International Journal of Information Science 2012, 2(6): 79-84
DOI: 10.5923/j.ijis.20120206.03

Very Large Information Services as Sources of Data to Subject Databases

Michael Trachtengerts

Joint Institute for High Temperatures, Russian Academy of Sciences, Moscow, Russia

Abstract The collecting of relevant data is most important and laborious activity in scientific thematic centres. Some approaches are applied to fulfil it more effectively. In this paper we consider how to use global information retrieval systems in science and knowledge as sources for data centres in a thematic field. The description of developed software is provided. It is shown as example that the use of such resources promoted the Thermophysical Centre in the Russian Academy of Sciences (thematic field — thermophysical properties of substances) to reduce essentially efforts to update its DB with high rank information. Some very large information systems (VLIS) are considered from the view of appropriate means for transfer the results of search into a subject DB.

Keywords CDS/ISIS, Subject Funds, Information Services, Information Input, Thermophysical Centre

1. Introduction

Collection of new published data is a permanent and labour-consuming process for most subject Data Bases. Improvement of data collection methods has some priority in internal activity of the subject data centres. Many of them in similar scientific fields perform much identical work. They are — data selection from the same sources, some operations of technical kind — setting information according to adopted rules (names and initials of authors, sources of publications, papers, keywords and other fragments of records), typing, precise control, and so on.

The collection improving should result in progress of scientific knowledge that depends in rapid revealing of new findings, facts, and so on. Quick access to them of many scientists is important to evaluate such o them that may be discarded if unreliable, or included in reliable knowledge. It is a difficult problem to allocate the relevant resources in a subject DB because they are spread widely through many sources of information.

Earlier solution of the problem was based on association of groups of the related centres in networks, development of uniform formats for information exchange between them. This practice showed that such approach gave rise to a set of hindrances in organization of such activity and often nullified results of such associations.

This paper focused on a method of convenient transfer the results of search procedures in VLIS into a local BD system of a subject Thermophysical Center as example. The most popular large information systems for science are considered.

* Corresponding author:
trachtengerts@yahoo.com(Michael Trachtengerts)
Published online at http://journal.sapub.org/ijis

2. Available Services

At the present time there are several very large information services (VLIS) available to scientists.

ISI Web of Knowledge (WoK) is an online academic database managed by Thomson Reuters (Thomson Reuters, Philadelphia, PA, USA)[1-5]. It provides access to scientific publishing databases and other resources: Web of Science (WoS), Science Citation Index, Current Contents, Medline, and some others. Its Web of Science database covers over 12 000 of the highest impact factor journals worldwide. The data bases cover also the most important books, conferences, and other editions.

The keystone DB is the highly selective multidisciplinary WoS which is maintained for over 50 years. The WoS consists of several multidisciplinary citation indexes with most valuable for us Science Citation Index Expanded, Index Chemicus and Current Chemical Reactions. Our attention is dedicated to the selection of the most important journals on thermodynamic and related fields included in the WoS. We found out that those journals are presented in the WoS journal list. Nonetheless, the highest standards for journals analysed by WoS lead to missing of some quality data that appear in second-ranked editions.

In addition to perfect key word search the comprehensive citation analysis is a remarkable feature of WoK. The cited references to articles in both sides, those in a paper and those on the paper, cover articles that are not found with chosen key words. They form a field of related links that

reveal missed information. Thermodynamics is a rather small field and the list of links may be checked with appropriate affords. In large fields the list numerous results may be refined in some different ways to receive a meaningful number of citations. This schema is particularly useful in evaluating new journals, for which a citation history does not yet exist.

Search of cited references in WoK is the most significant step in the new research process. Searching cited references for prior work on a subject is a way to avoid duplications and raise its effectiveness. As a whole, Thomson Scientific is the most important scientific information supplier. It focuses on the world's most important journals through the WoS, so that investigators will find the best, most relevant, and influential research.

The next important service for scientific information is Google Scholar [6, 7]. It indexes the full texts of academic and scholar publications and, opposite to Thomson Scientific, covers also low rank editions and some non peer reviewed journals. It has not such tools for information retrieve as Thomson's system suggests but it is free for all users. Google has features that are helpful for small and medium sized DB.

Google Scholar allows users to search for digital or physical copies of articles and indicates whether they may be downloaded in full text. Many of Google Scholar's search results link to commercial systems that charge the users for full text. In some cases Google Scholar is helpful to present links to alternative sources that do not demand fee and practice open access approach. The last issue of Google Scholar provides citation indexes that previously was the exclusive feature of WoS.

It is important to know that Google Scholar has the interface of a web browser. It permits to use many advance features of the Google Chrome browser which becomes the most popular. There are some of the most important:

— Search within a specific site that has valuable information for DB subject field;

— Search exact words in exact order by putting double quotes around a text;

— Use of + operator that blocks change of a single word;

— Use of * operator that shows substitutions only on whole words, not parts of words, as it applied in most browsers;

— Some rules of exceptions that help to throw out irrelevant documents;

— Search by author name, by terms in title, with data restriction to a period, and some other.

The returned results are in original format of document. So, they should be reformatted for import into a subject data base.

The next valuable source of data for science, especially for Russian readers, is the Science Electronic Library [9, 10]. It was started in 1998. Now it is the largest information portal in Russia for fundamental sciences, technologies, medicine, and education. Up to middle of 2012 it has about 16 millions articles, covers 32000 different journals collected worldwide, 2500 of them are Russian additions. More than 1500 Russian journals available Open Access.

The main section of user interface include:

— two lists of journal names for Cyrillic and Latin alphabets that link to a journal;

— Author Index with 4,5 millions names of authors, more than 550 thousands of them are scientists from Russia;

— link to full text search;

— the index of headings of science fields according to State classificatory of information;

— Subject Index with about 4 millions of key words;

— the index of more than 8000 publishers that can help authors to find a publisher for their manuscripts;

— New Journals Index that is important to find new sources in quickly rising world of science.

The registered user may receive personal cabinet. He can put there list of journal that he needs most often. The cabinet gives him place for saving some papers, cataloguing them in a different way, for storing queries applied before.

The system may inform a user about new journals received by eLIBRARY that have relevant data for queries saved in personal cabinet.

The most inconvenient feature of eLIBRARY for purposes of transfer data received is its output format that has no tags in presented records and makes reformatting to subject DB difficult. But sometimes there is no other way to receive rare data.

VLIS gives users not only valuable retrieved documents, but also provide results in various convenient standard formats. Text format is the most convenient for the further automated processing inside Data Centre. Therefore, there is a natural desire to use VLIS when it is possible as a pre-processor in a technological order for input information into a DB without being connected with attentions and plans of other data centres.

3. CDS/ISIS

Large-scaled numerical and bibliographic datasets in such subject field as thermophysics require a new work style. Today a typical scientist often copies numerous files to a local server and operates with datasets using his own resources. A subject DB with sufficient resources can significantly help him. But maintaining this DB update needs new approach too.

Increasing of the datasets is so large that it is much more economical to move the end-user's programs to the data in local systems. Science data centres provide access to both the data and the applications. Each of science centres manages massive datasets and makes affords to constantly adding and improving them.

The Thermophysical Center in Joint Institute for High Temperatures (JIHT) for a long period of time uses database management system CDS/ISIS developed in UNESCO[11, 12]. It was the main tool for developing of subject databases since its first versions.

CDS/ISIS is popular bibliographic information management software used in many countries, in Data Banks and government bodies. It may be integrated in the WWW with a number of tools now available. One of them is the freely available software, WWWISIS. This software is developed, maintained and distributed by BIREME, the Latin American & Caribbean Centre on Health Sciences Information. WWWISIS acts as a server for CDS/ISIS databases in a WWW client/server environment. It supports functions for searching, formatting and data entry operations over number of different databases.

Multi-base feature of CDS/ISIS is supported by menu driven tools and transforms it into generalized information storage and retrieval system. It allows to build and manage a number of structured non-numerical databases, i.e. databases whose major constituent is text. So, some numerical data can be included in a document being placed in separate fields. They are considered as text in alphabetical order. The search and retrieval system of CDS/ISIS is very powerful. It supports, among other search features field and subfield levels, proximity, right truncation and adjacency searching.

However, some users consider the search interface as not very user-friendly. A user has to get acquainted with the CDS/ISIS search and retrieval system to fully exploit the powerful features of the system. But it is quite natural for a system aimed to meet unexpected problems.

Table 1. Field descriptions in DB THERMAL

Tag	Field name	Subfield Index
001	Author	a
002	Title rus.	
003	Title ori.	
004	Journal	ab
005	Conference	
006	Abstract	
007	In archive	
008	Properties	a
009	Not used now	a
010	Phase	a
011	Not used now	a
012	Type of property	a
013	External field	a
014	Type of research	a
015	Chemical formula	a
016	Class of chemical	a
017	Internal number	
018	Paper type	
019	Language	
020	Year of publ.	
021		
022	Temperature low	
023	Temperature high	
024	Pressure low	
025	Pressure high	
026	Link to full text	

As a result, it was successfully applied to database on thermodynamic and thermophysical properties of substances which are of importance for power engineering and some other technologies.

According to ISIS scheme the local DB includes documents consisting of records in fields and subfields which can be repeated. For convenience of users the number of fields for search and display of the data was chosen small, and fields were considered as multipurpose. Therefore we chose information received with subscription from Web of Knowledge in plain text format with fixed set of fields as the main source of most important published data in the world.

Every DB managed by ISIS has distinct record schema described with Field Definition Table (FDT). The extract data schema for obtaining key words from specific fields and subfields described with Field Selection Table (FST). Both files are the most important keys in managing a DB in ISIS.

The records received from VLIS should be transformed in accordance with the FDT list of fields in our main DB THERMAL [13] which is shown in Table 1.

CDS/ISIS is capable to manage many different DBs with special field definitions of every one for various aims, for instance, as citation service [14].

4. Methods

Table 2. Tag Specifications for most VLIS output

Tag	Specification
PT	Publication Type (journal)
AU	Authors
TI	Document Title
SO	Publication Name
DE	Author Keywords
ID	Keywords Plus
AB	Abstract
PY	Year Published
VL	Volume
IS	Issue
PN	Part Number
SU	Supplement
SI	Special Issue
BP	Beginning Page
EP	Ending Page
AR	Article Number
DI	Digital Object Identifier (DOI)
UT	Unique Article Identifier
ER	End of Record

Transformation of the results received from VLIS for input into DB THERMAL is processing by recently developed program modules [15, 16]. The automatic analysis of records and their transformation are based on contents of fields that in VLIS identified by two-letter alphabetic tags. Many VLIS's use identical or close sets of tags in retrieval results. In Web of Knowledge for journal

articles and conferences there are 53 tags. Not all of them are of interest for our users, so we limited their list down to the following used fields in Table 2

The aim of transformation is to mark a VLIS record in a separate editor and to prepare it ready to transform it into international exchange format ISO 2709 which should be used every contemporary system for data import. ISIS has this feature and capable to import very large files of records.

We developed the routine for text transformation to ISO 2709 file named IsoWin [16]. It includes also a text editor that used by experts in scientific analysis of records. Comparing fields in Tables 1 and 2 one can see that a scientist in thermodynamics should be involved to extract properties of substances and chemicals and other subject's characteristics from general VLIS records and put them into right fields in THERMAL.

The retrieved result is pasted into editorial window of the IsoWin module "as is", that relieves of laborious manual processing. The unnecessary fields, some of them are shown below in the fragment from received data, are cut off automatically, see Table 3.

Table 3. Data fragment from VLIS search result

Tag	Field content
PTJ	
AU	Khaliullin, RZ; Eshet H; Kuhne, T D; Behler, J; Parrinello, M
TI	Graphite-diamond phase coexistence study employing a neural-network mapping of the ab initio potential energy surface
SO	PHYSICAL REVIEW B
ID	ELASTIC CONSTANTS; CARBON; EQUILIBRIUM; SIMULATION
AB	An interatomic potential for the diamond and graphite phases of carbon has been created using a neural-network (NN) representation of the ab initio potential energy surface. The NN potential combines the accuracy of a first-principles description of both phases with the efficiency of empirical force fields and allows one to perform a molecular-dynamics study, of ab initio quality, of the thermodynamics of graphite-diamond coexistence. Good agreement between the experimental and calculated coexistence urves is achieved if nuclear quantum effects are included in the simulation.
SN	1098-0121
PD	MAR
PY	2010
VL	81 IS 10
AR	100103
DI	10.1103/PhysRevB.81.100103
UT	ISI:000276248700003
ER	

The following step in data processing is transformation into an intermediate format [16] with the set of fields used in DB THERMAL. This fulfils by the "SCI to ISIS" module of IsoWin. As result appears in the same window of the editor, operator of data input or data expert can edit it if necessary, namely, to translate article title in Russian, to allocate additional characteristics, etc. Often information provided in the record is adequate for our purposes. After this step the ready records are turning onto ISO 2709 format by the specified module. Records in this format are ready for input in ISIS DB's and could not be changed. In spite of those different steps described here the whole data flow transformations can be easily fulfilled by the same routine IsoWin [16]:

The Table 3 shows a record received from Web of Knowledge.

Thus, some work of an expert evaluating general information received from VLIS nevertheless appears necessary according to features of the subject DB.

The detailed description of intermediate format record is provided in [16], an example is presented in Table 4. In short, every string in field content should end with sign "~" that points to jump into next field. Pairs of letters like "^A" allow to use subfields that is a tool to apply various print formats for different purposes. As the main aim of THERMAL is to supply scientists and technologists in Russia some of the fields assume Cyrillic alphabet. Fields 022-025 can be used for numerical values of temperatures and pressures in measurements mentioned in article body.

Table 4. The same data in an intermediate THERMAL format

Tag	Field content
@	
001	^AKhaliullin R.Z.^AEshet H.^AKuhne T.D.^ABehler J.^AParrinello M.~
003	Graphite-diamond phase coexistence study employing a neural-network mapping of the ab initio potential energy surface~
004	^APHYSICAL REVIEW B ^B2010, V.81, No 10, AR: 100103~
006	An interatomic potential for the diamond and graphite phases of carbon has been created using a neural-network (NN) representation of the ab initio potential energy surface. The NN potential combines the accuracy of a first-principles description of both phases with the efficiency of empirical force fields and allows one to perform a molecular-dynamics study, of ab initio quality, of the thermodynamics of graphite-diamond coexistence. Good agreement between the experimental and calculated coexistence curves is achieved if nuclear quantum effects are included in the simulation.~
020	10~
005	Конференция~
007	Находится в ~
010	^Афаза~
011	^Афазовый переход~
015	^Аформула в-ва~
016	^Акласс в-ва~
008	^Асвойство~
012	^Атип св-ва~
017	ТС Номер ТФЦ~
019	анг~
022	-00001.00 K,T_ниж~
023	-00001.00 K,T_вер~
024	-00000100.00 бар,P_ниж~
025	-00000100.00 бар,P_вер~
026	..\LibPDF\Имя файла~
027	^AAffiliation~
$$$	

The sign "@" shows the start point of new record of a document, string "$$$" shows the end of document.

The fields in the example below represent templates for editing in specific fields of THERMAL. They may be presented in Russian or in any other language as well. Expert can remove some of them if the relevant information is absent. The special attention is given to a field 026, in which link to the full text of article or other document placed if available. Full texts are collected in library section of the THERMAL DB. The user when sees "Full Text" sign can obtain and download the necessary article.

The library of full texts is in a special directory LibPDF because most of the papers presented in PDF format. The field 026 includes a key to write the link to full text.

One of the important features in using of VLIS by a subject data center is the need of elimination of duplicates in input information flow. At first it was supposed that duplications can be avoided by carrying out searches in VLIS with appointed time interval, for example, carrying out searches for the last week. However, it appeared that the considerable volume of useful information is lost in such a way, as retrieving formulas to VLIS usually somewhat incomplete than it is necessary to subject DB scope. These losses practically decrease with using such additional functions in VLIS, as search of citations and other relevant documents for key publications (the *related* function in Web of Knowledge system as example). In these functions time control of the publication is absent, and duplicating documents in input flow to DB becomes inevitable.

Table 5. Fragment of the auxiliary file

First Author	Article identifiers
Bruno T.J.	
	J. Res. Natl. Inst. Stand. Technol. 000549
	1994, V99, N 3, p. 263-266.
Brygoo S.	
	NATURE MATERIALS 003806
	2007, V.6, No 4, pp 274-277.
	NATURE MATERIALS 004029
	2007, V.6, No 4, pp 274-277.
Bryk Taras	
	THE JOURNAL OF CHEMICAL PHYSICS. 004180
	2010, 132, 074504.
Bucknum M.J.	
	JOURNAL OF THE AMERICAN CHEMICAL SOCIETY 004064
	1994, V.116, No 25, pp 11456-11464.
	MOLECULAR PHYSICS 004041
	2005, V.103, No 20, pp 2707-2715.

In the Thermophysical centre the duplication control is processed manually since we could not develop or find and apply a reliable automatic technique. Because of relatively small volume of DB documents the search for duplication was fulfilled with name of the first author in every input to document system, and duplication easily came to light. However such search was carried out inconveniently often and needed access to database. Therefore, the other method is now applied. For each current DB state an auxiliary file of identifying indicators of the documents is created. It is sorted by first author of publications. The fragment of such file for DB ТЕРМАЛЬ is shown in Table 5.

Every author in this list is followed by journal name, serial record number in local DB, and in the next line by year of publication and numerical features of the edition. This is enough to identify duplicate publications. In the shown fragment duplication in records 3806 and 4029 of first author S.Brygoo, one of which should be removed from DB, easily comes to light.

Text editors with *find* command are available in all operating systems that allow to easily establish duplications in input flow to a DB.

5. Conclusions

We demonstrated in this paper one of the means to achieve better results in data collecting for a scientific thematic Data Centre. The existing very large information services possess main bulk of information that is valuable to any special Data Centre and commonly could be received by search on subscribe. But it should be extended for specific features of a science field. The approach was applied in the Thermophysical Centre and reduced essentially efforts to maintain DB updates with documents from high rank editions. It seems to be effective for DB's with number up to 100 000 documents.

ACKNOWLEDGEMENTS

I wish to thank the members of the Thermophysical Center: Prof. G.Kobzev, Dr. V.Zitserman, Dr. A.Erkimbaev.

REFERENCES

[1] R. Yancey, "Fifty years of citation indexing and analysis", KnowledgeLink Newsletter from Thomson Scientific, September, 2005. Available on-line: http://www.scientific.thomson.com/newsletter

[2] J. Testa, "The Thomson Scientific journal selection process", International Microbiology, no. 9, pp. 135-138, 2006.

[3] K. O'Connor, "Web of knowledge", Australian Family Physician, vol. 36, no. 7, July, 2007.

[4] Available on-line: http://apps.webofknowledge.com/

[5] L.I. Meho, K. Yang, "New Era in Citation and Bibliometric Analyses: Web of Science, Scopus, and Google Scholar", Journal of the American Society for Information Science and Technology, vol. 58, no. 13, pp. 2105–2125, 2007.

[6] Finding articles with Google Scholar, Guide, Metropolitan State University, Available on-line: http://lgdata.s3-website-us-east-1.amazonaws.com/docs/1181/220467/google_scholar.pdf

[7] Google Scholar Basics, Walden University, January 25, 2012
Available on-line: http://www.waldenlibrarynews.com/blog/2012/1/25/google-scholar-basics.html

[8] Online Available: http:// http://portal.unesco.org/ci/en/ev.php-URL_ID=5330&URL_DO=DO_TOPIC&URL_SECTION=201.html

[9] Available on-line: http://elibrary.ru/

[10] Е.М. Полникова, С.М. Шабанова, "Научная Электронная Библиотека elibrary.ru: Руководство пользователя", ООО «РУНЭБ», Санкт-Петербургский государственный университет, 2010

[11] Available on-line: :http://www.unesco.org/webworld/isis

[12] Teh Kang Hai, Wong Sau Foong, "Developing a CDS/ISIS-based Online Cataloguing and Information Retrieval Interfaces for Use in Small Libraries", Malaysian Journal of Library & Information Science, vol. 1, no.1, pp. 1-20, July 1996.

[13] M.S. Trachtengerts, "The New Effective Tool for Text Data Bases — CDS/ISIS for Windows", VINITI Publisher, Scientific and Technical Information, Ser. 2, Information Processes and Systems, no. 6, pp. 30-33, (in Russian), 2006.

[14] S.N. Sinnarkar, "Development of an Institutional Database of Citations using CDS-ISIS", Software. Annals of Library and Information Studies, vol. 50, no. 4, pp. 153-155, 2003.

[15] M.S. Trachtengerts, "THERMAL — the database in Thermophysical Center, JIHT, Russian Academy of Sciences", in Proceedings of XI-th Russian Conference on Thermophysical Properties of Substances, St. Petersburg, October 4-7, (in Russian), 2005.

[16] M.S. Trachtengerts, "Technology for Preparation of Documents to Databases in the Exchange Format ISO 2709", VINITI Publisher, Scientific and Technical Information, Ser. 2, Information Processes and Systems, no. 7, pp. 28-31, (in Russian), 2006.