EXCELERATE Deliverable D9.3

Project Title:	ELIXIR-EXCELERATE: Fast-track ELIXIR implementation and drive early user exploitation across the life sciences	
Project Acronym:	ELIXIR-EXCELERATE	
Grant agreement no.:	676559	
	H2020-INFRADEV-2014-2015/H2020-INFRADEV-1-2015-1	
Deliverable title:	Report on implementation of value-added user applications and cohort integration	
WP No.	9	
Lead Beneficiary:	8 - CRG	
WP Title	Use Case D: ELIXIR framework for secure archiving, dissemination and analysis of human access-controlled data	
Contractual delivery date:	31 August 2018	
Actual delivery date:	31 August 2018	
WP leader:	Thomas Keane (EMBL-EBI), Jordi Rambla (CRG)	1: EMBL (EBI) 8. CRG
Partner(s) contributing to this deliverable:	CRG, EMBL (EBI)	

Authors and Contributors:

Sabela de la Torre (CRG – ES), Audald Lloret-Villas (CRG – ES), Jordi Rambla (CRG – ES)

Table of contents

1. Executive Summary

A range of RESTful APIs have been developed by the EGA in order to allow powerful users, consortia and research institutions to programmatically interconnect with our system. Whilst the majority of the users perform discrete and limited queries/submissions to the system, the majority of the queries/submissions are performed by a limited and advanced set of users. Taking these handful of teams as reference, we are herein describing the different approaches EGA APIs can be used for a better, faster and complete experience.

As described in the previous deliverable (9.2), the following programmatic endpoints are available and documented for all EGA users: submission API, public metadata API and private metadata API. The main difference between the later two is the requirement of authentication for complete discovery of user-related metadata.

EGA Submission API can be either directly used (https://ega-archive.org/submission/programmatic_submissions/submitting-metadata) or by using a tool or interface mounted in the top of it. EGA developed its own interface (Submitter Portal - https://ega-archive.org/submission/tools/submitter-portal) but it is great to see how other important partners are directly leveraging the API for their own solutions. We are focussing our attention on the ICGCsub, software provided by the ICGC consortia for all their projects worldwide for an harmonised and smooth submission to the EGA: https://github.com/icgc-dcc/egasub

Metadata can also be retrieved programmatically and in a custom manner from the EGA: https://ega-archive.org/metadata/how-to-use-the-api. This flexible possibility of obtaining and filtering the results allows (1) the generation of reports by the user, who do not need

to keep internal track of their submissions (and hence avoid duplications/mismatches), and (2) discovering crucial data information, a good filter before applying and downloading controlled access data. The first functionality can be either extended to the private metadata objects (i.e. in draft status) upon authentication (https://ega-archive.org/submission/programmatic_submissions/how-to-use-the-api). Worldwide known research institutions with a consistent system of data/metadata submissions (Broad Institute, DKFZ) are consistently considering such feature as a main information gatherer for their submissions.

2. Impact

Even though the EGA Submission API was officially launched by 2016, the Submitter Portal and API ensemble started production phase in January'18. Since then and until end of May'18, it has shown the following usage:

- 16% of analyses (916 submitted through the API in absolute numbers)
- 21% DACs (28)
- 12% datasets (32)
- 3% experiments (3,762)
- 21% policies (26)
- 3% runs (3,784)
- 8% samples (10,686)
- 13% studies (35)

The differences in the percentages are probably due that submission could be completed by different channels and by the fact some heavy users (e.g. The Sanger Institute) have existing mechanisms for managing highly repetitive parts, like the "run", "experiment" or "sample" objects. Besides, the implementation phase, which is mainly performed by informing and educating the users, is currently in place and make few more months to be fully completed. The goal is for new users to leverage these new tools and existing users to progressively migrate to them.

In parallel, the requests, suggestions and queries against our Metadata API have increased since the launch of version 2 on March 2018. Other than extensive filtering by metadata object, the users have found a great value on cross-querying in order to obtain data or metadata information based on a small set of knowledge. For instance, from the dataset ID any user can retrieve the biological information (samples) or the file types included in it. The received feedback so far has greatly helped us continue developing this tool so it becomes a natural complementation of metadata download - traditional approach for exploring and understanding the data archived at the EGA.

3. Project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives:

No.	Objective	Yes	No

Horizon 2020 grant n. 676559

1	To upgrade and make more portable -omics data collection and submission tools utilizing the European Genome-phenome Archive (EGA) as the core of an ELIXIR community secure data sharing network for - omics data.	X	
2	To enable value-added services at project specific, regional, or national resources by establishing ELIXIR-wide community facing tools that allow local resource owners and developers to add value to their systems through data and metadata services from the EGA.	X	
3	To extend and generalise the system of access authorization management and high volume secure data transfer developed in the EGA project to address the secure data access needs across ELIXIR resources and open new modes of secure data access such as through public and private clouds.		X

4. Delivery and schedule

The delivery is delayed: Yes ☑No

5. Adjustments made

None

6. Background information

Background information on this WP as originally indicated in the description of action (DoA) is included here for reference.

Work package number	9	Start date or starting event:	month 1
Work package title	**Use Case D: ELIXIR framework for secure archiving, dissemination and analysis of human access-controlled data**		
Lead	1 - EMBL-EBI		
Participant number and person months per participant **1 - EMBL (34 PM)**, 5 - UTARTU (16 PM), 6 - NBIC (0 PM), UMCG (LTP to NBIC) (5 PM), 8 - CRG (22.3 PM), 14 - UPF (23.5 PM), 20 - CSC (24 PM), THL (LTP to CSC) (3 PM), 24 - UiO (6 PM), 45 - UU (2 PM), SU (LTP to UU) (4 PM).			

Horizon 2020 grant n. 676559

Objectives

This Work Package has three main objectives:

To upgrade and make more portable -omics data collection and submission tools utilizing the European Genome-phenome Archive (EGA) as the core of an ELIXIR community secure data sharing network for - omics data. Tools developed here will support submission of all types of -omics data from human samples consented for biomedical research from disease consortia such as International Cancer Genome Consortium (ICGC), Rare Diseases (Rd-Connect), national cohorts, and biobanks. Emphasis is given to supporting investigator and locally driven research projects with human data consented for biomedical research. To enable these projects, the data submission tool chain will be made more portable and user-friendly with the goal of distributing a common toolset “in-a-box” to enable local and national groups to collect -omics data and meta data in a distributed manner which is consistent across European groups through ELIXIR coordination.

To enable value-added services at project specific, regional, or national resources by establishing ELIXIR-wide community facing tools that allow local resource owners and developers to add value to their systems through data and metadata services from the EGA. For example, local research projects would be enabled to make their data discoverable and searchable, and linked with available -omics data from various sources, by leveraging stable unique EGA identifiers. Further, locally developed project specific data portals will be enabled through defined standard APIs using real time secure data links which allow -omics big data archived in the EGA to be presented in combination with biobanks or cohort data.

To extend and generalise the system of access authorization management and high volume secure data transfer developed in the EGA project to address the secure data access needs across ELIXIR resources and open new modes of secure data access such as through public and private clouds. For example, a trusted ELIXIR Cloud service can receive local copies of selected datasets through a secure data mirroring system and provide access to data and compute to those users that already have data access permissions available from appropriate Data Access Committees stored in the EGA system. The WP will partner first with 2-4 large resource owners to gain the required expertise, document the process in multiple ELIXIR member states and finally to propose a way to scale up these services to match wider European requirements. This WP will also be used to drive creation of the ELSI framework that supports the workflow (WP12).

Horizon 2020 grant n. 676559

Work Package Leads: Thomas Keane, EMBL-EBI (since 1/3/2017); Jordi Rambla, CRG EGA (since 1/3/2017); Justin Paschall, EBI (up to 1/3/2017); Arcadi Navarro, ES (up to 1/3/2017)

Description of work and role of partners

WP9 - Use Case D: ELIXIR framework for secure archiving, dissemination and analysis of human access-controlled data [Months: 1-48]

EMBL, UTARTU, NBIC, CRG, UPF, CSC, UiO, UU

This WP delivers the core ELIXIR workflow for long term archive and re-use of human data consented for biomedical research requiring access-control based on a data access agreement and approval process. The workflow supports data submitters and ELIXIR Node coordination on data deposition into the EGA archive in a manner that will maintain data ownership in the hands of the original research data owner, enable data release to authorised individual users from the archive and to partner with downstream secure ELIXIR data analysis platforms. This workflow ad supporting infrastructure will allow the data owners to focus on their unique areas of data generation and analysis expertise while being able to rely on EGA and the ELIXIR infrastructure for their common –omics big data storage, coordination and distribution needs under appropriate legal frameworks. The work described here will leverage the work of other ELIXIR-EXCELERATE Work Packages, for example WP10 to scale each service structure to cover all ELIXIR Nodes and with WP4 for technical service support, and relies on WP12 to establish the necessary legal framework that supports workflows.

The Workflow can be summarized as:

1. Data preparation, validation, and submission to the EGA making use of common supporting tools and data models (e.g. through Node data Network, WP10). Focus on providing software tools and remote APIs enabling local leadership and customisation within context of specific projects, supported by common ELIXIR coordinated tools and data models.
2. Bidirectional linking and secure data streaming between -omics data archived in EGA and local repositories or data portals that hold further information about the project and samples.
3. Management of user access-rights for release of archived data to authorized researchers under Data Access

Agreements using ELIXIR tools, such as the REMS (WP4), that allow resource owners to manage data access rights.

4. Expanded access through ELIXIR partner secure clouds that can host EGA datasets, requiring the provision of metadata and authorization APIs
5. Data synchronization between the main EGA archive and authorized project specific resources and access points, such as compute clouds.

Task 9.1: Enhanced secure data submission tools. (49PM)

Horizon 2020 grant n. 676559

This task will update the existing EGA submission tools and documentation to facilitate large-scale data submissions operations, emphasizing local leadership and customization within a common framework.

Partners: ES, EMBL-EBI, FI

Subtask 9.1.1: Support for large scale submission of -omics data and sample metadata to the EGA. (30PM)

Support for large-scale submission of -omics data and sample metadata to the EGA through improved online tools, automated verification, and tools for the application of standard vocabularies to phenotype collection. These tools will make use of table “spread-sheet” based views of data for submitters less comfortable with technologies such as XML. Further tools and reports supporting global EGA stable identifier mappings will allow easier integration with local identifiers, in support of federated global tracking of submitted samples and their derived -omics data.

Subtask 9.1.2: Portable submission toolkit. (21PM)

This task is composed of data format definitions and software components, a “mini-EGA in-a-box” will allow increased local control and coordination of data collection, and allow early validation of standardized data and metadata formats.

This implementation provides the practical means for distributed projects to collect access-controlled human biomedical data in a manner that maintains a coordinated data model and dataset registry, enabling federated and a centralized single-point of discovery and access.

Task 9.2: Integrating centralized and distributed projects through transparent access to secure data: enabling local projects within a European wide framework. (40PM)

This task will enable local projects, such as study-specific data portals, local cohort resources, and national bioinformatics hubs by providing developer level APIs and services such that local efforts can efficiently build customized project branded solutions which make use of underlying ELIXIR and EGA tools and data archives.

Partners: ES, EMBL-EBI, FI, EE, NL, NO, SE

Subtask 9.2.1: Support secure integration of EGA data to downstream project client websites. (10PM)

Support secure integration of EGA data and metadata to downstream project client websites by providing new EGA programmatic interfaces that support standardized REST calls and provides results in ELIXIR endorsed formats (WP3 and WP6).

Subtask 9.2.2: Access management workflow support. (10PM)

Support access management workflows by data access committees through ELIXIR for EGA and other projects through developing applications of the Resource Entitlement Management Systems (REMS) expanding on an existing pilot project. This effort is focused on providing tools to delegate management to local projects and ELIXIR Nodes through new administrative roles.

Horizon 2020 grant n. 676559

Subtask 9.2.3: ELIXIR and EGA access integration. (20PM)

Specific efforts supporting controlled access-omics data infrastructure for use of partner national cohort studies in terms of submission, permissions management, and local and customized presentation of data under the cohort branding. Services will be tailored to respect the unique policy and data protection requirements of national cohorts, allowing single point of request and download from cohort branded web-pages. Support will be provided for distributed local hosting of datasets, within a common ELIXIR framework, where restrictions exist on the movement or hosting of data based on national borders.

Task 9.3: Federated authentication, large scale data management, and secure clouds in practice. (38PM)

This task is closely linked to the technically focused WP4 that provide the technical solutions required to deliver the outcomes of Task 9.1 and 9.2. In this task, technical components, including high volume secure data transfer and authentication and authorization management, are brought together to make -omics data from EGA and phenotypic data from cohort studies available for secure download, remote API access or from within public or private Cloud-based secure analysis environments. Cloud-based access to the EGA ecosystem provides a new access mode meeting a significant user need from research groups with limited local resources for compute and large-scale reference data storage.

Partners: EMBL-EBI, ES, FI, EE

Subtask 9.3.1: Large scale data mirroring support. (12PM)

Support for automated large scale data mirroring from the EGA archive to the authorized ELIXIR partner local services and cloud compute or HPC providers. This process instantiates concrete data flows based on data transfer technologies in WP4 to track domain specific files, versions of files, confirms transfer success, and tracks files available in different locations. Generic interfaces should provide transparent access to multiple underlying transfer and storage modules (e.g. gridFTP/irods/object store etc.)

Subtask 9.3.2: EGA data access authorization integration. (12PM)

Integrate EGA data access authorizations to local project data portals and Cloud access providers. This is a new service that allows authorized third-party services to programmatically check compliance with the current user data access authorizations from the ELIXIR coordinated repositories such as the EGA database each time user accesses a file in the cloud or other remote service. A first planned project using EGA data within the private, secure, cloud at CSC in Finland will provide our reference implementation.

Subtask 9.3.3: Data access APIs. (14PM)

Develop and implement standard data access APIs to be to used for inter and intra cloud communication and for secure remote REST API access in coordination with the Global Alliance for Genomics and Health (GA4GH).

For tasks 1-3 we expect to list a number of updates to the submission tools while we work with the first 2-4 chosen resources. These updates will be prioritised in the scope of

Horizon 2020 grant n. 676559

this WP. WP4 will provide AAI support for WP 9, and vice versa WP9 will work with WP4 to set the requirements for ELIXIR AAI services. WP9 needs to information on service component availability and this information is expected to be available from technical services registry such as cloud resource allocation, valid EGA data access authorizations, and file mirroring status if data are not yet ready to be used in the cloud. WP12 will Create a set of Legal Frameworks for ELIXIR-related operations that will be integrated within WP9 with the technical solutions devised for particular EGA needs.

Appendix 1: Report on implementation of value-added user applications

A1.1. Introduction

This deliverable, focussed on the exploration of EGA metadata and their integration to client systems, refers to subtasks 9.1.1 (Support for large scale submission of -omics data and sample metadata to the EGA), 9.2.1 (Support secure integration of EGA data to downstream project client websites) and 9.3.3 (Data access APIs).

We are herein describing the main API metadata points (for both submission and retrieval) and defining how users are already taking advantage of them with some concrete examples.

A1.2 EGA Metadata API

There is a series of necessary steps before accessing and downloading controlled human genomics data. These steps are likely to involve paperwork, a committee of specialists who review each application and a substantial amount of time. It is therefore essential to know the usefulness of a specific dataset for your project in advance.

Downloading the metadata tarballs, including the metadata files (XMLs) as well as cardinality maps (relationship across objects), used to be the best way to explore the information of a particular dataset. However, this option was limited to authorised users and thus the application challenge described in the above paragraph was not overcame.

Moving towards a FAIRification of our service, EGA developed this Metadata API for all the users to be able to (cross)query by specific and known metadata objects to fully understand the data before applying. A comprehensive tutorial of using the API and querying directly to the EGA can be found here.

Documentation

Common Aspects

Functionlity

The Metadata REST API alows EGA users to request metadata from the EGA.

Using this api you will be able to obtain the publicly available information from EGA study, sample, experiment, run, analysis, policy, dac and dataset. The API also allows for

Horizon 2020 grant n. 676559

objects to be cross-referenced in order to obtain for example all the datasets linked to a dac.

Identifiers

EGA objects can be identified by their **unique accession**. These are ID's displayed everywhere, shared among all EGA locations and specific for each data type (More information on the list below)

EGA Accession ID	EGA Object description
EGAS	EGA Study Accession ID
EGAC	EGA DAC Accession ID
EGAP	EGA Policy Accession ID
EGAN	EGA Sample Accession ID
EGAR	EGA Run Accession ID
EGAX	EGA Experiment ID
EGAZ	EGA Analysis Accession ID
EGAD	EGA Dataset Accession ID
EGAB	EGA Submission ID
EGAF	EGA File Unique Accession ID

Responses

Server responses

Understanding the server Responses

The server response is divided within 2 main sections they are the header and response.

- `"header"`
  - Contains information about the HTTP connection
  - `"code"`
    - Please see more about response codes on the table below
- `"response"`
  - Is the container for the response
  - `"numTotalResults"`
    - By default, only 10 results are shown.
    - This field tells you if there are more results hidden (e.g. if the user asks for all the samples, there might be hundreds or thousands)

Horizon 2020 grant n. 676559

- `"resultType"`
  - The type of the objects returned (e.g. sample, dac, dataset, etc.).
- `"result"`
  - The list of objects returned.

HTTP response codes

API responses are diverse but they always return a status call. Below you have a summary on what these codes mean and some tips that can help you troubleshooting them:

HTTP Status Code	Description	Resolution
200	OK	No error handling necessary
400	BAD REQUEST	Request incorrectly formulated
401	UNAUTHORIZED/FORBIDDEN	You do not have permissions over the object you are trying to access. / The authorized user is not permitted to make the given request.
404	NOT FOUND	The object /resource you are trying to access does not exist.
500	INTERNAL SERVER ERROR	This is a server-side error. Contact the Heldpesk to notify the issue.

Querying

Querying by Object

In the below documentation, worked examples can be found on how to use the API, to query about each unique EGA object.

https://ega-archive.org/metadata/v2/analyses/{id}

https://ega-archive.org/metadata/v2/dacs/{id}

https://ega-archive.org/metadata/v2/datasets/{id}

https://ega-archive.org/metadata/v2/runs/{id}

https://ega-archive.org/metadata/v2/samples/{id}

https://ega-archive.org/metadata/v2/studies/{id}

Worked example for the EGA data access committee(DAC) EGAC00001000514

https://ega-archive.org/metadata/v2/dacs/EGAC00001000514

Points to notice

Please, be advised that all EGA objects can be queried by using the same endpoint (this will return by default 10 results per page). For example:

https://ega-archive.org/metadata/v2/datasets

Returns 10 results.

https://ega-archive.org/metadata/v2/datasets?limit=5&skip=15

Returns 5 results skipping the initial 15 results. Notice that “skip” needs to be multiple of “limit” in order to display new results

Cros-Querying by Object

The EGA REST API Metadata also allows for the crossquerying of public metadata.

For example, should I want to query the datasets included in the ICGC DAC:

https://ega-archive.org/metadata/v2/datasets?queryBy=dac&queryId=EGAC00001000010

Below, you can find worked examples on how to perform the call to the API.

https://ega-archive.org/metadata/v2/analyses?queryBy=dataset&queryId=EGAD00001000645

https://ega-archive.org/metadata/v2/files?queryBy=dataset&queryId=EGAD00001000645

https://ega-archive.org/metadata/v2/runs?queryBy=dataset&queryId=EGAD00001000620

https://ega-archive.org/metadata/v2/samples?queryBy=dataset&queryId=EGAD00001000620

https://ega-archive.org/metadata/v2/datasets?queryBy=sample&queryId=EGAN00001092114

There are more endpoints included in this category. Please, find all the possibilities (green tick) in the table below.

Table

The table displays in a ordered fashion the objects that can be crossreferenced using the metadata API.

TO	Analysis	Dac	Dataset	Experiment	Policy	Run	Sample	Study	Submission	File
Analysis			✔					✔		

Horizon 2020 grant n. 676559

Dac			✔							
Dataset		✔					✔	✔		
Experiment			✔			✔		✔		
Policy		✔								
Run			✔	✔				✔		
Sample			✔							
Study	✔		✔	✔		✔	✔			
Submission										
File			✔							

Given that this API has been recently released, such functionality is not yet widely used across our regular users.Nonetheless, novel users as well as different services such as Repositive (https://repositive.io) or important research centers such as DKFZ (German Cancer Research Center - https://www.dkfz.de/en/index.html) have shown their interest, tried our system and suggested new queries or ways to display the information. In fact, a second version was launched in March 2018 with some changes in attributes upon user feedback. The mentioned services/institutions programmatically collect the metadata of interest in order to feed their own monitoring system.

A1.3. EGA Submission REST API

Along with the genomic and biomolecular files, submitters who wish to deposit their data to the EGA need to provide a series of information, so the final user of the these files can thoroughly understand the details. This set of required information is known as metadata and consists of a different set of objects:

- Study (this is the biological project)
- Samples (the information of the biological samples)
- Experiments (description of methodology details)
- Runs / Analysis (links the biological samples and the files within a certain study)
- Policy (terms and conditions under which the data can be accessed)
- Data Access Committee (DAC - team that governs the data and decides who should and should not have access to the data)
- Dataset (Downloadable unit that contains a series of runs/analysis, which are linked to a certain Policy and DAC).

Even though existent mechanisms for submitting the metadata existed (both for discrete submission and for large number of metadata objects) EGA developed a specific API for metadata submission in early 2016 (https://ega-archive.org/blog/rest-api-live/), which added multiple functionalities to this process. The new metadata submission was generated as alternative for the Programmatic submission (https://ega-archive.org/submission/sequence/programmatic_submissions) service and Webin interface (https://www.ebi.ac.uk/ena/submit/sra/#home) at the time. This new EGA-centric metadata submission point would allow all users to navigate through the metadata submission pipeline much more flexibly.

One the one hand, JSONs were included as second valid format to provide the information besides XMLs. Furthermore, it appeared intermediate statuses (such as draft or validated) and hence the possibility to provide metadata in batches instead of all at once, a very appreciated and requested functionality. See figure 1 for more details about the statuses. New range of notification and report statuses were also designed for a batter error interaction with the users.

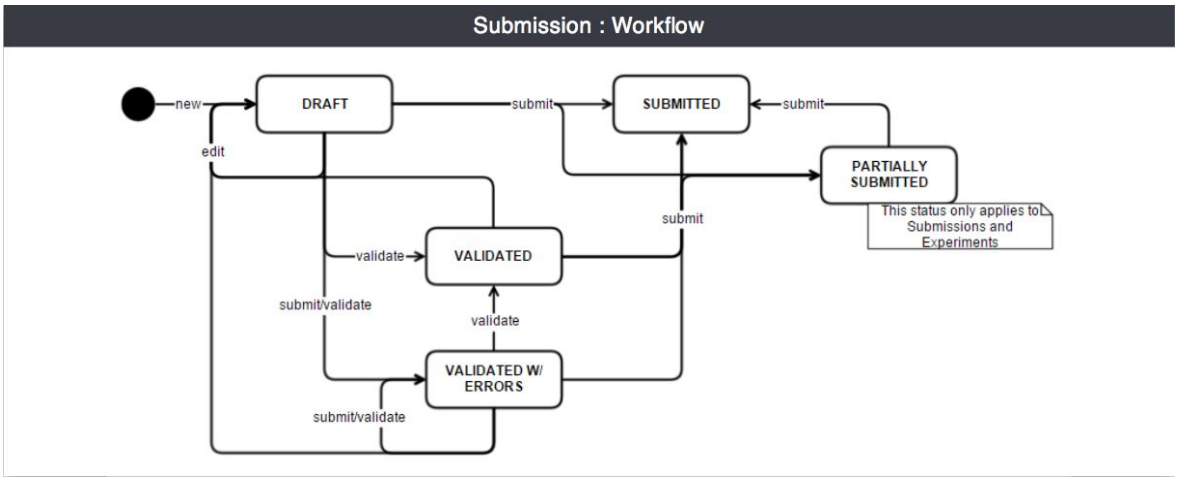

Figure 1. Different statuses for metadata objects

The release of the Submitter Portal in 2018 increased the visibility of such API, as the interface is mounted upon the Submission API, and became the default way to register the metadata at the EGA. More information about the API and the Submitter Portal can be found in Deliverable 9.2

There are two different usages for the Submission API:

- Retrieving metadata
- Submitting metadata

Retrieving metadata

Having a clear idea of the different metadata objects, details and statuses it is a key step for a successful submission to the EGA. Either because of not using the user-friendly interface or either because of massive number of objects, generating reports of the existing metadata has been challenging for some of EGA users.

It is now possible to query all the metadata objects and their particular details for a specific account upon authentication. The generation of personalised reports facilitates the whole process and avoids duplicates, confusions and mismatches with submitter own

Horizon 2020 grant n. 676559

records, very frequent otherwise. Final results can be filtered by object nature (sample, policy, dataset), status (draft, validated, submitted) and individual accession ID (when known). All documentation can be found online (https://ega-archive.org/submission/programmatic_submissions/how-to-use-the-api).

Submitting metadata

Defining all the information about the biological samples, experiments as well as its linkage to the files can be also performed programmatically via the Submission API. This new possibility (in continuous improvement for the last two years) has opened the range of possibilities seen before and comprehensively described in the documentation (https://ega-archive.org/submission/programmatic_submissions/submitting-metadata)

The user-friendly interface for such metadata registration (Submitter Portal) has become the recommended portal. Therefore, we expect the percent of metadata objects registered via the Submission API to dramatically grow for the next months. Nevertheless, users and institutions that have consistently submitter to the EGA by using alternative methods may have difficulties or concerns about changing the approach.

Having said so, we have gladly observed that intensive and recurrent submitters, such as the International Cancer Genome Consortium (ICGC - https://icgc.org) and the Broad Institute (https://www.broadinstitute.org) not only have been testing EGA submission API, but they have also adapted their pipelines and suggested improvements. Actually, it is worth mentioning that ICGC Data Coordination Centre, team that centralise worldwide ICGC-related submissions to the EGA and data applications, has developed a tool for all the collaborator countries: https://github.com/icgc-dcc/egasub

Even though some of the ICGC nodes prefer to keep their regular submission approaches, a substantial amount of them have perfectly integrated this tool within their current pipelines. Besides, EGA has been confirmed that EGAsub will be the recommended submission tool for all the participants in the next ICGC 10-years project. The development of the code for such bespoken tool has certainly encouraged us to keep improving the RESTful API and suggest broker institution and consortia to follow this path.

A1.4. Conclusions and next steps

The different APIs described so far have greatly empowered EGA most proficient users. While it was indeed possible to automatise metadata submissions to the EGA (mainly by generating compliant XML files), obtaining the information to fully understand each dataset as well as generating metadata reports for own monitoring was rather cumbersome. By taking full advantage of the metadata API (both submissions and retrieval - with or without authentication), the users are much more self-sufficient and are able to easily integrate their pipelines and EGA APIs. The improvement of the metadata richness and homogenization across projects is also worth mentioning.

According to the feedback EGA Helpdesk is receiving in a daily basis, and the feedback Helpdesk team is sharing back with the users, we envision an increase in the actual usage of the described APIs. A step forward in our communication and documenting could encourage submitters involved in small projects (the majority) to try these advanced

tools. A greater quantity and variability of user will end up, eventually, in a better understanding of our system limitations and subsequent improvement.

A1.5. Links

1. https://ega-archive.org/metadata/how-to-use-the-api
2. https://ega-archive.org/blog/rest-api-live/
3. https://ega-archive.org/submission/programmatic_submissions/how-to-use-the-api
4. https://ega-archive.org/submission/programmatic_submissions/submitting-metadata
5. https://github.com/icgc-dcc/egasub