B2SHARE: An Open eScience Data Sharing Platform

Sarah Berenji*, Carl Johan Håkansson*, Erwin Laure*,
Ilja Livenson†, Pavel Stranak ‡, Emanuel Dima§,
Dennis Blommesteijn ¶, and Mark van de Sanden ¶

*PDC Center for High Performance Computing
KTH Royal Institute of Technology, Stockholm, Sweden
Email: see https://www.kth.se/directory/skolor-D-DCP/

†NICPB, Estonia
Email: ilja.livenson@kbfi.ee

‡Institute of Formal and Applied Linguistics, Charlel University, Prague, Czech Republic
Email: stranak@ufal.mff.cuni.cz

§Department of Linguistics, University of Tübingen, Germany
Email: emanuel.dima@uni-tuebingen.de

¶SURFSara, Amsterdam, The Netherlands
Email: {dennis.blommesteijn / mark.vandesanden}@surfsara.nl

***Abstract*—Scientific data sharing services are becoming an essential tool in data driven science and can significantly improve the scientific process by making reliable and trustworthy data available, thereby reducing work redundancy, and providing insights on related research and recent advancements. For data sharing services to be useful in the scientific process, they need to fulfill a number of requirements that cover not only discovery and access to data but also ensure the integrity and reliability of published data.**

B2SHARE, developed by the EUDAT [1] project, provides such a data sharing service to scientific communities. For communities that wishes to download, install and maintain their own service, B2SHARE is also available as software. B2SHARE is developed with a focus on user-friendliness, customizability, reliability, and trustworthiness, and can be customized for different organizations and use-cases.

In this paper we discuss the design, architecture, and implementation of B2SHARE and show its usefulness in the scientific process with a couple of case studies from the biodiversity field.

Keywords*—European Data Infrastructure, Data Sharing, Data Repository, Research Data Management.

I. Introduction

Storing and sharing scientific data is an essential service that can benefit researchers greatly by improving the discoverability of research data and it has been targeted by various other projects, like Zenodo [2], CKAN [3], Figshare [4], DataCite [5], and CLARIN [6]. The nature of scientific data imposes special requirements that is not sufficiently covered by general data sharing platforms like Dropbox [7], Box [8], or Google Drive [9]. P. Doorn and H. Tjalsma discussed some differences between general data sharing and research data archives in [10].

According to the Kaptur evaluation of technical systems [11] and the review of options for the development of Research Data Management at the University of Sheffield [12], the requirements for scientific data repositories that distinguish them from general data sharing platforms are the needs to be trustworthy and reliable, providing sufficient access control, reporting, backup and recovery functionalities, as well as the possibility of restricting the physical data storage location to comply with applicable policies and regulations. They should also support copyright licensing and other legal requirements, preservation, extensive metadata, citation, large data support, and embargo periods. Additionally G. Pyrounakis et al. in [13] also discussed 14 expected features for scientific data repositories.

Many researchers are facing the problem of finding a simple, convenient and durable way of storing and sharing their data. They often have large amounts of small files as a result of massive crowd sourcing, derived data in form of spreadsheets, analysis results, and many other types, but do not belong to large research organisations with well defined data management policies. This is the so-called *long-tail data*. Currently, such data is often stored on notebooks and departmental servers, implying the risk of losing our scientific memory. Furthermore, these data are typically neither publicly available nor citable. B2SHARE, developed by the EUDAT project, is addressing this issue by

- Allowing registered users to upload typical "long tail" data objects into the EUDAT data store
- Enabling users to share such objects and collections with other researchers
- Relying on other EUDAT services (e.g. data replication and metadata) for bringing reliability and data retention into the picture

B2SHARE's vision is to offer the core functionalities of a data sharing service that not only is useful for managing data but also increases the visibility of data for researchers and scientists. Providing a modern attractive web interface, a user-friendly license chooser tool and different metadata models, B2SHARE can help all types of researchers to easily store, organize and share their data.

The remainder of this paper is organized as follows: Section II discusses the overall architecture of B2SHARE and its design choices, Section III presents details of our implementation, and in Section IV we give examples of the use of B2SHARE with a couple of use cases from the biodiversity area. After a discussion of related work in Section V we end the paper with an outlook on future work and some concluding remarks.

II. Architecture and Design

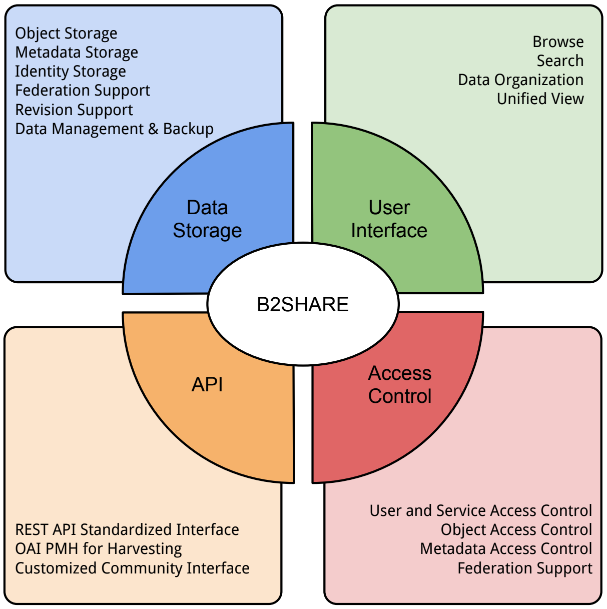


Fig. 1. Conceptual View of the B2SHARE Design and Architecture

B2SHARE has been designed as a modular service consisting of four main areas (cf. Figure 1):

- Data Storage
- User Interface
- APIs
- Access Control

This design allows for easy integration with external services through standardized interfaces, and customization of the service towards the specific needs of certain scientific communities while maintaining a unified view of the service.

Our design considerations for these four main areas were the following:

A. Data Storage

As one of its main objectives, B2SHARE is providing a data storage. Unlike a general data store, however, a sharing service for scientific data needs not only to manage the data proper, but also the associated metadata, and security-relevant identity data. We can thus identify three kinds of data B2SHARE needs to manage:

- *Scientific Objects* are the main data objects that scientists and researchers produce. They can be of any type of data: text, multimedia, experimental results or simulation outputs in key/value form, etc., and their size may vary from very small to massive scale. This in turn requires the possibility of adding quota systems and/or restricting the maximum file size. Due to browser limitations, the web interface of B2SHARE only accepts files of sizes smaller than 2GB. This limitation is not present, however, for files uploaded using the REST API. If communities need to upload much larger files they can contact the B2SHARE technical team, which will apply different ingestion methods on a case-by-case basis.
- *Metadata* is the explanatory data that gets filled in when object data are being uploaded. They vary from community to community, and are customized based on community requirements. The metadata, which is paramount for the discoverability of scientific data, needs to be accessible through standard interfaces such that it can be harvested and used by other services as well. As will be described below, B2SHARE is using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [14] for this purpose. For citability, a PID [15] will be assigned to data.
- *Identity Data* is composed of account information necessary to provide access control for individual users and research organizations, and for integration with other services.

B. APIs

In order to integrate the service in different workflows and other tools, a flexible set of APIs needs to be provided. Also, for mass uploads and downloads programmatic interfaces are typically more convenient than web interfaces. At the moment, B2SHARE provides two types of APIs: a REST-API for creating and retrieving objects and their metadata, and an OAI-PMH API for accessing the metadata. The REST-API can be used as mediator for external services or additional extensions required by users. Through the REST-API users and communities can have a JSON-based interface to interact with the B2SHARE service. B2SHARE's REST-API allows authenticated users to programmatically create, upload and share data.

B2SHARE also offers the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) API to the users.

Using this API, researchers can collect the complete metadata collection and build a meta-view on the content of the repository. OAI-PMH API provides an interface for third parties that allows unimpeded retrieval of the metadata. These APIs allow researchers and communities to integrate with other services and systems to re-use research data. Additionally, they provide a level of automation to speed up these tasks.

C. Web User Interface

In addition to the APIs discussed above, simple data deposit, search, and retrieval interfaces, ideally web-based, are required with a very low learning-curve to facilitate the uptake of the service by non-IT-experts. B2SHARE provides a very simple process for depositing data. The deposit workflow consists of three simple steps of uploading, choosing the scientific domain or research community, and describing data by adding related metadata. The graphical web interface for making deposits is customizable to meet metadata requirements specific to a research community or research topic area.

In order to have access to stored data in B2SHARE, a web-based search interface is provided. The search facility is based on the metadata entered into the system, uses a simple query language and supports regular expressions for fulltext search. Furthermore, it has the capability of searching the indexed metadata information through means of faceted search. Faceted search enables researchers to explore related works of their specified scientific fields within a controlled interface, and allows scientists to find data by walking through the indexed metadata.

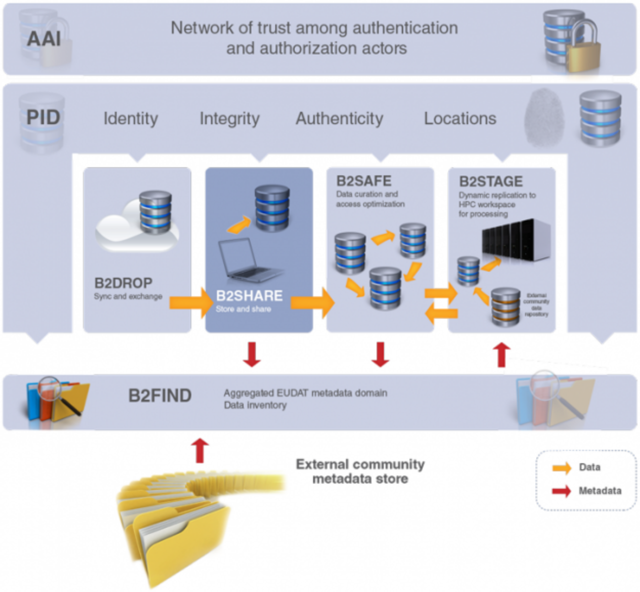

AAI
Network of trust among authentication and authorization actors
PID
Identity
Integrity
Authenticity
Locations
B2DROP
Sync and exchange
B2SHARE
Store and share
B2SAFE
Data curation and access optimization
B2STAGE
Dynamic replication to HPC workspace for processing
External community data repository
B2FIND
Aggregated EUDAT metadata domain
Data inventory
External community metadata store
Data
Metadata

Fig. 2. B2SHARE in EUDAT's B2 Service Suites

D. Access Control

A key feature of B2SHARE is sharing of data and metadata. However, not all data should be publicly available, but restrictions may be imposed that restrict the accessibility of data to certain individual and groups. Also, embargo periods may be imposed. For this purpose, B2SHARE provides access control lists (ACLs). Users can choose, at the time of submission of their data, to make the data publicly accessible or to keep access restricted. The metadata, on the other hand, is always public and searchable. Access control can be further adapted to various scenarios and for different communities (one example is the use of federated ACLs).

When publishing data, it is also advisable to associate an appropriate license with the data to protect the owner's interest. However, there is a myriad of licenses existing and it is often difficult to choose which of these is the most appropriate one. B2SHARE provides a License wizard (originally developed by the Institute of Formal and Applied Linguistics at the Charles University in Prague) to help the researchers to select the correct license for their data [16]. Researchers can use this user-friendly tool as a guide by going through it step by step and choose the appropriate license in the simplest possible way.

B2SHARE is one out of five integrated services of EUDAT [17] and through this integration with the other EUDAT services and with EUDAT's supportive infrastructure, can B2SHARE also provide reliable storage, enhanced discoverability, improved security and access control, as well as storage of dynamic data. As shown in Figure 2, B2SHARE interacts with the EUDAT services B2SAFE, B2FIND and B2STAGE in order to support safe replication, harvesting and cataloguing of metadata, and data staging. In addition, via the integration with B2DROP, two services together can also provide collaborative access to dynamic data, including data synchronization and exchange with other researchers, before the public sharing of the final results and data.

III. Implementation

The B2SHARE implementation strategy was not to build a new system from scratch but rather to adopt and modify an existing service. This had the short term benefit of having a pilot service available rather quickly and the long term benefit of taking advantage of community development efforts. We evaluated a number of services[1], particularly focusing on their suitability for the aspects mentioned above (extensibility, data storage, metadata support, access control) but also their licensing model and overall readiness [18].

The result of this evaluation was that Invenio [19] was chosen as base technology for B2SHARE, primarily because it is open source, has the possibility of adding metadata, provides a native web user interface, is extensible and scalable, supports the integration with PIDs, and supports large files.

B2SHARE is implemented as an overlay on top of Invenio and shares much of Invenio's source code for the base functionality. However, the B2SHARE development efforts focused particularly on the possibility for customisation, extendibility, and pluggability. Many organizations and particular users have different needs concerning security, performance, availability, and sustainability. The architectural diagram of B2SHARE is shown in Figure 3.

In this section we discuss the implementation of B2SHARE focusing on the four areas introduced in Section II.

[1] Beehub, Figshare, iDROP, Invenio, myExperiment, Scratchpad, VCD, and CKAN

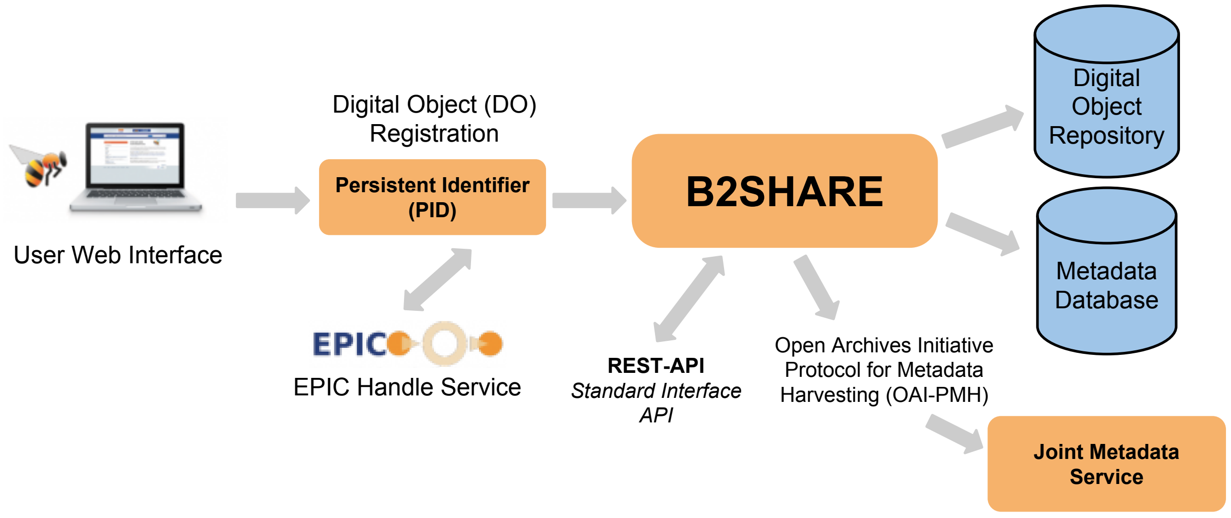


Fig. 3. Architectural diagram of B2SHARE

TABLE I. B2SHARE'S GENERIC METADATA MAPPING

UI field	Description	MARC XML field
Title	Name of the submission	245a
Description	Description of the submission	520a
Creator	Creator of the record	100a
Domain	Scientific domain	980a
Publisher	Publisher	260b
Publication date	When it was published	260c
OpenAccess	If the data should be public	542l
Contributors	Contributors	700
Language	Language of the submission	546a
License	License of the deposit	540a
Tags	Freeform tags	653a
Version	Submission version	250a
Domain fields	Domain specific fields	690a : 690b
System	EPIC PID record handle	024a
System	Checksum of the data	024a
System	user email	856f
System	site name	264b
System	creation time	264c
System	record id	001

A. Data Storage

As a data repository storing data is the main role of B2SHARE. B2SHARE uses a MySQL database for storing metadata information, and SQLAlchemy object-relational mapping [20] for the object-oriented run-time data processing.

We have defined a generic metadata model that covers a limited metadata set that can be used for any data deposit. Community or research area specific metadata models inherit from the generic metadata model. This ensures that every submission contains a certain level of basic information. The B2SHARE generic metadata model is based on MARC21 format [21]. The metadata model is described in table I. It is possible to export metadata to various formats like Dublin core, MARCXML, BibTex, MARC, EndNote, NLM, RefWorks.

To make data reusable and citable, B2SHARE is using EPIC PIDs [15] as a digital identifier. EPIC is compatible with the DOI system, and DOI servers can interpret EPIC PIDs as well. During the deposit process, B2SHARE issues a request to the EPIC PID service to create a new PID for the deposit, as well as updating EPIC with checksums calculated from the submitted files. The EPIC PID integration is optional, and can be made by setting the corresponding variable in the configuration file.

B2SHARE supports various filesystems for its storage back-end and can integrate with iRODS [22] [2] as a distributed storage sub-system. This allows choosing different backends with different characteristics (performance, backup, geographic distribution, etc.) according to the requirements of specific user communities. Using iRODS as back-end gives B2SHARE users the opportunity to have a safe replication service for stored data. This way, communities can have replication of their data and also support for other federation technologies.

B. APIs

As described above, B2SHARE has two APIs: a REST API and an OAI-PMH API. The REST API provides the basic functionality (CRUD) [23] for authenticated users to interact with B2SHARE. The REST-API can be used by research communities and users for interacting with B2SHARE via external programs. Furthermore, it gives users the ability to upload or download large datasets, which are not easily handled via web browsers. For simplicity and convenience, the REST-API is offering the same functionality as the web interface with exactly the same steps for deposit. It gives access to all related metadata as well as download links to the stored data file(s), and returns a response in the JSON format. It is also possible to get information for a specific domain or community through the REST-API. In the following we exemplify the usage of this API through `curl`, the common command line interface tool.

[2] iRODS(The Integrated Rule-Oriented Data System) is an open source data management software, which provides a means for managing large distributed collection of digital objects, maintaining metadata and applying data management policies.

The following steps need to be done in order to make a deposit in B2SHARE using the REST-API:

1) Create a new deposit:

```
curl -i -X POST  http://b2share.eudat.
    eu/api/depositions?access_token=
    LKR35GP7T
```

2) Upload (add) a file to a deposit:

```
curl -i -X POST -F file=
    @TestFileToBeUploaded.txt http://
    b2share.eudat.eu/api/deposition/23
    k85hjfiu2y346/files?access_token=
    LKR35GP7TF}
```

3) Adding related metadata and commit deposit:

```
curl -i -X POST -H "Content-Type:␣
    application/json" -d '{"domain":"
    generic",␣"title":"REST␣Test␣Title",
    ␣"description":"REST␣Test␣
    Description",␣"open_access":"true"}'
     http://.../api/deposition/
    DEPOSITION_ID/commit\?access_token=
    LKR35GP7TF}
```

Furthermore, by using the following command, authenticated users can retrieve deposited data for a specific community:

```
curl -i http://b2share.eudat.eu/api/records/
    nrm?access_token=LKR35GP7TF&page_size=10&
    page_offset=3
```

For metadata harvesting, B2SHARE supports the OAI-PMH protocol. The internal representation of metadata within B2SHARE is formatted in MARCXML, and can be used as mapping to OAI-PMH entries.

C. Access Control

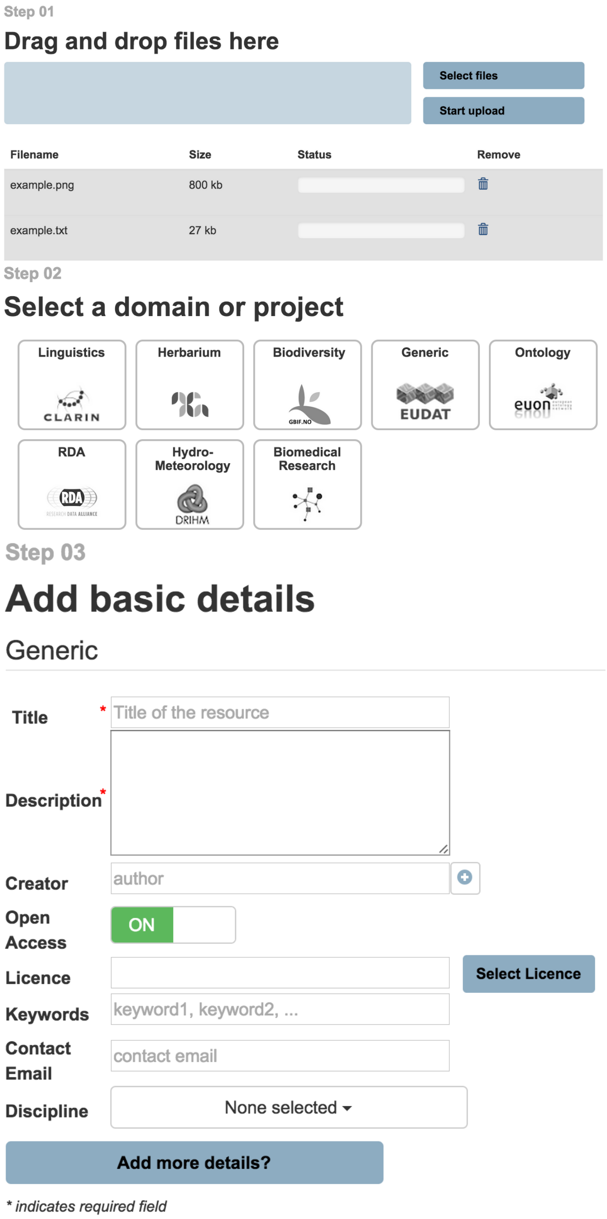
B2SHARE facilitates public read access to all the metadata in its repository, as it is fit for its mission of enabling scientific data sharing. Accessing the actual data, however, can be restricted by the depositor by using access control lists (ACLs). At the moment, when depositing, users can choose to make data items accessible publicly or restrict it as private data only. More sophisticated ACL schemes will be implemented in the future based on the requirements of the B2SHARE user communities.

Federation is a feature that allows users at one instance of B2SHARE to access data and metadata in another instance. Federated authentication and access control can be integrated with B2SHARE. This makes it possible to for example share data and metadata between storage systems or communities, while at the same time maintain a local user administration at each node.

D. User Interface

B2SHARE offers a web interface for easy upload/download of data together with related metadata, and provides different workflows based on specific domains. The focus of our development has been a clean and easy to use interface with a short learning curve. Federation and different interfaces can also be used by different organizations. In order to reach this goal B2SHARE's deposit module, B2Deposit, is providing the following simple deposit workflow:

1) Browsing or dragging the file(s) that user wants to deposit,
2) Choosing the related domain or community (Generic, BBMRI, etc.), which automatically generates the metadata form for the next step,
3) Adding metadata by filling in the form for selected domain or community.
4) And finally finishing by pushing the deposit button. (Figure 4)



Filename	Size	Status	Remove
example.png	800 kb		🗑
example.txt	27 kb		🗑

Open Access: ON

Discipline: None selected

Fig. 4. B2SHARE deposite workflow

These three steps give users an easy way to deposit and share their data. As default, B2SHARE offers the "Generic" domain that has a minimal number of metadata fields and can cover all data that does not belong to any other domain. In addition, there are optional fields for adding additional details for all domains, including Generic, for the users to fill in.

B2SHARE utilises Invenio's search facilities and allows to execute advanced queries using regular expressions, range of field values (e.g. dates) and selection of metadata. Facetted search can be used to narrow down the search results.

IV. USE CASES

In this section we discuss the usage of B2SHARE in two "long tail" data problems from the biodiversity field.

The Swedish Museum of Natural History (Naturhistoriska riksmuseet, NRM)[3], is investigating the use of B2SHARE as a data repository for digitized collections intended for interactions with "citizen scientists", particularly in the transcription of label data. By making the collection publicly available, not only the access to the collection should be enhanced but also the quality of the label information should be improved by collaborating with citizen scientists.

B2SHARE provides a suitable interface to store existing digital copies of NRM's collection items and store and share their new data and images. By uploading their images in B2SHARE, they will also be able to annotate label information for previously unprocessed images of collection items, study the quality of transcribed information and gain insights in the future design of a large-scale citizen science platform. Initially, a set of images of rare plants from a herbarium is suggested by NRM to become publicly available via B2SHARE so users can add metadata to the images. The user group could in this case include, or be restricted to, any set of potential users from NRM staff and local researchers up to all members of the general public who would like to participate in the transcription of label data. NRM's detailed use case for using B2SHARE is as follows:

1) *Image upload:* NRM members can upload scanned images of rare plants from a herbarium using the web interface or the REST-API and add related metadata. A unique identifier for the herbarium sheet will be assigned to the image which is typically corresponding to the herbarium sheet's catalogue number shown on the label.
2) *Iterative commenting and annotating images:* Uploaded images can be publicly visible and other users can easily find and reuse them. It is possible to make comments and add reviews to the image and share more information about deposited image. In future, by having editable metadata in B2SHARE, NRM users will be able to add or modify the entered metadata and annotate each image. Different version of annotations will be kept and after having at least two confirmations by two different users, they will be marked as complete. The overall annotation progress will be also display in an overview page.
3) *Metadata download:* By using B2SHARE as a repository, not only NRM staffs can upload and deposit data but also they can gather more information about each image from annotations and comments. They can download and use annotations for evaluating of the progress and metadata quality using the REST-API feature in B2SHARE.

The Global Biodiversity Information Facility (GBIF)[4] is an international open data infrastructure for sharing data about all types of life on earth. GBIF Norway is the Norwegian participant node of GBIF, hosted by the Natural History Museum at the University of Oslo.

The main task for GBIF Norway is to make primary data on biological diversity from natural history collections and observation databases in Norway freely available via the Internet. GBIF-Norway has so far published 81 Norwegian datasets which include a total of 13,3 million species occurrence records.

GBIF-Norway is intending to use B2SHARE as a data archiving platform for storing and sharing their datasets. They intend to share their datasets and metadata freely available under Creative Commons license CC0. Researchers and those who are interested in natural science can archive their biodiversity research data in B2SHARE. Table III presents their required metadata.

GBIF-Norway's use case for depositing data in B2SHARE is as follows:

1) *Dataset upload:* A natural science researcher creates a dataset and uploads it in B2SHARE. Respective metadata will be added to it as a description of data.
2) *Update dataset content:* When the data owner uploads new and updated versions of their archived datasets the previous version shall be archived before replaced by the new resource. The primary dataset identifier will resolve to the most recent version uploaded by the data owner.
   By supporting embargo in close future in B2SHARE, data owners will be able to embargo publishing their deposits for a limited time period. Using REST-API, researchers can upload batch datasets and have access to data easily.
3) *Darwin Core Archive version:* GBIF-Norway will be notified by automatic email when a new biodiversity dataset is uploaded or updated and will proceed to convert the respective datasets to the Darwin Core Archive format. The Darwin Core Archive format will be uploaded and connected to the same dataset metadata-level entry as the corresponding original dataset.

V. RELATED WORK

The need for providing easy to use tools for storing and sharing research data has resulted in a wealth of different approaches and tools. Larger, organized communities, like High Energy Physics or Climate, have often developed their own data management schemes, which we consider out of scope for this work as B2SHARE is explicitly targeting the "long-tail" science. We also consider simple file deposit services such as Dropbox out of scope as they do not provide essential features such as metadata and data discovery.

[3] http://www.nrm.se/

[4] https://www.gbif.org/

TABLE II. NRM'S METADATA DESCRIPTION

Field Name	Description	Mandatory
UUID	The unique identifier for the herbarium sheet shown in this image, typically corresponds to the herbarium sheets catalogue number shown on the label.	Y
Species name	Species name displayed on the herbarium sheet label.	Y
Collector name	Name of the collector shown on the label.	Y
Collection date	Collection date shown on the label. This may be incomplete and/or show only year or year/month.	Y
Locality	Location at which the item shown in the image was collected. This may range from a country name to specific place names and descriptions.	Y
Latitude	Only modern labels will typically carry coordinates.	N
Longitude	Only modern labels will typically carry coordinates.	N

TABLE III. GBIF'S METADATA DESCRIPTION

Field Name	Description	Mandatory
Version number	Version number	Y
GBIF ID	Refers to GBIF metadataset	Y
Country	Country	Y
Status	Endorsement status	Y

There are three highly related services that have emerged over the past few years and which are discussed in detail below: ZENODO, CKAN, and Figshare.

ZENODO [24] is an open source project, developed and hosted by CERN as an overlay on top of Invenio, similar to B2SHARE, which enables users to share data and publications on a web-based document repository with searchable metadata. Their objective is to provide a proper repository for preserving research data by using all of CERNs infrastructure. The file size limit is up to 2GB and it supports all type of files. Data in Zenodo can be published under different types of licences and it can have open or closed access. It encourage users to use a Creative Commons license for sharing their data. Excluding email addresses, all metadata is licensed under CC0. For citing data, Zenodo assigns a unique Digital Object Identifier (DOI) to each stored data. It provides two type of APIs: a REST-API for uploading data and an OAI-PMH API for harvesting metadata. Zenodo is storing data in CERN's Data Centres, with replication, to ensure that users data is safely stored.

B2SHARE and Zenodo have many common features, especially as they both are based on Invenio, and both projects are improving Invenio by adding their own modules and configurations. The main difference between B2SHARE and Zenodo is that B2SHARE supports federation and gives the communities the opportunity to have their own instances and installation. The modular design of B2SHARE does also allow for using different services as backends, including other EUDAT services.

Comprehensive Knowledge Archive Network (CKAN) [3] is an open source data platform, developed by the Open Knowledge Foundation, a non-profit organisation. It was originally developed for governments for sharing their data publicly in an "open government" spirit. CKAN is in use by numerous governments, organisations and communities around the world and according to their website it has more than 100 instances. Researchers can choose license for their data from a drop down menu. CKAN also supports permanent URIs for citation, e.g. DOIs, by extension packages [25]. Also, one of the useful features of CKAN is having version control for data and metadata[26] and facilities for data preview. It has a good support for APIs: RESTful JSON API for querying and accessing data. By supporting harvesting functionality, CKAN gives the organisations that already have data in their repositories, to pull their data into CKAN.

Many organisations already have their data in repositories with well-defined process and procedures for publishing and managing them. In this case these data can be simply pulled regularly into CKAN from their existing repositories. To facilitate this model we have developed a sophisticated and customisable "harvesting" mechanism which can fetch and import the mentioned data.

In contrast to B2SHARE, which is provided as a software package but also as a service, CKAN is a software solution for sharing data and it does not provide the service itself. It basically provides the required tools for sharing and finding data and organizations can use it as a platform for managing their information.

Figshare [4] is another platform for sharing data, although it is not open source code. Its development was started by Mark Hahnel and is now supported by Digital Science - a Macmillan Publishers company. There is a 250MB limit for file size in free plan and has different limits up to 1GB for premium users. It gives users 1GB of private storage for free. According to [27], data can be stored privately or publicly under CC license: CC-BY for different type of files, CC0 for datasets and, MIT for code. It also provides a desktop uploader and badges to be added to websites, blogs and, etc. This feature makes it quicker for researchers to upload files and get information about new sharings. Figshare assigns DataCite DOIs to every public data and citation, and can export to Endnote, RefMan and Mendeley formats. There is an API tool recently developed in Figshare which is still a beta version and helps users to push or pull data to or out of Figshare. It has different functionalities dependant on being authenticated user or not [28]. It gives users a preview of their files, and also additional commenting feature for them. Using Amazon Web Services (AWS), data on Figshare is guarantied to be safe and sound and always available.

Figshare is providing the service for sharing data and is available as a software package for large organisations (it is not end-user oriented). It is a commercial service, as opposed to B2SHARE which is free to use for the end users.

VI. Conclusions and Future Work

In this paper we have presented B2SHARE, a user-friendly, customizable data sharing service for scientific data. Thanks to its modular design, B2SHARE can easily be modified and customized for different user communities, using different metadata schema, storage backends, or embedding it in community workflow tools.

A publicly available instance of B2SHARE is made available to pilot users at
tt https://b2share.eudat.eu.

Future work on B2SHARE follows two stands: ease of deployment and additional functionality. With recent advancements in operating system containers like Docker [29], we are investigating having an orchestrated containerized version of B2SHARE for deployment. Such a version would particularly facilitate the deployment of B2SHARE as a cloud service.

A number of new features are currently under development. An "embargo" function will allow researchers to delay the public access to their data. This way, they can upload their data in B2SHARE and select any date for publishing in future. B2SHARE will also support version control for uploaded data files, such that the researchers can update their data files whenever they want. Furthermore, automatic preview of data will be added to B2SHARE. With this functionality, users can check the files before downloading. In the near future there will be more improvements for REST-API functionalities like adding querying and searching feature to it.

B2SHARE is a part of an extensive scientific data infrastructure provided by EUDAT and integration with other EUDAT services is under way. As mentioned earlier, B2DROP, B2SAFE, B2STAGE and B2FIND are the services of EUDAT's B2 suit that will integrate with B2SHARE soon.

Finally, one of the most important features of scientific data repositories is the potential insight they provide. For example, an online analytics system could provide services like auto completion when metadata is being entered, related work suggestions when browsing or more advanced reports on various researches. B2SHARE has the potential ability to make data available to data analytics systems and there are plans in EUDAT for supporting such features in the future.

Acknowledgment

This work was partly supported by the EC through the EUDAT project. We would like to thank our colleagues from the NRM Stockholm and GBIF-Norway for their respective use-cases. The authors wish to thank Mohamad Rezaei who assisted in the proof-reading of the paper.

References

[1] "EUDAT." [Online]. Available: http://eudat.eu/

[2] "Zenodo." [Online]. Available: https://zenodo.org/

[3] "CKAN." [Online]. Available: http://ckan.org/

[4] "Figshare." [Online]. Available: http://figshare.com

[5] "DataCite." [Online]. Available: https://www.datacite.org/

[6] "Common Language and Resource Technology Infrastructure." [Online]. Available: http://www.clarin.eu/

[7] "Dropbox." [Online]. Available: https://www.dropbox.com/

[8] "Box." [Online]. Available: https://www.box.com/

[9] "Google Drive." [Online]. Available: https://drive.google.com/

[10] P. Doorn and H. Tjalsma, "Introduction: Archiving research data," *Archival Science*, vol. 7, no. 1, pp. 1–20, 2007.

[11] L. Garrett, C. Silva, and M.-T. Gramstadt, "KAPTUR: technical analysis report," 2012. [Online]. Available: http://www.research.ucreative.ac.uk/1239/1/Kaptur_technical_analysis.pdf

[12] J. A. Lewis, "Research Data Management Technical Infrastructure: A Review of Options for Development at the University of Sheffield," 2014. [Online]. Available: http://files.figshare.com/1579074/RDMTI.pdf

[13] G. Pyrounakis, M. Nikolaidou, and M. Hatzopoulos, "Building digital collections using open source digital repository software: A comparative study," *International Journal of Digital Library Systems (IJDLS)*, vol. 4, no. 1, pp. 10–24, 2014.

[14] "Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)." [Online]. Available: http://www.openarchives.org/OAI/openarchivesprotocol.html

[15] "PIDs in EUDAT." [Online]. Available: http://eudat.eu/User+Documentation+-+PIDs+in+EUDAT.html

[16] "lindat license selector." [Online]. Available: http://ufal.github.io/lindat-license-selector/

[17] W. Gentzsch, D. Lecarpentier, and P. Wittenburg, "Big Data in Science and the EUDAT Project," in *Global Conference (SRII), 2014 Annual SRII*, Apr. 2014, pp. 191–194.

[18] M. van de Sanden, R. B. Baxter, C. Cacciari, G. Fiameni, E. Laure, D. Broeder, H. Thiemann, F. Iozzi, and J. Jensen, "D5.2.2 EUDAT Early Candidate Services," Tech. Rep., 2013. [Online]. Available: http://eudat.eu/system/files_force/EUDAT-DEL-WP5-D5%202%202-EUDAT%20Early%20Candidate%20Services.pdf?download=1

[19] "Invenio." [Online]. Available: http://invenio-software.org/

[20] "SQLAlchemy, a Python SQL Toolkit and Object Relational Mapper." [Online]. Available: http://www.sqlalchemy.org/

[21] "MARC 21 Format for Authority Data ." [Online]. Available: http://www.loc.gov/marc/authority/

[22] "Integrated Rule-Oriented Data System (iRODS)." [Online]. Available: http://irods.org/

[23] "B2SHARE HTTP REST API." [Online]. Available: https://github.com/EUDAT-B2SHARE/b2share/wiki/REST-API

[24] "Zenodo FAQ." [Online]. Available: https://zenodo.org/faq

[25] "CKAN extension for assigning a DOI to datasets." [Online]. Available: https://github.com/NaturalHistoryMuseum/ckanext-doi

[26] J. Winn and Others, "Open data and the academy: an evaluation of CKAN for research data management," 2013. [Online]. Available: http://eprints.lincoln.ac.uk/9778

[27] "Figshare Licensing." [Online]. Available: http://figshare.com/licensing

[28] "Figshare API." [Online]. Available: http://api.figshare.com/docs/intro.html

[29] "Docker." [Online]. Available: https://www.docker.com/