所外実験データ転送システム BENTEN

公益財団法人高輝度光科学研究センター　情報処理推進室
松本　崇博、横田　滋、松下　智裕

Abstract

SPring-8 で計測した実験データを所外からアクセスするための基盤として、実験データ転送システム BENTEN（Beamline ExperimeNTal stations oriENted data transfer system）を開発し、2019 年 3 月より運用を始めました。BENTEN はユーザー認証機能を持ち、データを一般に公開する機能（オープンデータ）、および実験課題の共同メンバーのみにデータ共有を行うアクセス制限機能を実装しています。現在、BENTEN は BL14B2 において X 線吸収微細構造（XAFS）標準試料のデータ公開や、ユーザー実験の計測データへの所外からのアクセスで利用されています。今後は硬 X 線光電子分光（HAXPES）標準試料のデータ公開を進めるとともに、共用ビームラインの複数の実験ステーションへと展開していき、様々な実験データの利活用を進めていく予定です。

1. はじめに

SPring-8 で計測した実験データを所外に転送するための実験データ転送システム BENTEN を開発し、2019 年 3 月より運用を開始しました[1,2]。本稿では、BENTEN によるデータ公開を中心に紹介いたします。

オープンデータとは、データをインターネット上に公開し、誰でも無料で利活用できるようにすることです。近年、科学分野ではマテリアルズ・インフォマティクスなどデータ科学が注目されています。データ科学ではデータから知を創出するため、SPring-8 など放射光施設で計測された実験データに関してもオープンにして皆が利活用できる形にすることが強く求められています。

社会的にも公的資金を用いて計測した実験データについては、一定期間後にオープンにすべきとのデータポリシーの考えがあります。ESRF など海外の放射光施設では計測してから 3 年後にデータを公開する動きが出てきています[3]。

SPring-8 におけるオープンデータの取り組みとしては、BL14B2 における XAFS の標準試料のデータ公開を 2013 年より JASRI 産業利用推進室を中心に行ってきました。現在公開している XAFS の測定データ数は 800 程度であり、世界第 2 位の統計量になっています[4]。公開された XAFS 標準試料データは実験計測時の参照資料などで活用されています。

データ公開にあたっては、実験データ転送システム SP8DR を整備し運用を行ってきました[5]。SPring-8 では、遠隔実験や測定代行において実験の共同メンバーのみにデータ共有範囲を制限する利用ケースもあります。このため、SP8DR ではシステム利用時に SPring-8/SACLA 電子申請システムのアカウント（SPring-8 ID）認証を行い、アクセス制限付きのデータ転送にも対応しています。この SPring-8 ID のアカウント登録は一般に公開されているため、オープンデータに関して誰でもアクセスできるようになっています。

このように実験データ転送システムを運用してきましたが、SP8DR では 1 ビームライン単位でシステムを構築する必要性があるなど、導入のためのハードルが高く管理コストがかかること、使い方が難しいなど課題がありました。これらの課題を解消するため、実験データ転送システム BENTEN を新規に開発して対応しました。

2. 実験データ転送システム BENTEN について

実験データ転送システム BENTEN は、SPring-8 の複数の共用ビームラインの実験におけるデータ転送で汎用的に利用でき、かつ簡易に使えるソフトウェアとして開発しました。

BENTEN はオープンデータの基盤としても利用できますが、公開したデータを利活用するためには、データ

自身が FAIR 原則を満たすことが推奨されています[6]。FAIR とは Findable（見つけられる）、Accessible（アクセスできる）、Interoperable（他施設ともデータの相互運用ができる）、Re-usable（再利用できる）の頭文字をとったものです。よって、データを単に公開するだけでは不十分であり、人がデータを理解して利用できるように、測定条件など充分なメタデータを付加し、データを正規化する必要があります。また、データは機械学習などの AI での利用も想定されるため、機械可読性を高めることも重要になります。他にも、システム運用時には適切なデータマネージメントを行う必要があります。

このため、BENTEN では以下の要件を満たすように設計しました。

- データ登録やデータ検索、ダウンロードなどの各機能が簡易に利用できること。
- 複数のビームラインの実験で利用でき、多種多様な実験データフォーマットに対応できること。
- 登録したメタデータ項目を用いて実験データが柔軟に検索できること。
- データのライフサイクル管理（登録、更新、一般公開や限定公開などデータ共有範囲の設定、およびデータ消去）が簡易に行えること。
- データを引用するため、各データセットにはユニークで永続的な PID（Persistent ID）が割り当てられており、また、データの責任者が把握できること。

BENTEN ソフトウェアは SPring-8 以外でも独立にインストールして利用することができます。将来、BENTEN は OSS（オープンソースソフトウェア）として提供することで、他の放射光施設でも利用できるようにする予定です。

BENTEN システムの概要を図 1 に示します。BENTEN システムでは、BENTEN agent のサーバーが、認証、データ登録、データ検索、ダウンロードなど全ての実験データ転送機能へのインターフェースを提供するように設計しました。BENTEN agent での通信には Web サービスでよく使われている REST API[7] と呼ばれる http に基づくプロトコルを用いており、応答は JSON 形式で行います。BENTEN agent は Python による Web フレームワークである Django[8] を用いて構築しました。インターフェースを REST API で統一することで、Web ポータルや、他のユーザーアプリケーションで簡易に利用することができます。REST API の機能を簡易に利用するため Python API も開発しました。データ登録は Python API を用いて作成されたコマンドを用いて簡易に行うことができます。所内外からのデータアクセスは Django で構築された Web ポータルから行います。データアクセスに Python API を利用することもできますが、セキュアなデータアクセスのため、現在のところ利用は所内に限定しています。

BENTEN システムでは、ユーザー毎に適切なアクセス制限を行うため、利用にあたっては始めに認証が必要になります。データ登録時にはビームラインのアカウント、データアクセス時には SPring-8 ID のアカウント認証が必要です。認証には、近年クラウドでの認証でもよく使われている OpenID Connect 1.0[9] を用いています。

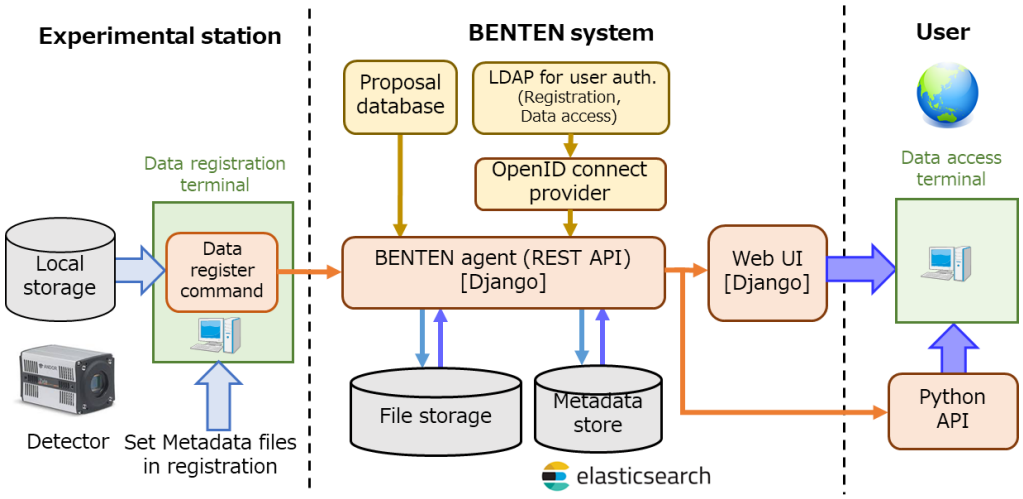

図 1　BENTEN システムの概要

表 1　BENTEN におけるメタデータ項目の例

項目名	説　明	値の例
@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

2.1 データ登録

データ登録の際は、図 1 に示すように、実験データに対してメタデータが記述されたファイルを作成し、これらのファイルを纏めて BENTEN システムに登録します。

様々な実験データの利活用を効率的に行うためには、共通データフォーマットを定義し、そのフォーマットに従って実験データを記録することが望まれます。海外施設では NeXus[10] と呼ばれる HDF5[11] によるデータコンテナをベースとして作られたデータフォーマットで標準化する方向で進んでいます。しかし、日本ではデータフォーマットの標準化は進んでおらず、各実験で様々なデータフォーマットが使われています。

このため、実験データはそのままの形で扱い、メタデータ記述を JSON 形式のデータフォーマットで統一する手法を採用しました。JSON は人の可読性や機械可読性に優れたテキストベースのデータフォーマットです。メタデータ項目はサンプル、測定パラメータ、装置などカテゴリに分類して定義しています。メタデータ項目の例を表 1 に示します。階層構造を持つデータにも対応できるように、異なる文節を“@”でつなげる形でメタデータ項目名を定義しています。

メタデータ項目は、実験により様々な項目があり、これらを柔軟に定義する必要があります。このため、メタデータを管理するデータベースには Elasticsearch[12] を採用しました。Elasticsearch はスキーマレスのデータベースであり、必要に応じてメタデータ項目を随時追加することができます。また、全文検索にも対応しており、柔軟にデータ検索をすることができます。

メタデータの記述にあたっては、必須の記述項目をいくつか定義しています。最も重要なメタデータ項目は課題番号です。課題番号は実験課題毎に割り当てられますが、実験課題の共同メンバーも課題番号と関連付けて課題データベースに登録しています。よって、課題番号を用いることで、実験課題の共同メンバーに限定したデータアクセスが実現できるように設定できます。このように、課題番号は実験データの共有範囲を決める側面もあるため、極めて重要なメタデータ項目です。

その他は、データ公開・非公開のアクセス条件を示すフラグも必要になります。データを公開する際には、データの責任者とその所属を設定することが必要になります。

このシステムでは、1 データセットとして、複数のデータファイルと複数のメタデータのファイルで構成されることを想定しています。以下に、データ登録における 1 データセットのファイル構成例を示します。

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

ここで、<X>はデータ登録時に指定する登録名です。この際、“<X>.”で始まる同名ファイルが 1 つのデータセットとみなされます。メタデータを示すファイルには json の拡張子が付いています。メタデータは用途毎に生成することもあるため、複数のメタデータファイルが登録できるようにしています。また、<X>のディレクトリを作成し、ディレクトリ以下のファイルを登録対象とすることもできます。ディレクトリは複数の実験データファイルをまとめて登録する際に利用できます。

データ登録後、データセットのコンテンツを更新する際は実験データやメタデータファイルの内容を更新して再登録します。

このように、データ登録・更新はファイルをベースに手続きが可能であり、データベースを直接編集する必要はないため、簡易に利用することができます。

2.2 データアクセス

データ登録後は、図２に示すように、所外から Web ポータル経由でデータにアクセスできます。利用の際は、SPring-8 ID のアカウントでの認証が必要になります。なりすましによる不正アクセスを防ぐため、２要素認証を実装しています。アカウントでのログインの後、利用にあたってはメールアドレスによる本人同意が必要になります。

図２の左側には認証したアカウントでアクセス権のあるデータのディレクトリツリーを表示しています。ここでディレクトリ構造は以下の形をとっています。

・/<施設名>/<分類名（ビームライン名など）>/<ディスク名>/ …

図２の例では、施設名が SPring-8、分類名が BL14B2 になっています。このようなディレクトリ構成をとることで、複数施設、およびビームラインのデータを一元的に扱うことができます。分類名の下は、用途毎にストレージ領域を分けてデータ管理するため、ディスク名のディレクトリを設置しています。図 2 では XAFS 標準試料を扱うため、ディスク名に Standard を設定しています。

ディスク名の下は、各ビームラインでのストレージ領域がそのまま見える形にしています。これにより、ユーザーが管理しやすい形でデータ公開をすることができます。

データの検索は、ディレクトリツリーをたどることにより行うこともできますが、登録されたメタデータ項目を用いて横断的にデータ検索することもできます。

図２では Zr の試料名を指定して全文検索を行っています。検索にマッチしたファイルは図２の右上にリストされます。それぞれのファイルを選択すると、図２の右下にデータに基づくメタデータ項目とその値のリストを閲覧することができます。データは、ファイルやディレクトリを指定し、zip ファイルでまとめてダウンロードできます。

3. SPring-8 における BENTEN 利用

BENTEN は 2019 年３月より SPring-8 で運用を開始しました。所外からの実験データアクセスのための Web portal も設置しています[13]。

BENTEN は SPring-8 の共用ビームラインで汎用的に活用できますが、BL14B2 における旧実験データ転送システム SP8DR の更新を最初のターゲットとして利用整備を進めました。現在、BL14B2 での BENTEN 利用は試験中ですが、既に本番環境としても利用できる形で運用を行っています。

BENTEN 運用を行うにあたり、データ共有範囲を管轄する課題番号をどのように正確に定義するかが大きな課題になりました。これを解決するため、課題番号発行機を開発し対応しました。

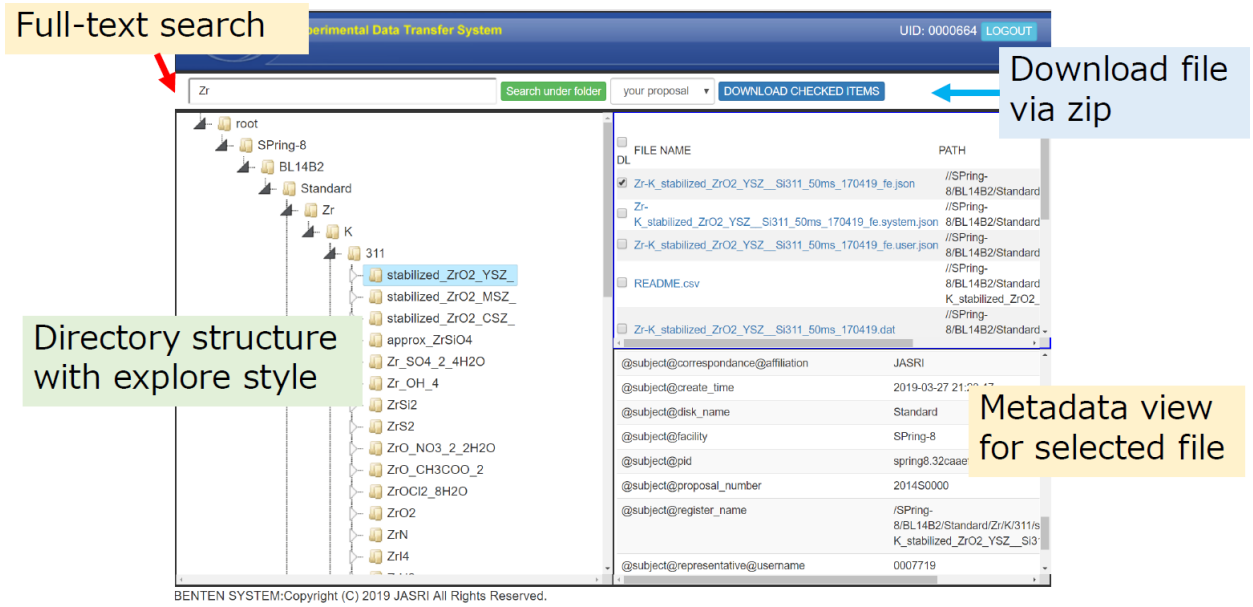

図２　BENTEN Web ポータルでのデータアクセス例

課題番号発行時は、ビームラインスタッフがユーザーに USB デバイスを貸与します。ユーザーは課題番号発行機に USB デバイスを差し込むとともに、ユーザーカードをカードリーダーにかざします。その後、課題番号発行機ではユーザーに紐付く課題番号リストが表示されます。ユーザーが対象の課題番号を選択すると USB デバイスに課題番号が保存されます。次に、ユーザーは課題番号発行機から USB デバイスを取り出し、実験計測の計算機に差し込むことで、課題番号をメタデータとして入力することができます。このように、USB デバイスに物理的に課題番号情報を保存し、利用することで間違って課題番号を設定することを防ぐようにしました。

BL14B2 では、ユーザー実験において BENTEN システムを自動測定で利用できるように調整し、運用を開始しました。自動測定で利用するため、メタデータは課題番号の他はほとんど定義されていませんが、所外から実験課題の共同メンバーのみに制限してデータアクセスができるため、便利に活用されています。

また、XAFS 標準試料のデータ公開も開始しました。XAFS 標準試料に関しては、オープンデータで活用するため、充分なメタデータを入力してオフラインでデータ登録しています。なるべく効率よくメタデータ入力を行うため、メタデータ項目の多くは自動抽出しています。サンプル情報や測定条件、測定器のパラメータの一部など自動抽出が難しい項目に関しては手動でも登録して対応しました。

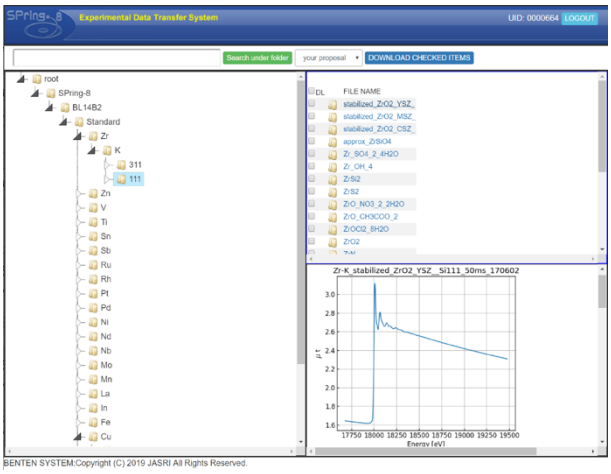

図 3　BENTEN Web ポータルにおける XAFS スペクトルのサムネイル表示例

図 3 に BENTEN Web ポータルにおける XAFS 標準試料のデータアクセスの例を示します。データセット毎にサムネイルを付加することで、メタデータ項目のリストとともに閲覧することができます。

4. まとめと今後の予定

本稿では、BENTEN における所外実験データ転送について紹介しました。BENTEN は放射光実験のデータ転送において汎用的、かつ簡易に利用できるソフトウェアとして設計しました。

2019 年 3 月より SPring-8 で BENTEN の運用を開始し、XAFS 標準試料のオープンデータや BL14B2 でのユーザー実験におけるアクセス制限付きのデータ転送で活用されています。

今後は SPring-8 の共用ビームラインでの利用展開を進めていく予定です。現在は、HAXPES の標準試料のデータ公開や、CT 計測での画像データの遠隔からのアクセスなど整備を進めています。

謝辞

BENTEN の開発および運用を進めるにあたって、JASRI 産業利用推進室の方々には多大なご協力を頂きました。この場を借りてお礼を申し上げます。

参考文献

[1] T. Matsumoto *et al.*: *AIP Conference Proceedings* **2054** (2019) 060076.
[2] T. Matsumoto et *al.*: *Proceedings of ICALEPCS* (2019), to be published.
[3] https://www.esrf.eu/datapolicy
[4] K. Asakura *et al.*: *J. Synchrotron Rad.* **25** (2018) 967-971.
[5] H. Sakai *et al.*: *Proceedings of ICALEPCS* (2013) 577-579.
[6] https://www.force11.org/group/fairgroup/fairprinciples
[7] https://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm
[8] https://djangoproject.com
[9] http://openid.net/connect/
[10] https://www.nexusformat.org
[11] https://www.hdfgroup.org
[12] https://www.elastic.co/products/elasticsearch
[13] https://benten.spring8.or.jp

松本　崇博　MATSUMOTO Takahiro

（公財）高輝度光科学研究センター　情報処理推進室
〒679-5198　兵庫県佐用郡佐用町光都 1-1-1
TEL : 0791-58-0980　ext 3270
e-mail : matumot@spring8.or.jp

横田　滋　YOKOTA Shigeru

（公財）高輝度光科学研究センター　情報処理推進室
〒679-5198　兵庫県佐用郡佐用町光都 1-1-1
TEL : 0791-58-0980　ext 3912
e-mail : yokota@spring8.or.jp

松下　智裕　MATSUSHITA Tomohiro

（公財）高輝度光科学研究センター　情報処理推進室
〒679-5198　兵庫県佐用郡佐用町光都 1-1-1
TEL : 0791-58-0868
e-mail : matusita@spring8.or.jp