EGI-Engage

Security toolset release for BBMRI-ERIC

D6.11

Date	13 November 2016
Activity	WP6
Lead Partner	KTH
Document Status	FINAL
Document Link	https://documents.egi.eu/document/2981

Abstract

This deliverable describes the security toolset for BiobankCloud and the extensions performed in EGI-Engage since February 2016, the release of milestone M6.2 (Security and privacy requirements and secure storage architectural design are agreed). BiobankCloud builds on a Hadoop distribution, called Hops (www.hops.io), which was developed in the BiobankCloud project, to provide scalable storage and processing for genomic and Biobank data. BiobankCloud provides a web-based user interface for accessing and processing data stored in Hops that already provides 2-factor authentication. In this deliverable, we present extensions to support federated authentication with Shibboleth, which will enable easier integration of BiobankCloud in large organizations that run Shibboleth (such as EGI, universities, and Biobanks). Our solution is based on implementing an Apache webserver as a service provider (SP) that facilitates authentication with a Shibboleth Identity Provider (IDP) and as a proxy-frontend to BiobankCloud. We also integrate our shibboleth extensions with both the UI (Hopsworks) and our platform for automated installation based on Karamel and Chef.

The EGI-Engage project is co-funded by the European Union (EU) Horizon 2020 program under Grant number 654142

COPYRIGHT NOTICE

DELIVERY SLIP

	Name	***Partner/Activity***	***Date***
From:	Jim Dowling, Kamal Hakimzadeh	KTH / WP6 (BBMRI Competence Centre)	2016-11-12
Moderated by:	Małgorzata Krakowian	EGI.eu / WP1	
Reviewed by	Gergely Sipos	EGI.eu / WP6	2016-11-17
	Gautier Berthou	SICS Swedish ICT	2016-11-29
Approved by:	PMB and AMB		2016-11-12

DOCUMENT LOG

Issue	***Date***	***Comment***	***Author/Partner***
v.1	10/11/2016	First draft. Document writing was shared.	Jim Dowling & Kamal Hakimzadeh / KTH
v.2	13/11/2016	First version of the document for review.	Jim Dowling & Kamal Hakimzadeh / KTH
v.3	08/12/2016	Updated based on reviewer comments.	Jim Dowling / KTH
FINAL	12/12/2016	Final version	Jim Dowling & Kamal Hakimzadeh / KTH

TERMINOLOGY

A complete project glossary and acronyms are provided at the following pages:

- https://wiki.egi.eu/wiki/Glossary
- https://wiki.egi.eu/wiki/Acronyms

Contents

Executive summary

This deliverable describes security tools developed by the BBMRI Competence Centre of the EGI-Engage project, within the context of the IT infrastructure of BBMRI-ERIC and the Common Service IT instrument, to which all the full-member countries of BBMRI-ERIC contribute. It follows up on experience from the BBMRI Preparatory Phase 1 as well as the security and privacy requirements defined in milestone M6.2[1] of the EGI-Engage project's Competence Center for BBMRI.

BiobankCloud builds on a new distribution of Hadoop, called Hops (Hadoop Open Platform-as-a-Service)[2], developed within the EU-financed BiobankCloud project. BiobankCloud supports 2-factor authentication for users, supports user and study management, allows users to upload and manage files in studies, and run data-parallel programs, based on frameworks such as Spark, Flink or Adam. In the first part of this document, we summarize security features of BiobankCloud. This deliverable addresses a significant missing security feature that was missing from the IT infrastructure of BBMRI-ERIC: support for federated authentication with Shibboleth. The rest of this document describes how we added support for Shibboleth to BiobankCloud, and the BiobankCloud access control model, that has been updated through the course of this project.

[1] EGI-Engage M6.2: https://documents.egi.eu/document/2677
[2] HOPS: www.hops.io

1 Introduction

Tool name	Shibboleth for BiobankCloud
Tool url	http://www.hops.io/?q=shibboleth
Description	Support for Shibboleth Authentication for BiobankCloud/Hopsworks
Value proposition	Enable easier integration of BiobankCloud/Hops in large organizations that run Shibboleth (such as EGI Federated Cloud, universities, and Biobanks)
Customer of the tool	BBMRI-ERIC as a Resource Provider
User of the service	Both individual researchers and site admins
User Documentation	http://hops.readthedocs.org/en/latest/
Technical Documentation	http://hops.readthedocs.org/en/latest/
Product team	KTH
License	Apache v2
Source code	https://github.com/hopshadoop http://www.hops.io/?q=shibboleth

2 Service architecture

2.1 High-Level Service architecture

BiobankCloud (Bbc) is a web application deployed in a Glassfish application server with 2-factor authentication, provided by our own custom authentication realm. Using Glassfish allows us to leverage container-managed authentication in the Java Enterprise Edition (EE) platform. For Shibboleth support, we added the ability to disable support for 2-factor authentication in our custom authentication realm. Typically, users will authenticate themselves against a Shibboleth Identity Point (IDP) and a Shibboleth Service Provider (SP), implemented as an Apache webserver, will provide secure access to Hopsworks on Glassfish. (See Figure 1.)

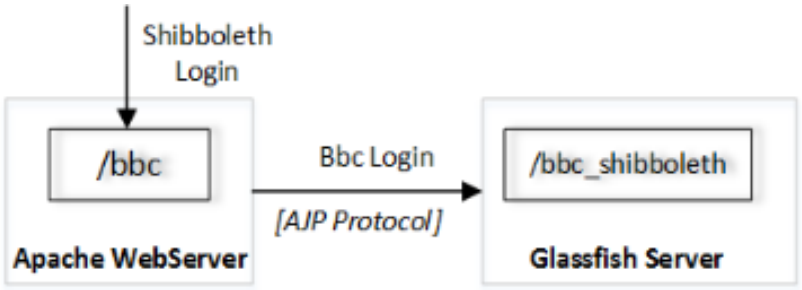

Figure 1. Shibboleth Login to BiobankCloud by fronting the Glassfish Server (running BiobankCloud) with a Apache Server that proxies requests to Glassfish using the AJP Apache module.

Error! Reference source not found. illustrates how the login process will be performed ith Bbc. Each Bbc user has two options to log in. If a user has her own Bbc account already, she may choose to login with it as the first option. Once she logs in, she will be authorized with the privileges of that user. The second option is to use a Shibboleth account. Here we must note that, the complete list of institutions from which a user can login via Shibboleth accounts have not been determined yet. If the Shibboleth login is granted, the login process continues with further steps.

1. If this user has used Bbc before and also logged in by using her Shibboleth account, it is likely that her Shibboleth account has been mapped to her local Bbc account already. If this is the case, the user is granted with the privileges of the mapped user.
    - If no mapping exists between the Shibboleth account and any of the local Bbc accounts, Bbc checks if there is any candidate account in its database which can be mapped to the user's Shibboleth account. Candidates are searched through e-mail addresses.
2. If a matching e-mail address is found, the mapping is done automatically by Bbc by adding that user's unique Shibboleth ID to the Bbc user database. Once the user is mapped to the Bbc account, she is authorized as the mapped user.
    - If no candidate is found, Bbc asks the user whether he has already a Bbc account under a different email address?
        - If yes, Bbc requires the user to login into Bbc with his Bbc account and redirects the user to a form where the user can manually couple his Bbc account with his Shibboleth account.
        - If the user has no Bbc account yet, he is asked whether he wants to have one?
            - If yes, Bbc creates an account for the user by taking the received attributes of the authenticated Shibboleth account and does the mapping between the Bbc and Shibboleth accounts. At this point, user might also enter further information about himself which does not exist among the Shibboleth attributes.
            - Otherwise, the user will be authorized with the privileges of the built-in guest account. The guest account only allows access to public data.

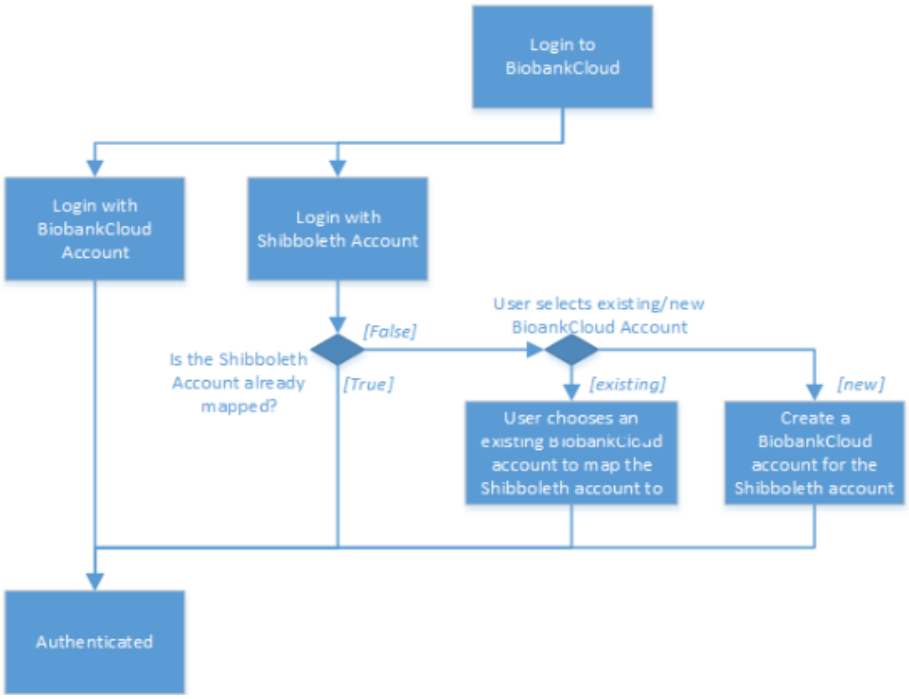

Figure 2. BiobankCloud Login Process with Shibboleth Enabled

2.2 Integration and dependencies

New BiobankCloud accounts can be created by the Apache SP using the new registration page. This required adding a new REST service for registering the new account via the Apache SP, who is a trusted service who can call the REST endpoint (See Figure 3). Nevertheless, it is recommended that in a Shibboleth setup, network traffic to the BiobankCloud server should be restricted to only come from the Apache SP. This can be done by adding IPTables rules to the host operating systems for the SP that only allow network traffic to come from the Apache SP.

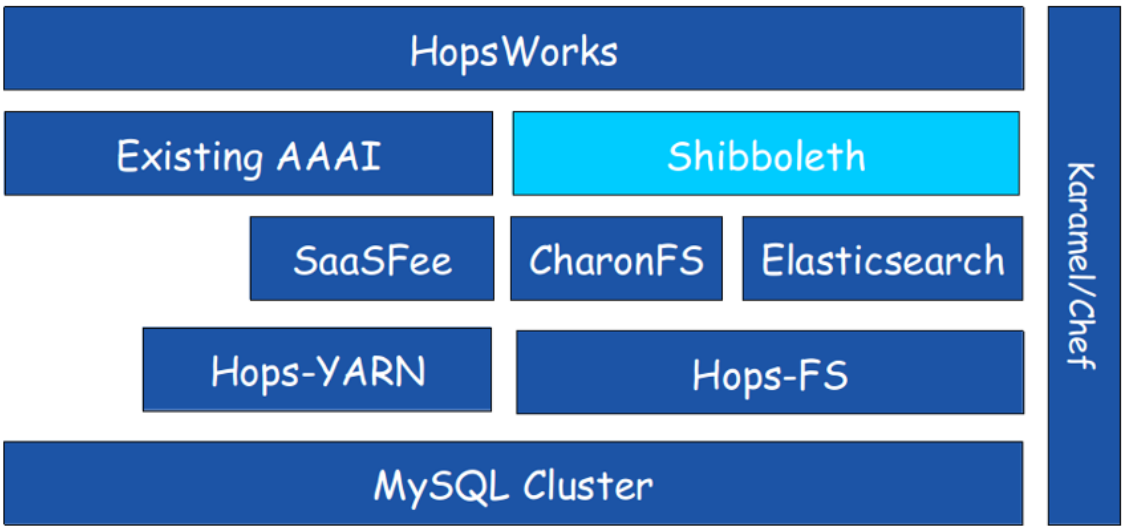

Figure 3. BiobankCloud Architecture with Added support for Shibboleth

2.3 Configuration of Shibboleth

Several configuration files are generated when our new chef cookbook, shibboleth-chef, is used to install the Apache SP that acts as the frontend to Glassfish, containing BiobankCloud. We should enforce HTTPS on Apache, /etc/httpd/conf.d/shib.conf, that allows a front-end machine to proxy a virtual host through to a server running on another machine: /etc/httpd/conf.d/ssl.conf should contain the FQDN of your hostname like this: ServerName shibtest.hops.io:443

Near the bottom of /etc/httpd/conf.d/ssl.conf but before the closing </VirtualHost> directive add the following:

```
# don't pass paths used by rApache and TwoRavens to Glassfish
ProxyPassMatch ^/custom !
ProxyPassMatch ^/bbc !
ProxyPassMatch ^/hopsworks !
# don't pass paths used by Shibboleth to Glassfish
ProxyPassMatch ^/Shibboleth.sso !
ProxyPassMatch ^/shibboleth-ds !
# pass everything else to Glassfish
ProxyPass / ajp://localhost:8009/
<Location /shib.xhtml>
 AuthType shibboleth
  ShibRequestSetting requireSession 1
  require valid-user
</Location>
```

Note that /etc/httpd/conf.d/shib.conf and /etc/httpd/conf.d/shibboleth-ds.conf are expected to be present after installing Shibboleth. The shibboleth configuration file, /etc/shibboleth/shibboleth2.xml, should look something like the sample shibboleth2.xml file below. The cookbook should substitute your hostname in the entityID value. The cookbook should ensure that attributePrefix="AJP_" is added under ApplicationDefaults, otherwise the required Shibboleth Attributes will be null and users will be unable to log in.

```
<!--
This is an example shibboleth2.xml generated originally by http://testshib.org
See also:
- attribute-map.xml
- bbc-idp-metadata.xml
https://wiki.shibboleth.net/confluence/display/SHIB2/NativeSPConfiguration
-->

<SPConfig xmlns="urn:mace:shibboleth:2.0:native:sp:config" xmlns:md="urn:oasis:names:tc:SAML:2.0:metadata"
    clockSkew="1800">
    <!-- FIXME: change the entityID to your hostname. -->
    <ApplicationDefaults entityID="https://shibtest.hops.io/sp"
        REMOTE_USER="eppn" attributePrefix="AJP_">
    <!-- You should use secure cookies if at all possible.  See cookieProps in this Wiki article. -->
        <!-- https://wiki.shibboleth.net/confluence/display/SHIB2/NativeSPSessions -->
        <Sessions lifetime="28800" timeout="3600" checkAddress="false" relayState="ss:mem" handlerSSL="false">
                <SSO>SAML2 SAML1</SSO>
            <!-- SAML and local-only logout. -->
            <!-- https://wiki.shibboleth.net/confluence/display/SHIB2/NativeSPServiceLogout -->
            <Logout>SAML2 Local</Logout>
            <!--
                Handlers allow you to interact with the SP and gather more information.  Try them out!
                Attribute values received by the SP through SAML will be visible at:
                http://shibtest.hops.io/Shibboleth.sso/Session
            -->
<!-- Extension service that generates "approximate" metadata based on SP configuration. -->
```

```
        <Handler type="MetadataGenerator" Location="/Metadata" signing="false"/>
        <!-- Status reporting service. -->
        <Handler type="Status" Location="/Status" acl="127.0.0.1"/>
        <!-- Session diagnostic service. -->
        <Handler type="Session" Location="/Session" showAttributeValues="true"/>
        <!-- JSON feed of discovery information. -->
        <Handler type="DiscoveryFeed" Location="/DiscoFeed"/>
    </Sessions>
 <!-- Error pages to display to yourself if something goes horribly wrong. -->
    <Errors supportContact="root@localhost" logoLocation="/shibboleth-sp/logo.jpg"
        styleSheet="/shibboleth-sp/main.css"/>
<!-- Loads a metadata file that describes only the Testshib IdP and communicate with it. -->
  <!-- IdPs we want allow go in /etc/shibboleth/bbc-idp-metadata.xml -->
    <MetadataProvider type="XML" file="bbc-idp-metadata.xml" backingFilePath="local-idp-metadata.xml"
legacyOrgNames="true" reloadInterval="7200"/>
    <!-- Attribute and trust options you shouldn't need to change. -->
    <AttributeExtractor type="XML" validate="true" path="attribute-map.xml"/>
    <AttributeResolver type="Query" subjectMatch="true"/>
    <AttributeFilter type="XML" validate="true" path="attribute-policy.xml"/>
  <!-- Your SP generated these credentials.  They're used to talk to IdP's. -->
    <CredentialResolver type="File" key="sp-key.pem" certificate="sp-cert.pem"/>
  </ApplicationDefaults>
<!-- Security policies you shouldn't change unless you know what you're doing. -->
  <SecurityPolicyProvider type="XML" validate="true" path="security-policy.xml"/>
 <!-- Low-level configuration about protocols and bindings available for use. -->
  <ProtocolProvider type="XML" validate="true" reloadChanges="false" path="protocols.xml"/>
</SPConfig>
```

attribute-map.xml

By default, some attributes /etc/shibboleth/attribute-map.xml are commented out. Edit the file to enable them.

bbc-idp-metadata.xml

The configuration above looks for the metadata for the Identity Providers (IdPs) in a file at/etc/shibboleth/bbc-idp-metadata.xml. You can download a sample bbc-idp-metadata.xml file and that includes the TestShib IdP from http://testshib.org.

Disable SELinux

You must set SELINUX=permissive in /etc/selinux/config and run setenforce permissive or otherwise disable SELinux for Shibboleth to work. "At the present time, we do not support the SP in conjunction with SELinux, and at minimum we know that communication between the mod_shib and shibd components will fail if it's enabled. Other problems may also occur." –https://wiki.shibboleth.net/confluence/display/SHIB2/NativeSPSELinux

2.4 BiobankCloud Access Control Changes

Since M6.2 (Security and privacy requirements and secure storage architectural design are agreed) from February 2016, there have been some modifications to the access control model for BiobankCloud. The full access control model can be read at *http://hops.readthedocs.org/en/latest/*.
Only the updates to BiobankCloud's access control model are detailed below.

X.509 Certificates for Secure Network Communication and Authentication of Clients

When a user creates a new Study in BiobankCloud, we now create a X.509 certificate (*project-specific user certificate*) which contains the unique ID assumed by the user within this Study. Services in BiobankCloud such as Kafka, HDFS, and YARN, also use a certificate, a *host-local certificate*, generated once on installation for every host. Both project-specific certificates and host-local certificates are signed by the same CA. BiobankCloud now makes use of certificates to provide client-based SSL authentication with services as well as secure network communications. The process of generating certificates is described here and illustrated in Figure 4:

- User Alice registers to BiobankCloud with an email *alice@gmail.com*. This email is used to authenticate Alice into the system with Shibboleth.
- When Alice creates a Study, the Project Manager module generates a certificate, which is signed by either the Root CA or an intermediate CA.
- The certificate is then persisted in BiobankCloud's database so that is made available when the Study invokes a service.
- Services also have their certificates signed by the same Root CA.

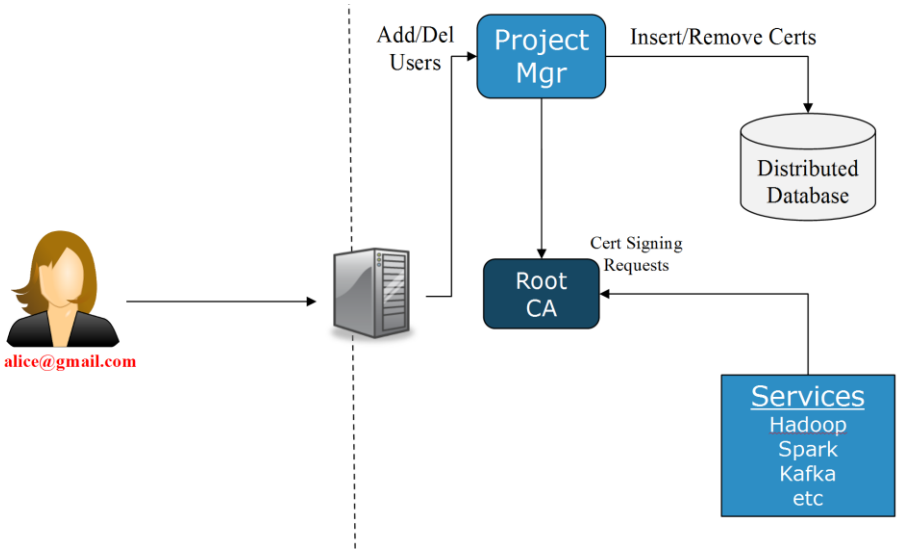

Figure 4. BiobankCloud (Hopsworks) has been extended to provide an X.509 certificate for each project-specific user as well as a host-specific certificate that is shared by services, such as Hadoop YARN, the Hadoop filesystem (HDFS), Spark, and Kafka.

3 Release notes

3.1 Requirements covered in the release

This release provides support for federated authentication to BiobankCloud. A new Apache SP component must now be operated as part of BiobankCloud and access to a Shibboleth IdP from within the target installation location is required.

New Functionality in Hopsworks

The database schema for users has been extended to capture the additional information regarding Shibboleth accounts. The corresponding Java Entity Beans "User" and Session Bean "UserManager" have been adapted to handle the additional information.

Hopsworks has been extended to create and edit users in a way that users can also couple/decouple their Hopsworks/BiobankCloud accounts with/from their Shibboleth accounts, if they have any.

Chef Cookbook Changes

We wrote a new cookbook (shibboleth-chef) to automate the installation and support the configuration of the Apache webserver and the Shibboleth daemon for the SP. The Apache modules we need to add were: mod_shib, ajp_mod, and mod_proxy.

We also had to patch to our hopsworks-chef cookbook to enable Shibboleth to work correctly with the AJP protocol, GRIZZLY-1787. We need to replace the glassfish4/glassfish/modules/glassfish-grizzly-extra-all.jar with a patched version 'glassfish-grizzly-extra-all.jar'.

A jk-connector network listener should be set up with Chef, and SSL warnings should be suppressed (due to a bug, GLASSFISH-20694):

```
asadmin create-network-listener --protocol http-listener-1 --listenerport 8009 --jkenabled true jk-connector
```

```
asadmin set-log-levels org.glassfish.grizzly.http.server.util.RequestUtils=SEVERE
```

4 Feedback on satisfaction

Gautier Berthou, a researcher at SICS Swedish ICT, tested the BiobankCloud SP against the Shibboleth IDP provided by the TestShib website. The platform, worked fine.

5 Plan for Exploitation and Dissemination

Name of the result	*Shibboleth for BiobankCloud*
DEFINITION	
Category of result	*Software & service innovation*
Description of the result	*Extensions to BiobankCloud to support Shibboleth for authentication. The extensions include updates to BiobankCloud's UI (called Hopsworks), a new Shibboleth SP (service provider) running as an Apache Webserver, and Chef cookbooks with orchestration for the automated installation of all of the above.*
EXPLOITATION	
Target group(s)	Research Infrastructures such as BBMRI nodes, the IT infrastructure of BBMRI-ERIC, and EGI Federated Cloud that already provide Shibboleth authentication. Many universities across Europe provide federated authentication services using Shibboleth.
Needs	These groups would like the seamless integration of BiobankCloud into their security infrastructures. By supporting Shibboleth, BiobankCloud users will no longer have to create new password-backed accounts, as existing accounts can be used to authenticate against a Shibboleth IDP.
How the target groups will use the result,	The target groups will be able to provide both BiobankCloud as a service and Hadoop-as-a-Service at their organizations, while integrating with existing federated authentication mechanisms, reducing the barrier to entry for BiobankCloud/Hops at such organizations.
Benefits	The benefits will be (1) for users there will be reduced overhead in managing separate login accounts and (2) for operators, there will be no longer the need to manage separate accounts on both BiobankCloud and the host institution's existing Shibboleth infrastructure.
How will you protect the results?	We will provide the software under an Apache v2 license to promote adoption of the tool along with the Hops platform that is similarly licensed under the Apache v2 license.
Actions for exploitation	We implemented support for automating the deployment of BiobankCloud along with Shibboleth. This involved implemented a Chef cookbook with support for Orchestration in Karamel.
URL to project result	http://www.hops.io/?q=shibboleth
Success criteria	We expect by the end of the BBMRI Competence Center project, BiobankCloud/Shibboleth will be deployed and running at least by 2 Biobanks of the BBMRI network.
DISSEMINATION	
Key messages	Now, BiobankCloud can integrate with your existing authentication infrastructure using Shibboleth.
Channels	Talks at conferences. Blogs.
Actions for	Jim Dowling presented the support for Shibboleth in BiobankCloud (Hopsworks)

dissemination	at Europe Biobank Week in Vienna, September 2016. A blog entry will be posted on , along with the existing documentation.
Cost	*0*
Evaluation	We will monitor the web statistics on both www.hops.io and www.karamel.io. Karamel will track the number of installations of BiobankCloud (Hopsworks) with shibboleth enabled.

6 Future plans

The BiobankCloud platform with Shibboleth support will deployed and tested at least at 2 biobanks during EGI-Engage, to gain experience at running the platform in production. The future of the software within BBMRI will be defined based on their feedback.