Research note

Federation of Ecological Data Repository of EAP-ILTER

Chau-Chin Lin,[1,3] Guan-Shuo Mai,[2] Sheng-Shan Lu[1]

【Summary】

Metadata Catalog (Metacat) is a tool that has been adapted independently by EAP-ILTER member networks since 2004. Metacat is a repository for data and metadata, which helps scientist find, understand and effectively use such data sets. The research data sets are currently documented in these Metacats in a standardized way and provide the scientific community these data within each individual network. The Metacat framework enables efficient network-based discovery and access to such data. Although the current Metacat servers are installed individually by member networks of EAP-ILTER, they can be connected to be a data confederation system through web services. We proposed a Metacat federation framework to connect EAP-ILTER regional data-sharing using two web services technologies, SOAP and REST. We tested this framework using Metacat servers of KNB, TERN, JaLTER and FRIM. The results indicates that this framework not only allows EAP-ILTER member networks that have a Metacat server to readily join the services with their current infrastructure but also can share the data of other regions. In addition, a multilingual controlled vocabulary translating engine provides the possible usage of retrieving metadata documents stored with different local languages in the future.

Key words: LTER,metadata,data sharing,Metacat,federation.

Lin CC, Mai GS, Lu SS. 2016. Federation of ecological data repository of EAP-ILTER. Taiwan J For Sci 31(4):337-42.

[1] Forest Protection Division, Taiwan Forestry Research Institute, 53 Nanhai Rd., Taipei 10066, Taiwan. 林業試驗所森林保護組，10066台北市南海路53號。

[2] Biodiversity Research Center, Academia Sinica, No. 128, Sec. 2, Academia Rd. Nankang, Taipei, Taiwan. 中央研究院生物多樣性研究中心，11529台北市南港區研究院路二段128號。

[3] Corresponding author, e-mail:chin@tfri.gov.tw 通訊作者。

Received April 2016, Accepted May 2016. 2016年4月送審 2016年5月通過。

研究簡報

亞太長期生態研究網之生態資料聯合查詢

林朝欽[1,3] 麥舘碩[2] 陸聲山[1]

摘 要

元數據目錄自2004年起即為亞太長期生態研究網的成員選為工具之一，它是一套生態研究數據倉儲的系統，協助研究人員有效搜尋及了解研究數據集。數據以標準化格式詳細描述以儲存在此系統中，目前亞太長期生態研究網的成員各自分別架設獨立的系統提供研究人員使用。此系統可以透過內建的網路服務功能共享資料，因此本研究提出聯合各獨立的元數據目錄架構，並以美國、日本、台灣及馬來西亞四個倉儲系統測試此架構。此外，本研究並建立多語言的關鍵字控制詞，讓不同的研究人員可以使用自身母語查詢而獲得各元數據目錄的數據。

關鍵詞：長期生態研究、元數據、數據分享、元數據目錄、聯合。

林朝欽、麥舘碩、陸聲山。2016。亞太長期生態研究網之生態資料聯合查詢。台灣林業科學31(4):337-42。

Scientists who work in Long Term Ecological Research (LTER) use a wide variety of protocols to collect data on complex topics such as forest dynamics and carbon flux (Jones et al. 2001, Lin et al. 2008). These heterogeneous data sets are stored in autonomous database systems dispersed through out the LTER community. In order for these data to be networked and preserved of for future studies, including their reuse in replicating and validating scientific conclusions, a metadata-driven framework that enables rapid, powerful access and discovery of such data has been developed by the National Center for Ecological Analysis and Synthesis (NCEAS) in the US since 2001 (Leinfelder et al. 2010). In this framework, called Metacat (short for "metadata catalog"), scientists describe data syntax and semantics using metadata vocabularies defined by domain communities. The system serializes the metadata using eXtensible Markup Language (XML) and stores the documents in a schema-independent XML database (Berkley et al. 2001). One advantage of this metadata-driven model employed by the Knowledge Network Biocomplexity (KNB) project is that it easily supports data storage without prescribing a particular serialization mechanism or imposing structural constraints on data files.

Since 2005, the Eastern Asia and Pacific International Long Term Ecological Research (EAP-ILTER) has proposed to share metadata within the community by using Ecological Metadata Language (EML), which is an XML format metadata specification that allows data owners to preserve their original data format by describing it rather than conforming to a standardized schema (Lin et al. 2006). Scientists from the Taiwan Forestry Research Institute (TFRI), the agency that hosts the Taiwan Ecological Research Network (TERN), have worked with their US counterparts since 2004 to adapt Metacat and EML for EAP-ILTER and organized three workshops to promote the usage of these ecoinformatics tools. Currently, Metacat has been established and used by Japan LTER, Taiwan LTER, and Malaysia

LTER, respectively.

Although Metacat solves key challenges that impede data management efforts in an individual LTER community, it still has a problem with data-confederation within the regional scale. Fortunately, The KNB project has designed the EcoGrid which uses web services technologies to provide access to disparate data on different networks and storage systems (Michener et al. 2005). The EcoGrid allows scientists access to a wide variety of data and analytic resources such as data, metadata, analytic workflows and processors networked at different sites and at different organizations via the internet. Because Metacat uses HyperText Transfer Protocol (HTTP), it supports the distributed EarthGrid to query and retrieve metadata and data from different Metacat servers. Here we describe a Metacat federation framework that addresses the use of web service technologies to consolidate disparate long-term ecological data.

The Metacat system is controlled by a Java servlet that communicates using HTTP protocol. The servlet acts as a dispatcher, passing commands and data from client applications to the various subsystems that handle Metacat's functions. There are five main subsystems of the servlet: storage, replication, query, validation, and transformation (Jones et al. 2001). With the exception of these five main subsystems, Metacat handles authentication through a generic interface that runs the Lightweight Directory Access Protocol (LDAP). Data storage, query, replication, and transformation functions are mediated via the Metacat server, which has interfaces to metadata housed in the Relational Database Management System (RDMS) , and authentication services (LDAP).

Currently, Metacat is an open-source component, and individual networks or sites can set up independently, extend and customize the system to support their data and metadata needs. Since Metacat uses HTTP protocol, many different styles can be used to integrate or share Metacat data such as replication of database, remote procedure call, and message bus and file transfer. Among these four alternatives, remote procedure call and message bus are referred to as web services (Pautasso et al. 2008).

Conceptually, web services are software components provided through a network-accessible endpoint. The service consumer and provider use messages to exchange invocation request and response information in the form of self-contained documents. Technically, there are two schools of web services: Simple Object Access Protocol (SOAP) and Representative State Transfer (REST) (Vinoski 2002, Pautasso 2008). SOAP is a standards-based approach developed at Microsoft in 1998. The SOAP systems must include the following items:

1. A formal contract must be established to describe the interface that the web services offers. A Web Services Description Language (WSDL), an XML language for defining interface file, must be provided.
2. The architecture must address complex nonfunctional requirements.
3. The architecture needs to handle asynchronous input and output.

In contrast, REST is a simpler and trendier new approach in that each unique Universal Resource Locator (URL) is a representation of some object. Users can obtain the contents of that object through an HTTP command.

SOAP has been designed as a grid-service programming interface in Metacat that accomplishes seamless access to data via standardized service. In addition, the Metacat HTTP interface supports REST via a variety of actions that facilitate information query,

retrieval and storage. HTTP requests can be sent from any client application that communicates using the Web's HTTP protocol. For example, the Metacat query subsystem provides two kinds of query: structured query action (squery) and query action (query). The difference between squery and query is that squery uses custom query syntax in xml format named pathquery. Query action, however, uses parameters that pass as url-encoded form. Therefore, the federation of different Metacat systems can be set up through these two approaches of web services.

Figure 1 displays the conceptual model which applies SOAP and REST approaches for Metacat data federation. Once Metacat is set up, it directly enables both SOAP and REST functions. Metacat builds a REST function automatically. For SOAP function, Metacat requires manual set up and publishing of the services. Once Metacat is set up and releasing services, path-query syntax can be built for precise queries against arbitrary metadata and a controlled vocabulary file for language translation. Using the single query approach of Metacat's query subsystem, a multi-thread simulation class has been adapted to retrieve metadata stored in different Metacat servers simultaneously. Then a wrapper brings the results from either type of services together to one interface.

Software implementation of the Metacat federation framework was developed by using PHP programming language. PHP originally meant Personal Home Page, and is a self-referential acronym for Hypertext Processor. It is an open-source, server-side, HyperText Markup Language (HTML) embedded scripting language to create dynamic web pages (Tsenov 2006).

Due to the lack of multi-thread and multi-process for the Windows system of the basic PHP system, we introduced the multi-thread simulation class developed by Alex Lau for parallel query and data retrieving. Figure 2 shows the interface of our PHP design. The interface collected KNB and five defaults of Metacat servers in Taiwan, leaving all possible Metacat servers to be added by users. We tested KBN, TFRI, JaLTER and Malaysia LTER metacat servers with the following results.

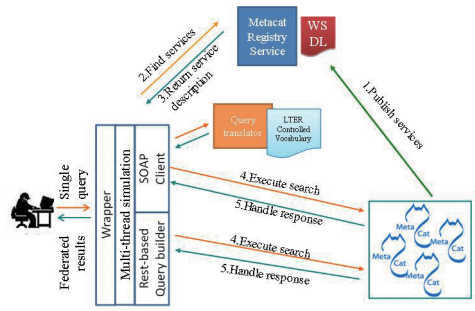

Fig. 1. The conceptual framework of Metacat federation. The Metacat server provides either SOAP or REST services and is registered to be found by execute search.

Figure 3 shows the query result by using "Forests" as the keyword. The result returned 449 EML documents that are currently stored in these five Metacat servers either in English, Chinese characters, or Japanese characters. The 449 EML documents include 198 from KNB, 35 from TFRI, 21 from JaLTER, and 0 from FRIM. In addition, KNB also harvests TFRI's EML documents. One EML document was also harvested in KNB .

In order to determine the nature of the problem of not searching no FRIM's EML document, we first checked FRIM's server individually using URL-encoded form and found 12 EML documents under the keyword "Forests"; however, from our interface, 0 records could be queried. The problem was caused by the old version of the Metacat server.Apparently, the old version doesn't support

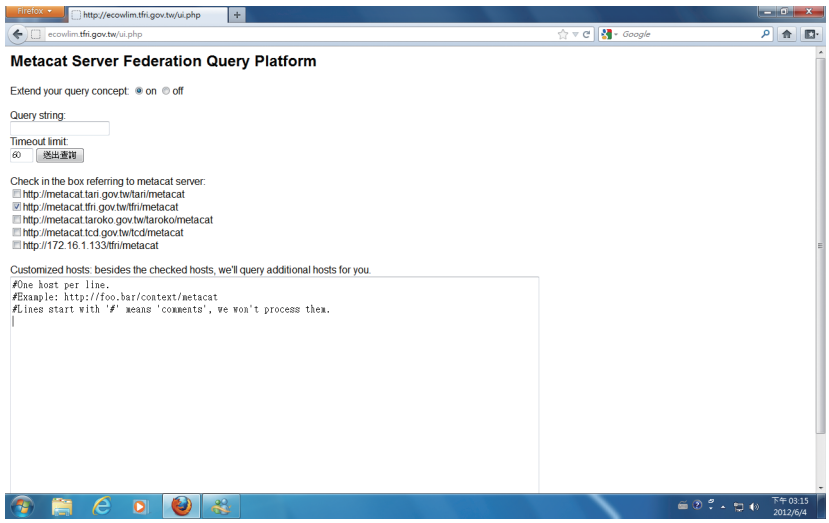

Fig. 2. The PHP interface of the Metacat federation query platform.

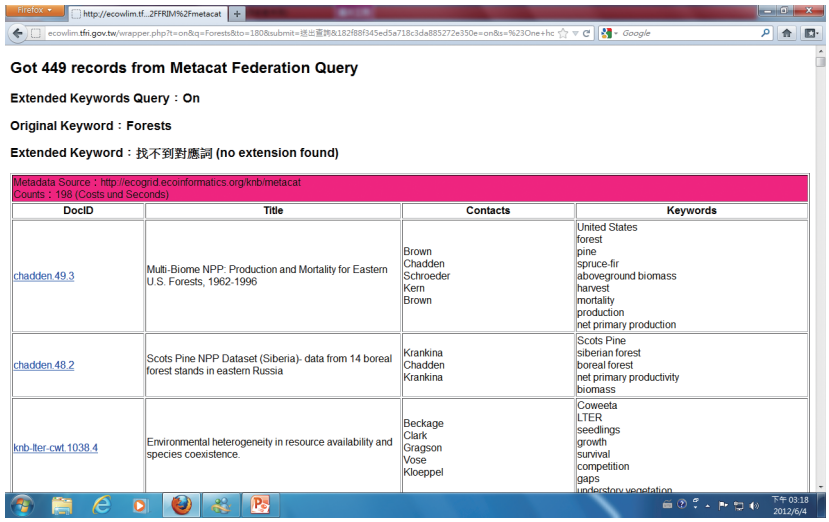

Fig. 3. A snapshot of a query result from five Metacat servers.

structured query action. We then harvested the 12 EML documents and inserted them into our local Metacat server. They can now be searched in the updated Metacat server. Therefore, the interface we designed needs to add the limitation of use to notify users who try to query the old version of Metacat lower than version 1.9.x.

We have proposed a data sharing among distributed network systems that use the same ecoinformatics tools. Our proposed method is based on the open-source standards REST and SOAP. Software solution using PHP programming language based on the conceptual web services model is presented. Four EAP-ILTER data catalogs have been tested. The result shows that data sharing within the community can be easily achieved. Furthermore, the federation platform can also query other regions of Metacat servers if exiting online.

Future work will focus on the problem of integrating and analyzing the data sets that were queried and retrieved from different LTER networks.

LITERATURE CITED

Berkley C, Jones MB, Bojilova J and Higgis D. 2005. Metacat: a schema-independent XML database system. Proceedings of the 13th International Conference on Scientific and Statistical Database Management. IEEE Computer Society.

Jones MB, Berkley C, Bojilove J and Schildhauer M. 2001. Managing scientific metadata. IEEE Internet Computing 5:59-68.

Leinfelder B, Tao J, Costa D, Jones MB, Servilla M, O'Brien M and Burt C. 2010. A metadata-driven approach to loading and querying heterogeneous scientific data. Ecological Informatics 5:3-8.

Lin CC, Porter J and Lu SS. 2006. A metadata-based framework for multilingual ecological information management. Taiwan Journal Forestry Science 21(3):377-82.

Lin CC, Porter J, Lu SS, Jeng MR and Hsiao CW. 2008. Using structured metadata to manage forestry research information: a new approach. Taiwan Journal Forestry Science 23(2):133-43.

Michener W, Beach J, Bowers S, Downey L, Jones MB, Ludascer B et al. 2005. Data integration and workflow solution for ecology. Proceedings of the Second international conference on Data Integration in the Life Sciences p. 321-4.

Pautasso C, Zimmermann O and Leymann F. 2008. RESTful web services vs. big web services: making the right architectural decision. 17th International World Wide Web Conference (WWW 2008), Beijing, China.

Tsenov M. 2006. Web services example with PHP/SOAP. International Conference on Computer Systems and Technologies.

Vinoski S. 2002. Putting the web into web services. IEEE Internet Computing 6:90-2.