Chemical Database: a case of chemical data integration via unique identifier

Zhao Yuehong, Xu Junbo, and Wen Hao

(Institute of Process Engineering, Chinese Academy of Sciences, Beijing 100190)

Abstract

To integrate the distributed, heterogeneous chemical data resources purposed-built for collecting data within a particular subdomain in chemistry is urgently needed to meet comprehensive data requirement by research and engineering. However, the warehousing and federation models are impractical under the condition of different database ownership and various DBMS. Instead, the mediator approach is a good choice, in which the data integration is built on top of the data access services provided by database owners, therefore, differences in DBMS, data type and formats are hidden behind the services. The essential of this method is to access chemical information with proper compound identifier accepted by all the databases[1].

Currently, chemical information stored in our chemical databases, owned by different institutes of Chinese Academy of Sciences, is annotated by the unique internal identifier, normally index number, which only works in the local database system as a clue for data access and identification. On the other hand, some unique external identifiers, i.e. CAS registry number, and unique structural representations are collected as attached information of a compound which can be easily accessed via local identifier. Considering these external identifiers are widely used as clue for compound data access and integration for distributed chemical data resources[2-4], they are good choice for open data access over Internet even incomplete or no external identifiers for some compounds in our databases. In order to make it easy to use over Internet, sizes of identifiers should not be too large, then CAS registry number (max. 10 digital numbers) [5] and standard InChIKey (27 characters)[6] could be the good candidates, especially for standard InChIKey, because it is derived from the structural information of a compound by the independent program. However, for a chemical database with long history, it often means that much effort needed to

modify original data access program and to update databases if other compound identifiers were used to replace the local identifier. To overcome this obstacle, an approach involving setting up a table containing the local and two external identifiers was proposed in this paper, wherein the table works as a bridge between the external identifiers and the local identifier. By this approach, the open data access service using unique external identifiers as parameter can directly call the previously developed data access functions only with a cost of query of the identifiers mapping table.

In this paper, the integration of distributed chemical data resources within Chinese Academy of Sciences was carried out on the basis of above mentioned approach, and a series of open data access services and a table for identifiers mapping were developed for this objective. By combining these services with the basic data grid services, such as resource registration and management, user management, etc., a grid-like infrastructure as shown in Figure 1 was set up to facilitate the distributed data integration. To solve the problem of a compound without CAS register number and structural information as well, the compound registry service is used to help obtain such message by the formula and names of a compound. The key feature of the system is to retrieve the properties data of the same compound contained the distributed, heterogeneous databases via data access services, which make it easy for user to build new data model and high-level services based on them. Based on such infrastructure, an integrated chemical data query system, Chemical Database (http://www.chemdb.csdb.cn), was developed. It provides an interface for user to query with widely used CAS register number, Chemical formula, name, properties or structure as query word to get unique identifier for retrieving properties of a compound from multiple distributed databases via data access services. Now information of a compound from three institutes of Chinese Academy of Sciences has been integrated successfully, covering the major chemistry research subdomains, such as engineering chemistry data from IPE[1], organic chemistry data from SIOC[2] and applied chemistry data from CIAC[3]. It enables us to seamlessly query the related data of a compound all from a single interface. The query result of phenol (C7H6O2, 65-85-0) was shown in Figure 2 as an example. The result shows that the distributed

[1] IPE Institute of Process Engineering, Chinese Academy of Sciences
[2] SIOC Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences
[3] CIAC Changchun Institute of Applied Chemistry, Chinese Academy of Sciences

chemical data resources can be effectively integrated in a scientific meaningful way, whereas very little additional work is needed for data services providers. At same time, open data access will help realize direct data access if XML format were used for data transfer, making data integration and interoperability more flexible between different platforms. Furthermore, cooperative work environment for chemistry research can be set up by integrating tools and services in the domain of chemistry based on the current work.

Key words: Chemical data; Compound identifier; Data integration

References

[1] T.I. Oprea, A. Tropsha. Target, Chemical and bioactivity databases – integration is key. Drug Discovery Today: Technologies, 2006, 3:356.

[2] L.L. CHepelevl, M. Dumontier. Chemical Entity Semantic Specification: Knowledge representation for efficient semantic cheminformatics and facile data integration. Journal of Cheminformatics, 2011, 3:20.

[3] R. Grossman, P. Kasturi, D. Hamelberg, et al. An empirical study of the universal chemical key algorithm for assigning unique keys to chemical compounds. Journal of Bioinfromatics and Computational Biology, 2004, 2:155-171.

[4] Taylor K., Gledhill R., Essex J.W., et al. A semantic Datagrid for Combinatorial chemistry Grid computing workshop 2005. From http://eprints.ecs.soton.ac.uk/11778/1/semanticdatagrid.pdf, 4/20/2012.

[5] American Chemical Society. CAS Registry and CAS Registry Numbers. From http://www.cas.org, 5/2/2012.

[6] IUPAC. IUPAC International Chemical Identifier. From http://old.iupac.org, 5/2/2012.

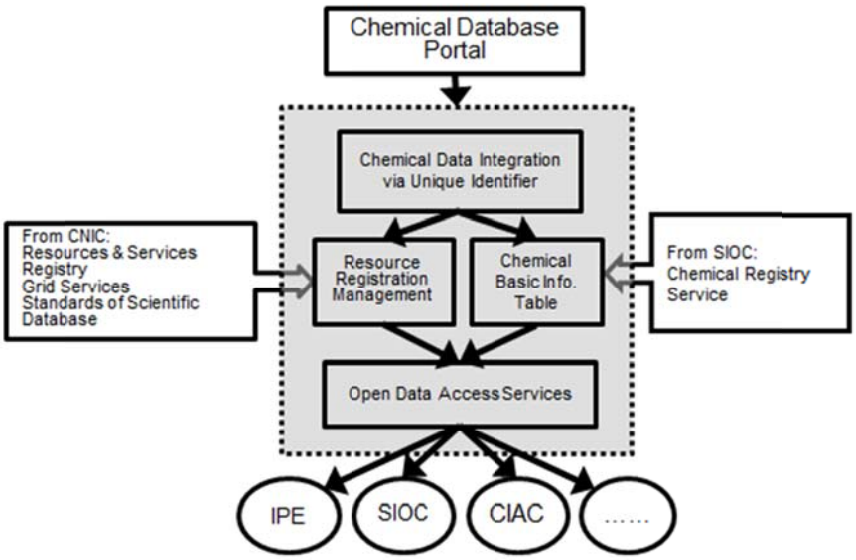

Fig. 1 Basic infrastructure of Chemical Database System

Fig.2 Integrated Chemical Data for benzoic acid (C7H6O2, 65-85-0) via Unique Identifier