ISBN: 978-3-95450-209-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2019-TUBPL03

EXPERIMENTAL DATA TRANSFER SYSTEM BENTEN AT SPring-8

T. Matsumoto†, S. Yokota, T. Matsushita, K. Nakada, Y. Hiraoka, M. Kodera, Y. Furukawa, A. Yamashita, Japan Synchrotron Radiation Research Institute Hyogo, Japan

Abstract

Recently, there have been high demands of open data to promote data science such as material informatics. At SPring-8, we have been operating the data transfer system (SP8DR) for open data of the X-ray Absorption Fine Structure (XAFS) standard sample since 2013. However, it proved to be difficult to use during various experiments at SPring-8. To overcome these problems, we recently developed the BEamline ExperimeNTal stations oriENted data transfer system (BENTEN) for generic use in synchrotron radiation experiments. BENTEN provides an easy-to-use and unified interface with REST API for data access from both inside and outside SPring-8. BENTEN implements user authentication and can also provide restricted data access among the members of the experiment. Data registration is performed with metadata files that describe data such as experimental conditions and samples. To manage multiple metadata in the experiments, Elasticsearch was used as a metadata store. Data can be accessed flexibly via a full-text search. We launched the BENTEN system in March 2019 and provided open access to the XAFS standard sample and restricted data access in user experiments at BL14B2. We plan to use BENTEN on public experimental stations to promote data science as well as other experimental data.

INTRODUCTION

SPring-8 is a third-generation synchrotron radiation facility in Japan. Brilliant synchrotron radiation X-rays are produced from 8 GeV electron beams and are used for various experimental measurements for scientific research and industrial applications. Experimental stations are equipped for each of the 57 beamlines and a large amount of data is produced from these experiments.

Recently, there have been notable advances in data science, such as materials informatics. As data science produces knowledge from data, it is highly desirable that data be opened up to experimental data. Open data from the point of view of social responsibility are also desired, especially in the case of data obtained through public funds. In the case of other synchrotron radiation facilities, such as ESRF, the embargo period for the publication of experimental data is set at 3 years as data policy [1].

At SPring-8, we have been providing open data access for XAFS standard sample since 2013 with the experimental data transfer system called as SP8DR [2]. The amount of XAFS data is approximately 800, which corresponds to the second place in the world [3]. These XAFS data were utilized as reference for the measurements in user experiments.

There are also demands for restricted data access for remote experiments and proxy measurements at SPring-8. To satisfy these demands, SP8DR implements authentication with the SPring-8/SACLA user information account (SPring-8 ID) and provides authorized data access. Although SP8DR requires authentication, everyone can access to the open data because the SPring-8 ID account registration is open to public users.

However, there were several problems with SP8DR. For example, SP8DR required building a system for each beamline. Therefore, it was difficult to extend the system into other beamlines. It was also not easy-to-use nor flexible to manage metadata in various experiments. To overcome these problems, we recently developed the experiment data transfer system, BENTEN. In this paper, we report about BENTEN and its operation at SPring-8.

EXPERIMENTAL DATA TRANSFER SYSTEM BENTEN

We developed BENTEN as a generic software for experimental data transfer to provide data access through the Internet for synchrotron radiation experiments with an easy-to-use interface. As BENTEN covers open data access, it is recommended to follow the FAIR principle [4]. FAIR stands for Findable, Interoperable, Accessible and Reusable. Therefore, it is not enough to open data as it is. Data must be regulated for human comprehension by attaching metadata in the experiments, such as samples and measurement parameters. Machine readability is also important because the data will be used with artificial intelligence (AI), such as machine learning.

To operate the remote data access system in our facilities, we also need to have flexible data management. Therefore, we required BENTEN to comply with the following conditions:

- Easy-to-use for data transfer functions such as authentication, metadata creation, data registration, and data access.
- Metadata of raw data and derived data can be easily described and flexibly managed for measurement data in several experiments.
- Flexible data search can be performed with registered metadata items.
- The data cycle, such as creation, access to and deletion of open/closed, can be easily controlled.
- To refer dataset, each dataset is linked with persistent ID (PID), and the contact person or organization is associated.
- The data reliability in data transferring is guaranteed by the checksum verification.

† matumot@spring8.or.jp

ISBN: 978-3-95450-209-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2019-TUBPL03

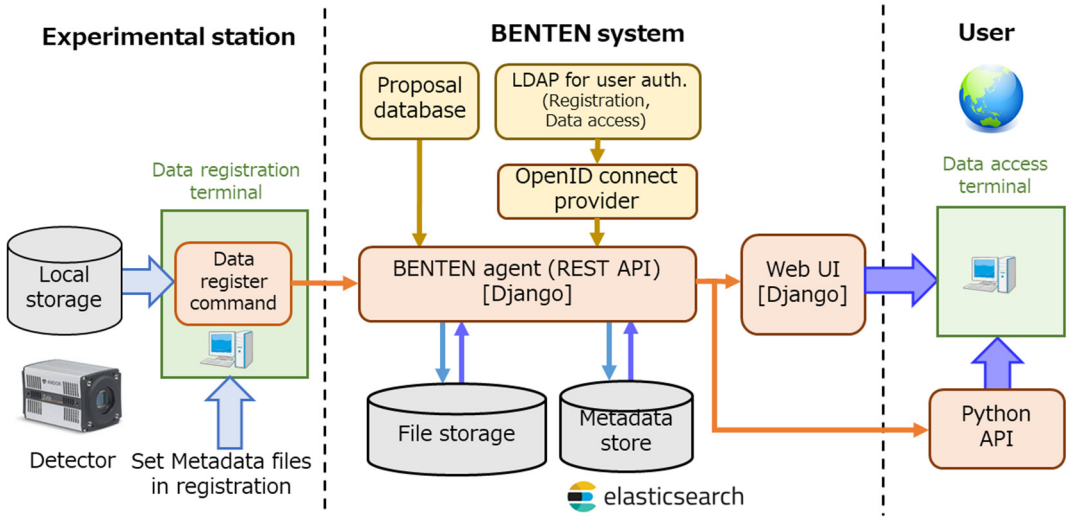

Figure 1: Schematic view of the BENTEN system.

We developed the BENTEN software for SPring-8. But it can also be used independently anywhere through proper installation and settings. We plan to manage BENTEN as Open Source Software (OSS) to promote uses in other facilities.

In Fig. 1, the schematic view of the BENTEN system is shown. We designed the BENTEN agent to provide all the functions in the data transfer system using REST API [5] with JSON response. Django [6] was used to develop the BENTEN agent. As the REST API is a call service based on http protocol, it can be flexibly coordinated with other software for Web and native applications. We also developed Python API for the BENTEN agent to make it easy to use. Data registration can be easily done with command scripts based on the Python API. For data access from the Internet, a web portal was developed using Django. The Python API can also be used for data access, but it is currently limited to use within the facility for secure data access.

To use BENTEN, authentication with the login procedure is required first. We adopted OpenID connect 1.0 [7] for the authentication because it can manage both user attributes as well as authentication, and it was widely used in cloud systems as OSS. We required authentication with beamline account for data registration, and SPring-8 ID account for data access. The user attributes associated with each account were used to control the range of access privileges.

Data Registration

In the data registration, we need to describe metadata in the experiments into metadata files and these are registered with experimental data files into the BENTEN system using the data register command as shown in Fig. 1.

For the metadata description, it may be possible to attach information to existing experimental data files. However, in Japan, various data formats are used for synchrotron radiation experiments, and no progress has yet been made in standardizing data formats. Although NeXus [8], which adds meaning to HDF5 container format [9], is widely used as a common data format for scientific facilities including synchrotron radiation experiments around the world.

Therefore, we have adopted the use of various data formats as they are. However, we required attaching metadata files with unified data format using JSON, which is a flexible text-based data format suitable for reading by both humans and machines. In the metadata files, metadata items are described with key-value pairs.

In BENTEN, registered metadata items were classified into 7 categories; subject (basic data information), facility, sample, measurement, instrument, dataset, and system. Example of these metadata items are shown in Table 1. To express several metadata hierarchically, we used "@" to combine different phrases in metadata keys.

Table 1: Example of Metadata Items

key	Description	Example value
@subject@correspondance	Contact name	Takahiro Matsumoto
@subject@correspondance@affiliation	Affiliation of contact name	JASRI
@subject@proposal_number	proposal number	2014S0000
@subject@pid	Persistent ID	spring8.784d08a8-f39a-4ba0-ac13-6440688b54fd
@measurement@method	Measurement method	XAFS

ISBN: 978-3-95450-209-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2019-TUBPL03

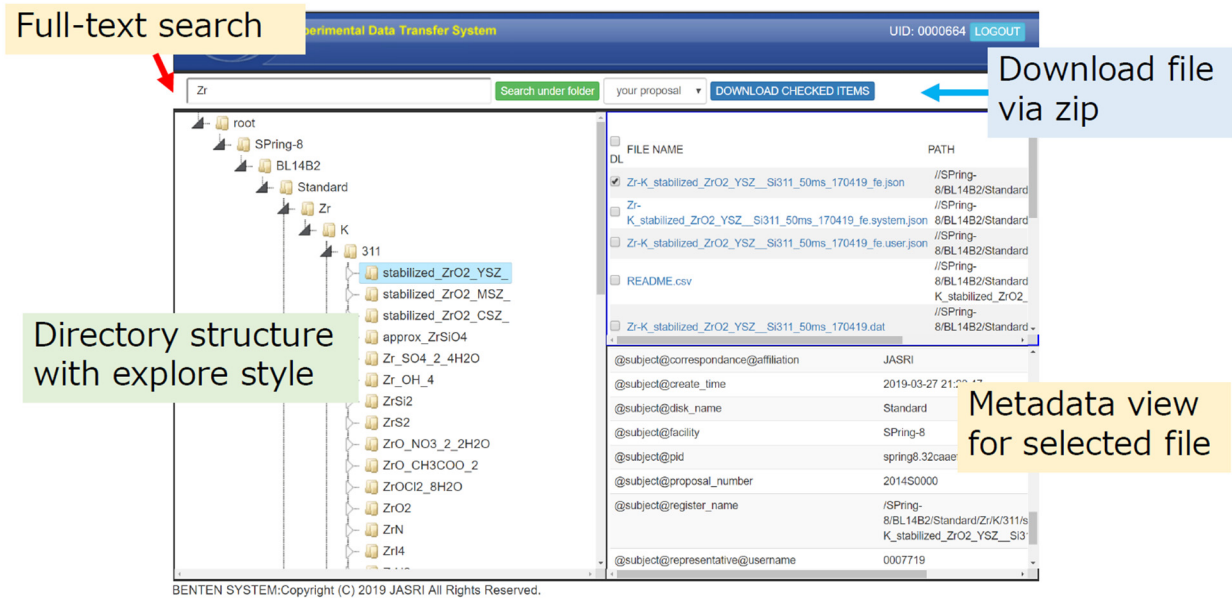

Figure 2: An example of data access for XAFS standard sample within the BENTEN web portal is shown.

In the experiment, we needed to define several metadata items. For flexible metadata management, we adopted Elasticsearch [10] as a metadata store. Because Elasticsearch is a schema-less database, we can easily add metadata items required for experiments. Elasticsearch also implements full-text search, and this function can be used for flexible data access.

For the metadata description, several required metadata items were defined that are essential to describe data. The most important required item is the proposal number. The proposal number is provided for each experiment and stored in the proposal database with the members of each proposal. Therefore, we can use the proposal number to derive restricted data access from the proposal in the experiment. Thus, the proposal number has the function of providing a range of shared data and is extremely important.

For the required metadata, we also defined the access indicator (open or close) to determine accessibility. To refer the data, the Persistent ID (PID) was also defined, which is uniquely assigned for each dataset. Contact name and its affiliation was also defined to inquire the data.

To reduce cost of describing these required metadata, metadata items other than proposal number can be automatically defined when not filled in. For example, the contact name can be derived from the proposal database by specifying the proposal number.

In BENTEN, one dataset can be composed from several files. An example of file structure for one dataset is shown below.

- <X>.json, <X>.system.json, <X>.user.json, …
- <X>/AAA.csv, <X>/BBB.tiff, …

First, we defined <X> to specify a name for registration. We then included files starting from "<X>.". Files with JSON extension were used to describe metadata. To define metadata depending on the purpose, many JSON files can be attached. To include many experimental data files, we can create the directory "<X>" to locate the files under the directory. After the preparation of these files, the data register command was used to transfer the dataset to the BENTEN system.

To update the dataset, we requested to update the contents of the data files including the metadata files and performed the re-registration. In the re-registration, the updated files are only registered by checking the consistency in the timestamp and hash of the files. Therefore, the update procedure can be easily performed based on file management and does not require editing the metadata store directory.

Data Access

After the data registration, we can access the data from the Web portal as shown in Fig. 2. To access the data through the Internet, we are first required authentication with SPring-8 ID. To prevent unsecured access with an invalid login attempt, we have imposed a two-factor authentication. After authentication with the login account, we send an email associated with the account and request consent to use the system.

In the left panel of Fig. 2, we can check the accessible data (open data and account data) and the directory tree is shown with structure shown below.

- /<facility name>/<class name (such as beamline name)/<disk name>/…

In the case of Fig. 2, the facility name is set as SPring-8 and class name is set as BL14B2. With the directory structure, we can have a centralized data management for multiple facilities and beamlines. Under the class name directory, the directory for the disk name was also prepared, and

ISBN: 978-3-95450-209-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2019-TUBPL03

this directory was used to classify experimental data for each purpose. In the previous case, the disk name is set as Standard to manage data for the XAFS standard sample.

Under the disk name directory, the data is allocated with the same structure in the beamline storage. Therefore, the transferred data can be managed with an easy-to-understand structure for users.

Access to the data can be done in several ways. The most naive way to access the data is to follow the directory structure. It is also possible to access the data with the Lucene query by specifying metadata items. As described above, full-text search is also possible for flexible data access. In Fig. 2, the Zr sample is searched with full-text search and matched files are shown in the right upper panel. Here, we can select each file to check the associated metadata items and these are displayed in the right lower panel. Data download can be easily performed by selecting target files or directories and obtained with a zip file.

OPERATION OF BENTEN AT SPring-8

We started the operation of BENTEN at SPring-8 in March 2019. The BENTEN web portal was prepared for data access from the internet [11].

To promote the BENTEN system in the public SPring-8 experimental stations, we first attempted to use BENTEN at BL14B2 to update the previous data transfer system SP8DR. We have experienced several issues with the operation of BENTEN, but we have already solved most of them, and BENTEN is now used at BL14B2 in close proximity to the production environment.

The most important issue in the operation was to define correctly the proposal number to the metadata. As the proposal number is used to define the range of data sharing, it is extremely important.

To overcome the problem, we developed an issuing machine for the proposal number. The issuing machine has a card reader interface and a USB device. When the user places the user card into the card reader, the SPring-8 ID is read and the list of available proposal numbers is displayed in the issuing machine, then the proposal number selected by the user is stored in the USB device. To define the proposal number, the beamline staff provides the USB device to the experimental user, and the user configures the proposal number to the USB device using the issuing machine. The USB device is then adjusted to the measuring machine to provide metadata for the measurement. Thus, we protected against errors the configuration of the proposal number by using physical USB device to transfer the data.

At BL14B2, the BENTEN system was introduced with the issuing machine for the proposal number. and user experiments can be automatically performed with data transfer. Due to the automated measurements, metadata other than proposal number have not been set yet. However, transfer data can be accessed through the Internet with restricted members in the experiment, and experimental users can conveniently use the system.

We also used BENTEN for open data access to XAFS standard sample. These data were registered offline with enough metadata. For efficient metadata preparation, metadata items were automatically extracted when possible. For example, metadata items were extracted from the XAFS data format (PF9809). We also prepared metadata tools to obtain facility information such as the accelerator energy, current, and filling pattern.

However, manual preparation of metadata remains necessary to provide enough metadata. We manually input metadata items such as sample, measurement conditions, and instrument in text data for XAFS standard sample, and used this for the data registration.

An example of data access for XAFS standard sample is shown in Fig. 3. In the right lower panel, we can check XAFS spectrum plot as well as metadata values as shown in Fig. 2. To facilitate the understanding of the data, BENTEN allows to set a thumbnail image for each dataset and this image can be viewed with the Web portal. The content of the thumbnail can be customized to fit the purpose.

We also experienced management issues in the operation.

Initially, we needed to have metadata management. Although BENTEN can allow us to define several metadata in the experiments, we need to have a consistent metadata set for each beamline. Therefore, we ask the beamline staff to manage the metadata items used in the experiments. If the required metadata items are not defined, the user can assign the metadata items to the manager.

We also needed to have a policy for storage management in transfer data. At SPring-8, there are 57 beamlines and a large volume of data may be required for storage in transfer data. In this case, we would incur a high maintenance cost if we want to keep the data forever. Therefore, BENTEN does not guarantee the data in the transferred storage, and imposes the data guarantee for each beamline. If the data is missing in the BENTEN storage, the data must be recovered by re-registration of the data in the beamline storage. When the data in the beamline storage is missing, the data can be restored from BENTEN, but it is limited when the data is available in the BENTEN side.

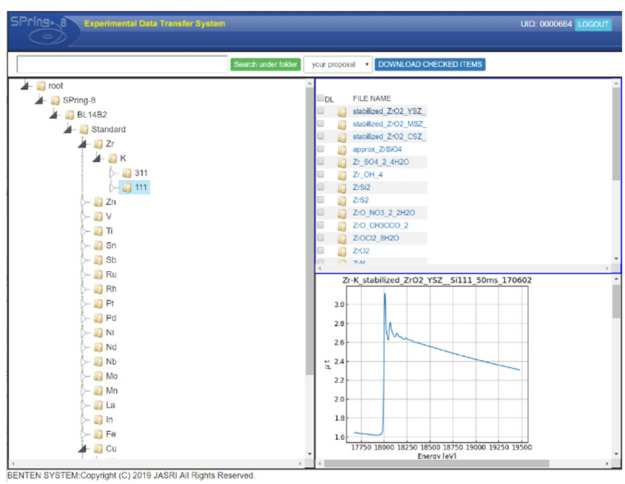

Figure 3: Screenshot of the XAFS spectral plot image using the BENTEN web portal.

ISBN: 978-3-95450-209-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2019-TUBPL03

CONCLUSION AND FUTURE PLANS

In this paper, we have reported the experimental data transfer system BENTEN. BENTEN has been designed for generic use in experimental data transfer for synchrotron radiation experiments with a user-friendly interface. BENTEN implements authentication and can be used to provide restricted data access as well as open data access. Metadata is managed with Elasticsearch, and various metadata from experiments can be flexibly managed. Data access can be flexibly done through full-text search. We started the operation of BENTEN at SPring-8 since March 2019, and provided open data access of the XAFS standard sample and restricted data access in user experiments at BL14B2.

In the future, we plan to use BENTEN for experiments in public experimental stations. We are now preparing for open data of hard X-ray photoemission spectroscopy (HAXPES) standard sample, and remote access to image data from Computed Tomography (CT).

The federation of data with other facilities is also an important issue for proceeding with open data. In Japan there are many synchrotron facilities (SPring-8, KEK-PF, Aichi-SR etc.). Therefore, if the data format used in BENTEN was standardized and can be utilized among facilities, we can facilitate the data utilization to promote data science.

We began discussing open data access to XAFS data in the XAFS society in Japan, and considered the possibility of standardizing the data format and deposit the facility data into the data repository being prepared by the materials data platform center of the National Institute for Materials Science (NIMS) in Japan. By utilizing a shared data repository in NIMS, we could easily access experimental data in different facilities with a unified data repository, and we could also have a linked data access for other materials databases.

To promote the use of data in synchrotron radiation experiments, we need to use several data repositories in addition to the BENTEN system in our own facilities. We are trying to cooperate closely with other facilities for improvements.

ACKNOWLEDGEMENTS

We would like to thank the members of JASRI's industrial applications division for their useful discussions and suggestions on the development and operation of BENTEN.

REFERENCES

[1] ESRF Data Policy, https://www.esrf.eu/datapolicy

[2] H. Sakai, Y. Furukawa, and T. Ohata, "Development of SPring-8 Experimental Data Repository System for Management and Delivery of Experimental Data", in *Proc. 14th Int. Conf. on Accelerator and Large Experimental Control Systems (ICALEPCS'13)*, San Francisco, CA, USA, Oct. 2013, paper TUPPC014, pp. 577-579.

[3] K. Asakura *et al*., "The challenge of constructing an international XAFS database", *J. Synchrotron Rad*., no. 28, p. 967, 2018.

[4] FORCE 11, https://www.force11.org/group/fairgroup/fairprinciples

[5] U.C., Irvine, https://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm

[6] Django, https://djangoproject.com

[7] OpenID, http://openid.net/connect/

[8] Nexus, https://www.nexusformat.org

[9] The HFD Group, https://www.hdfgroup.org

[10] elastic, https://www.elastic.co/products/elasticsearch

[11] BENTEN web portal, https://benten.spring8.or.jp