adding unique identifiers for external data source data and modifying metadata value types can quickly, efficiently, and cost-effectively integrate external data. In the current environment where metadata description fields vary across literature databases, the ability to map metadata between them is significant for achieving data interaction and transfer. The more metadata fields that can be mapped, the more fully data can be utilized. This paper, based on mapping journal article metadata among WOS, Scopus, and NSTL, describes the process and methods for NSTL to integrate and utilize third-party source metadata and proposes issues requiring attention during metadata mapping and integration. Currently, NSTL has already applied purchased WOS data to its data processing workflow and will gradually add data from other sources, which will greatly benefit data processing speed and system automation levels.

References

[1] Zhang Jianyong, Zeng Yan. *NSTL Literature Data Processing Specification* [M]. Beijing: Intellectual Property Publishing House, 2009. (Zhang Jianyong, Zeng Yan. NSTL Literature Data Processing Specification [M]. Beijing: Intellectual Property Publishing House, 2009.)

[2] Song Linlin, Li Haitao. Metadata Interoperability in Mass Digitization Project: A Survey and Suggestions [J]. *Journal of Library Science in China*, 2012, 38(5): 27-38.

[3] Shen Xiaojuan, Gao Hong. Proceed from Metadata Mapping to Discuss Metadata Interoperability Problem [J]. *Journal of the National Library of China*, 2006(4): 51-55.

[4] Sa Lei. Study in Metadata Interoperability [J]. *Information Science*, 2014, 32(1): 36-40.

[5] Web of Science [EB/OL]. [2014-05-08]. http://www.webofknowledge.com/WOS.

[6] Scopus [EB/OL]. [2014-06-18]. https://www.scopus.com/.

[7] The Singapore Framework for Dublin Core Application Profiles [EB/OL]. [2015-05-08]. http://dublincore.org/documents/singapore-framework/.

[8] NSTL_journalarticle.xsd [EB/OL]. [2015-05-20]. http://spec.nstl.gov.cn/specification/namespace/NSTL_j

[9] Issue [EB/OL]. [2015-05-22]. http://spec.nstl.gov.cn/specification/index.php?title=Issue.

[10] Journal Article Metadata Specification [EB/OL]. [2015-06-05]. http://spec.nstl.gov.cn/specification/index.

Author Contributions

Yu Qianqian: Analyzed metadata specifications of the three databases, conducted mapping comparisons, wrote and revised the paper.
Zhang Jianyong: Proposed the basic framework and implementation scheme for integrating and utilizing third-party data, provided paper revision suggestions.

Received: July 27, 2015
Revised: September 6, 2015

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