TECHNICAL REPORT   **Open Access**

Progress of the IUGONET system - metadata database for upper atmosphere ground-based observation data

Shuji Abe[1*], Norio Umemura[2], Yukinobu Koyama[3], Yoshimasa Tanaka[4], Manabu Yagi[5], Akiyo Yatagai[2], Atsuki Shinbori[6], Satoru UeNo[7], Yuka Sato[4] and Naoki Kaneda[7]

Abstract

Background: The Interuniversity Upper atmosphere Global Observation NETwork (IUGONET) project is a 6-year research project which started in 2009. The objective of this project is to establish a metadata database of various ground-based observation data covering a wide region from the Sun to the Earth; this will encourage more studies on the mechanisms of long-term variations in the upper atmosphere.

Findings: For archiving purposes, the metadata database system for cross-searching various data distributed across many universities and institute was developed based on the existing repository software called DSpace as the core component and the Space Physics Archive Search and Extract (SPASE) data model as the metadata format. The IUGONET metadata database is still in operation since it was released in March 2012. The system is continuously examined, tested, and updated to improve its quality. The OpenSearch interface in the IUGONET metadata database allows the user to use external applications easily for exchanging metadata and/or for analyzing data.

Conclusions: We conducted self-examination of our product, which was added for planning future directions of the IUGONET project.

Keywords: Metadata database; Analysis software; Upper atmospheric physics; Earth and planetary sciences; Ground-based observation

Findings

Introduction

In order to understand long-term changes in the Earth's atmosphere, it is essential to discuss the various atmospheric layers as a coupled system and not to regard them as separate layers. The upper atmosphere, the focus of this paper, is defined as the region above about 50 km altitude and consists of six layers, namely the mesosphere, thermosphere, ionosphere, plasmasphere, magnetosphere, and heliosphere. This region is affected by the input of materials, momenta, and energies from the upper region (e.g., ultraviolet radiation from the sun and the electromagnetic energy from the solar wind) and from the lower region (e.g., atmospheric waves from the stratosphere and

troposphere). In addition to the vertical coupling processes, it is also important to consider the meridional coupling in the region that covers the equatorial, low, middle, and high latitudes.

The upper atmosphere is characterized by the coexistence of both ionized plasma and neutral gas and also by the drastic changes in the physical quantities across the layers (i.e., density, pressure, temperature, etc.). To clarify the physical mechanisms of the phenomena in the upper atmosphere, therefore, it is necessary to comprehensively analyze various types of physical quantities observed in the multiple layers. However, it is often difficult for researchers of different fields to get from a single source the information of the observed data, for example, physical quantities, instruments, observatories, contact persons, location, and format of data files.

---

* Correspondence: abeshu@icswse.kyushu-u.ac.jp
[1]International Center for Space Weather Science and Education, Kyushu University, 6-10-1, Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan
Full list of author information is available at the end of the article

In order to resolve the difficulty and promote an interdisciplinary study, the Interuniversity Upper atmosphere Global Observation NETwork (IUGONET, http://www.iugonet.org/, Accessed 7 Oct 2014) project was initiated in 2009 by five Japanese universities and institutes, namely, Tohoku University, Nagoya University, Kyoto University, Kyushu University, and National Institute of Polar Research (Hayashi et al. 2013). To date, there are some institutes involved in collaborations with the IUGONET project, for example, National Institute of Information and Communications Technology, National Astronomical Observatory of Japan, Kakioka Magnetic Observatory Japan Meteorological Agency, and more. These universities and institutes have developed a worldwide ground-based observation network of the upper atmosphere. Figure 1 shows the locations of the observatories related to the IUGONET project, distributed all over the world. This project has developed two tools: one is a metadata database for cross-searching the atmospheric data distributed across different universities and institutes and the other is an analysis software for visualizing and analyzing data (Tanaka et al. 2013). It is obvious that many projects have specialized data processing and products for their own data (e.g., Olsen et al. 2013). On the other hand, IUGONET is a challenging approach because it handles a variety of observational projects and related components, for example, instrument, observatory, person, data itself, and so on, in single metadata format and database structure.

This paper particularly focuses on the metadata database system developed by IUGONET. Background of the IUGONET metadata database system describes the fundamental policy. In the 'Operation and improvement of the IUGONET metadata database system' section, the daily operation, maintenan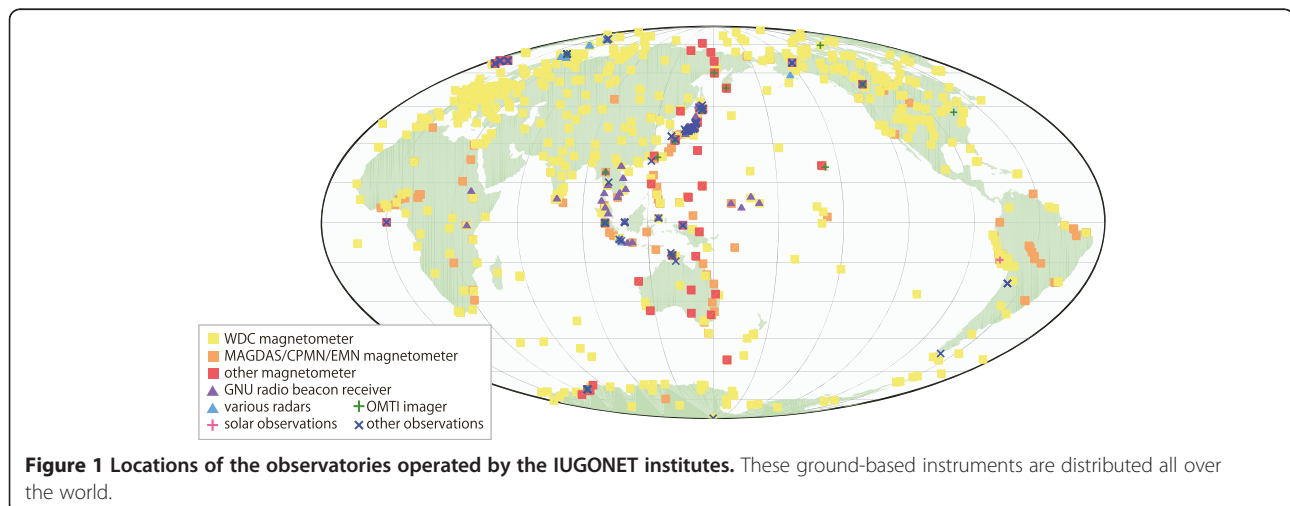ce of the system, some evaluations of our product, and the conjunction with data analysis software are described. In the 'Discussion and future efforts' and 'Conclusions' sections are the discussion and summary, respectively.

Background of the IUGONET metadata database system

To avoid problems like ownership of data, authentication, and authorization, the observational data is managed without a central server. The metadata database is built as a virtual integrated database environment to share the metadata of ground-based observational data including the uniform resource locator (URL) of data file. However, problems were faced before each organization released the data to the public, such as lack of human resources for the implementation of the database and the development members' composition mostly inclined toward the specialists in upper atmospheric physics only.

Another problem was the project timeline. The project timeline was not based on the detailed plan of the system designer. Therefore, there was not enough time to design and implement a metadata database system suitable for natural science.

Under such restrictions, we adopted DSpace (DuraSpace, DSpace, http://www.dspace.org/, Accessed 5 Oct 2014), a free software, which wrapped Apache httpd, PostgreSQL, Tomcat, etc., as a metadata database system. We also paid attention to the following minimum requirements to build the metadata database system. Easy technical information sharing on system installation, customization, and management is one of the reasons for DSpace adaptation. The total number of DSpace worked as an institutional repository around the world is about 2,500 (DuraSpace, DSpace User Registry, http://registry.duraspace.org/registry/dspace/, Accessed 5 Oct 2014) and in Japan it is about 300 (National Institute of Informatics, NII

Figure 1 Locations of the observatories operated by the IUGONET institutes. These ground-based instruments are distributed all over the world.

Institutional Repositories Program, http://www.nii.ac.jp/irp/list/, Accessed 5 Oct 2014). Most institutional repositories are managed by university libraries. Therefore, the institutional repository is also managed in the library of the institution which has participated in the IUGONET project except for NIPR. High interoperability is also one of the reasons for DSpace adaptation. DSpace supports the common interoperability standards used in Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) (Open Archives Initiative, The Open Archives Initiative Protocol for Metadata Harvesting, http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm, Accessed 5 Oct 2014), Search/Retrieve via URL/Search/Retrieve Web Service (SRU/SRW) (Library of Congress, Search/Retreval via URL, http://www.loc.gov/standards/sru/, Accessed 5 Oct 2014), OpenSearch (A9.com, Inc., OpenSearch, http://www.opensearch.org/Home, Accessed 5 Oct 2014), etc. These web application programming interfaces (APIs) are compatible with external systems like databases and data analysis software. Scalability is another reason for DSpace adaptation. If the metadata records become large, it is difficult to deal with them by a single server. Therefore, we are investigating the system structure again (e.g., distributed model). The cross-searching using external interface, such as OpenSearch, makes it possible to form multi-database connections like the relationship between National Diet Library (NDL) Search (National Diet Library, NDL Search, http://iss.ndl.go.jp/, Accessed 5 Oct 2014) and Institutional Repositories in Japan (DuraSpace, DSpace User Registry, http://www.nii.ac.jp/irp/list/, Accessed 5 Oct 2014). The IUGONET metadata database is running under DSpace 1.7.0.

Concerning the metadata format, there was not enough time to define an original metadata format for the project. Several existing metadata formats were investigated, and as a result, the Space Physics Archive Search and Extract (SPASE) was chosen for the base metadata format (Thieman, J. R., Welcome to the SPASE Group, http://www.spase-group.org/, Accessed 5 Oct 2014). SPASE is suitable for the IUGONET project because it is closely related to Solar Terrestrial Physics (STP) and upper atmosphere researches. In addition, SPASE has a scalability format. We can append new metadata elements and terms for our data. In fact, we appended some modifications to the SPASE format, for example, additional terms to represent non-digital archives, additional terms to represent heliospheric coordinates, and new metadata elements to describe observation location and range. SPASE is written in XML format, but DSpace 1.7.0 cannot handle XML format directly. To solve this problem, a SPASE Dublin-Core converter was developed for the IUGONET metadata database (see 'Operation and improvement of the IUGONET metadata database system' section).

## Operation and improvement of the IUGONET metadata database system

Routine operation and maintenance of the system

Since February, 2011, the main system and stand-by system of the IUGONET metadata database have been running at Kyushu University and Nagoya University, respectively. The metadata XML files provided by the IUGONET members are also stored in these universities.

The metadata is released through a procedure described in the following steps: (1) create the metadata and upload it to the temporary repository by the metadata providers, (2) automatically check the metadata by the metadata checker, and (3) register the metadata to the metadata database. As for step 1, the metadata providers create the metadata by using Eclipse (The Eclipse Foundation, Eclipse, https://www.eclipse.org/, Accessed 16 Oct 2014), text editor, batch program, etc., and put them in a temporary repository called as 'draft repository'. We adopted Git (Hamano, J. C., Git, http://git-scm.com/, Accessed 16 Oct 2014), which is a free and open-source distributed version control system, for the repository. The draft repository is divided into each metadata provider to prevent the mixing of the XML files. Regarding step 2, the metadata checker program named 'md_checker' validates the XML structure and values and checks the existence of the other metadata linked to them. If the result from the metadata checker is valid, the metadata are transferred to the pre-registration repository. If invalid, the metadata are retained in the draft repository, and error reports are sent to the providers by email. As for step 3, a batch program 'git2dspace' reads the XML files in the pre-registration repository, converts the metadata written in the SPASE format to Dublin-Core, and registers them to DSpace.

Figure 2 shows a variation of the number of registered metadata. Based on the SPASE ontology, the metadata are classified into data set, instrument, observatory, contact person, granule (i.e., individual data files), etc. As of February 2014, the number of registered metadata is 3,151 for data set and 10,285,577 for granule, and it has been increasing about a hundred per day. Of these metadata, 84% of these metadata is registered by the IUGONET members while the remaining 16% is by project collaborators.

Our metadata database has functions of browsing with an internet browser, and of XML interfacing with external programs (see 'Performance evaluation and its improvement' section). The browsing is done as follows: (1) open the metadata search page using the internet browser; (2) search the data by specifying the metadata type, keywords, observation date and time, observation location (latitude and longitude), and so on; and (3) select one item from the search results to obtain details of the metadata.
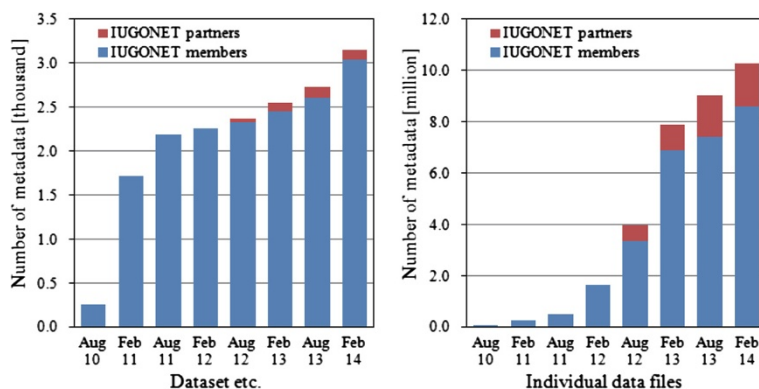
Figure 2 Variation of number of the registered metadata. Left-hand shows the number of data set, instrument, observatory, contact person, etc., and right-hand shows the number of granule (i.e., individual data files).

The details of the metadata include description of data, instrument, observation location (latitude and longitude), location of the data files (URL), contact person, data usage policy, etc., so it is possible for users to not only get the information but also download data files from the remote data server.

Listed below are the examples of search terms used in the metadata database by the users. The search words include some terms related to satellites, planets, materials, etc., as well as many terms related to the Earth's upper atmosphere, which implies that the IUGONET has a potential to extend to cover various fields of science other than the upper atmosphere.

Examples of search terms used in the metadata database by the users:

Earth's upper atmosphere
MF radar, Super DARN, MAGDAS, EISCAT, smart, magnetogram, dst, aurora, ionosphere, geomagnetic field, etc.
Others
ceilometers, electron, ozone, X-ray, Jupiter, climate, CO_2, O_3, GOES, cloud, carbon, etc.

Performance evaluation and its improvement
Figure 3 shows the processing time required for the registration of metadata. The metadata are registered by the git2dspace program as stated in the following: (1) convert the XML files and issue commands to register the metadata to DSpace, (2) execute the commands to register the metadata to PostgreSQL ('Import to PostgreSQL' in Figure 3), and (3) update the index for Apache Lucene ('Index-update' in Figure 3). Item 1 was developed uniquely by IUGONET, and items 2 and 3 are originally included in DSpace.

The result showed that it takes about 200,000 s (55.5 h) to register a million of metadata to PostgreSQL, regardless of the total number of registered metadata. We

have also confirmed that the registration time constantly increases about 1,200 s per 10,000 metadata, so it can be calculated simply from the registration number at that time. On the other hand, the time required for updating the search index increases in proportion to the total number of the registered metadata. For example, it takes 15,820 s (4.4 h) to increase the metadata from zero to a million, 74,432 s (20.7 h) from 4 to 5 million, and 157,524 s (43.8 h) from 9 to 10 million. It means that even if we add only one metadata to the database, it takes more time corresponding to the total number of the registered metadata.
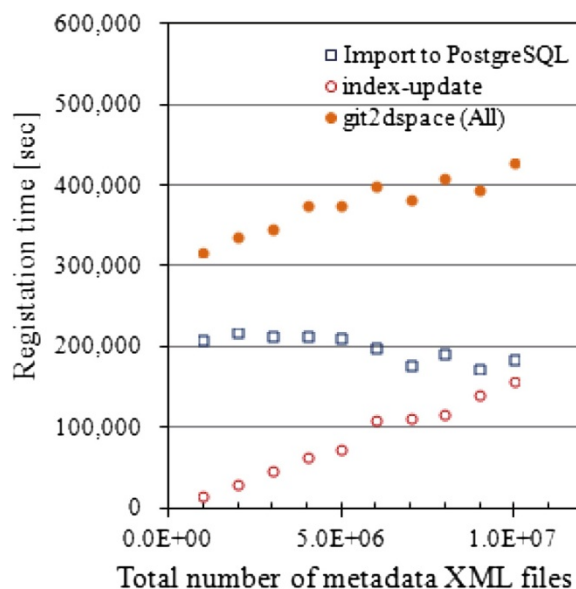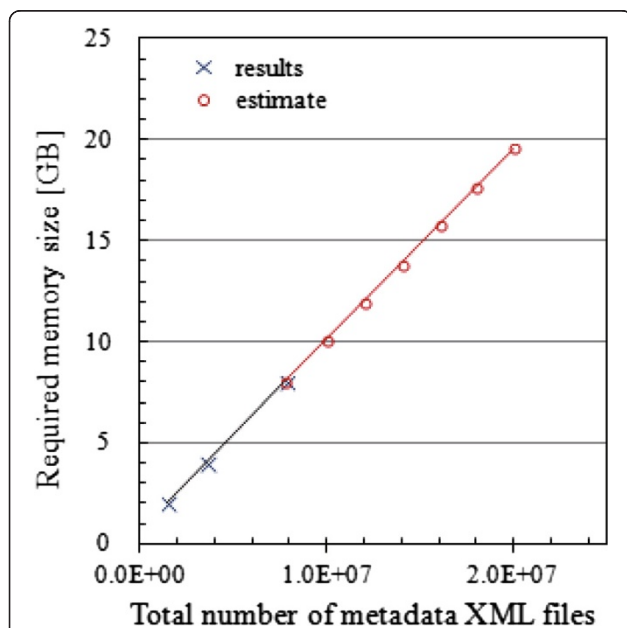

Figure 3 Variation of the necessary time to put metadata in the IUGONET metadata database. The number of metadata was changed from 0 to 10 million, with 1 million increments in between. The amount of time required for registering to PostgresSQL stays constant without reference to the total number of metadata. On the other hand, the time required for updating search index is increasing proportionally.
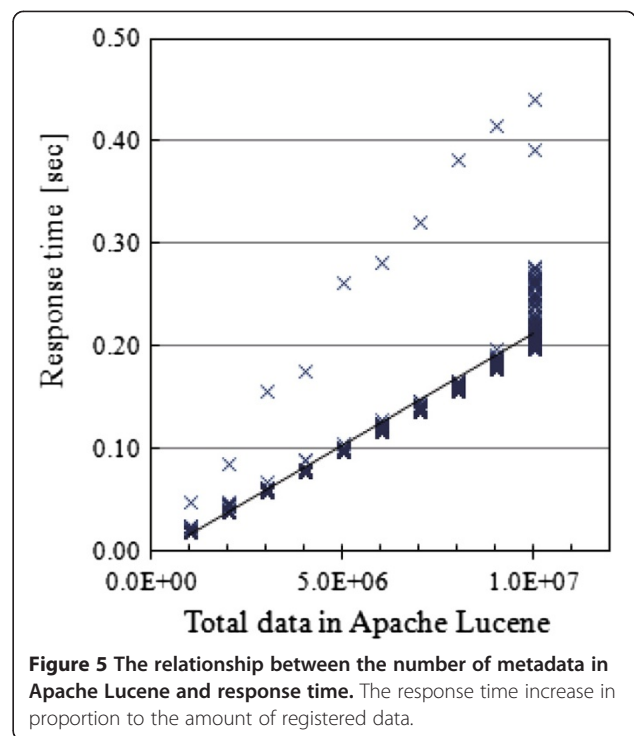
It takes about 2 days for the existing IUGONET metadata database to complete the registration process, as expected from the test results. Thus, it is difficult to register the metadata once a day, and more challenging to keep up with the increase in the volume of observed data every day.

In addition to the registration time, another issue is the index-update process (Figure 3) which requires a lot of computer memory. Figure 4 shows the memory usage of updating the search index. This figure indicates that the memory usage increases in proportion to the total number of the metadata. For example, 4 GB (10 GB) memory is used to register 4 million (10 million) metadata. Therefore, we use a 64-bit computer for the current metadata database and allocate the memory of 16 GB for the Java heap space.

Figure 5 shows the response time of the search engine (Apache Lucene). It is evident from this figure that the response time increases in proportion to the total number of the metadata, i.e., 0.02 s for a million metadata, 0.11 s for 5 million metadata, and 0.21 s for 10 million metadata. We have confirmed that this constant of proportionality is the same for the case of 100 million metadata (2.17 s). Figure 6 shows the response time of Apache Lucene for various search terms. The response time is proportional to the number of hits, for example, 0.03 s for 1,043,221 hits ('STEL'), 0.15 s for 6,218,429 hits ('RISH'), and 0.21 s for 10,000,000 hits ('IUGONET'). It should be noted that the response time in practical operation is generally greater than for these test cases because the metadata database might receive search queries from multiple


Figure 5 The relationship between the number of metadata in Apache Lucene and response time. The response time increase in proportion to the amount of registered data.

users (i.e., multiple requesting) and sort options might be added by users.

In order to overcome above-mentioned issues, we first modified a part of git2dspace that issues the commands for registering the metadata to DSpace. As a result, the

Figure 4 Memory usage for updating the search index on DSpace. The memory usage increases in proportion to the total number of the metadata.
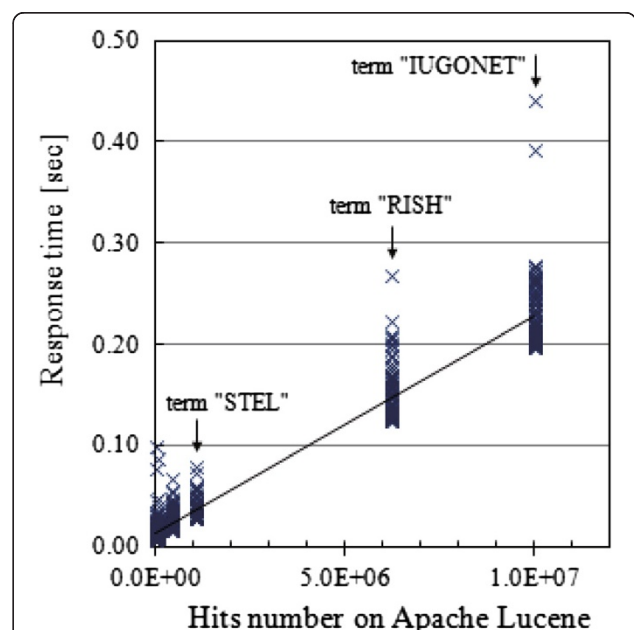
Figure 6 The response time of the IUGONET metadata database related to different terms. The amount of registered data is fixed (10 million). The response time increase in proportion to the amount of retrieved data.

processing time for this part was reduced to one fourth of the original time (before the improvement). Note that Figure 3 shows the result before the above improvement. Then, we examined if it is possible to reduce the processing time by updating the version of DSpace, since it is responsible for most of the processing time. However, since the program code for the metadata registration has not been revised for DSpace version 1.8.3 (i.e., the final version of DSpace 1) and DSpace version 3, it is not expected to reduce the registration time and the required memory usage. Another issue of the search response time is caused by using the single index, thus, it might be necessary to use the multiple index. Therefore, it is necessary to consider the fundamental changes of the metadata database system, including the server structure and software.

Cooperation with data analysis software

As mentioned in the 'Background of the IUGONET metadata database system' section, IUGONET metadata database is available not only by internet browser (http://search.iugonet.org/iugonet/) but also by using external applications. The metadata database accommodates queries with OpenSearch and act as a back-end of applications. The external programs can utilize the metadata database as follows: (1) make a URL that has a query including the search parameters, (2) search data by GET method in the HTTP protocol, (3) get the search result in the XML format, (4) parse the XML file to obtain the necessary information, and (5) use the information for visualizing and analyzing data. The following is an example of the format for OpenSearch query: http://search.iugonet.org/iugonet/open-search/request?(parameter=value)&(parameter=value)&…&(). The query terms that could be used are described in http://www.iugonet.org/en/opensearch.html. When the database receives a query, the database returns the XML file which includes all elements of appropriate metadata for ATOM1.0 format.

Parts of the routine in iUgonet Data Analysis Software (UDAS) refer the metadata database to get some information by OpenSearch. For example, the routine for loading the Solar Magnetic Active Research Telescope (SMART) data gets the URL of the data file from the metadata database. For the other load routines, the URL of the data files is hard-coded in the routines, so we need to modify the load routines whenever the file location changes In the case of the load procedure for SMART telescope, the procedure to change the URL is not necessary. Only an update of the metadata for the URL of the database is needed, and the change will be reflected immediately.

We also provide some procedures (also included in UDAS) to get information of observatories, or plot the location of observatories on the map using 'latitude' and 'longitude' elements of 'Observatory' metadata on the metadata database. The number of registered observatories is still increasing, and thus, it is good to refer the metadata database instead of including it in UDAS.

Discussion and future efforts

One of the purposes of IUGONET is the promotion of cross-cutting research. Therefore, it is important that IUGONET metadata database is used in various research fields. In order to achieve the above purpose, it is necessary to provide users a method to operate our metadata database in their own server. In addition, the products developed on our project may be no longer used and/or maintained since the IUGONET project is scheduled to end. To avoid such a situation, it is necessary to open our software to the general public. Therefore, we developed some support software which can assist to construct and manage our IUGONET metadata database. In addition, we put our working products in the shared web service of the Internet. We use a hosting service called GitHub for the software development project. By using GitHub, our operational costs can be drastically reduced. In addition, any user can try to install and operate our product via GitHub. The IUGONET metadata database is already used outside the IUGONET institute, for example, as a metadata management system of the imager and medium frequency (MF) radar of National Institute of Information and Communications Technology. Furthermore, our metadata database is also considered as a base model for managing radiation data of Fukushima Prefecture. These result shows our product can be expected to be accepted as a of the cross-cutting research system.

In recent days, discussions about open data are increasing in many research fields. The STP community which almost all IUGONET institutes belong to is no exception. One of the topics of open data is an approach towards data citation, for example, appending the digital object identifier (DOI) to data. It is natural that the IUGONET rides this worldwide flow which promotes to utilize the data. In order to deal with this framework on the IUGONET metadata database, we are trying to renew a metadata schema for IUGONET common metadata format. Moreover, in this update, we are considering reexamination of a namespace in the metadata schema. This renewal gives the IUGONET metadata format XML schema the interoperability and compatibility and contributes to the advancement of IUGONET metadata database.

In order to confirm what kind of contribution the IUGONET metadata database has made to the community until now, we interviewed five institutes (seven organizations) inside the IUGONET. As a result, we found that our activities are widely respected for the quality of metadata archive system, for example, as a starting point

for data search. On the other hand, many users request us to support new datasets, such as satellite data, and to improve data analysis functions. We will fulfill these requests in the next phase of the IUGONET project. In addition, we understand the need to improve the function of data visualization and associative searching for beginners, which a part of the functions can be ready in our system, in the near future.

Conclusions

In this paper, we have discussed the progress and the future vision of IUGONET metadata database. To develop a system for the upper atmosphere data from ground-based observation accumulated over 50 years since the first IGY by Japanese universities/institutes, and accelerate cross-cutting researches by using the system, we released the metadata database. It was a big challenge in our communities. The system is based on DSpase software and SPASE metadata format. We examined some evaluations of our product and made numerous improvements in it. One of the applications in our system is a linkage of data analysis software. For scientists in natural science, data visualization is an important basic tool for their researches. Our product can support their requests by several methods. Our reliable self-assessment helps to improve our product and define actions for future efforts.

Availability and requirements

Project name: Inter-university Upper atmosphere Global Observation NETwork (IUGONET) project
Project home page: http://www.iugonet.org/
Operating system(s): Linux
Programming language: Java, Ruby
Other requirements: DSpace, Apache, Tomcat, PostgreSQL
License: BSD licence
Any restrictions to use by non-academics: none

Author details
[1]International Center for Space Weather Science and Education, Kyushu University, 6-10-1, Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan. [2]Solar-Terrestrial Environment Laboratory, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan. [3]Data Analysis Center for Geomagnetism and Space Magnetism, Graduate School of Science, Kyoto University, Kitashirakawa-Oiwake-cho, Sakyo-ku, Kyoto 606-8502, Japan. [4]National Institute of Polar Research, 10-3, Midori-cho, Tachikawa 190-8518, Tokyo, Japan. [5]Planetary Plasma and Atmospheric Research Center, Graduate School of Science, Tohoku University, 6-3 Aramaki Aza-Aoba, Aoba-ku, Sendai 980-8578, Miyagi, Japan. [6]Research Institute for Sustainable Humanosphere, Kyoto University, Gokasho, Uji 611-0011, Kyoto, Japan. [7]Kwasan and Hida Observatories, Graduate School of Science, Kyoto University, Kurabashira, Kamitakara-cho, Takayama 506-1314, Gifu, Japan.

References
Hayashi H, Koyama Y, Hori T, Tanaka Y, Abe S, Shinbori A, Kagitani M, Kouno T, Yoshida D, UeNo S, Kaneda N, Yoneda M, Umemura N, Tadokoro H, Motoba T, IUGONET project team (2013) Inter-university Upper Atmosphere Global Observation Network (IUGONET). Data Sci J 12:WDS179–WDS184. doi:10.2481/dsj.WDS-030
Tanaka Y-M, Shinbori A, Hori T, Koyama Y, Abe S, Umemura N, Sato Y, Yagi M, UeNo S, Yatagai A, Ogawa Y, Miyoshi Y (2013) Analysis software for upper atmospheric data developed by the IUGONET project and its application to polar science. Adv Polar Sci 24:231–240. doi:10.3724/SP.J.1085.2013.00231
Olsen N, Friis-Christensen E, Floberghagen R, Alken P, Beggan CD, Chulliat A, Doornbos E, da Encarnação JT, Hamilton B, Hulot G, van den IJssel J, Kuvshinov A, Lesur V, Lühr H, Macmillan S, Maus S, Noja M, Olsen PEH, Park J, Plank G, Püthe C, Rauberg J, Ritter P, Rother M, Sabaka TJ, Schachtschneider R, Sirol O, Stolle C, Thébault E, Thomson AWP, et al. (2013) The Swarm Satellite Constellation Application and Research Facility (SCARF) and Swarm data products. Earth Planets Space 65(11):1189–1200