



EXCELERATE Deliverable D3.2

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:	A report describing the semantically enriched Europe PMC that links literature to ELIXIR Core resource data, exemplified by GeneRif annotations	
WP No.	3	
Lead Beneficiary:	1 - EMBL	
WP Title	Data Resources and Services	
Contractual delivery date:	31 August 2018	
Actual delivery date:	31 August 2018	
WP leader:	Jo McEntyre, Christine Durinx	1- EMBL - EBI, 25 - SIB
Partner(s) contributing to this deliverable:	1- EMBL - EBI, 9 - CIPF, 12 - BSC, 14 - UPF, 15 - IMIM, 25 - SIB	

Authors and Contributors:

Jo McEntyre, Aravind Venkatesan, Julien Gobeill, Damiano Piovesan, Silvio Tosatto, Patrick Ruch

Table of contents

1. Executive Summary	3
2. Impact	3
3. Project objectives	4
4. Delivery and schedule	4
5. Adjustments made	4
6. Background information	5
Appendix 1: D3.2 A report describing the semantically enriched Europe PMC that links literature to ELIXIR Core resource data, exemplified by GeneRIF annotations	8
A1.1. Introduction	8
A1.2. Technical and infrastructural components to support scalable curation	9
A1.2.1 Europe PMC Annotations platform	9
A1.2.2 Annotations API	11
A1.2.3 SciLite	12
A1.2.4 Automated annotation submission system	12
A1.2.5 Fuzzy matching of GeneRIF statements	13
A1.3. Use Cases	14
A1.3.1. Literature-data linking exemplified by GeneRIF annotations	14
A1.3.2 Integrating annotations from Europe PMC Annotations platform into DisProt Database	16
A1.4. Conclusion	17
A1.5. References	17

1. Executive Summary

The advances in high-throughput technologies have resulted in tremendous growth in biological data, consequently increasing the number of published research papers. This provides challenges for curators and researchers alike in finding, and assimilating scientific facts described in articles. Therefore, services that support browsing of scientific content and identifying key biological concepts with minimal effort would be beneficial for the community. To this end, the task 3.3 of Work Package 3 under the Elixir-Excelerate grant is concerned with exploring technical solutions and the infrastructure needed to support scalability of curation tasks.

This document reports the work delivered so far in laying the foundation for an infrastructure that facilitates literature-data integration and enrichment. The work can be split into two main aspects: (1) establishing infrastructural elements, that includes, an automated system to ingest annotations, a platform to integrate annotations from various sources and APIs to redistribute the annotations; (2) applications to display annotations on articles and handle complex GeneRIF annotations. Finally, this work covers the implementation of a mechanism to create deep links between literature and data and the integrations of annotations as part of the database curation workflow.

The infrastructural elements delivered under Work Package 3 has demonstrated how text-mining annotations from different sources can be ingested and made available for reuse. Going forward, these components needs to be tested for its usability. This requires engagement the curation community to derive actionable insights that may feed into improving the various components, for example, text-mined annotations, browsing and article triage.

2. Impact

The key elements described in this report are as follows:

- **Europe PMC Annotations platform** - One of the primary goals of this task was to facilitate the aggregation and redistribution of text mining results from multiple text mining groups. This is supported by the Europe PMC Annotations platform, a community platform that integrates text-mined annotations from various sources, making them available for the wider scientific community. The platform currently hosts over 480 million annotations covering different annotation types, such as named entities (genes/proteins, diseases, organisms), accession numbers, protein-protein interactions and gene-disease associations, to name a few.
- **Annotations API** - The Annotations API provides programmatic access to the annotations hosted by Europe PMC Annotations platform. The goal of developing the API is to share the text-mining outputs to a wider community of biomedical scientists.
- **SciLite** - Previously we had reported the development of SciLite, an application that overlays annotations from the Annotations platform on research articles. Recently, the application has been improved to mitigate instances where annotations are not displayed due to HTML page rendering.
- **Annotation submission system** - The submission system was recently developed to streamline publishing of annotations to the Annotations platform. The system offers a private cloud storage account for annotation providers to periodically load and maintain data. Additionally, the system validates all incoming annotations for its compliance with the W3C recommended [Web Annotation Data Model](#).
- **Improved string matching algorithm to handle complex GeneRIF statements** -

An application was previously developed to map GeneRIF statements to the corresponding source articles. While the initial set of statements published via the Annotations platform were cases of direct mapping, the application has now been improved by implementing a string approximation algorithm (Levenshtein algorithm) to handle more complex GeneRIF statements.

- **Literature-data linking exemplified by GeneRIF annotations** - We have implemented a mechanism to facilitate bidirectional linking of annotations. All GeneRIF annotations were issued a URI. These URIs can be used to cite GeneRIF annotations by databases that include them in their data records. While on the one hand, readers can follow the link to the data record for a given GeneRIF on Europe PMC (via SciLite), on the other hand, data records citing the GeneRIF will be linked back to the sentence the GeneRIF was extracted from. This mechanism offers a solution making granular deep links between the literature and data for users.
- **Integrating annotations into database curation workflow** - The DisProt database is a community resource for intrinsically disordered protein regions, curated from literature. The DisProt group have now developed an interface to fetch annotations (e.g. PDB accessions, gene/proteins) from the Annotations platform for all articles to be curated. Furthermore, a new field will be added to include text snippets describing disordered proteins regions and will be made available via the platform.

3. Project objectives

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

No.	Objective	Yes	No
2	Increase the sustainability of manually curated resources, which, while of high value and essential to the life- science community, are very labour-intensive to operate	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	3	Start date or starting event:	month 1
Work package title	Data Resources and Services		
Lead	Jo McEntyre (EMBL-EBI, Christine Durinx (SIB)		
Participant number and person months per participant 1 – EMBL 85.00, 9 - CIPF 1.33, 12 - BSC 12.00, 14 - UPF 14.10, 15 - IMIM 5.50, 25 - SIB 66.00			
<p>The core mission of ELIXIR is to build a sustainable infrastructure for biological information across Europe. Data resources and services (hereinafter referred to as “resources”) are a key part of this infrastructure and can vary; from submission databases that contain research data outputs such as DNA sequences (e.g. European Nucleotide Archive), to highly dynamic resources that aggregate, process and visualise research data, often adding layers of value through manual curation by highly qualified personnel. (e.g. UniProtKB/Swiss-Prot). .</p>			
Objectives <p>The overall objective of this Work Package (WP) is to build a framework to inform and drive the sustainable development of Europe’s core life-science data resources. The goals of WP3 are to: • Promote excellence in resource development and operation through providing a unified framework for the identification and monitoring of key bioinformatics resources across Europe. • Increase the sustainability of manually curated resources, which, while of high value and essential to the life- science community, are very labour-intensive to operate. This will be done by integrating the literature with data, with particular emphasis on maximizing value added by curation</p>			
Description of work and role of partners Task 3.1. Promote and implement good practice in data resource and service management through the formalization of metrics and quality criteria enabling the identification of ELIXIR Named and Core Resources, and informing their life-cycle management (24PM) The first requirement for the development of a unified framework for the management of key bioinformatics resources across Europe is to identify which resources (a) meet a variety of quality criteria with respect to scientific impact and level of			

service, and (b) which of these are of fundamental importance to the life-sciences community. Therefore, ELIXIR resources will be identified and classified into two categories: - ELIXIR Named Resource will be attributed to ELIXIR Resources from the project partners (ELIXIR Nodes) that are compliant with a set of metrics/criteria that guarantee their quality. - ELIXIR Core Resources will be the subset of ELIXIR Named Resources that, based on metrics/quality criteria, are of fundamental importance to the life-science community and that are considered to be an authority in their field with respect to one or more characteristics.

Definition of clear metrics/quality criteria that measure current and projected use of ELIXIR resources as well as their scientific impact, and the reliability of the service, will underpin the identification of ELIXIR named and core resources and provide data to inform life-cycle management on an ongoing basis. The initial set of metrics and quality criteria for ELIXIR resources will be identified based on prior resource management experience of WP partners, on the work completed by the ELIXIR technical coordinators group, and experiences from other disciplines such as the Data Seal of Approval project 56. Formal opportunities for ELIXIR-wide review of the proposed criteria will be conducted through presentations and workshops aligned with project management meetings. Metrics and quality criteria will evaluate both the scientific impact of the resources on the life-science community and the reliability of the service. They include, but are not limited to: uptime and download speeds, usage statistics (IPs, page views, downloads), citations in the literature, data submission rates, international collaborations, programmatic access, and curational effort.

In defining measures of quality it is important to recognise the context in which the service is being provided and to base categorization on a range of criteria. For example, a resource that serves a small community may not have as many page views as a large resource, yet reach 90% of the community it supports. Other may play a foundational role to derived services. It will be important to differentiate between submission databases and “added-value” databases that organize, curate, or otherwise represent submitted data, as the profile of use of these types of resource may be very different. Equipped with an agreed set of criteria, it will be possible to effect a number of actions: - Identify new resources for inclusion in the ELIXIR set. - Set quality standards for emerging resources and inform their development. Page 25 of 96 - Build confidence among users through the identification of ELIXIR resources directly (such as a “badge” on the resource itself) and through a variety of portals such as the Tools and Data Services Discovery Portal (WP1). - Monitor usage trends and manage resource life cycles effectively using objective criteria. - Build understanding of the impact of ELIXIR resources both within the ERA and within global research infrastructures.

Resource development based on Metrics and Quality Criteria Alongside the definition of the metrics and quality criteria, coordinated management processes will be required to review candidate resources, encourage use of ELIXIR-approved badges (or similar), and monitor resource life cycles. We expect the organizations running the resources to actively contribute to this process, and that this in itself will provide feedback mechanisms to improve and refine the criteria. This coordinated feedback model will have the added benefit of providing opportunities for peer-peer capacity building (WP10) in the areas of life-cycle management and sustainability, and metrics/quality criteria implementation as we share expertise between ELIXIR Nodes.

Partners: EMBL-EBI, CH

Task 3.2. Inform ELIXIR Resources life-cycle management and improve the ELIXIR Resource portfolio through the implementation of an active and computer-assisted infrastructure for the monitoring of ELIXIR Named and Core Resources based on the metrics and quality criteria formalized in Task 3.1 (76.1PM) In the interest of transparency and to build excellence across resources, metrics and quality criteria for ELIXIR named and core resources will be held centrally at the ELIXIR Hub (see also WP12.3). Access to this collated data will be made available to all Nodes and resources involved, and potentially more widely as aggregated data. In this task, technical processes will be developed to generate and collate the metrics and quality criteria agreed in Task 3.1. Operating in active mode over a period of time, the emerging trends will inform ELIXIR Resources life- cycle management and improve the ELIXIR Resource portfolio overall. The processes developed will gather, report and upload metrics and other quality criteria in agreed formats and to an agreed timescale to the ELIXIR Hub.

The need to collate metrics/quality criteria centrally for analytical and comparative purposes raises questions regarding the technical implementation of such a system. There are a number of challenges in doing this, not least the willingness of the resource providers to share detailed metrics and quality criteria regarding their resource. Subsequent to this will be the need to provide confidence, particularly in the case of metrics, that what is being measured/reported from different resources is comparable in a fair manner; this will require sharing of methodological approaches (such as how robot traffic to websites is treated) through a shared understanding of what is considered a page view across different resources. Finally, agreement on a timetable and format for quality and metrics information will be required so that it can be easily collated in one place. These challenges may give rise to a need for technical effort in the participating resources and such requirements will be supported through the ELIXIR Hub core budget if required.

Partners: EMBL-EBI, EMBL-ELIXIR, CH

Task 3.3. Increase the sustainability of curated resources through literature-data integration and resource crosslinking (86.1PM) The integration of the literature with data is critical for understanding the biological context of new results, for showing clear provenance of scientific assertions, and for discovering new information. While these are important activities for all of the scientific community using online resources, the requirement is most intense within scientific curation processes. The excellent quality of many European bioinformatics resources relies on manual curation, a process in which trained experts review experimental data reported in publications and extract relevant information for inclusion in data resources. This requires searching, reading, filtering, verifying and recording information; labour-intensive, and therefore costly, processes. However, curation saves time and adds significant value for researchers, obviating the need for potential users to individually seek out and synthesise threads of scientific information. Technological advancements in the past few years provide new opportunities to expedite the work of curators and also provide novel approaches to integrating the literature with data for the wider scientific community. For example, when a curator adds a new piece of information to a data record, the source article is cited in the record. However, it would be useful to link from that specific annotation directly to the precise point in the article that was extracted by the curator, for example, a figure legend. This will allow researchers and curators alike to understand exactly where that piece of information came from when viewing the data record, or conversely, to follow a link to

see more data when reading the article - a connection that is currently not possible to traverse. Such developments will provide efficiency savings in resource and tool interfaces, reduce repetition, and when published, will provide granular deep links between the literature and data for users.

In this task, a roadmap for infrastructure that integrates the literature with data through a variety of novel approaches, including text mining, will be developed. Elements of this roadmap will be demonstrated by a collection of pilot developments that provide deep links between the literature and established or emerging ELIXIR data resources. Automated approaches, such as text mining, that identify and extract useful biological concepts will be a necessary part of this activity, from generating granular links to suggesting articles to curate in the longer term. Harnessing the Page 26 of 96 expertise of the text and data mining community as a whole would maximize the impact of this aspect. This task aims to engage with existing database providers and novel Use Cases (WP6 to 9) to develop a roadmap that combines the above elements to develop an infrastructure for literature-data integration and enrichment, and furthermore to demonstrate this through a collection of pilot developments. To do this we will use known high-quality annotations such as GeneRifs (sentences extracted from articles that have been included in gene database records) and the Europe PMC database of life science research articles.

Partners: EBI, CH, ES

Appendix 1: D3.2 A report describing the semantically enriched Europe PMC that links literature to ELIXIR Core resource data, exemplified by GeneRIF annotations

A1.1. Introduction

The [Work Package 3](#) (WP3) of the ELIXIR-Excelerate grant consists of three tasks. While two tasks deal with establishing indicators and processes to identify data resources critical to life science research. The third task is specifically concerned with the scalability of curation. Given that curation is a cornerstone for providing accessible resources that benefit the wider research community, this task explores possible technical solutions that support curation, and the infrastructure needed to support such activities within the context of ELIXIR.

This report provides an overview of the current developments undertaken within Task 3.3 of WP3, which lays the foundation for supporting scalable curation. In the subsequent sections we provide: an update on the Europe PMC Annotations platform; description on how the annotations can be accessed via the Annotations API; description of the improvements made to the SciLite application for annotation displays; description of a submission system for automatic ingestion of annotations and the application of fuzzy matching algorithm to improve GeneRIF annotation mapping on articles. Finally, we present two use cases: a) a

mechanism to provide deep links between annotations and the corresponding data resources and b) an initial implementation to integrate annotations into the curation workflow of DisPort database.

A1.2. Technical and infrastructural components to support scalable curation

This section briefly outlines the infrastructural components developed under WP3. [Figure 1](#) illustrates the workflow by which annotations can be aggregated and accessed from the Annotations platform.

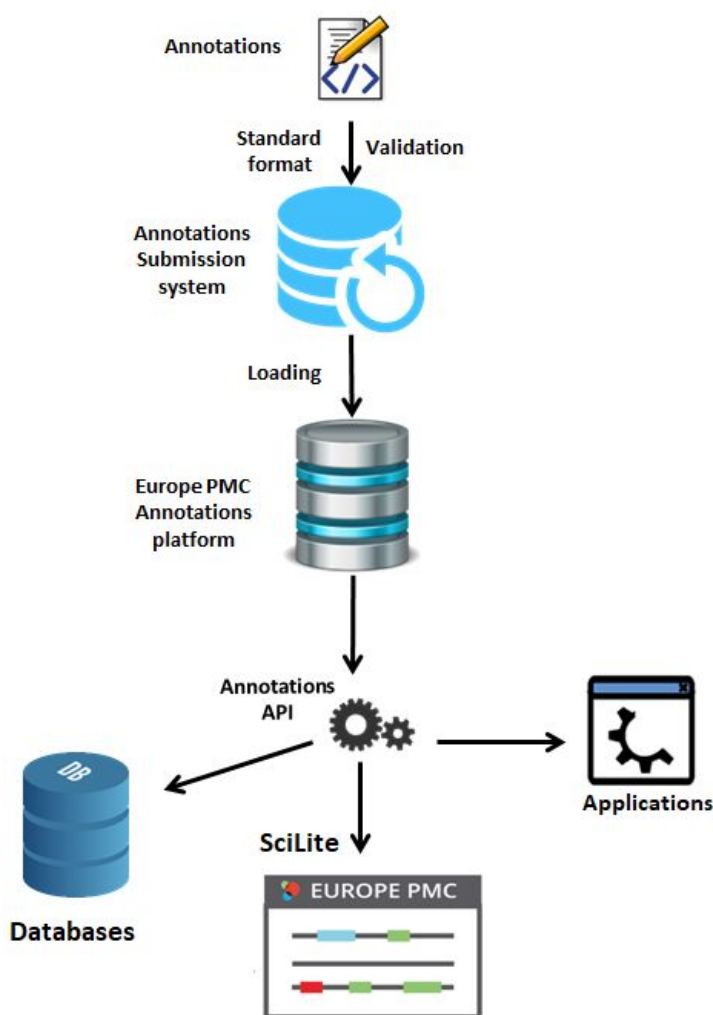


Figure 1: An overview of the workflow for the annotations infrastructure

A1.2.1 Europe PMC Annotations platform

The Europe PMC Annotations platform is a community platform established to capitalise on the advances made by the text-mining community, bringing the results to a broader scientific community. The open architecture of the platform, based on standard data format ([Web Annotation Data Model](#)), allows text-mining outputs from different sources to be integrated.

In this way, the outputs can be reused in ways that can benefit a wide audience of researchers and curators, as well as other interested stakeholders.

The platform is flexible with regard to the frequency of annotation deposition from providers: anything from daily updates to a static dump can be accommodated. Currently, the platform hosts annotations made on abstracts and full-text articles (the open access set and articles with the license type: CC0, CC-BY, or CC-BY-NC), from 8 providers. Building this community of providers, via workshops and collaboration, has been a key part of the work funded by ELIXIR ExceleRATE. The annotation types included thus far are as follows:

- Named entities
 - Gene/protein names
 - Gene variations
 - Organisms
 - Diseases
 - GO terms
 - Chemicals
- Accession numbers for over 20 major data resources, including the ELIXIR Core Data Resources.
- Relationships
 - Gene-disease associations
 - Transcription factor - gene target interactions
- Text phrases
 - Protein-protein interactions
 - Gene function annotation (GeneRIF, or Gene Reference into Function)
 - Biological events (protein phosphorylation)

[Table 1](#), lists the types of annotations and the corresponding providers.

Table 1: List of content providers and annotation types for the Europe PMC Annotations platform

Annotation type Providers	Named entities	Accession numbers	Relationships		Text phrases		
			Gene-disease relationship	Transcription factor-gene target relationship	Protein-protein interaction	Gene function annotation	Biological events
SIB*						✓	
Europe PMC*	✓	✓					
DisGeNET*			✓				
Open Targets Platform			✓				
IntAct					✓		
NaCTEM							✓
PubTator (NCBI)	✓						
NTNU/BSC				✓			

*WP3 (task 3.3) partners

The platform has so far aggregated over 480 million annotations. [Table 2](#) provides the breakup of the annotations counts per provider.

Table 2: The table lists the article and annotation counts per provider

Provider(s)	No. of Articles	No. of Annotations
SIB	226,179	350,770
DisGeNET	20,701	80,048
Europe PMC	19,288,983	478,585,118
IntAct	136	323
Open Targets platform	1,738,842	9,159,123
PubTator (NCBI)	24,006	105,479
NTNU/BSC	788	105,479
NaCTEM	468	8,714

A1.2.2 Annotations API

The Annotations API (www.europepmc.org/AnnotationsApi) was developed recently by Europe PMC to provide programmatic access to the annotations hosted by Europe PMC Annotations platform. The API is RESTful and has modular structure, offering the following functionalities:

- **Get annotations by article ids** fetches the annotations associated with a specified list of articles. The user can further define desired annotation type, provider, or article section (such as methods or introduction).
- **Get annotations by entity** retrieves all mentions of a given entity (e.g. p53, diabetes, or *Staphylococcus aureus*).
- **Get annotations by provider** extracts annotations supplied by specific provider, e.g. SIB or DisGeNET.
- **Get annotations by relationship** retrieves annotations tagging relationships, such as gene-disease or transcription factor-target gene associations.
- **Get annotations by section and/or type** fetches annotations of a given type from a specified article section, for example all chemicals mentioned in “Materials and Methods”.

The API returns results in a number of formats, including JSON, XML, and ID_LIST (article identifiers). Additionally, the annotations are also available in JSON-LD format, which produces a linked data representation of the annotations.

A1.2.3 SciLite

SciLite[1] is an application developed as part of Europe PMC to display annotations hosted by the platform directly in content. For the reader SciLite makes it very easy to skim-read articles. Furthermore, those annotated entities are linked to the corresponding resources, so the reader can get to the underlying data in a straightforward way. This is achieved by making API requests using the Annotations API, to fetch all relevant annotations for a given article.

Displaying annotations (aggregated from different sources) on HTML pages is challenging. This stems from a number of factors related to the nature of HTML page rendering. These factors include: a) presence of subtags in a given sentence, b) mismatch between the text in the annotation and text appearing in the HTML page and c) special characters that are not properly encoded in the annotated text. In such cases an exact match approach of locating annotations on HTML pages fail, resulting in annotations not being displayed. To overcome these issues, we recently applied a fuzzy match algorithm to re-map those missed annotations. Due to the nature of fuzzy match we apply the approach only on sentence based annotations, such as gene-disease associations. Additionally, fuzzy matching algorithms are computationally expensive, therefore the algorithm is applied conditionally when exact match fails to locate the annotations. This approach has proved to be a useful addition to the ScLite application, significantly improving annotation display. A detailed description of the approach can be found [here](#).

A1.2.4 Automated annotation submission system

During the initial implementation phase, annotations from providers were loaded to the Annotations platform by bespoke ETL (Extraction, Transformation and Loading) processes. To make this procedure scalable, in ELIXIR ExceleRATE we developed a submission system that automatically loads all incoming annotations to the Annotations platform.

Each annotation provider is given access to a private Cloud Storage account to upload the data they wish to be included in the Annotations platform. The user account consists of two directories: 'Submissions' and 'Results'. On depositing the annotations in the 'Submissions' directory, the data are automatically validated for its compliance with the standard format (refer [section A1.2.1](#)) and loaded to the Annotations platform. The validation step produces an email informing the provider the result of the specific submission. Additionally, a validation report is saved in the 'Results' directory, allowing the provider to check the results of their submission and make necessary corrections in case of errors. Only submissions with no errors are loaded to the Annotations platform. [Figure 2](#), illustrates the workflow of the submission system.

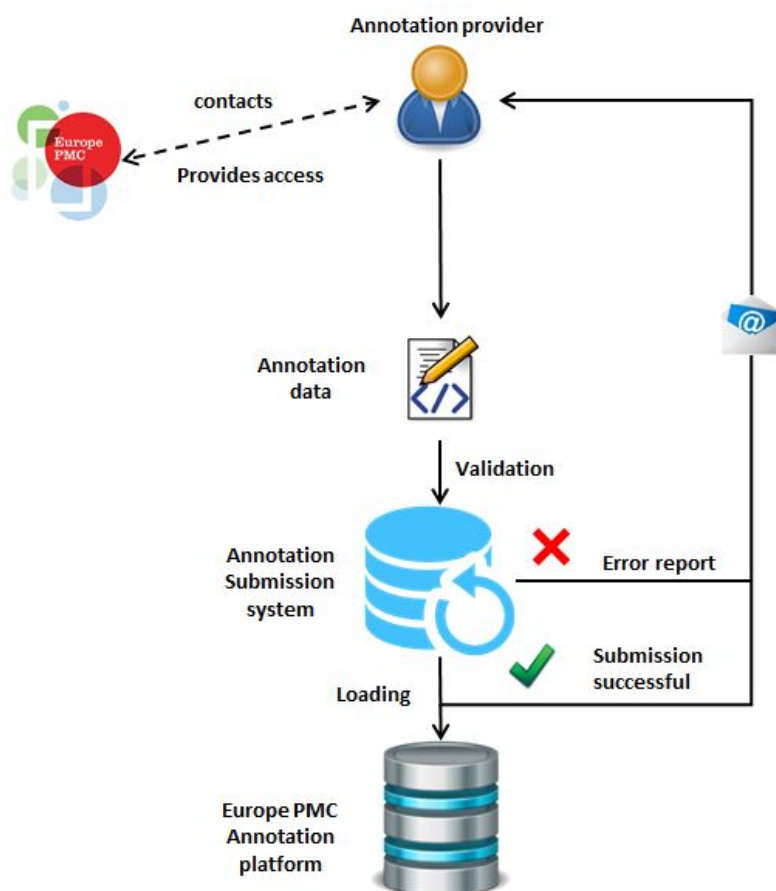


Figure 2: Annotations submission system workflow

Furthermore, to help providers submit compliant annotations, a validation script has been made available via [GitHub](#). This script may be used to pre-validate the data, making the submission process quicker.

The submission system has undergone extensive testing with Europe PMC data and is currently open to the other WP partners to submit data. Upon further testing the system will be made public. Details of the submission system can be found [here](#).

A1.2.5 Fuzzy matching of GeneRIF statements

Statements on gene functions, known as Gene References into Function (GeneRIF) are collected by the National Center for Biotechnology Information (NCBI). They are short statements extracted by domain experts from scientific articles. GeneRIFs are intended to facilitate access to publications documenting experiments on gene and its functions.

GeneRIFs can range from a simple “copy-paste” or a more complex synthesis of text snippets from articles.

The nature of GeneRIFs often render standard text searches ineffective for the readers. Moreover, GeneRIFs are linked to an article and not to the specific sentence in the given article, this presents a challenge as the readers have to scan entire paper to retrieving the evidence. Thus, the task of re-mapping GeneRIF statements to its corresponding sentences in the article was investigated.

While approaches like machine learning are popular for automatic extraction of scientific facts from biomedical literature. This approach relies on the availability of gold standard datasets, to handle modified text snippets and map them accurately on content. Currently there is a lack of gold standard datasets and gathering this knowledge is time consuming and expensive. In contrast, the approximate string matching (or fuzzy matching) proves to be a simpler approach to estimate the similarity between an evidence (a GeneRIF statement, in this case) and a passage in a given article.

The first batch of GeneRIFs submitted to the Annotations platform were based on ‘exact match’ of the statement to corresponding the text in the article. To account for complex GeneRIF statements, a dedicated application was developed based on the Levenshtein algorithm. The algorithm quantifies the similarity between two strings, by computing the number of single-characters operations (e.g. insertions, deletions or substitutions) required to transform one string into the other. The aforementioned algorithm was applied on the entire MEDLINE Abstract and full text collection via Europe PMC. This resulted in a dataset of 337,000 and 13,770 GeneRIFs re-mapped on abstracts and full texts respectively. The annotations are now part of the Europe PMC Annotations platform.

A1.3. Use Cases

A1.3.1. Literature-data linking exemplified by GeneRIF annotations

When curators update a data record with a new piece of information, the source article is often cited in the record. However, it would be useful to link the data record directly to the precise sentence in the article that was extracted by the curator. This will allow the consumer of the data to understand exactly where that piece of information came from. The method of sentence-level linking was hitherto not possible. One of the aims of this task was to provide a mechanism that enables deep links between the literature and established or emerging ELIXIR data resources.

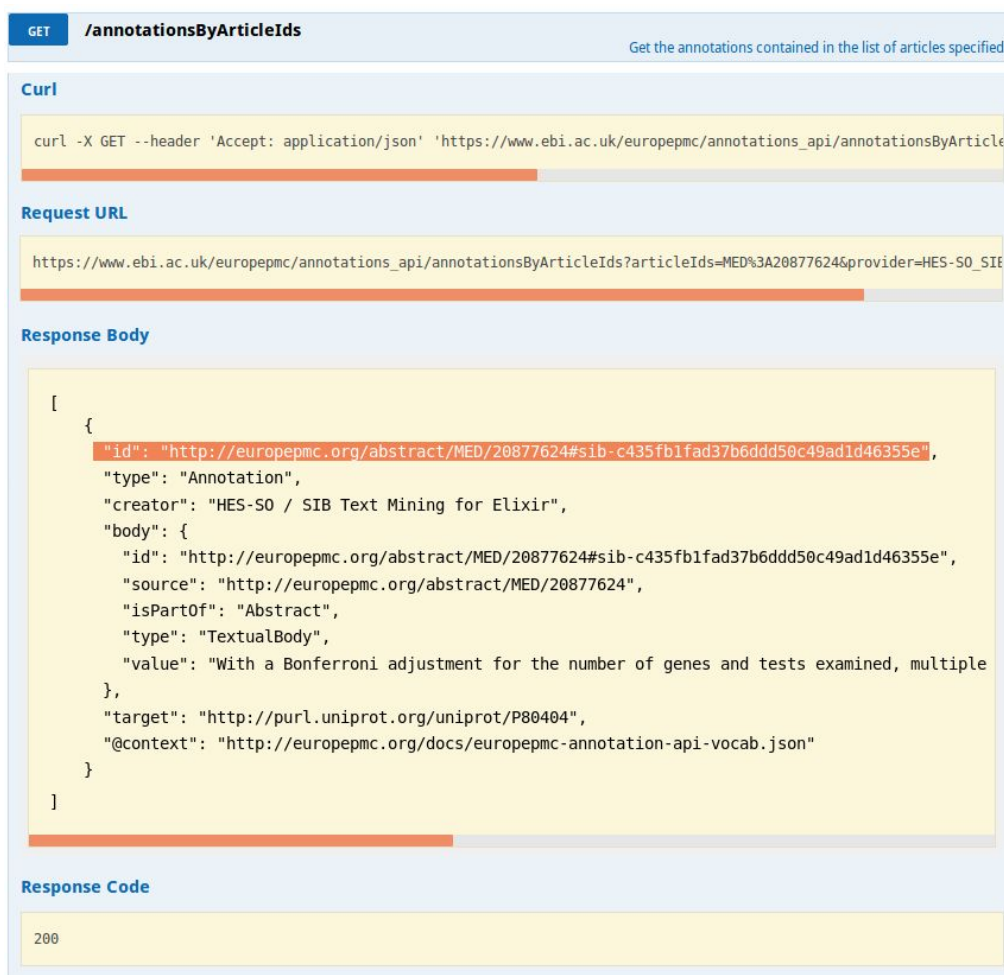
The [Web Annotation Model](#) specification allows annotations to be uniquely identified using URIs, offering a mechanism to cite annotations across multiple platforms. To this end, for all GeneRIF annotations we have minted resolvable annotation URIs. For instance, the GeneRIF annotation from the given article (PMID: [20877624](#)) will have an annotation URI of the form:

<http://europepmc.org/abstract/MED/20877624#sib-6c6b0616e5e3cd78620cfb911b953551>.

This URI can be used to cite the sentence in the article, the GeneRIF was extracted from. The GeneRIF annotation URIs are made available via the Annotations API. For a given

article ID the corresponding GeneRIF annotation can be retrieved in JSON-LD format, that will contain the annotation URI (refer [section A1.2.2](#) and see [Figure 3](#)).

Figure 3: The screenshot of the GeneRIF annotation (in JSON-LD format) retrieved from the Annotations API for the article - PMID: 20877624. The annotation URI for the GeneRIF is highlighted.



With this effort we have demonstrated the possibility of bidirectional linking of annotations. Readers will be able to follow the link to the data record for a given GeneRIF in an article on Europe PMC, or conversely, when viewing the data record, the cited GeneRIF will link to the particular sentence in the article. [Figure 4](#), presents an example of the deep links for a given GeneRIF annotation.

With the deep linking mechanism in place, we plan to engage with databases, such as UniPort and neXtProt to include the GeneRIF annotation links to the corresponding data records. Similarly, the annotation links will be minted for all the data sets hosted by the Annotations platform, enabling consumers of annotations to be able to link back to the entity mentions in the article on Europe PMC.



Abstract


In this study, we analyzed the prevalence and clone size of BRAF V600E mutation in 209 patients with multiple myeloma and related the results to clinical phenotype, response and survival. Biopsies were screened for BRAF V600E by allele-specific real-time PCR (AS-PCR). Positive results were confirmed by immunohistochemistry, Sanger sequencing and, in three patients from whom we had stored purified myeloma cells, whole-exome sequencing. Eleven patients (5.3%) were BRAF V600E mutation positive by AS-PCR and at least one other method. The fraction of mutated cells varied from 4 to 100%. BRAF V600E-positive patients had no characteristic clinical phenotype except for significantly higher levels of serum creatinine (125 versus 86 $\mu\text{mol/l}$). Seven of eleven patients responded with at least very good partial response to alkylators, immunomodulatory agents or proteasome inhibitors. Progression-free and overall survival were similar in patients with and without the mutation. By this integrated approach, we found that patients with BRAF V600E mutation responded very well to broad acting drugs and there was no relation to prognosis in early-stage myeloma. In particular, a large mutated fraction did not correlate with aggressive disease.

Gene Function

By this integrated approach, we found that patients with BRAF V600E mutation responded very well to broad acting drugs and there was no relation to prognosis in early-stage myeloma

Annotation source: HES-SO / SIB Text Mining for Elixir



UniProtKB - P15056 (BRAF_HUMAN)

Display [BLAST](#) [Align](#) [Format](#) [Add to basket](#) [History](#) [Help video](#) [Other tutorials](#)

Entry	Protein	Serine/threonine-protein kinase B-raf
Publications	Gene	BRAF
Feature viewer	Organism	Homo sapiens (Human)
Feature table	Status	Reviewed - Annotation score: ★★★★★ - Experimental evidence at protein level ¹
Function	<h3>Function¹</h3> <p>Protein kinase involved in the transduction of mitogenic signals from the cell membrane to the nucleus. May be the postsynaptic responses of hippocampal neuron. Phosphorylates MAP2K1, and thereby contributes to the signal transduction pathway. ★ 1 Publication ▼</p>	
Names & Taxonomy		
Subcell. location		

Figure 4: Example of literature-data linking using GeneRIF annotations

A1.3.2 Integrating annotations from Europe PMC Annotations platform into DisProt Database

The DisProt database [2] is a community resource for annotating protein sequences for intrinsically disordered (ID) regions from the literature. The resource is maintained at the University of Padua (member of the Elixir Italy node). The DisProt group have integrated annotations (bio-entities) fetched via the Europe PMC Annotations API (refer [section A1.2.2](#)) in the [DisProt annotation interface](#), accessible to authorized curators. DisProt now retrieves information available for a given PMCID (using Annotations API and [Europe PMC Article API](#)), such as UniProt accessions, gene name and cofactors, as well as cross-references to other databases like PDB in real-time.

The new DisProt annotation interface also includes new fields to track article sentences (and the corresponding article section) or curator statements. This will provide a corpus of ID annotations from article that will be made available via Europe PMC Annotations platform. Further, these annotations could also serve as a new training set for text mining tools for information extraction with respect to ID.

A1.4. Conclusion

The work delivered under Elixir-Excelerate WP3 has established the foundation for an infrastructure to support curation tasks. However, the progress made so far requires road-testing, we need to ensure that these annotations and infrastructure are indeed useful and reusable in distributed curation platforms. To this end, we are conducting user research to understand curation practices and gain specific understanding of commonalities in curator workflows. This work will feed into optimising the current infrastructural components, for example, improving text-mined annotation quality and coverage, exploring the reuse of text-mined annotations via Europe PMC in article triage.

A1.5. References

1. Venkatesan A, Kim JH, Talo F *et al*. SciLite: a platform for displaying text-mined annotations as a means to link research articles with biological data [version 2; referees: 2 approved, 1 approved with reservations]. *Wellcome Open Res* 2017, **1**:25 (doi: [10.12688/wellcomeopenres.10210.2](https://doi.org/10.12688/wellcomeopenres.10210.2)).
2. Piovesan D, Tabaro F, Mičetić I *et al* DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res.* 2017 Jan;45(D1) D1123-D1124 doi:10.1093/nar/gkw1279.