

Specification for Construction of Technology and Product Thematic Database of Aluminum Industry

Zhijun Guo¹ and Yi Huang²

¹ Data Center

The Open University of China

No. 75, Fuxing Road, Beijing 100039, P. R. China

guozhijun@crtvu.edu.cn

² Information Technology Support Center

Institute of Scientific and Technical Information of China

No. 15, Fuxing Road, Beijing 100038, P. R. China

huangyi@istic.ac.cn

Received December 2014; revised December 2014

ABSTRACT. Thematic database with industry characteristics can provide important information quickly and accurately for scientific research and enterprise production that will promote innovation greatly. This study proposed a specification for construction of technology and product thematic database of aluminum industry. This specification includes management mechanism, data specification and data collection and processing workflow which are important aspects in the building process of thematic database. The specification has guiding significance for the construction of technology and product thematic database of aluminum industry.

Keywords: Aluminum Industry; Thematic Database; Construction Specification

1. **Introduction.** Developing the thematic database with industry characteristics is a major issue which the large and medium-sized enterprises and innovative clusters currently face. With integration of digital information and development of individual user requirements, there is an urgent demand for building a large scale, distinctive, authoritative thematic database, especially in the information age. The necessity for construction of the thematic database bases on three reasons. First, the boundlessness, diversity and personalization of needs for social information make that the current comprehensive bibliographic database still cannot meet the requirements of a particular subject or field personnel to retrieve information. However, the thematic database can satisfy users to obtain literature in a timely and accurate approach. Second, the rapid development of information network and automation enables that the thematic databases can efficiently meet the basic requirements for specific information storage. Third, the thematic database can be served as the

cornerstone of large-scale integrated databases. It is no doubt that the construction of a large number of high quality thematic databases will lay a good foundation for building, updating and developing a comprehensive database. Furthermore, the comprehensive database will play a key role in promoting the development of thematic database due to its features of intersecting and prospective.

Thematic databases can provide accurate, adequate, and effective important information quickly for scientific research. It enables researchers to save time and avoid duplication of efforts and will promote the development of innovation in scientific research.

2. Research Target. The construction of technology and product thematic database for aluminum industry is to create an enterprise innovation resource database which is around the demand for information resources using in enterprise innovation. The database based on the engineering knowledge system overall framework and norm. The major goal is collecting aluminum and related industries information of relevant guidelines, policies and financial support on technologies, products and applications. After being processed, the collected information is used to create multiple thematic databases which are eventually merged into integrated enterprise innovation knowledge base. The integrated enterprise innovation knowledge base provides resources and services for enterprise innovation. The contents of construction are as following.

(1) Research and develop management mechanism for the construction of industry thematic database.

This content is to research organization and management system and enact realistic plan for thematic database construction based on the overall requirement of project. The main content consists of content system (data scope, content classification, etc.), management system for database construction, quality control system, data processing and updating system, communication mechanism. It is designed for ensuring the successful completion of the project

(2) Research and develop data specification.

This content is to research and enact data specification for industry thematic database based on the overall requirement and the standard specifications making by relevant subjects. The data specification is composed of determining of data source, quality assurance of data source, data format specification, metadata specification, data storage mechanism, data update mechanism, content related to data processing and workflow. The data format specification and metadata specification mainly reference to the content of the subject. It is used in ensuring the authoritative, accuracy and applicability of the thematic database content.

(3) Carry out data collection and processing work for thematic database.

This content is to carry out the collecting, processing, standardizing, reviewing, storing-for data based on the management mechanism and data specification. The process needs to record the raw data, intermediate data and final data in order to ensure that the final data can be traced back to the original data source. The processing operation log is recorded so as to keep track of the data processing.

3. Research Scheme and Overall Roadmap. The research scheme and overall roadmap are applied to insuring the quality of the data through standard operations and strictly checking in the basis of the requirement of thematic databases construction. First, the research needs to determine the data source and establish a classification system and build up a data format according to the materials which are made up of the general requirement of the project, the mission statement, the key vocabulary of key equipment and technology of aluminum smelting, and then to build a data collection and processing procedure by means of network-based automated crawlers and complemented with artificial collection. Finally, the result is to create a special thematic database through collecting, filtering, classifying, processing, reviewing and storing data.

(1) Determine the data source.

This step is to collect the data by a scientific way after determining the topic. First, make it clear of the main content, direction and range of the data collection. Second, using information databases and Internet collect data. The information databases include Wanfang database, Vip database, CNKI database, DRCNET database, etc. The Internet data source includes portal sites for aluminum industry such as www.cnal.com, www.alu.cn, www.ometal.com and well-known Internet search engines, such as Baidu and Google.

(2) Establish a classification system.

A classification system is established for the technology and product thematic database of aluminum industry after pre-checking the topic. The classification system is made of three levels based on the general requirement of the project, the mission statement and the key vocabulary of key equipment and technology of aluminum smelting. There are a total of 105 categories in all three levels, and the categories in the third level are described in order to classify the collected data.

(3) Establish and standardize data format.

Depending on the type of the database, a specification for metadata format is designed. The specification is composed of field names and field descriptions. The collected data include journal articles, dissertations, conference papers, standards, project outcomes, policies and regulations, equipment information, product samples, as well as professional website. The fields consist of title, English title, keyword, abstract, subject, classification number, language, author, author unit, source, release time, full-text, etc.

(4) Find tools for data collecting and processing.

By identifying and comparing, the CKI system is chose to collect and process data. This tool uses search engines and text mining techniques, along with the classification and keywords set by users, to automatically process web content including automatically collecting, processing, recognition, duplication removal, similarity analysis, sensitive words filtering. In addition to the automatic collection, the system can also add information through artificial operations such as selection, editing, and other means.

(5) Establish and regulate the workflow for data collecting and processing.

The overall workflow for data collecting and processing is showed as figure 1. For the sake of making the collecting and processing work smoothly, the information source is divided into two parts by the study group. The first part is the industry website which can

be automatically processed by the CKI system. The second part is the large information databases such as Wanfang database, which need to retrieve and crawl manually. The collected data is stored in an original repository so as to facilitate proofread and tracking, and will be further processed by technical personnel. The further processing includes: (1) Filter data. It removes redundant and useless literature. The same literature may be included in multiple sources, so that the duplicates should be removed. The useless literature, including weak correlation, low level, missing key information such as no source, title, should be deleted. (2) Improve the technical content. It reflects the essence of the literature with a certain depth and precision through adding summary, refining keywords, revealing basic information. (3) Improve the data content. It normalizes the data according to a unified format. The insufficiency field contents should be made perfect. All the downloaded data should be formatted and imported into database in accordance with the design of the thematic database. As the information databases include a mass of fields, it can export the useful fields in the light of the needs to form a personalized database.

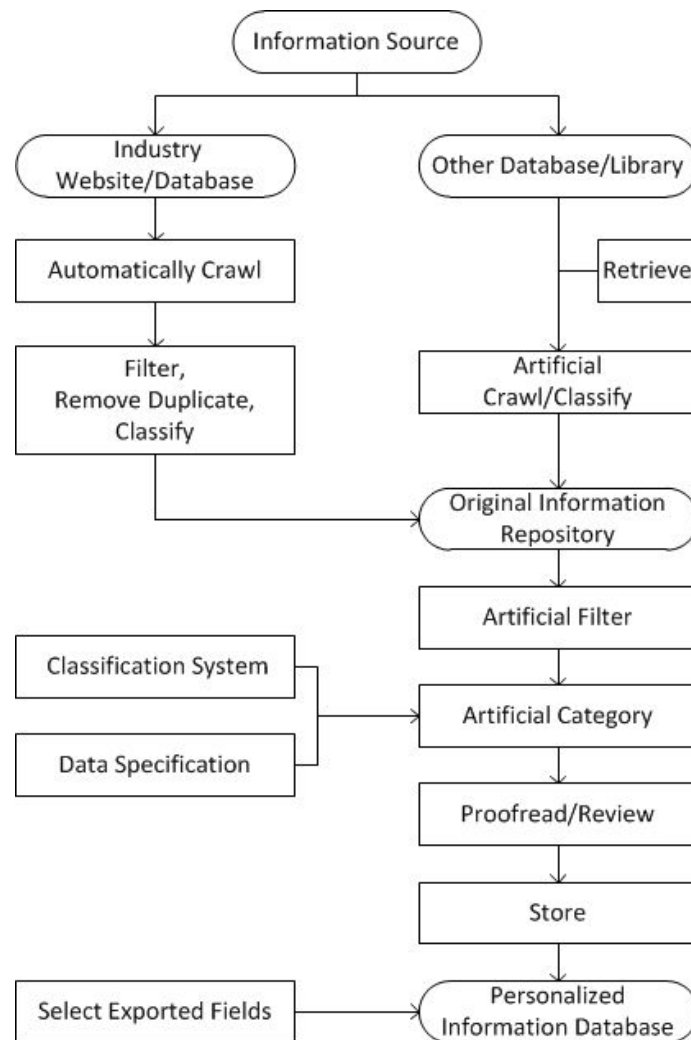


FIGURE 1. INFORMATION COLLECTION AND PROCESSING PROCEDURE

4. **Research Content.** Following certain technical standard and specification is one of the basic requirements for thematic database construction. In accordance with the overall project requirement and standard specifications making by relevant studies, the study group researches and develops the data specification of thematic database for aluminum industry, which includes data source specification, database classification specification, metadata specification, data updating mechanism, workflow control for data processing and other related content.

4.1. **Data source specification.** Data source is shown in table 1:

TABLE 1. RELATED WEBSITES FOR ALUMINUM INDUSTRY

Information Type	Name	URL of information source
Domestic large-scale integrated database	Wanfang	http://g.wanfangdata.com.cn
	Vip	http://www.cqvip.com
	CNKI	http://www.cnki.net
	NSTL	http://www.nstl.gov.cn
	GDP	http://gpd.sunwayinfo.com.cn
Domestic aluminum industry website	Aluminum Fabrication	http://www.alu365.com
	China Aluminum	http://www.chinalvwang.com
	South China Aluminum	http://www.hnlvye.net
	CCEN	http://www.ccen.net
	Shandong Aluminum	http://www.qoloo.com.cn
	Global Metal	http://www.ometal.com
	Global Aluminum	http://www.alu1886.com
	ALU.CN	http://www.alu.cn
	CNAL	http://www.cnal.com http://www.al-market.com
	China Metal	http://www.metalchina.com
	China Aluminum alloy	http://www.zglhjw.com
	Albiz	http://www.albiz.cn
	SMM	http://www.smm.cn
	Aluminum - hexun	http://futures.hexun.com/aluminum
CHALCO	http://www.chalco.com.cn	
Foreign aluminum industry website	Alcoa	http://www.alcoa.com
	Alumar	http://www.alumar.nl (English) http://www.alumar.com.br (Portuguese)
	Hindalco Industries	http://www.hindalco.com
	NALCO	http://www.nalcoindia.com
	HYDRO	http://www.hydro.com
	Rio Tinto Alcan	http://www.riotintoalcan.com

	RUSAL	http://www.rusal.ru
	VALE	http://www.vale.com (Portuguese) http://www.vale.cn (Chinese)
	CVG Bauxilum	http://www.bauxilum.com (Spanish)
	BHP Billiton	http://www.bhpbilliton.com
	Glencore	http://www.glencore.com
Other	Baidu Library	http://wenku.baidu.com
	Docin	http://www.docin.com
	Iask.Sina	http://ishare.iask.sina.com.cn

4.2. Database classification specification. The database classification system should set up classes on demand and emphasis. The main purpose of the classification system is to organize resources more effective and make them convenient for retrieving and using by users. Therefore, it is necessary for the classification system to demonstrate not only the form features of the resources, but also to reveal the content features of the resources. According to the characteristics of data resources of aluminum industry, combined with the classification system design principle, the study team established a classification system for the thematic database construction which including a technology classification system and a product classification system.

- (1) Technology classification system for aluminum industry (technical topics). The technology classification (technical topics) is shown in table 2.

TABLE 2. TECHNOLOGY CLASSIFICATION (TECHNICAL TOPICS)

Code	First Level Category	Second Level Category
	Bauxite	
	Alumina	
	Electrolytic Aluminum	
	Secondary Aluminum	
	Aluminum Fabrication	
	Auxiliary	
		Energy Conservation & Environment Protection
		Energy & Power
		Automatic Control
		Safety Production
		Other

- (2) Technology classification system for aluminum industry (production processes). The technology classification (production processes) is shown in table 3.

TABLE 3. TECHNOLOGY CLASSIFICATION (PRODUCTION PROCESSES)

Code	First Level Category	Second Level Category
	Bauxite	
		Exploration
		Mining
		Beneficiation
		Reclamation
	Alumina	
		Bauxite ore crushing and pre-homogenization
		Lime firing
		Slurry Preparation
		Bayer process for autoclave leaching
		Sintering Clinker
		Sintering clinker dissolution
		Isolation and washing of red mud
		Control filter
		Sintering crude liquid desilication
		Sodium aluminate solution decomposition
		Isolation and washing of aluminum hydroxide
		Liquor evaporation
		Aluminum hydroxide roasting
	Electrolytic Aluminum	
		Dissolve
		Electrolysis
		Anode gas purification
		Purification of liquid aluminum
	Secondary Aluminum	
		Scrap aluminum pretreatment
		Ingredients smelting
		Purification metamorphism

- (3) Technology classification system for aluminum industry (production methods). The technology classification (production methods) is shown in table 4.

TABLE 4. TECHNOLOGY CLASSIFICATION (PRODUCTION METHODS)

Code	First Level Category	Second Level Category
	Bauxite	
		Open pit mining
		Underground mining
		Open - Underground Combined Mining
	Alumina	

		Bayer process
		Sintering process
		Combination process
	Electrolytic Aluminum	
		Pre-baking
		Self-baking
	Aluminum Fabrication	
		Smelt
		Extrusion
		Rolling
		Heat tinting
		Die forging
		Powder

(4) Product classification system for aluminum industry. The product classification is shown in table 5.

TABLE 5. PRODUCT CLASSIFICATION

Code	First Level	Second Level	Value space
01	Application		1. Packaging 2. Chemical Industry 3. Power Industry 4. Machinery Manufacturing 5. Building Industry 6. Aerospace industry 7. Transportation 8. Light industry 9. Other
02	Process		
0201		Primary product	1. Alumina 2. Aluminum Alloy
0202		Semifinished product	1. Sheet 2. Strip 3. Foil 4. Powder 5. Ingot 6. Pipe 7. Profile 8. Bar 9. Wire
0203		Final product	

03	Price range		<ol style="list-style-type: none"> 1. 0-999¥ 2. 1000-9999¥ 3. 10000-99999¥ 4. >=100000¥
04	Place of origin		<ol style="list-style-type: none"> 1. East China 2. South China 3. North China 4. Central China 5. Northwest China 6. Southwest China 7. Northeast China 8. Hong Kong, Macao and Taiwan 9. Other countries
05	Aluminum type		<ol style="list-style-type: none"> 1. Aluminum 2. Aluminum-copper alloy 3. Al-Mn alloy 4. Al-Si alloy 5. Magnesium alloy 6. Aldrey 7. Aluminum-zinc- magnesium alloy 8. Aluminium - other elements alloy 9. Spare alloy

4.3. **Metadata specification.** In terms of the characteristics of the aluminum industry database, the study team designed a metadata format specification including the collected field names and field descriptions which is shown in table 6 and table 7. The collected data cover journal articles, dissertations, conference papers, standards, project outcomes, policies and regulations, equipment information, product samples, as well as professional website. The fields contain title, English title, keywords, abstract, subject, classification number, language, author, author unit, source, release time, full-text, etc.

TABLE 6. METADATA FORMAT SPECIFICATION FOR ALUMINUM INDUSTRY 1

Metadata Name	Type Requirement
Product name	
Product introduction	
Keyword	
Processing method	Refer to specification
Application	Refer to specification
Process	Refer to specification
Place of origin	Refer to specification
Production date	yyyymmdd

Price	CNY
Price range	Refer to specification
Aluminum type	Refer to specification
Picture	
Country	
Province	
Manufacturer	
Contacts	
Phone	
Fax NO.	
E-mail	
Address	
Source name	
Source website	

TABLE 7. METADATA FORMAT SPECIFICATION FOR ALUMINUM INDUSTRY 2

Metadata Name	Type Requirement
Title	
Keyword	
Summary	
Text	
Technical topic	Refer to technology classification specification
Production process	Refer to production process classification specification
Production method	Refer to production method classification specification
Source name	
Source website	

4.4. Data updating mechanism. Building a thematic database is by no means an easy thing, so it is necessary to insist in keeping tracking information in a certain period of time, and to update, maintain the database regularly. Furthermore, it is very important to continually add content during the follow-up database maintenance process in order to realize the value of the thematic database by providing continuous and complete information.

5. Conclusions.

5.1. Achievements and applications. There are some achievements through the efforts of the study group according to the mission statement. The achievements are listed below:

(1) Drew up a list of data sources for the project *Construction of Technology and Product Thematic Database for Aluminum Industry*. The list has four data source categories which are domestic large-scale integrated databases, domestic industry websites, foreign industry websites and search engines.

(2) Established a classification system for project *Construction of Technology and*

Product Thematic Database for Aluminum Industry.

(3) Established a metadata format specification for project *Construction of Technology and Product Thematic Database for Aluminum Industry*. The metadata format Includes papers, standards, policies and regulations, product samples, equipment information, project information, web pages and other types of information.

(4) Designed and established a workflow for information collection. This workflow is use to regulate the data sources, data collecting, data filtering, data processing, data reviewing, data storing and data exporting.

(5) Established a technology and product thematic database for aluminum industry which has up to 10500 records.

The technology and product thematic database for aluminum industry is based on the project requirement and special features of aluminum industry, which has the merits of well-targeted, information centralized, easy retrieval, etc. It is conducive to fully understand and analyze existing technologies and products in aluminum industry so as to further enhance the technological innovation and capabilities.

During the project implementation process, in order to establish the specifications for data source, data collection, data processing, data storage and so on, the study team explored the construction of specification of industry thematic database and has achieved initial success. These exploration and the results achieved in the study have played a positive role in the field of setting up a business innovation knowledge base and in the study of demand for information resources in innovation process.

5.2. Problems and suggestions.

(1) The construction of enterprise innovation knowledge base needs to strengthen the overall plan and uniform the standard.

The construction of enterprise innovation knowledge base is not a simple application development task. It is a complex, long-term system engineering which has associated with many aspects. The overall plan and uniform standard are very important. If there is no unified plan and guidance, each department will set up specifications based on their own requirements, that might cause it unable to form a standard global specification system. Hence, it need to strengthen the overall plan, uniform the standard, forbid to go in their own way and promote cooperation across departments and regions so as to advance the construction of enterprise innovation knowledge base.

(2) Management and sharing mechanisms for construction of enterprise innovation knowledge base needs to be further improved.

The construction of enterprise innovation knowledge base is a project which requires multi-unit cooperation. It is important to build up scientific and rational management and sharing mechanisms in the cause of giving full play to the various units together. At present, each unit has carried out research in their respective fields and has made some achievements, but because of lack of communication and sharing, cannot form an overall effect and impact. This requires to establish an effective management system and sharing mechanism and to promote more depth and extensive cooperation and sharing in each unit,

thus creating a powerful force to jointly boost the construction of enterprise innovation knowledge base.

Acknowledgment. This work is partially supported by National Key Project of Scientific and Technical Supporting Programs No. 2011BAH30B01; the authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

- [1] Yao Liu, Zhifang Sui, Qingliang Zhao, Yongwei Hu, Ruijia Wang, On Automatic Construction of Medical Ontology Concept's Description Architecture, *International Journal of Innovative Computing, Information and Control*, vol.8, no.5, pp.3601-3616, 2012.
- [2] Yao Liu, Xuefei Chen, Li Sujian, Ruijia Wang, A Semantic analyzing Method in the field of Technological Literature, *ICIC Express Letters*, 5(9): 3225-3230, 2011.
- [3] Sui Zhifang, Kang Wei, Tian Ye, Synchronously Extracting Instances and Attributes for the Concepts from the Web, *International Journal of Knowledge and Language Processing*, vol3, no.3, pp.1-17, 2012.
- [4] Yao Liu, Xuefei Chen, Zhifang Sui, Yongwei Hu, Qingliang Zhao, Research on semantic method of library resources' organizing, *ICIC Express Letters*, 5(4): 1011-1017, 2011.
- [5] Henglong Li, Songqing Gu, *Technology Handbook of Aluminum Production*, Metallurgical Industry Press, pp184, 2011.