ICIC Express Letters
Part B: Applications
Volume 4, Number 4, August 2013
ICIC International ©2013 ISSN 2185-2766
pp. 897–904

RESEARCH ON SEMANTIC METADATA ONLINE AUXILIARY CONSTRUCTION PLATFORM AND KEY TECHNOLOGIES

Yao Liu and Ruijia Wang

Information Technology Support Center
Institute of Scientific and Technical Information of China
No. 15, Fuxing Road, Beijing 100038, P. R. China
liuy@istic.ac.cn

Received November 2012; accepted February 2013

Abstract. *In this paper, we propose methods to promote library resources semantization process by using natural language processing technology and machine learning, and developing semantic metadata online auxiliary construction platform of Chinese information resources to do semantic label of related literature. The system proposed in this paper can help to construct the semantic metadata of scientific and technical information based on the Web, and utilize up-to-date semantic technologies to realize domain thesaurus and effectively integrate various kinds of metadata. This system can also use the semi-structured text to automatically or auxiliarily construct the semantic metadata system, and realize the edit, development and maintenance of a certain semantic metadata project.*

Keywords: Semantic metadata, Natural language processing, Knowledge engineering

1. **Overview.** A great deal of literature resource is stored in the library, so it has become key problems of achieving leap development of next generation literature services in library and reaching the stage of knowledge service based on literature information retrieval services to help users to find knowledge in the literature comprehensively, quickly and accurately, and to showcase the knowledge from different knowledge dimensions and find the relevance between the knowledge, in order to help users to create knowledge innovation more effectively. So the semantization of library resources is the general trend in a networked environment. Semantization means choose suitable semantic label to change the resource into the format that computer can recognize with the semantic features reflected from label content in the resource, and make the computer understand and master the resource content to some extent.

In this paper, we consider that semantization of library resources can be regarded as shallow labeling of library resource [1-5], and is the interactive implementation of content semantization and form semantization. Therefore, based on NLP (natural language processing) theories and methods, we propose this method to utilize the traditional library resource organization mode, construct semantic metadata system, and construct intergrated platform of auxiliary construction and annotation [6,7]. Two key problems of this research are semantic metadata generation and semantic annotation.

2. **Ideas and Methods.** In this paper, we use NLP technologies and machine learning methods to develop the semantic metadata online auxiliary construction platform of Chinese information resources, semantically label the related literature, and generate initial semantic metadata by utilizing traditional organizational resources (such as thesaurus) with machine learning method based on the relatively semantization of a large amount of content, and simultaneously realize implementation of semantic metadata system construction and organizational resources semantization. The process and structure are shown in Figure 1.

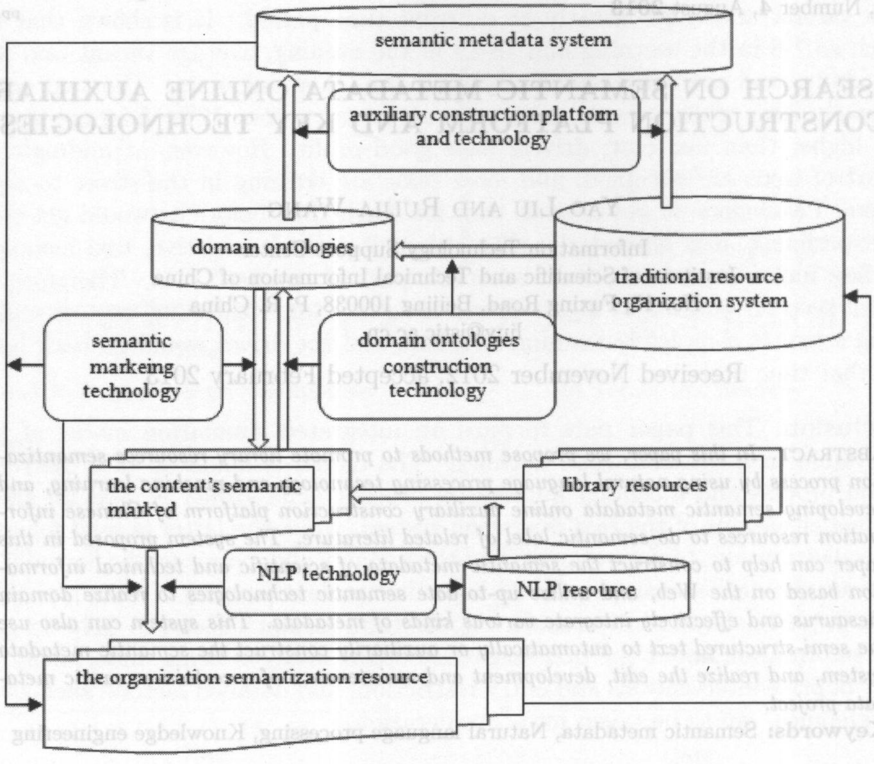

FIGURE 1. Process and structure

3. Development and Implementation of Auxiliary Construction Platform. Intergrated platform system of semantic metadata auxiliary construction and annotation can support multiuser online to auxiliarily construct and edit semantic metadata. This system can realize joint edit, development and maintenance of a semantic metadata project by letting multiuser expediently visit, create and maintain semantic metadata based on B/S (browser/server) technology, and managing users in Web page way. The main features include users' authorization management, semantic metadata projects management, semantic metadata online edit, semantic metadata projects comparison, conversion processing of semantic metadata in different formats, semantic metadata project format management and import and export of semantic metadata.

3.1. Import of domain thesaurus. Firstly upload the domain thesaurus with neat format and clear hiberarchy to this system through the Web, then the system would convert it to semantic metadata base model and show it as tree form in the Web (as shown in Figure 2).

3.2. Import of metadata. Import the structured metadata or semi-structured text (data file or network) into base semantic metadata model generated by thesaurus, in order to enrich the contents of domain semantic metadata.

3.2.1. *NLP automatic analysis and processing functions.* NLP automatic analysis and processing functions include structured vocabulary processing function (as shown in Figure 3), textbooks and other text processing function and professional dictionary processing function (as shown in Figure 4). Structured vocabulary processing functions mainly process resources with structured information, such as traditional Chinese medicine thesaurus, thesaurus, classification vocabulary, classification thesaurus; textbooks and other text processing functions mainly process relatively standard electronic text, such as text-books; professional dictionary processing functions mainly import and process professional dictionary.

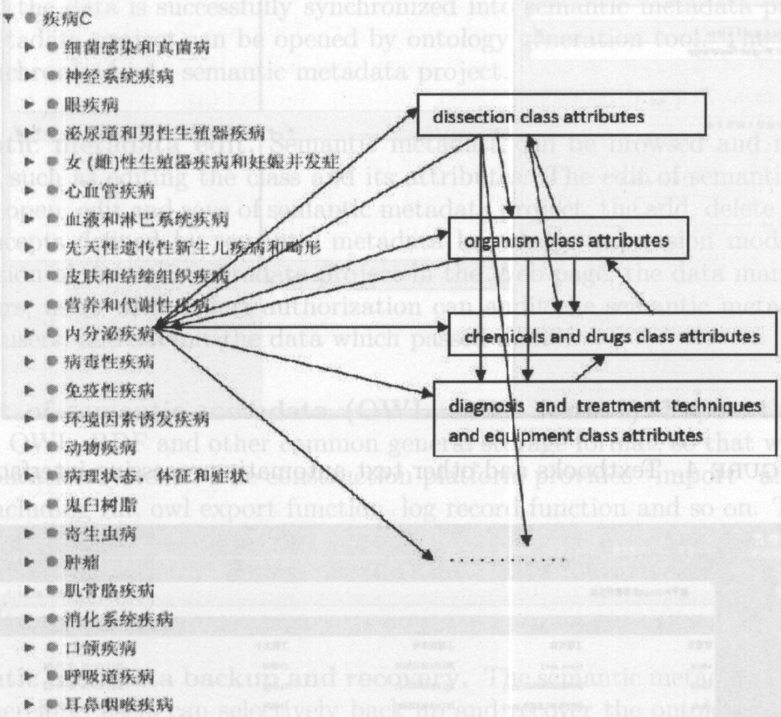

FIGURE 2. Schematic diagram of the knowledge description frame

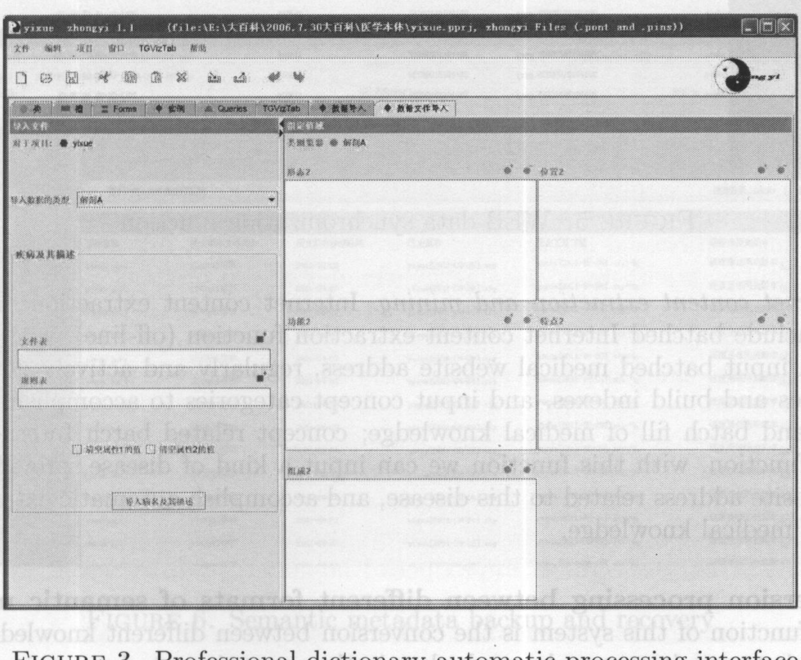

FIGURE 3. Professional dictionary automatic processing interface

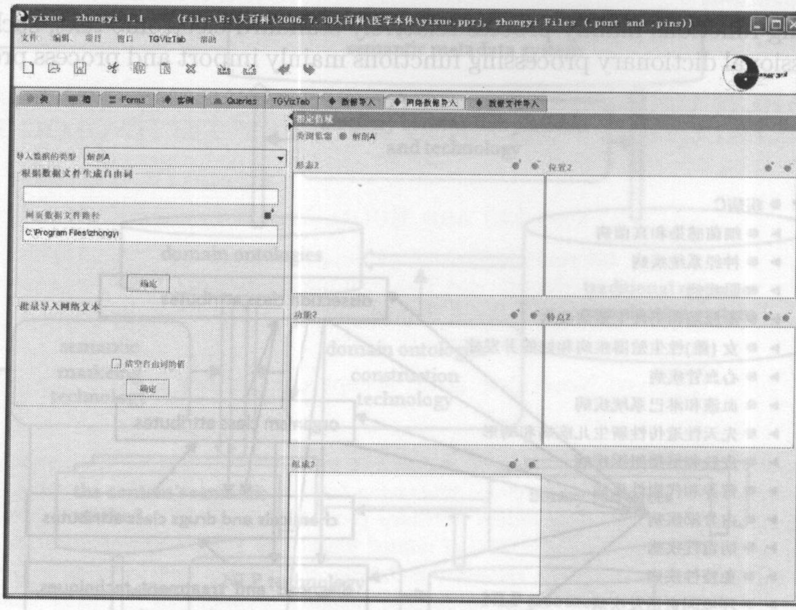

FIGURE 4. Textbooks and other text automatic processing interface

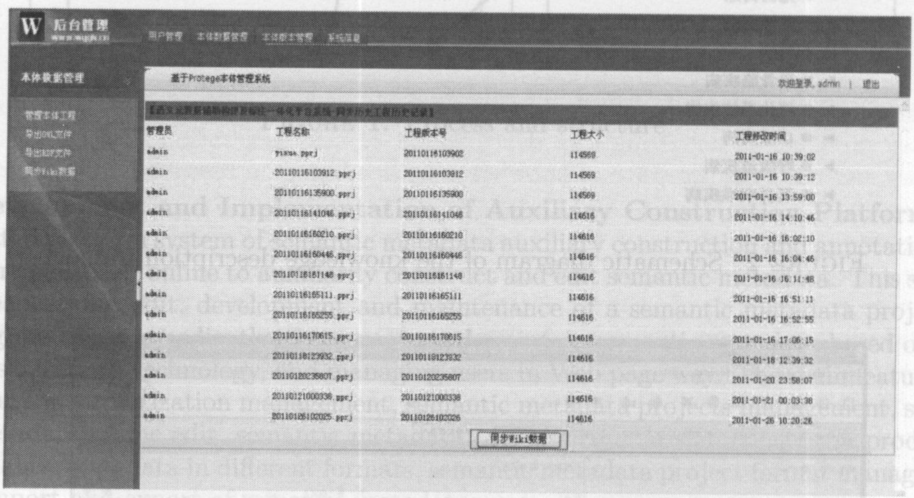

FIGURE 5. WEB data synchronization function

3.2.2. *Internet content extraction and mining.* Internet content extraction and mining functions include batched Internet content extraction function (off-line), with this function we can input batched medical website address, regularly and actively download all the webpages and build indexes, and input concept categories to accomplish automatic extraction and batch fill of medical knowledge; concept related batch Internet content extraction function, with this function we can input a kind of disease, provide batched medical website address related to this disease, and accomplish automatic extraction and batch fill of medical knowledge.

3.3. **Conversion processing between different formats of semantic metadata.** The other function of this system is the conversion between different knowledge storage ways. That is to say, it can resolve each class in the semantic metadata project into text file that could be recognized, and import it into the data base for users to edit and update its content, and then reflect the updated content in the semantic metadata project files.

Open the synchronous Web data option, and we can see the interface in Figure 5.

Click the option "synchronize Web data" to start synchronizing Wiki data into semantic metadata project. After successfully synchronizing Web data, there will be a successful prompt, and the data is successfully synchronized into semantic metadata project. The semantic metadata project can be opened by ontology generation tool. The data is successfully synchronized into semantic metadata project.

3.4. **Semantic metadata edit.** Semantic metadata can be browsed and modified by this system, such as editing the class and its attributes. The edit of semantic metadata includes the open, edit and save of semantic metadata project; the add, delete and modify of basic concepts defined by semantic metadata knowledge expression model; the tree form expression of semantic metadata project in the Web page; the data management of different users, users with expert authorization can audit the semantic metadata edited by different users, and submit the data which passed audits.

3.5. **Export of semantic metadata (OWL, RDF format).** Export the semantic metadata in OWL, RDF and other common general storage format, so that we can use it in other application systems. The construction platform provides "import" and "export" functions, including rdf, owl export function, log record function and so on. This system exports the semantic metadata project as rdf files for users to download, and exports the semantic metadata project as owl files for users to download, and writes the log in the log directory.

3.6. **Semantic metadata backup and recovery.** The semantic metadata is constantly modified; therefore, users can selectively back up and recover the ontologies, in order to manage the historical data.

The other main function of this system is that it can recover the semantic metadata into previous version.

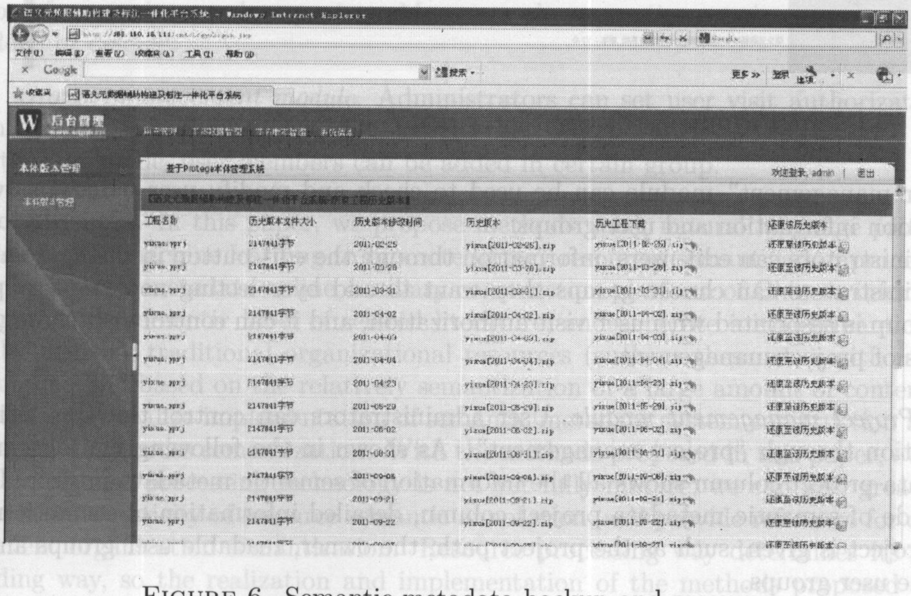

FIGURE 6. Semantic metadata backup and recovery

3.7. Realization of develop authorization management of semantic metadata. This system has strict authorization control. Users can create semantic metadata initial system, and they can also add users or groups, and designate read and write authority for the users and groups. Users can take action to the content within their extent of authorization.

3.7.1. *User management module.* User authorization is used to control different semantic metadata visit authorization of different users, and it can be divided into three classes.

Normal users. Normal users can open and edit the semantic metadata project within their domains.

Expert users. Except the authorization of normal users, the expert users can also audit the semantic metadata modified and edited by normal users.

System administrator users. Except the authorization of normal users and expert users, the system administrator users can also create users with different authorizations.

3.7.2. *User authorization management module.* User administrators can see the interface as follows after login in:

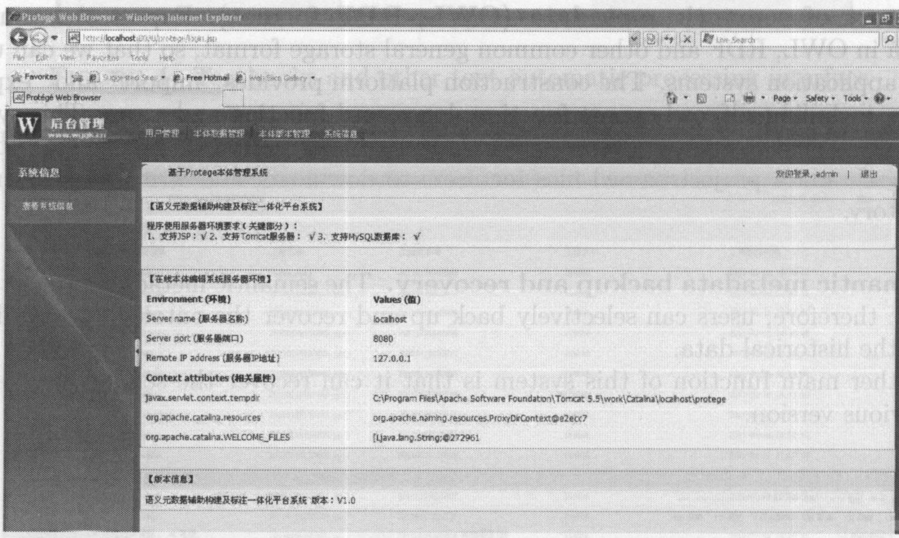

FIGURE 7. User authorization management module

"User management" module can be used to check and modify user name, password, description information and user groups.

Administrators can edit users' information through the edit button in the figure above.

Administrators can choose groups they want to add by selecting new pages popping up. Group is associated with user visit authorization, and it can control visit through the settings of project management.

3.7.3. *Project management module.* User administrators can control the user visit authorization through "project management". As shown in the following figure, semantic metadata project column shows all the information of semantic metadata project. In the right side of semantic metadata project column, detailed information of semantic metadata project is given, such as the project path, the owner, readable user groups and the writable user groups.

Edit the basic information of semantic metadata project: edit the description information, the path and the name of semantic metadata project.

3.7. Realization of develop authorization management of semantic metadata. This system has strict authorization control. Users can create semantic metadata initial system, and they can also add users or groups, and designate read and write authority for the users and groups. Users can take action to the content within their extent of authorization.

3.7.1. *User management module.* User authorization is used to control different semantic metadata visit authorization of different users, and it can be divided into three classes.

Normal users. Normal users can open and edit the semantic metadata project within their domains.

Expert users. Except the authorization of normal users, the expert users can also audit the semantic metadata modified and edited by normal users.

System administrator users. Except the authorization of normal users and expert users, the system administrator users can also create users with different authorizations.

3.7.2. *User authorization management module.* User administrators can see the interface as follows after login in:

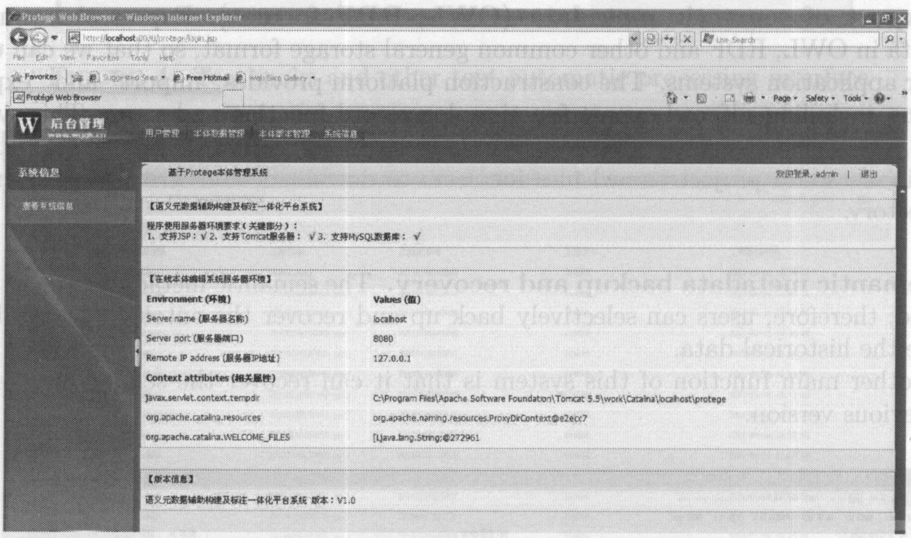

FIGURE 7. User authorization management module

"User management" module can be used to check and modify user name, password, description information and user groups.

Administrators can edit users' information through the edit button in the figure above.

Administrators can choose groups they want to add by selecting new pages popping up. Group is associated with user visit authorization, and it can control visit through the settings of project management.

3.7.3. *Project management module.* User administrators can control the user visit authorization through "project management". As shown in the following figure, semantic metadata project column shows all the information of semantic metadata project. In the right side of semantic metadata project column, detailed information of semantic metadata project is given, such as the project path, the owner, readable user groups and the writable user groups.

Edit the basic information of semantic metadata project: edit the description information, the path and the name of semantic metadata project.

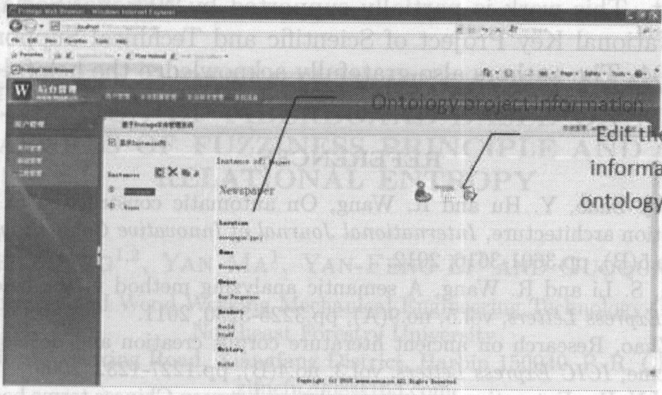

FIGURE 8. Project management module

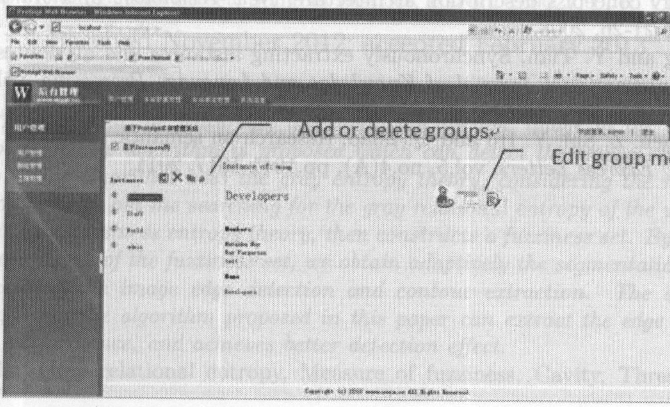

FIGURE 9. Groups management module

Add or delete reader and writer groups of semantic metadata project, and owner information of semantic metadata project. Moreover, the semantic metadata project can also be added or deleted.

3.7.4. *Groups management module.* Administrators can set user visit authorization through "groups management".

In the edit mode, user members can be added in certain group.

4. **Conclusions.** In this paper, we propose methods using natural language processing technology and machine learning to promote library resources semantization process. We develop semantic metadata online auxiliary construction platform of Chinese information resources to do semantic label of related literature, and generate initial semantic metadata by utilizing traditional organizational resources (such as thesaurus) with machine learning method based on the relatively semantization of a large amount of content, and simultaneously realize implementation of semantic metadata system construction and organizational resources semantization. With the methods proposed in this paper, even the natural language processing technology is not yet fully mature, we can also greatly improve the automaticity of resource semantic annotation, and provide reference for quickly implement organization semantization. Resource processing way determines its service providing way, so the realization and implementation of the methods proposed in this paper can effectively promote the process of library resources semantization, and make the semantic annotation processing and knowledge service possible.

Acknowledgment. This work is partially supported by National Social Science Fund No. 12BTQ006, National Key Project of Scientific and Technical Supporting Programs No. 2011BAH10B04. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] Y. Liu, Z. Sui, Q. Zhao, Y. Hu and R. Wang, On automatic construction of medical ontology concept's description architecture, *International Journal of Innovative Computing, Information and Control*, vol.8, no.5(B), pp.3601-3616, 2012.

[2] Y. Liu, X. Chen, S. Li and R. Wang, A semantic analyzing method in the field of technological literature, *ICIC Express Letters*, vol.5, no.9(A), pp.3225-3230, 2011.

[3] Y. Liu and Y. Zhao, Research on ancient literature corpus creation and development of Chinese traditional medicine, *ICIC Express Letters*, vol.3, no.4(B), pp.1227-1232, 2009.

[4] Z. Sui, Y. Liu and Y. Hu, Extracting hyponymy relation between Chinese terms based on term types' commonality, *ICIC Express Letters*, vol.3, no.4(B), pp.1233-1238, 2009.

[5] Y. Liu, Z. Sui, Y. Zhou and Z. Wang, Research on automatic construction of Chinese traditional medicine ontology concept's description architecture, *New Technology of Library and Information Service*, vol.5, pp.21-26, 2008.

[6] Z. Sui, W. Kang and Y. Tian, Synchronously extracting instances and attributes for the concepts from the web, *International Journal of Knowledge and Language Processing*, vol.3, no.3, pp.1-17, 2012.

[7] Y. Liu, X.-F. Chen, Z. Sui, Y. Hu and Q. Zhao, Research on semantic method of library resources' organizing, *ICIC Express Letters*, vol.5, no.4(A), pp.1011-1017, 2011.