Grant agreement no. 675451

CompBioMed

Research and Innovation Action
H2020-EINFRA-2015-1
Topic: Centres of Excellence for Computing Applications

D6.2 - Deployment of Project Informatics Platform

Work Package:	6	
Due date of deliverable:	Month 12	
Actual submission date:	29 / September / 2017	
Start date of project:	October, 01 2016	Duration: 36 months

Lead beneficiary for this deliverable: *UCL*
Contributors: *UCL*

Project co-funded by the European Commission within the H2020 Programme (2014-2020)		
Dissemination Level		
PU	Public	YES
CO	Confidential, only for members of the consortium (including the Commission Services)	
CI	Classified, as referred to in Commission Decision 2001/844/EC	

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

Disclaimer

The content of this deliverable does not reflect the official opinion of the European Union. Responsibility for the information and views expressed herein lies entirely with the author(s).

Table of Contents

List of Figures

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

1 Version Log

Version	Date	Released by	Nature of Change
V1.1	04/09/2017	Stefan Zasada	First Draft
V1.2	18/09/2017	Stefan Zasada	After comments from reviewers

2 Contributors

Name	Institution	Role
Stefan Zasada	UCL	Author
Narges Zarrabi	SURFsara	Reviewer
David Wright	UCL	Reviewer
Peter V Coveney	UCL	Reviewer
Emily Lumley	UCL	Reviewer

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

3 Definition and Acronyms

Acronyms	Definitions
B2ACCESS	The EUDAT user management service
B2DROP	The EUDAT collaborative data workspace
B2FIND	The EUDAT metadata search service
B2HANDLE	The EUDAT PID assignment service
B2SAFE	The EUDAT data replication service
B2SHARE	The EUDAT data archiving and searching interface
B2STAGE	The EUDAT service to move data to HPC resources
CDI	Collaborative Data Infrastructure
EUDAT	An EU funded collaborative data infrastructure
PID	Persistent Identifier

*“This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Grant Agreement No **675451**“*

4 Introduction

The work described in this deliverable is a result of CompBioMed task **6.4 - Develop and Deploy an Informatics Platform** which will store all the data collected and processed within CompBioMed.

Originally, the task had planned to lead to the deployment of an instance of the p-medicine informatics platform to allow biomedical researchers to analyse their data, perform queries and extract content to initiate their modelling and simulation activities. All data within the data warehouse that reside at the core of this informatics platform are fully anonymised. However, this activity was subject to change based on the user needs analysis conducted by work packages 2, 5 and 6 to assess the data requirements of the project researchers and associate partners. The results of this analysis, presented in the next section, have greatly influenced the decisions regarding the deployment of this platform, described in the final sections of this document.

The system described in this deliverable is also designed to meet the detailed requirements outlined in deliverable **D 1.3 – Data Management Plan**, to which the reader is referred.

5 Requirements

CompBioMed has four separate tasks/deliverables that depend on understanding the project's data requirements. To understand the requirements of the project, within the project a data working group was formed to survey the requirements of the consortium.

This survey was delivered via SurveyMonkey and comprised fifty questions split into four categories:

Background project questions
Non-simulation data (e.g. used to build models)
Simulation data
General data questions

The survey received 27 responses from core project partner workflows and also associate partners, and this information has been used to inform the deployment of the data management platform.

Twelve responses were received from core workflow partners, and 15 from associate partners who have joined the consortium. Most respondents were working with file based data rather than structured databases, and identified an opportunity for CompBioMed to provide storage and collaboration services for data sharing.

Typical datasets ranged from 100 MB-10 GB, and the total volume of data project partners want to store is around 20-25TB with significant growth expected. Many respondents already had their own data management arrangements in place, but 25% of respondents did not want to continue using their existing data storage systems and were looking for CompBioMed to provide a service. Additionally, by adopting a robust best practice system, CompBioMed is likely to be able to assume leadership within the community in terms of data storage and

"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"

archiving. The basic data requirements in terms of preservation and sharing are summarized in Figure 1.

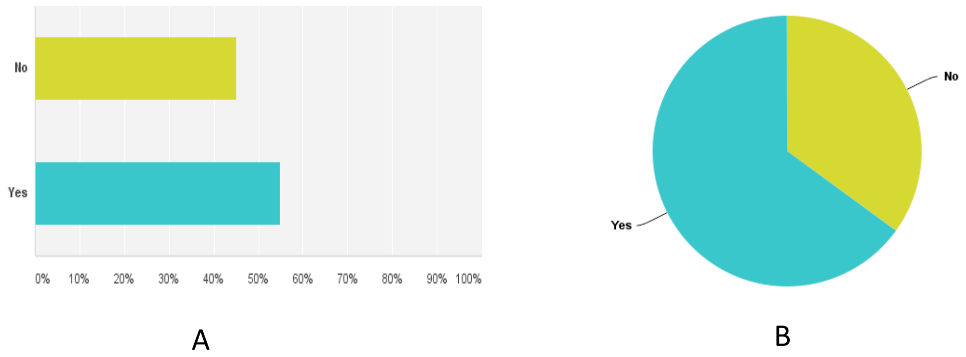


Figure 1. (A) Percentage of people who want to be able to share data outside of the project and (B) percentage of people who require a long term preservation service

Our survey also asked users about the current state of the art regarding their data management processes and procedures. We found that most respondents did not currently adopt tools such as persistent identifiers (PIDs) in order to reference data objects. The details can be found in Figure 2 - Figure 4.

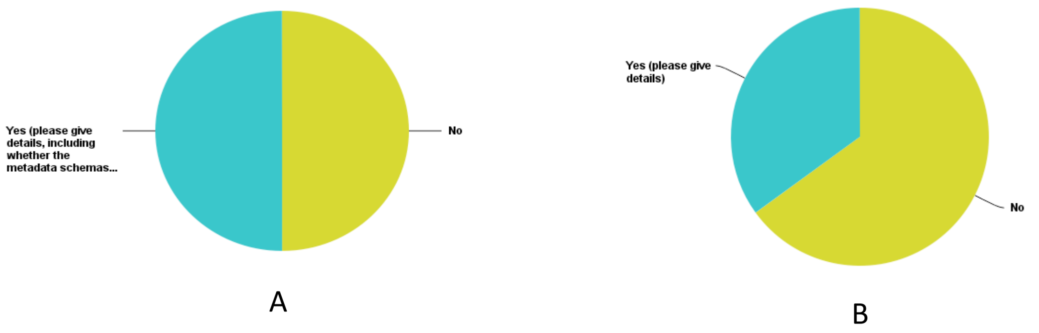


Figure 2. Percentage of projects that associate (A) metadata and (B) persistent identifiers with their data

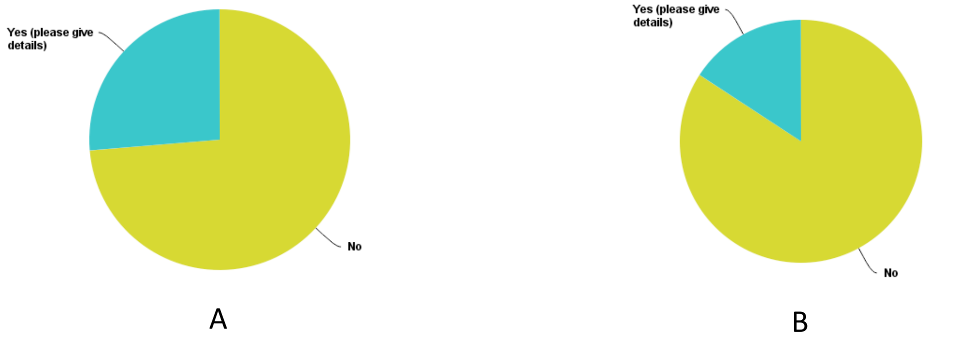


Figure 3. Percentage of projects that (A) have a coherent data architecture and (B) replicate their metadata

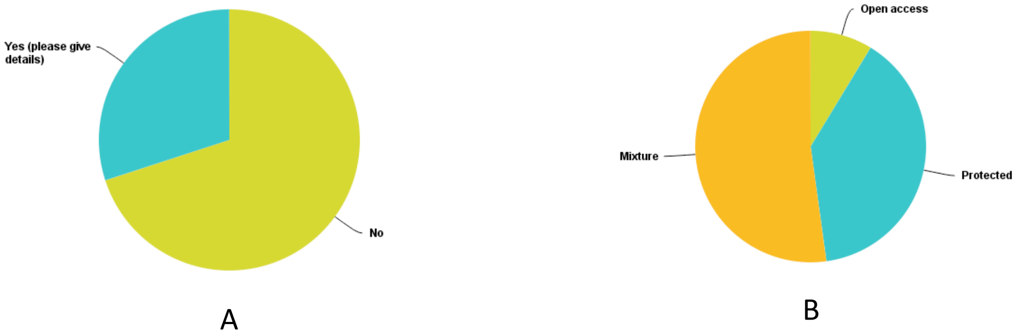


Figure 4. Percentage of projects that (A) have sensitive data and (B) can make their data available via open access

The fact that very few of the research communities represented by CompBioMed are making use of sensitive data with strong privacy requirements means that we took a design decision to leave the management of consent, anonymisation and privacy issues to the data owners, and have not attempted to implement a data management platform that can enforce such policies as mandatory.

6 Platform Description

The requirements outlined in the previous section point to a system capable of storing arbitrary data in many different file formats, which can be shared with other users and preserved for the long term.

It's beyond the scope of the CompBioMed project to develop such a system from scratch, but fortunately it is also unnecessary. The EUDAT project[1], started in 2011, aims to provide Europe's scientific and research communities with a sustainable pan-European infrastructure for improved access to scientific data. Burgeoning volumes of valuable and complex data create new challenges related to data management, access and preservation. EUDAT aims to address these challenges and exploit the opportunities using its vision of a Collaborative Data Infrastructure.

EUDAT, therefore, exists to work with projects such as CompBioMed, and provides an attractive means for such projects to meet their data management and preservation obligations. EUDAT comprises a set of services that can be composed in different ways to meet the objectives of the project. Some of these services are run centrally by EUDAT, whereas others exist as software that can be downloaded and deployed by projects that use EDUAT. This means that a project can adopt a multi-service ecosystem where some services are deployed within the project and linked to other services in the wider EUDAT infrastructure. That is the model that we have adopted for the CompBioMed data management platform. We describe the main EUDAT services we use in this section, and the local deployment of the B2SHARE service in the next. Figure 5 shows how these various services interact.

[1] http://www.eudat.eu

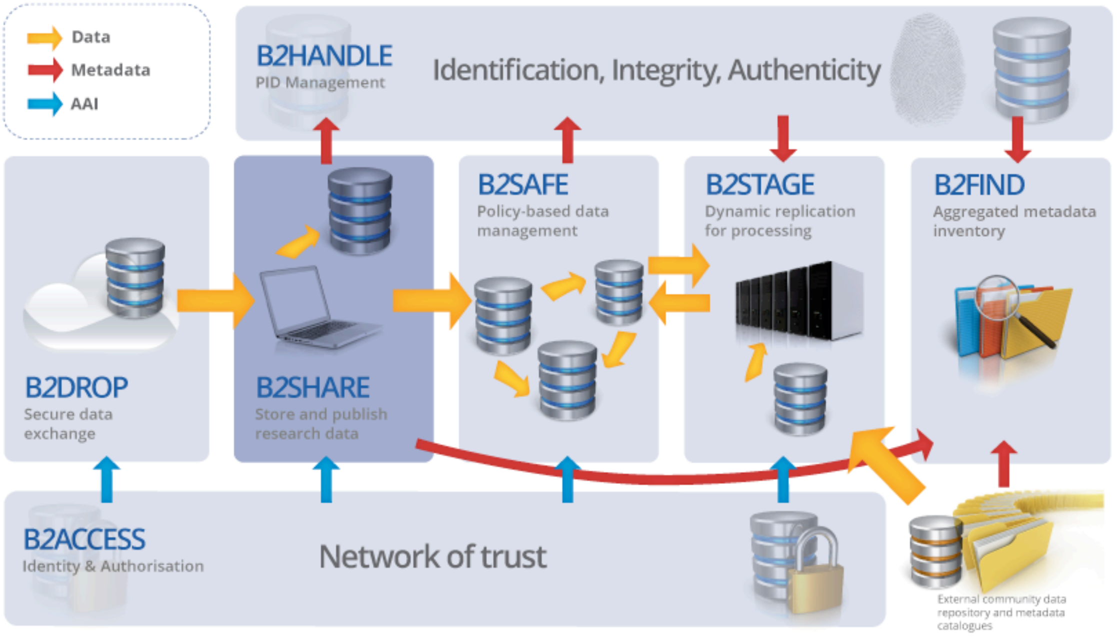

Data
Metadata
AAI
B2HANDLE
PID Management
Identification, Integrity, Authenticity
B2SAFE
Policy-based data management
B2STAGE
Dynamic replication for processing
B2FIND
Aggregated metadata inventory
B2DROP
Secure data exchange
B2SHARE
Store and publish research data
B2ACCESS
Identity & Authorisation
Network of trust
External community data repository and metadata catalogues

Figure 5. Architecture of the EUDAT system. The B2SHARE service is deployed on CompBioMed resources, and is integrated and interacts with other services in the wider EUDAT infrastructure.

6.1 EUDAT Services

Assigning a unique dataset reference makes it possible to refer to data in citations and enhances discoverability. It also provides the user with clear versioning capabilities for datasets. We use the centralised B2HANDLE service provided by EUDAT to assign persistent identifiers to all of the datasets that are archived in our project.

CompBioMed also makes use of the B2DROP service provided by EUDAT for sharing live data internally in the project, which will ease the transition of making data openly available in future. B2DROP is a tool to store and exchange data with collaborators and to keep data synchronized and up-to-date. CompBioMed takes advantage of the free storage space provided for research data within the B2DROP framework.

All data hosted within the EUDAT CDI is advertised through the central B2FIND catalogue and assigned a persistent identifier. The B2FIND service is a web portal allowing researchers to easily find and access collections of scientific data, and allowing them to access the data using a web browser. CompBioMed is in the process of developing a community metadata schema to describe the datasets generated by the project.

6.2 Locally Deployed Services

The EUDAT B2SHARE service allows data shared openly or kept private. Regardless of whether deposited data are made open or kept private, metadata records submitted as part of a data deposit are made freely available for harvest via OAI-PMH protocols. Accessible data are made available directly to users of EUDAT CDI services through graphical user interfaces and

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

application programming interfaces. We have deployed our own instance of the B2SHARE software, linked to the wider EUDAT infrastructure through the B2HANDLE and B2FIND services, to publish data for third-party use.

7 Deployment

An instance of the B2SHARE service has been deployed on resources at CompBioMed partner UCL, with 100TB of storage allocated to the project from UCL's resources.

B2SHARE is a graphical, web-based tool, which is designed to be easy to use, and B2SHARE also exposes an HTTP REST API. The basic operations of the interface are described below.

The service is available via http://b2share.compbiomed.eu.

The service is secured via EUDAT's OpenID provider (via B2ACCESS), as CompBioMed at present does not run its own identity provider service. This means, at present, that CompBioMed must register for an account with EUDAT.

7.1 Searching for Data

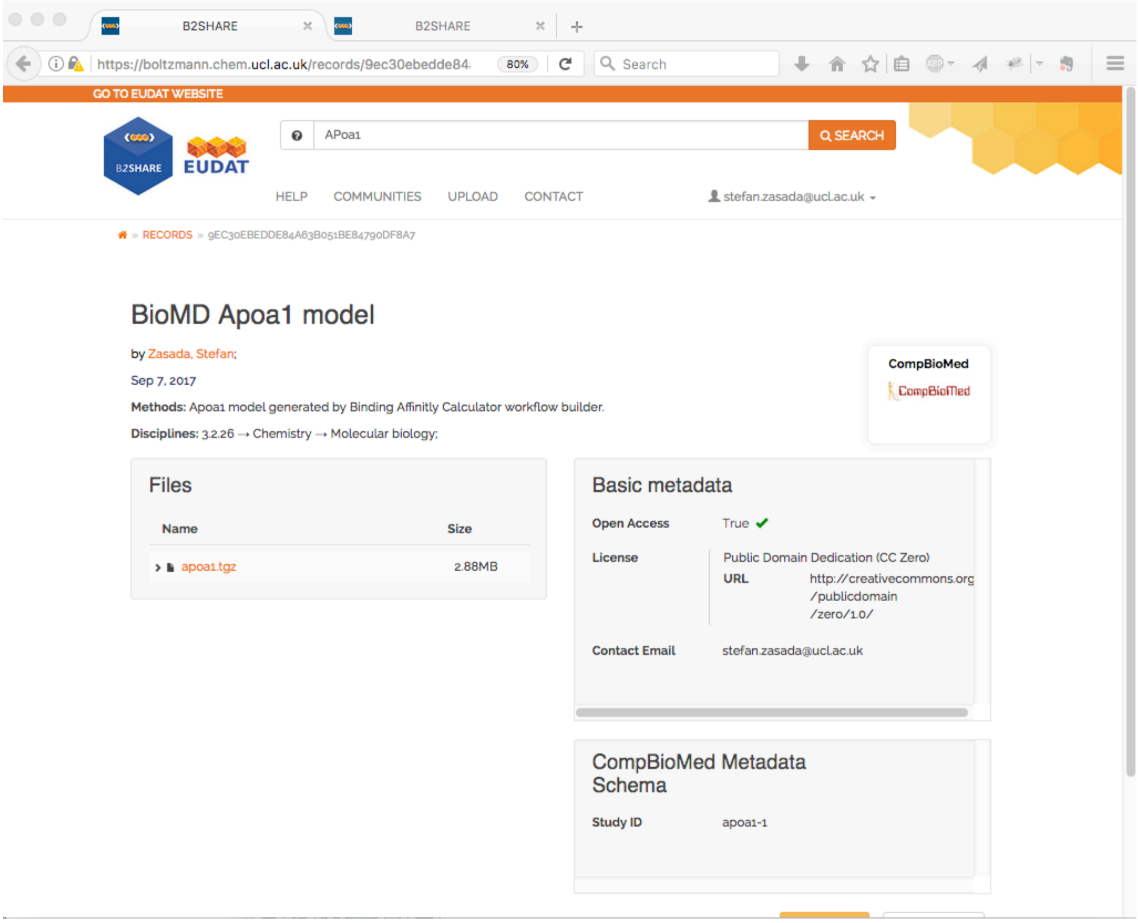
Both registered and unregistered users can search for data. The text entered can be part of a title, keyword, abstract or any other metadata. Unregistered users can only search for data sets that are publicly accessible.

Advanced searches can be performed by clicking the **Search** button, then entering the additional search criteria on the page that is shown.

Once a record has been found using the search facility, the user can click on its name to display the data's page. This page shows the metadata and files associated with the data object. For each file, the file size, checksum and PID are shown (see Figure 6).

The owner of a record is able to edit the metadata by clicking on the **Edit record** button.

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*



BioMD Apoa1 model

by Zasada, Stefan;

Sep 7, 2017

Methods: Apoa1 model generated by Binding Affinitly Calculator workflow builder.

Disciplines: 3.2.26 → Chemistry → Molecular biology;

Files

Name	Size
apoa1.tgz	2.88MB

Basic metadata

Open Access	True ✓
License	Public Domain Dedication (CC Zero) **URL** http://creativecommons.org/publicdomain/zero/1.0/
Contact Email	stefan.zasada@ucl.ac.uk

CompBioMed Metadata Schema

Study ID	apoa1-1

Figure 6. Each data object is assigned its own page, which displays its details and allows the user to download the associated files. The search process allows users to quickly find objects.

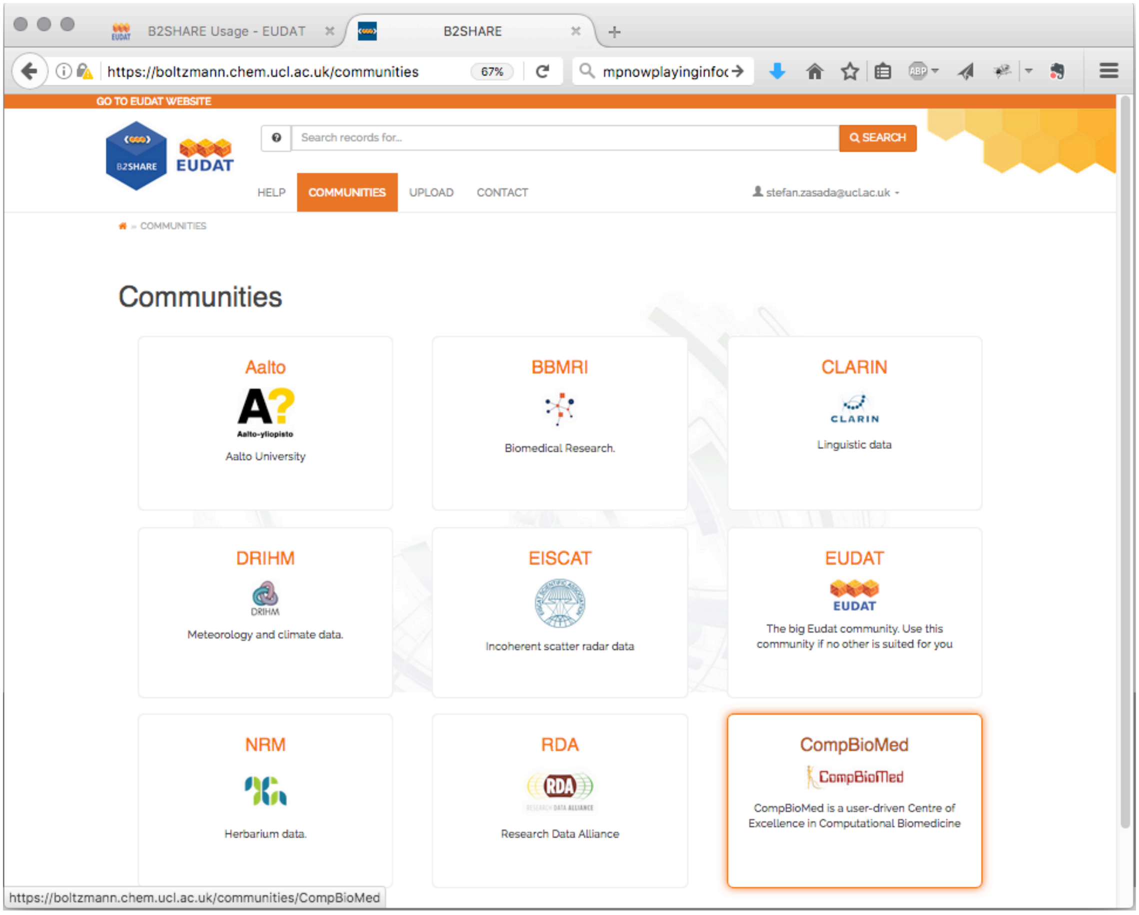
A user can also click on their username or email address on the front page and select **Profile** to go to the profile view of their account. From here they can find links to an overview of all their published or draft data records. At the end of the page current API tokens and new tokens can be generated on request.

7.2 Community

Data uploaded to B2SHARE is organised by community, where a community represents a specific metadata schema used to annotate the data objects stored within it.

A button is generated within the landing page of the interface for the community (see Figure 7), which allows users who are logged in to browse all data objects that have been uploaded to that community.

"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No ***675451****"*



Communities

Aalto
Aalto University

BBMRI
Biomedical Research.

CLARIN
Linguistic data

DRIHM
Meteorology and climate data.

EISCAT
Incoherent scatter radar data

EUDAT
The big Eudat community. Use this community if no other is suited for you

NRM
Herbarium data.

RDA
Research Data Alliance

CompBioMed
CompBioMed is a user-driven Centre of Excellence in Computational Biomedicine

Figure 7. The home page of the web interface, showing the available communities

7.3 Data Upload

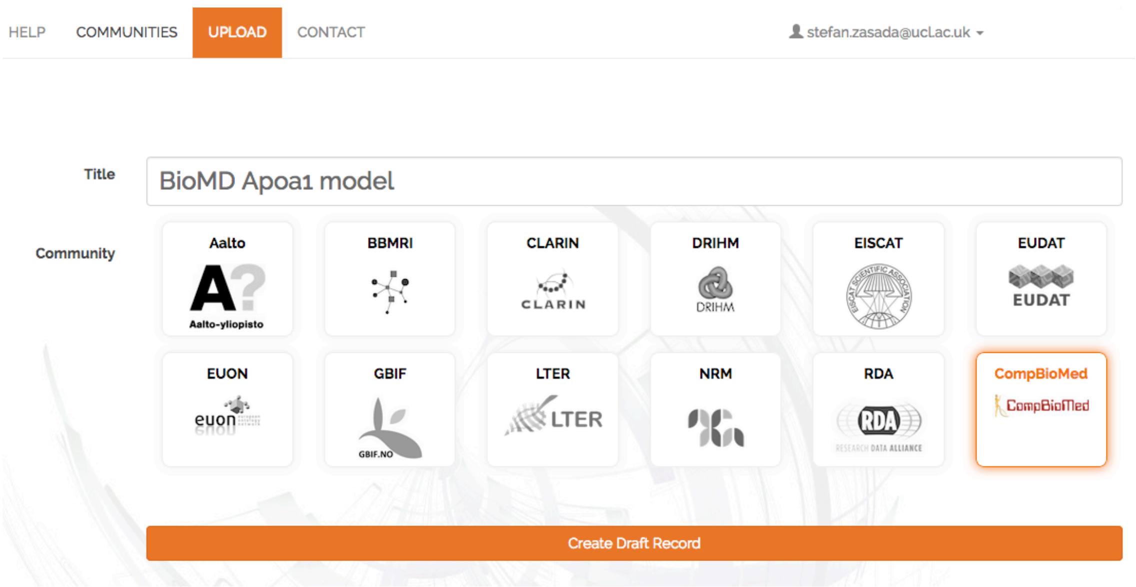
Only registered users are permitted to upload new records to B2SHARE. Clicking the Upload link in the main interface opens the first of a three stage process required to upload data. The steps are as follows:

1. Enter the title for the data object that is being uploaded, then select the Community that the data belongs to. In the case of CompBioMed, users should select the CompBioMed option (see Figure 8). Click on **Create Draft Record**.

2. Next the user can either drag and drop files from their local machine to the web interface, or import files stored in their B2DROP account.

3. Finally the user has to fill in the basic metadata fields. These fields depend on the community chosen. In the case of the CompBioMed community, basic information such as title, description (and type) and whether the data is open access are mandatory fields, while other information such as the creator, licence, URL, embargo date, keywords and study ID are optional. Hovering the mouse-pointer over the text field will show a description of the purpose of the field. The licence can also be selected through a built-in wizard. See Figure 9 for details.

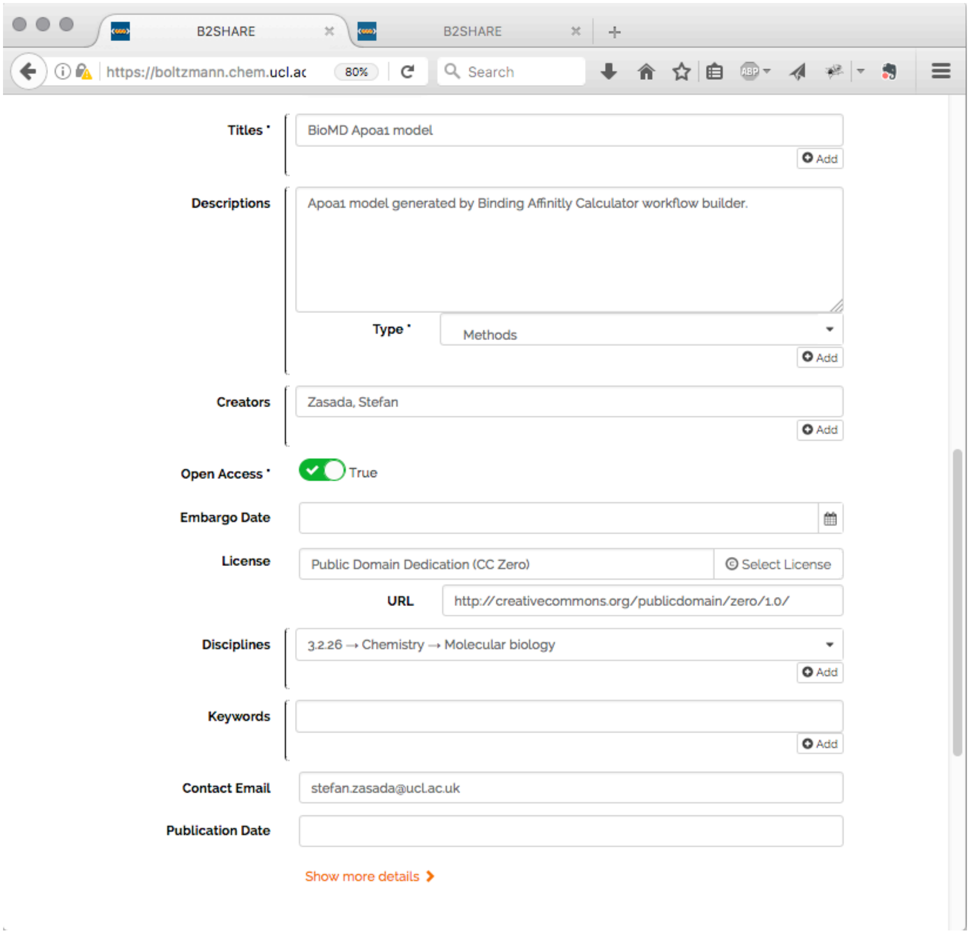
*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*



Title: BioMD Apoa1 model

Community: CompBioMed

Figure 8. The first stage of the data upload process – entering a title and selecting a community



Titles *: BioMD Apoa1 model

Descriptions: Apoa1 model generated by Binding Affinitly Calculator workflow builder.

Type *: Methods

Creators: Zasada, Stefan

Open Access *: True

Embargo Date:

License: Public Domain Dedication (CC Zero)

URL: http://creativecommons.org/publicdomain/zero/1.0/

Disciplines: 3.2.26 → Chemistry → Molecular biology

Keywords:

Contact Email: stefan.zasada@ucl.ac.uk

Publication Date:

Figure 9. The CompBioMed community metadata capture form in B2SHARE

Ticking Submit draft for publication will ensure that the uploaded data is assigned a persistent identifier (PID). This can then be used to refer to the data in future.

*"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No **675451**"*

7.4 REST API

The B2HARE HTTP REST API can be used to interact with B2SHARE via external services or applications, for example for integrating with other websites (research community portals) or for uploading or downloading large data sets that are not easily handled via a web browser. This API can also be used for metadata harvesting.

This is particularly useful in the context of CompBioMed workflows, since it means that workflows are automatically able to ingest data in the CompBioMed B2SHARE service as they are created, thus alleviating the burden on the user.

Only authenticated users can use the API. Each HTTP request to the server must pass an access_token parameter that identifies the user. The access_token is an opaque string which can be created in the user profile section of the B2SHARE web user interface. B2SHARE's access tokens follow the OAuth 2.0 standard, and allow API calls to be made on a user's behalf.

To get an access token, a user must login to the B2SHARE web interface and click on his or her username. On the User Profile page, go to the **API Tokens** section, enter a token identification name (e.g. **my-workflow**) and click **New Token**. This will create an access_token, which can then be used when making API calls.

8 Conclusions

The deployment of B2SHARE meets a specific requirement of the CompBioMed project to be able to archive and share research data. The benefits are twofold: the service is effectively run by CompBioMed, on CompBioMed managed resources. However, it integrates with the wider EUDAT infrastructure and can leverage EUDAT services such as B2HANDLE and B2FIND, greatly improving the sustainability efforts of the project to publish and preserve its data outputs.

We are in the process of developing a community metadata standard to describe the data produce by the diverse research strands of CompBioMed.

We also plan to engage in further work to automate the ingestion of data into B2SHARE directly from project workflows, via the B2SHARE API.

"This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No ***675451****"*