



EXCELERATE Deliverable D1.3

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:	Registry release with comprehensive coverage of ELIXIR Node resources, including resource data format curation and analysis	
WP No.	WP1	
Lead Beneficiary:	38 - DTU	
WP Title	Tools Interoperability and Service Registry	
Contractual delivery date:	30 September 2018	
Actual delivery date:	11 October 2018	
WP leader:	Søren Brunak and Alfonso Valencia	38 - DTU; 12 - BSC
Partner(s) contributing to this deliverable:	DTU, IP, UiB , UTARTU, MU, CNRS, BSC	

Authors and Contributors:

Jon Ison (DK), Hans Ienasescu (DK), Piotr Chmura (DK), Veit Schwämmle (DK), Hervé Ménager (FR), Bryan Brancotte, Kenzo-Hugo Hillion (FR), Matúš Kalaš (NO), Ahto Salumets (EE), Erik Jaaniso (EE), Hedi Peterson (EE), Séverine Duvaud (CH), Heinz Stockinger (CH), Egon Willighagen (NL) Jonathan Melius (NL), Magnus Palmblad (NL), Tomáš Raček (CZ), Dan Polanský (CZ), Radka Svobodová Vařeková (CZ), Josep Gelpi (ES), Adam Hospital (ES), Björn Grüning (DE), Peter Løngreen (DK), Søren Brunak (DK)

Reviewers:

N/A

Table of contents

Executive Summary	3
Impact	4
Project objectives	5
Delivery and schedule	5
Adjustments made	5
Background information	6
Appendix 1: Registry release with comprehensive coverage of ELIXIR Node resources, including resource data format curation and analysis	10
1. Technical overview	10
1.1 Milestone M1.7.2 (novel user interfaces)	11
1.2 Milestone M1.1.3 (EDAM + tooling)	12
2. Community build-up process	13
2.1 Scope of the registry	13
2.2 Contributor growth	13
3. Software description model	14
3.1 Model of tool function	14
3.2 EDAM ontology	15
3.2.1 EDAM continuous integration	16
3.2.2 EDAM Formats sub-ontology	16
3.3 Controlled vocabularies	16
4. Tool Information Standard	16
5. Portal content	17
5.1 Content summary	17
5.2 Content annotation	17
5.2.1 EDAM annotations	19
5.3 Content review & clean-up	19
5.4 Import of data service API metadata via OpenAPI	20
6. Portal software	21
6.1 Front-end developments	21
6.2 Back-end developments	24
7. bio.tools integration with other resources	24
8. Utilities	25
8.1 biotoolsSum (collections view)	25
8.2 edamMap	27
8.3 edamBrowser	29

9. Future Work	30
10. Licensing	31

1. Executive Summary

The objective of EXCELERATE Deliverable 1.3 is the development of a discovery portal (bio.tools¹) built upon a federated curation of a registry of key software resources for bioinformatics worldwide. The core aim is to provide a practical portal that will help scientists with resource discovery and interoperability.

Four reports (D1.1 - D1.4) describe the portal:

- M12: **prototyping** of portal software and registry data model, addition of seed content
- M24: **consolidation** of software features and content to a stable model
- M36: **expansion** of registry content towards comprehensive coverage, with new registry features
- M48: **integration** with other software systems, registry applications and portal impact evaluation

This report (D1.3 at M36) describes further *consolidation* and *expansion* of the registry and portal described in D1.2, which presented 5,952 bio.tools entries. We have further improved the model, ontology and software, whilst cleaning and more than doubling the registry content volume, achieving “near enough” comprehensive coverage of prevalent tools as envisioned in D1.2.

This deliverable report describes work done with ELIXIR-EXCELERATE resources. Design and development aspects include:

- production of a major new release of the stable data model (biotoolsSchema 3.0.0) which simplifies the previous stable version 2.0 (described in D1.2) whilst extending the functionality
- the application of the “respectable beta” Tool Information Standard (see D1.5), consolidated from the prototype (see D1.2), to drive content quality improvement in bio.tools
- 3 new EDAM ontology releases
- revision of the client and server-side software including conformance to biotoolsSchema 3.0.0

Operational aspects include:

- registry growth to 11,479 entries (was 5,952 in D1.2) refactored to the model and including major clean-up and improvements guided by the Tool Information Standard
- 214,771 aggregated annotations on tools (was 97,262 in D1.2)
- content contributor growth to 800 (was 537 in D1.2), following the community build-up process (described in D1.7)
- the portal software has been made available under open license (GPL-3.0).

¹ <https://dev.bio.tools> (latest development version) or <https://bio.tools>

In addition, we describe specific developments as envisioned for milestones below (all due in M36), including new user interface features, utilities, content developments and documentation:

- M1.1.3 “EDAM release with coverage of different resource categories and RIs. Implementation of tooling for sustainable community development”
- M1.7.2 “Implementation of novel highly usable interfaces from analysis of user experience and usability requirements”

Including:

- **edamBrowser** web application for browsing resources (in bio.tools, TeSS etc.) based on their scientific (EDAM) annotation and community-development of EDAM. An article describing the work has been accepted for publication in *The Journal of Open Source Software*².
- **biotoolsSum** utility for search and browsing of local tool collections, embeddable on external Web sites
- **edamMap** utility to assist curators in the creation of bio.tools entries; this text mining utility for EDAM has been improved in light of its application to bio.tools content import *en masse*, and is provided as a Web application and Web API

We described in D1.2 how focus was shifting to *content quality* and (somewhat longer term) an increasing priority to foster the community of tool developers and tool end-users. We also summarised an envisioned “endgame”: to provide a persistent reference to high-quality (curated and verified) “canonical” descriptions of *unique* tools, with information about their provision *via* online services and various downloadable artefacts, and including entries for different versions of a tool, where these have major functional differences. In this report we summarise our progress towards this endgame.

Detailed information on new, or improved technical components is available online. All aspects of the project are interdependent and work is ongoing in all areas, to drive content quality improvement, ensure sustainable growth, develop useful features for end-users and to support emerging integration scenarios and applications.

2. Impact

The bio.tools publication³ is in the top 5% of all research outputs ever tracked by AltMetrics, with an AltMetric attention score ranked #56 out of 19,618 outputs from Nucleic Acids Research, and #3 out of 357 outputs from Nucleic Acids Research of similar age (Impact Factor of 9.2 in 2015), with 25 (Scopus), 34 (Dimensions) or 48 (Google Scholar) citations. There were 31,896 users and 99,371 page views (from Google Analytics) of bio.tools pages in the 2nd EXCELERATE Periodic Report period, (was 14,268 and 57,955 in 1st period), with 4,868 users/month (June 2018) (was 937 in Mar 1, 2017). Notably, countries in the first (19% of users) and second (9%) rank are USA and India respectively; the site has international impact. EDAM is the 11th most used of all 718 ontologies listed at NCBO BioPortal (in last report it was 14th of 544), with typically well over 1000 visits/month (from BioPortal). The EDAM publication has an AltMetric attention score ranked #367 out of 7,616 outputs from Bioinformatics, and #1 out of 14 outputs from

² <https://doi.org/10.21105/joss.00698>

³ <https://doi.org/10.1093/nar/gkv1116>

Bioinformatics of similar age. Registry content growth to 11,479 entries and 214,771 aggregated annotations (was 5,952 and 97,262 in D1.2), including 67,851 scientific (EDAM) annotations; a step towards a comprehensive catalogue of consistent, precise “canonical” descriptions of the prevalent bioinformatics tools. Content contributors (people with bio.tools user accounts) growth to 799 contributors (was 470 in D1.2). With sustained developed, bio.tools can have a profound impact, as evidenced by the re-use of the infrastructure by other major projects and infrastructures internationally (*e.g.* BD2K AZTEC.bio, EMBL-ABR, Euro/GlobalBioimaging) and our recent publication⁴ demonstrating directly the practical value of bio.tools (with EDAM) to workflow composition (a hard and important research problems in bioinformatics).

3. Project objectives

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

No.	Objective	Yes	No
1	Deliver a discovery portal built upon a federated curation of a wide range of key resources for bioinformatics resources world-wide.	x	
2	Service monitoring, resource integration, interoperability aspects, and community centred benchmarking efforts.		x
3	Deliver impact for end-users across academia, health organizations, and industry.	x	

4. Delivery and schedule

The delivery is delayed: • Yes ☒ • No

5. Adjustments made

The deliverable is only slightly delayed due to the drafting of the final report.

6. Background information

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

⁴ <https://doi.org/10.1093/bioinformatics/bty646>

Work package number	1	Start date or starting event:	month 1
Work package title	Tools Interoperability and Service Registry		
Lead	Søren Brunak (DK) and Alfonso Valencia (ES)		
Participant number and person months per participant 1 – EMBL 12.00; 2 – UOXF 6.00; 5 – UTARTU 43.00; 10 - IRB 13.00; 12 - BSC 11.00; 17 - INESC-ID 1.24; 21 – UiB 18.00; 25 – SIB 9.50; 26 – CNRS 9.00; 29 – IP 12.00; 35 – MU 18.20; 38 - DTU 26.00 (LTPs: UCPH 25.00 + AU 25.00)			
Objectives WP1 will deliver a discovery portal built upon a federated curation of a wide range of key resources for bioinformatics resources world-wide. It will involve service monitoring, resource integration, interoperability aspects, and community centred benchmarking efforts. All activities, including intensive user support, are focused around delivering impact for end-users across academia, health organizations, and industry. The ELIXIR Tools and Data Services Registry is the cornerstone of the WP. Work Package Leads: Søren Brunak (DK) and Alfonso Valencia (ES)			
Description of work and role of partners			
WP1 - Tools Interoperability and Service Registry [Months: 1-48] DTU, EMBL, UOXF, UTARTU, IRB, BSC, INESC-ID, UiB, SIB, CNRS, IP, MU Based on its first release in January 2015, WP1 will further develop the ELIXIR registry mechanism, interfaces and content upkeep strategy. The WP contains plans for the development and extension of its functionality and scope (Tasks 1.1, 1.2 and 1.5). The federated curation of the registry will ensure comprehensive content and high quality annotations, both of which are essential for the sustainable impact of the registry in the community. Scientific and technical consistency and utility will be achieved by using the EDAM controlled vocabulary. Exposing the results of efforts addressing tool benchmarking and monitoring of the resources listed in the registry will provide the end-user with a robust, scientifically relevant measure of tool quality and performance. Furthermore, the work on workbench integration and interoperability will lower the cost to developers of integrating their resources in key workflow environments, and assist the users with establishing and updating their day-to-day workflows. Finally, WP1 contains plans for comprehensive, registry related user support, which will ensure impact for users, and a dynamic management element, including marketing and community development to build the federated organization behind the registry. The user-centric approach will thus stand as the guiding principle for the entire portal and guard its relevance to the community.			

Task 1.1: Federated Registry Curation (96PM)

This task will deliver essential scientific and technical coverage in the registry and the vocabulary (EDAM) that underpins registry consistency and utility. A major community curation effort is required, including vocabulary development, resource annotation and registration. To ensure that the curation is high quality and sustainable, it must be federated across registry stakeholders, hence a major priority is building and supporting the community of federated curators. In tandem, the curation will be accompanied by focused software and other technical developments, that automate, validate and embed the curation process in relevant software systems; the essential underpinning of sustainability.

The registry has two primary purposes; to help discover tools and services and use them. Discovery means to find, understand, compare and select. It is a prerequisite to (inter)operability, which demands a precise understanding of software dependencies. Our approach is based on the acceptance that software interoperability will, for the foreseeable future, be implemented primarily by developers rather than intelligent software agents. We will therefore, once a comprehensive set of ELIXIR Node resources are described in basic detail, extend the curation of the registry to annotate, using EDAM Format URIs (unified resource identifiers), the data formats that are supported by tools and data services.

From this, we will analyse the format-usage landscape to provide a basis for targeted software developments to improve interoperability of registered resources. We foresee these developments, which might include conversion of tools to use common formats, and development of format- converter software where needed, to be facilitated via the Matchmaking Service mechanism (D1.5).

The registry scope will be:

1. Comprehensive coverage of ELIXIR Node resources, including tools, data services (APIs) and host databases, prioritising ELIXIR-badged services and new resources from the Use Cases.
2. Coverage of other biomedical science Research Infrastructures (RIs), and key resources beyond ELIXIR (European and non-European).

A task force will be comprised of ontology developers, curators, scientific domain experts and relevant technical experts. It will run Curation and Usability hackathons with the recurrent theme of curation: resource annotation and registration, with necessary EDAM development. To facilitate networking and community build-up, two types of social event will be combined with the hackathons:

1. Knowledge Exchange Workshops, including representatives of relevant infrastructures, institutes and projects, on themes related to the registry suggested by the community.
2. Cross-domain Strategy Workshops to gather technical officers from ELIXIR Nodes, RIs, key resources, and other key initiatives, to discuss and develop common approaches for registry curation across RIs internationally.

EDAM provides the registry with a consistent vocabulary for topics (general scientific and technical disciplines), operations (tool functions), types of data, and specific data formats and data identifiers. Task 1.1 will work with the existing EDAM community, develop its open governance and contribution mechanisms and deliver essential utilities to ensure that maintenance, validation and community development is

sustainable in the long term. We will assess and validate coverage by correlating EDAM concepts to terms used for curation, which will then inform and drive necessary additions and desirable clean-ups (removal of concepts). We will develop focused essential utilities for EDAM maintenance including automation of the release process, basic validation of content, reporting of changes between versions, deployment to ontology browsers such as BioPortal and OLS, technical integration of EDAM with applications including the registry and others, mapping of provider-supplied terms and phrases to EDAM, and revise annotation upon new EDAM releases.

To underpin the sustainability of the federated curation, this task will deliver focused software and other technical developments that will automate the registration and update of provider-supplied information, leveraging their own local software infrastructure where possible. We will work with providers to support them in doing this, and, where possible, adapt technically the local solutions to make them more broadly applicable to others. Further, in order to facilitate coverage, all relevant resource providers will be given smooth and convenient access to resource registration. This will be achieved by a combination of simple-to-obtain local login accounts and opening for using eduGAIN authentication to register resources.

Finally, this task will ensure that registered resources are citable, discoverable by the major search engines, and are placed in scientific context. It will also include technical mark-up to support “Semantic Web” applications, e.g. Schema.org-compatible microdata or RDFa to support Google “rich snippets” and other structured search results in the major browsers. Hence, the registry will promote the registered resources and deliver impact for developers and institutes by making resources rank higher in search results and hence more findable.

Task 1.1 partners: DK, NO, FR, CH, CZ, EMBL-EBI, PT

Task 1.2: Benchmarking and Monitoring (15PM)

This task will support the monitoring and community benchmarking of analytical tools, in a systematic and sustainable way e.g. based on the efforts in WP2. Firstly, it will review the existing service quality and performance metrics and assess their usefulness in the context of a registry. This may require development of a light-weight controlled vocabulary capturing the concepts distilled from the preparatory activities above and those of WP2.

Task 1.2 partners: DK, ES, CZ, CH

Task 1.3: Workbench integration and interoperability (36PM)

There is general trend towards the use of workflows as a preferred environment for the convenient use of tools and data access, especially when resources must be used in combination with one another. This task will boost convenience and resource interoperability by implementing a Workbench Integration Enabler service that will develop the vision “register your software once - get it supported everywhere”. Technically, this service will translate the description of any tool or service that is registered in the Tools and Data Services Registry into the metadata format required by the existing major workbenches, including Mobyle, Galaxy and Taverna. Furthermore, we will develop a new, lightweight Service Launchpad for running tools and services

which have programmatic access and which can be invoked using information available in the registry.

To develop the Enabler Service, we will align the registry software description model and the schemas used by the workbench systems or required by the Launchpad, and subsequently revise the model and schemas to facilitate the metadata transfer. Furthermore, to prove the principle, new high priority tools and services, including those developed in the Use Cases.

Task 1.3 partners: DK, EE, FR, CH, PT

Task 1.4: User support and derived registry development (36.7PM)

This task will provide direct and indirect user support to deliver impact for ELIXIR end-users. Direct support will be achieved primarily by leveraging the existing and highly popular user bioinformatics forums (BioStars, BioPlanet etc.).

A User-support specialist will patrol such forums and respond to questions in one of four ways:

- 1) Where resources answering to the Users needs exist in the registry, a link to them in the registry will be provided via our API.
- 2) Where resources exist in the registry, but the registry API cannot be used to answer the question directly, they will request new features of the API and in so doing drive development of the Query Interface.
- 3) Where an appropriate resource exists but has not been registered, they will request the appropriate registry curator add it to the registry.
- 4) Where a registered resource exists that is close, but not quite what is required, they will forward feature requests to the appropriate developers, possibly via the Matchmaking Service (D1.5).

Indirect user support will be achieved primarily by ensuring the registry interfaces are highly usable and match very closely the needs of the user. To achieve this, we will run user experience sessions during the Curation and Usability community. Scientific and technical consistency and utility will be achieved by using the EDAM controlled vocabulary.

Exposing the results of efforts addressing tool benchmarking and monitoring of the resources listed in the registry will provide the end-user with a robust, scientifically relevant measure of tool quality and performance. Furthermore, the hackathons (see Task 1.1) in order to evaluate usability. We will develop comprehensive Good Practice Guidelines for the curation of the registry in all aspects, but in particular the annotation of common types of resources using EDAM.

We will also participate in the development of an ELIXIR Experts Registry where users can discover relevant expertise within the ELIXIR network, and an ELIXIR User Helpdesk to answer general questions concerning use of the registry, forwarding specialised scientific and technical enquiries to relevant experts.

Task 1.4 partners: DK, CH

Task 1.5: Management, marketing and community build-up (46PM)

This task will build the federated organisation primarily by identifying and facilitating key collaborations between registry stakeholders. This will be achieved by organising

'Resource Synergy Meetings', where we will identify and encourage targeted software developments, e.g. to coordinate curation and data sharing. We will also promote resource integration and usability, e.g. by cross-linking resources and through API harmonization. As a prerequisite to these Synergy Meetings, a Resource Metadata Catalogue, listing all relevant resources, their scientific and technical scope, and information fields (schema), will be compiled and used to compare providers and identify redundancies. We will also use these meeting to cross-link the Tools & Data Services Registry with other key ELIXIR registries, for example the Training Materials Registry, the ELIXIR Events Registry, and the Experts Registry.

This task will also develop an oversight and management strategy and leverage partners within and beyond the ELIXIR organisation to implement strategy. To drive delivery, it will identify and encourage collaboration, monitor actions, identify delays, and intervene where necessary. It will raise community awareness and therefore impact by contributing to a forceful marketing campaign via all appropriate marketing channels, including popular social media. It will provide support to funders, publishers and others at the EU and national level, that policy is aligned with the aims of the registry organisation.

Task 1.5 partners: DK

7. Appendix 1: Registry release with comprehensive coverage of ELIXIR Node resources, including resource data format curation and analysis

1. Technical overview

We have further developed the core technical components of the portal:

- **software description model** (Section 3.3). The second stable version (biotoolsSchema 3.0.0) was developed and moved to production. The registration interface, query interface, back-end architecture, and API were all refactored, and the registry content (Section 3.5) migrated for compliance. biotoolsSchema 3.0.0 simplifies the previous version whilst extending the functionality, enabling complete support for it in bio.tools and completing, or superseding the three phases of schema work outlined in D1.2.
- **controlled vocabularies** (Section 3.3.3). The 16 vocabularies described in D1.2 have been extended and improved.
- **EDAM ontology** (Section 3.3.2). Three versions (1.19 - 1.21) were released, with additional Quality Control (QC) checks added to the continuous integration process and tooling.
- **tool information standard** (Section 3.4). The candidate standard for tool information standard (described in D1.2) based upon biotoolsSchema and EDAM

has been extensively revised and released (see D1.5 report), and has been used to drive content quality improvements (Section 3.5).

- **registration interface** (Section 3.6.1). Redesigned, and now includes embedded usage guidelines and crosslinks to relevant sections of the Curation Guidelines (see D1.7). Development of the new tool annotator prototype (described in D1.2) continues, and will provide greatly improved features for adding & editing content once moved into production.
- **query interface** (Section 3.6.1). Cosmetic improvements were made to the Tool Cards and grid view, which are now more convenient, compact and clean.
- **back-end architecture** (Section 3.6.2). Now includes plug-in framework to support multiple serialisation (input/output) formats of biotoolsSchema 3.0.0-compatible data.
- **tool IDs and persistent URLs** (Section 3.6.2). The simplified “cool URI” scheme (described in D1.2) has been implemented, alongside a major refactoring of tool names for this purpose (Section 3.5.3).
- **API** (Section 3.6.2). Refactored for biotoolsSchema 3.0.0.

Content growth and quality improvement (Section 3.5) include:

- registry growth to 11,479 entries (was 5,952 in D1.2) refactored to the model and including major clean-up and improvements guided by the Tool Information Standard
- 214,771 aggregated annotations on tools (was 97,262 in D1.2)
- content contributor growth to 800 (was 537 in D1.2), following the community build-up process (described in D1.7)

Specific developments contributing to the milestones M1.1.3 and M1.7.2 are described below.

1.1 Milestone M1.7.2 (novel user interfaces)

M1.7.2 as envisioned in the EXCELERATE proposal would boost bio.tools usability through appropriate user interface (UI) developments. We have implemented various improvements to the experience of users when annotating, registering, editing and browsing bio.tools entries, including 1) Revision of the bio.tools user interface from mock-ups or prototypes developed for M1.7.1 (some developments are still pending – see below). 2) The biotoolsSum utility. 3) The edamBrowser utility.

edamBrowser (Section 3.8.3) is a web application that provides a novel graphical user interface for browsing resources (in bio.tools, TeSS *etc.*) based on their scientific annotation. It is optimised specifically for visualisation of EDAM and of EDAM usage (annotations). It also allows for community suggestions and additions to the ontology, via GitHub integration (see D1.3).

biotoolsSum (Section 3.8.1) utility provides a “collections view” (a summary of a local tool collection) for use on external Web sites. Following a prototype summarised in M1.7.1, it has been released and is used by the Czech ELIXIR node.

UI prototypes (from M1.7.1) now implemented in bio.tools include redesigned Tool Cards and grid view (see above). The major revision to the look and feel of the bio.tools UI, described in mock-ups or prototypes (from M1.7.1 or M1.1.2) is still pending:

- **bio.tools landing page**
- **topics view** for scientific topic-based navigation based on the EDAM ontology (although this is provided non-natively by edamBrowser)
- **curators interface** for validating bio.tools entries as per the Tool Information Standard
- **tool annotator** for greatly improved tool registration

1.2 Milestone M1.1.3 (EDAM + tooling)

EDAM development has continued apace with bio.tools content growth, remains use-case driven, primarily by bio.tools but increasingly by other projects including Galaxy, Biocontainers, Euro/GlobalBioimaging BISE portal, to name a few. The potential to provide a common vocabulary with bio.tools and provide users with additional information for tool discovery and workflow composition, as stated in D1.2, has been fulfilled by the publication “Automated workflow composition in mass spectrometry based proteomics.” (accepted for publication in Bioinformatics journal⁵ (see D1.8).

- **EDAM ontology** releases with continuous integration (see above)
- **edamMap** utility (term mapping software) has been developed further, including improvement in light of its application to bio.tools content imports *en masse*, its provision as a Web application, and a Web API (Section 3.8.2)
- development of **edamBrowser** (see above)
- **EDAM Formats sub-ontology curation** (Section 3.3.2.2). Progress has been made towards developing EDAM into a comprehensive data formats catalogue (as planned in D1.2)
- **edam2json** tooling developed to pre-parse EDAM into JSON tree for use by bio.tools and other
- comprehensively documented issue templates for concept requests in each sub-ontology

In D1.2 we summarised the development, testing and application of a tool to **import database API metadata** via the openAPI specification into bio.tools: a critical development to expose data services in bio.tools in a sustainable way (Sections 3.5.4). This work has now been written up as a pre-publication “Automatic OpenAPI to Bio.tools Conversion”⁶, but much more is needed to move this into production.

In D1.2 we explored in detail a generic technical **strategy for integration of bio.tools with other resources**, in a way that supports these resources, underpins the sustainable growth of bio.tools and helps to establish bio.tools as the primary archive of basic tool metadata for unique tools. We have progressed implementing this strategy, and have a

⁵ <https://doi.org/10.1093/bioinformatics/bty646>

⁶ <https://doi.org/10.1101/170274>

clearer picture (Section 3.7) how to proceed, including a generic technical mechanism for integration with other resources, and GitHub integration.

2. Community build-up process

The community build-up process supporting high quality content and growth within the highly complex, distributed and dynamic software landscape is described in D1.7. Establishing stable mechanisms for synchronisation of content, where necessary, with major tool providers (stated in D1.2) remains a priority.

2.1 Scope of the registry

The scope remains unchanged since major clarification in D1.2. Workflows, in scope from the outset (see D1.1) are of increasing prominence: we will prioritise their registration whilst maintaining a “light touch” to their modelling.

2.2 Contributor growth

There are 799 (was 537 in D1.2) unique contributors (people with bio.tools user accounts) of content to the registry. The contributor growth (Fig. 1) has been approximately linear since 2016 Q2 with a slight uptick in 2018. The continued organic (*i.e.* unsolicited) growth is encouraging and shows the natural pull of new content to the registry is increasing.

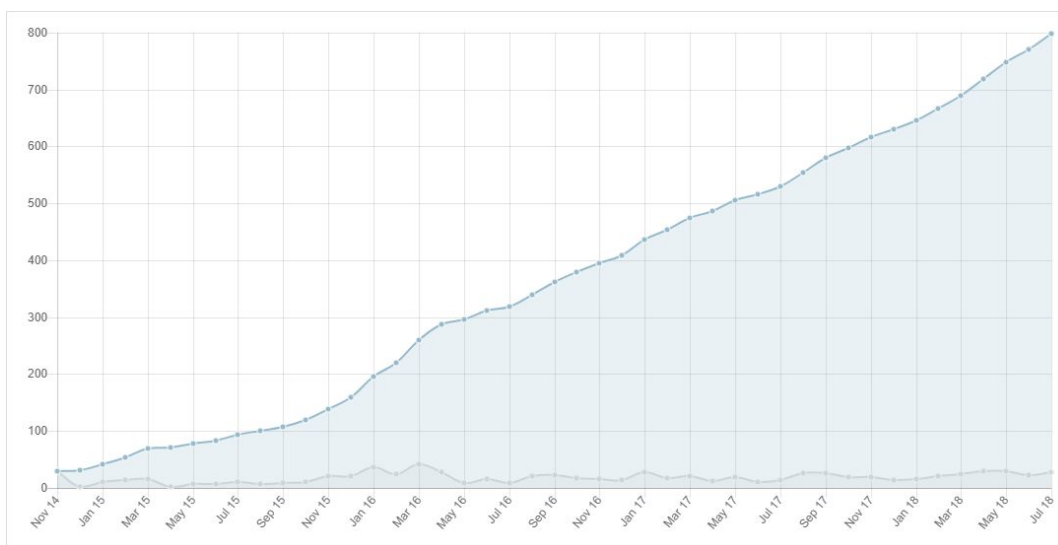


Figure 1. Contributor growth. The blue line is the total number of contributors, grey line is new contributors that month.

The distribution of contributors by top-level domain (Fig. 2) shows that email addresses in “.com” (commercial) is the primary domain, by a margin that has increased significantly since D1.2. France, Germany, United Kingdom, Denmark, and then “.edu” (United States-affiliated institutions of higher education) follow.

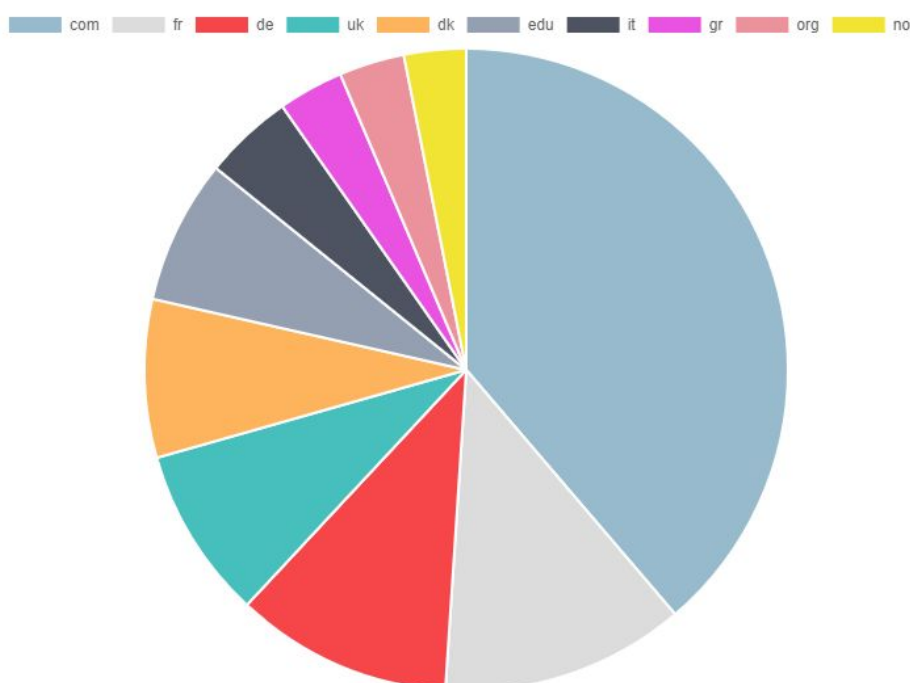


Figure 2. Contributors by top-level domain

3. Software description model

The second stable version (biotoolsSchema 3.0.0) was developed and moved to production. The registration interface, query interface, back-end architecture, and API were all refactored, and the registry content (Section 3.5) migrated for compliance. biotoolsSchema 3.0.0 simplifies the previous stable version 2.0 (described in D1.2) whilst extending the functionality, enabling complete support for it in bio.tools and completing, or superseding the three phases of schema work outlined in D1.2. Improvements made are summarised online⁷. The model will be subject to future improvements, but we do not anticipate any breaking changes. Major releases are restricted to once per year to provide stability for the bio.tools developers and external developers and dependencies.

3.1 Model of tool function

We described in D1.2 the need for better guidelines for annotating a tool's functionality. Specific guidelines⁸ have been developed as part of the Curators Guide (described in D1.7). For example, they clarify what to do in cases where a tool provides an option between doing one operation or another, or always performs certain operations. There is scope for further improvement to the modelling, but these are not yet timely.

3.2 EDAM ontology

EDAM development has continued apace with bio.tools content growth, remains use-case driven, primarily by bio.tools but increasingly by other projects including Galaxy⁹, Biocontainers¹⁰, Euro/GlobalBioimaging BISE portal¹¹, to name a few. The objective, as

⁷ <https://github.com/bio-tools/biotoolsSchema/blob/master/CHANGELOG.md>

⁸ http://biotools.readthedocs.io/en/latest/curators_guide.html#toolfunctions

⁹ <https://usegalaxy.org/>

¹⁰ <http://biocontainers.pro/>

stated in D1.2, is to provide a common vocabulary with bio.tools and provide users with additional information for tool discovery and workflow composition.

All EDAM versions are available for download¹² and browsing from BioPortal¹³ or EMBL-EBI OLS¹⁴. Highlights of changes follow; see the online summary of changes¹⁵ and detailed changelog¹⁶ for exact details of changes. Four new releases were made:

- **EDAM_1.19:** major clean-up (deprecations, synonyms, and rearrangements) in the Operation sub-ontology, clean-up of technical artifacts (definition syntax *etc.*) and provenance data including deprecations, additions for electron microscopy, new data formats
- **EDAM_1.20:** addition of data formats, improvement / additions of Operations and Topic synonyms and labels, incorporation of new CI validations and reporting to changelog
- **EDAM_1.21:** clean-up of Identifier sub-ontology enforcing internal logics for “placeholder” and “concrete” concepts (see below), clean-up of Data branch distinguishing human and machine-readable types, addition of alternative, human-readable IDs to Topic branch (experimental), revisions to support annotation of EBI OLS¹⁷, miscellaneous additions (mostly formats) and other improvements, including requests received via edamBrowser (Section 3.8.3)

In addition, EDAM 1.19 - 1.21 included many additions and improvements for proteomics, specifically mass spectrometry data analysis, with oversight from a Thematic Editor (see D1.7). EDAM 1.22 (in preparation) and subsequent releases will include contributions for other communities including cheminformatics, metabolomics and structural bioinformatics.

Technical developments were made to ease the development and application of EDAM:

- **edam2json** tooling¹⁸ developed to pre-parse EDAM into JSON tree for use by bio.tools and others
- comprehensively documented¹⁹ issue templates²⁰ for concept requests in each sub-ontology
- formalised definition²¹ of concepts that are “placeholders” (primarily for structuring EDAM) and “concrete” (primarily for annotation purposes) with corresponding best practice guidelines²² for EDAM development

¹¹ <http://biii.eu/>

¹² <https://github.com/edamontology/edamontology>

¹³ <http://bioportal.bioontology.org/ontologies/EDAM>

¹⁴ <https://www.ebi.ac.uk/ols/ontologies/edam>

¹⁵ <https://github.com/edamontology/edamontology/blob/master/changelog.md>

¹⁶ <https://github.com/edamontology/edamontology/blob/master/changelog-detailed.md>

¹⁷ <https://www.ebi.ac.uk/ols/index>

¹⁸ <https://github.com/edamontology/edam2json>

¹⁹ https://github.com/edamontology/edamontology/tree/master/.github/ISSUE_TEMPLATE

²⁰ <https://github.com/edamontology/edamontology/issues/new/choose>

²¹ http://edamontologydocs.readthedocs.io/en/latest/technical_details.html#concept-types

²² http://edamontologydocs.readthedocs.io/en/latest/developers_guide.html#hierarchy

3.2.1 EDAM continuous integration

EDAM Continuous Integration (CI) - an automatic control that checks the EDAM ontology at every modification for a number of potential errors - has been further developed:

- **edamxpathvalidator** (<https://github.com/edamontology/edamxpathvalidator>) additional checks of consistency rules and best practices in the EDAM ontology have been added, including for concept deprecation, and cardinality of labels and definitions
- a comprehensive list of future checks was compiled, and is encapsulated (currently) in the EDAM Editors Guide²³ and EDAM Developers Guide²⁴ (summarised in D1.7) - they will be incorporated into the CI system in due course

3.2.2 EDAM Formats sub-ontology

In D1.2 we stated the aspiration for EDAM to provide a comprehensive and practical catalogue of data formats used in the life sciences, to support applications in tool interoperability and workflow composition, and described planning to that end. In this reporting period:

- We have begun to implement the plans described in D1.2, including adding many new formats in the reporting period, and anticipate to write the work up, in due course, in an article “*Towards a comprehensive catalogue of bioinformatics data formats*” (to be submitted to *Bioinformatics* journal).
- The potential value for workflow composition, as stated in D1.2, has been fulfilled by the publication “*Automated workflow composition in mass spectrometry based proteomics.*” (accepted for publication in *Bioinformatics*²⁵ journal (see D1.8).

3.3 Controlled vocabularies

Minor extensions were made to the sixteen controlled vocabularies (described in D1.2) to increase usability and specificity. Comprehensive documentation including definitions of each term in applicable vocabularies is available online²⁶ and in the XSD file itself²⁷.

4. Tool Information Standard

In D1.2 we described a candidate information standard for description of tools to provide a framework for the bio.tools content quality improvement, quality metrics and labels, and integration with other resources. This has now been developed into a “respectable beta” and published online²⁸ (see D1.5 report). Taken together, biotoolsSchema, EDAM and the Tool Information Standard, provide a rigorous and consistent specification of the syntax, semantics and information requirement for tool metadata.

²³ http://edamontologydocs.readthedocs.io/en/latest/editors_guide.html

²⁴ http://edamontologydocs.readthedocs.io/en/latest/developers_guide.html

²⁵ <https://doi.org/10.1093/bioinformatics/bty646>

²⁶ <https://bio.tools/schema>

²⁷ <https://github.com/bio-tools/biotoolsSchema/blob/master/versions/biotools-2.0.0/biotools-2.0.0.xsd>

²⁸ <https://bio-tools.github.io/Tool-Information-Standard/>

5. Portal content

5.1 Content summary

As of July 2018, the registry includes a total of 11,475 entries (was 5952 in D1.2). The impressive content growth (Fig. 3) is the result of a combination of further imports *en masse* providing improved coverage of key collections (from D1.2), centralised curation via bio.tools studentships, and unsolicited contributions, *i.e.* organic growth. The “wobble” in the number of entries during 2017 Q3 and Q4 is an artefact entry de-duplication (described in D1.2). We stated in D1.2 our confidence that the registry contains a representative subset of all tools. It is now closer to comprehensive coverage of all the major prevalent tools, which motivate our curation priorities, primarily to ensure new tools are captured and improve content quality.

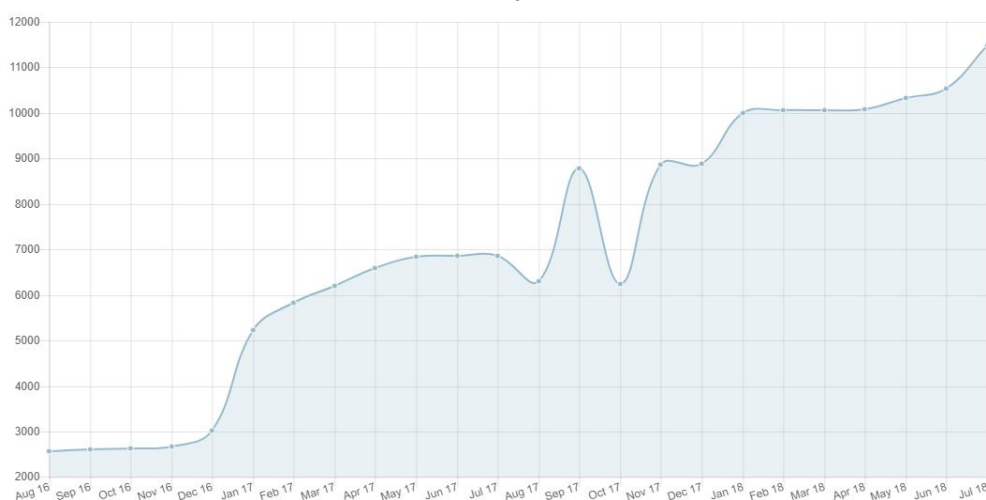


Figure 3. Content growth. Number of bio.tools entries vs. time.

5.2 Content annotation

Content is annotated as per attributes defined within biotoolsSchema (Section 3.3) including in terms from the EDAM ontology (Section 3.3.2) and other controlled vocabularies (Section 3.3.3). The registry now includes a total of 214,771 (was 97,262 in D1.2) individual annotations.

The graph of number of annotations over time (Fig. 4) follows a similar pattern (and explanation) to content growth (Fig. 3).

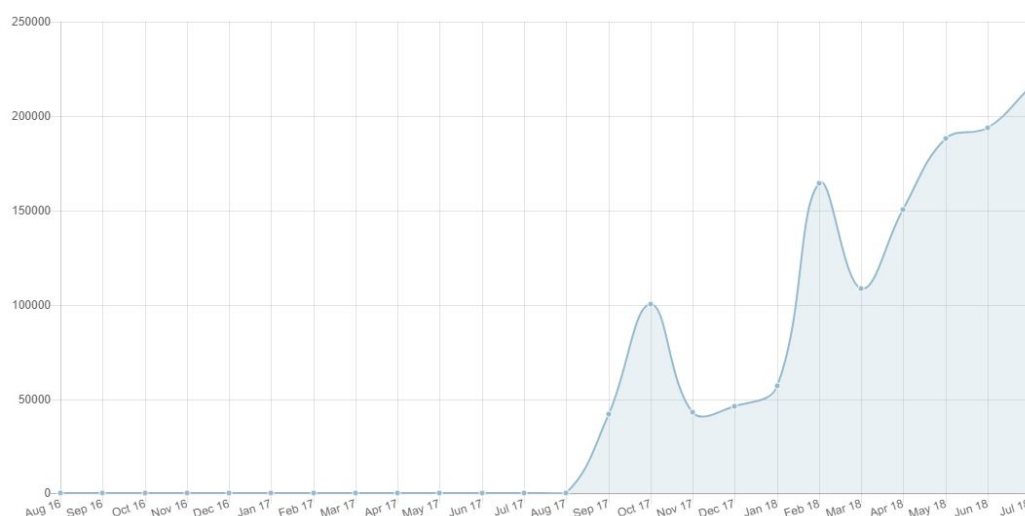


Figure 4. Annotation growth. Number of annotations on bio.tools entries versus time.

A breakdown of the annotations by type (Table 1) reveals that the priorities have been on practical usage information including documentation, contact information, operating system and code availability.

Annotation type	Number of entries	Annotation type	Number of entries
Name	11475	Name	5952
Description	11475	Description	5952
Homepage	11475	Homepage	5952
Tool Type	13378	Tool Type	7297
Unique ID	11475	Unique ID	5952
Topic	31646	Topic	14397
Publication	13209	Publication	7477
Contact	12915	Contact	4246
Operation	21907	Operation	14466
Documentation	11452	Documentation	2933
Operating System	27372	Operating System	2056
Input Output	7337	Input Output	6724
Code availability	13949	Code availability	5014
Accessibility	546	Accessibility	393
Data format	6823	Data format	5112
Community	5640	Community	1436
Downloads	2697	Downloads	1903
Total	214771	Total	97262

Table 1. Annotations by type. The number of aggregated metrics of types defined in the Tool Information Standard is shown: now (left) and from D1.2 (right). Note that for purposes of comparison, the data are formatted to the (older) version of the standard available when D1.2 was published.

Note: Annotations are calculated according to the Tool Information Standard (Section 3.4), where attributes are grouped for purposes of determining adherence to the standard. For example, a single annotation for "Documentation" group implies that one of "General documentation", "API documentation" or "API specification" attributes were specified.

5.2.1 EDAM annotations

The registry includes a total of 67,713 (was 40,699 in D1.2) individual EDAM annotations. The “wobbles” in the number of annotations (Fig. 5, Fig. 6) in 2017 Q3 and Q4 are a reporting artefact during a period of intensive clean-up (Section 3.5.3) following *en-masse* imports.

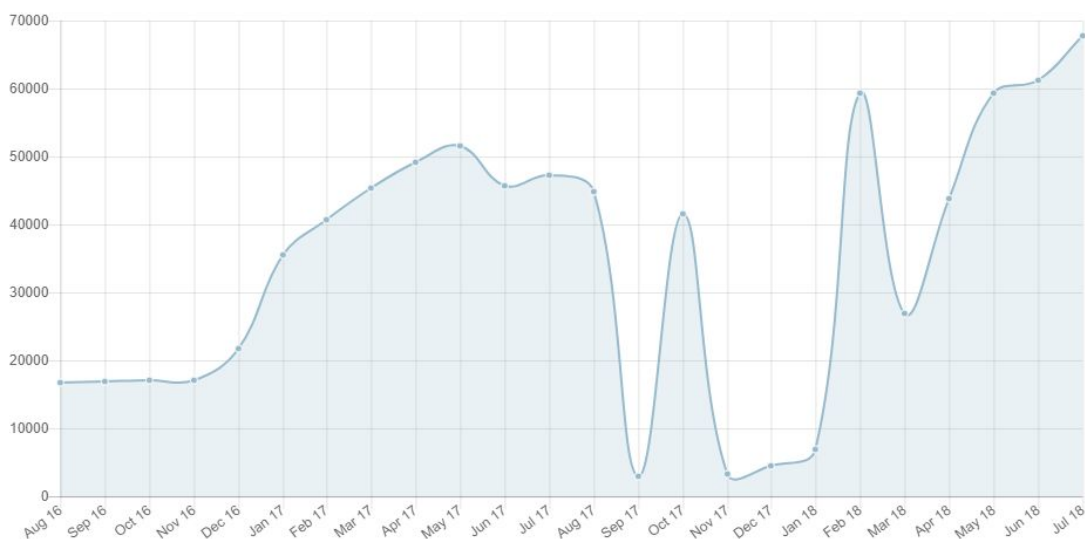


Figure 5. EDAM annotation growth. Total number of EDAM Topic, Operation, Data or Format annotations versus time.

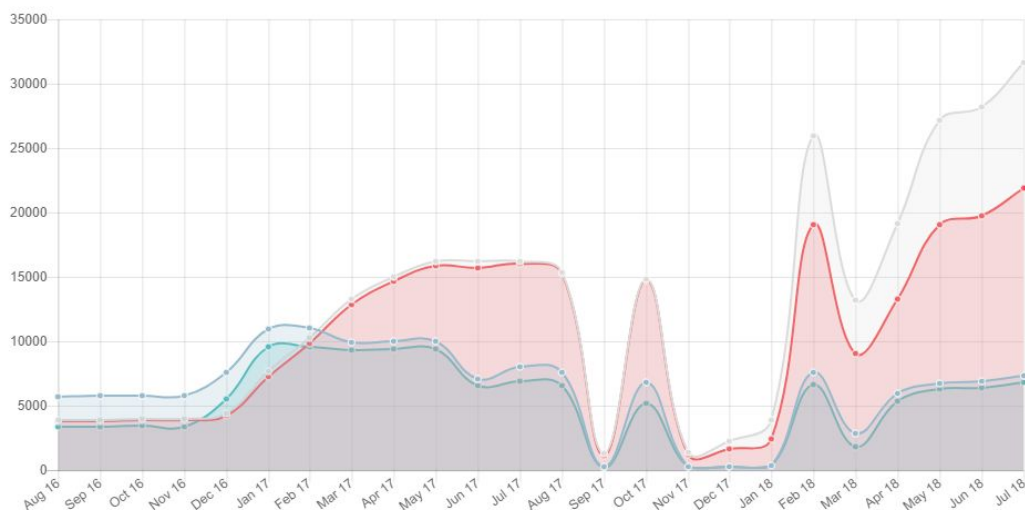


Figure 6. EDAM annotation growth (breakdown). Total number of EDAM Topic (grey line), Operation (red), Data (blue) and Format (cyan) annotations versus time.

5.3 Content review & clean-up

D1.1 and D1.2 stated how the priority was rapid content expansion, but this occurred at the expense of the inclusion of some low or variable quality entries, including duplicate entries with redundant or inconsistent information. D1.2 described a review of the content, resulting in QC checks enumerated in the Curation Guidelines (see D1.7), and how we

embarked upon a major refactoring and clean-up of the content, with view to accomplishing the desired “end-game” (Section 1):

- **consolidation of duplicate entries**
 - the de-duplication process (see D1.2) was completed
 - guidelines²⁹ for users to check the registry and take appropriate action, before registering a tool, were added to the Curators Guide - we aim to embed these into the bio.tools user interface in due course
 - consolidation of duplicates upon detection, as part of routine housekeeping
 - duplicate detection and reporting mechanism to prevent duplicates in the first place has been partially implemented (tools with the same name are not allowed) but needs further work (it cannot be 100% reliable)
- **refactoring for clean tools names and identifiers** to provide “cool URIs” (Section 3.6.2)
 - guidelines for tool names³⁰ and transformation to toolIDs were improved, as part of a broader set of Curation Guidelines (see D1.7)
 - systematic refactoring of existing tool names (and thus IDs) conforming to standards above was completed³¹
- **refactoring entries to describe unique tools**³² was mostly complete, but with some work remaining for tools currently described by an entry for some online service (primarily Galaxy services) only (some entries reflect the Galaxy instance, not the underlying unique tool, and this remains to be fixed)
- **assignment of entry ownership**³³ has been occurring organically via the entry ownership/edit request functionality, but remains to be done more systematically (more curator capacity is needed for this)
- **Improving EDAM Topic & Operation annotations**³⁴ as originally stated (removing “top-level” / vague EDAM terms) is complete, however, much more can be done to improve the scientific annotations, primarily by engaging Thematic Editors.

5.4 Import of data service API metadata via OpenAPI

We described in D1.2, as a work in progress, development of the the OpenAPI-Importer utility³⁵, and an example application to the registration in bio.tools of biological database APIs. This has been written up as a pre-publication “*Automatic OpenAPI to Bio.tools Conversion*”³⁶. The approach has potential for providing in a sustainable way rich descriptions of data services in bio.tools, but more work is still needed (see D1.2, Appendix J.2) to bring the developments into production. Funding was sought for this from the ELIXIR Tools Platform for this (but not forthcoming).

²⁹ http://biotools.readthedocs.io/en/latest/curators_guide.html#before-you-start

³⁰ http://biotools.readthedocs.io/en/latest/curators_guide.html#name-tool

³¹ <http://tinyurl.com/cleantoolids>

³² <https://biotools.sifterapp.com/issues/40>

³³ <https://biotools.sifterapp.com/issues/171>

³⁴ <https://biotools.sifterapp.com/issues/156>

³⁵ <https://github.com/bio-tools/OpenAPI-Importer/>

³⁶ <https://doi.org/10.1101/170274>

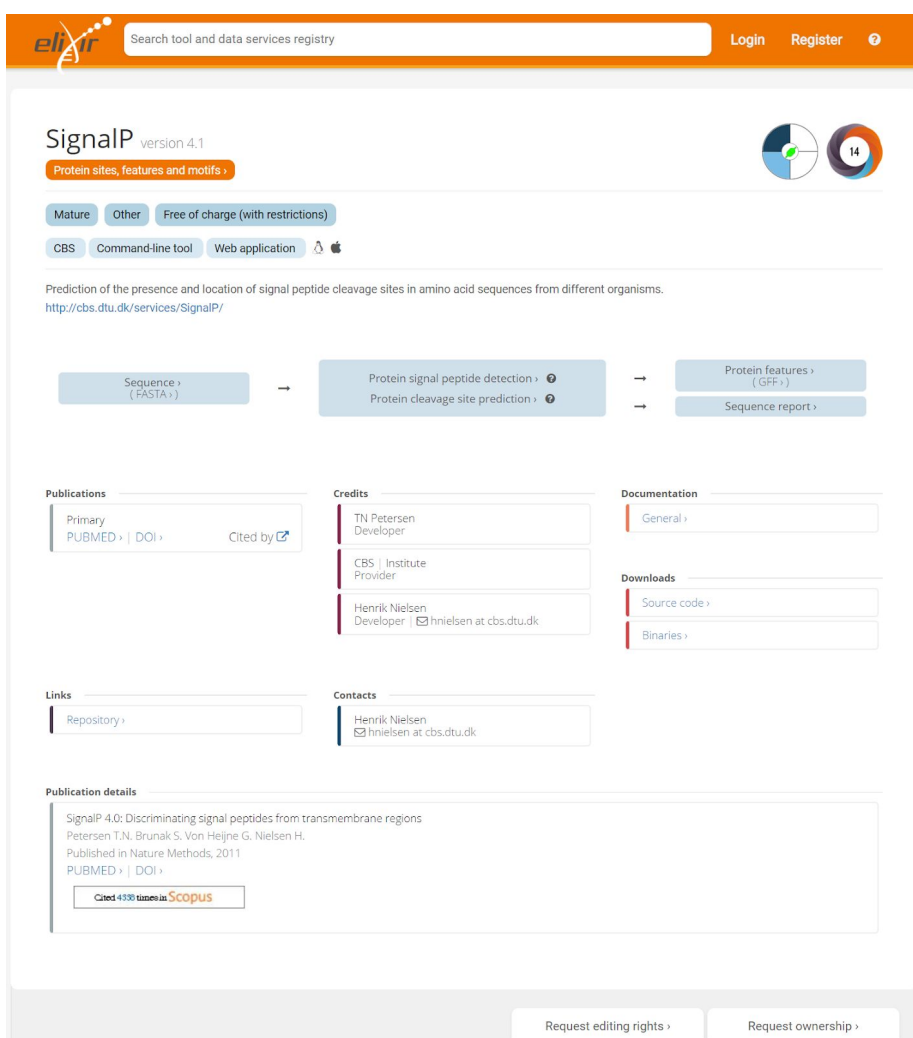
6. Portal software

The registration interface, query interface, back-end architecture, and API were all refactored for compliance to the second stable version (biotoolsSchema 3.0.0) of the data model (Section 3.3). Various additional improvements were made (see below).

6.1 Front-end developments

These include:

- Developments (described in M1.3) for deeper **literature integration** were implemented:
 - rendering basic information for tool publications, in copy-pastable form
 - inclusion of citation counts
 - sorting bio.tools entries by citation count and AltMetric attention score
- UI prototypes (from M1.7.1) now implemented in bio.tools :
 - redesigned **Tool Cards** (Fig. 7) including the search bar at the top of the page
 - redesigned **grid view** (Fig. 8) for comparing tools, to provide users with more control over the displayed content. All the possible displayable attributes are now accessible via a menu, along with sorting options, allowing a user to easily customize what information is being displayed.
- redesigned **registration interface** (Fig. 9) now includes embedded usage guidelines and crosslinks to the bio.tools Curators Guide.



SignalP version 4.1

Protein sites, features and motifs

Mature Other Free of charge (with restrictions)

CBS Command-line tool Web application

Prediction of the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms.
<http://cbs.dtu.dk/services/SignalP/>

Sequence (PASTA) → Protein signal peptide detection Protein cleavage site prediction → Protein features (GFF) Sequence report

Publications

Primary PUBLISHED | DOI Cited by

Credits

TN Petersen Developer

CBS | Institute Provider

Henrik Nielsen Developer | hnielsen@cbs.dtu.dk

Documentation

General

Downloads

Source code Binaries

Links

Repository

Contacts

Henrik Nielsen | hnielsen@cbs.dtu.dk

Publication details

SignalP 4.0: Discriminating signal peptides from transmembrane regions
 Petersen T.N. Brunak S. Von Heijne G. Nielsen H.
 Published in Nature Methods, 2011
 PUBMED | DOI

Cited 433 times in Scopus

Request editing rights Request ownership

Figure 7. bio.tools Tool Card. The search bar is now available in the Tool Cards. AltMetric widget (from D1.2) and OpenEBench widget (from D1.5) are shown in the top right.

elixir Search tool and data services registry 11495 tools Menu json ?					
Sort by Updated Added Name Citation Count Publication Date Display as Compact Detailed					
Name	Description	Function	Topic	Documentation	
kpLogo	Probability-based logo tool for integrated detection and visualization of position-specific ultra-short motifs from a set of aligned sequences.	<ul style="list-style-type: none"> Sequence motif recognition Enrichment analysis k-mer counting Visualisation 	<ul style="list-style-type: none"> Nucleic acid sites, features and motifs Protein sites, features and motifs Functional, regulatory and non-coding RNA 	<ul style="list-style-type: none"> Manual Training material 	
COFACTOR	Structure-based multiple-level protein function predictions. By structurally threading low-resolution structural models through the BiOLIP library, it infers three categories of protein functions including gene ontology, enzyme commission and Read More	<ul style="list-style-type: none"> Protein modelling Protein binding site prediction Protein threading Protein function prediction 	<ul style="list-style-type: none"> Structure prediction Protein binding sites Software engineering Sequence sites, features and motifs Function analysis 	<ul style="list-style-type: none"> General 	
SpartaABC	Simulate sequences based on indel parameters inferred using an approximate Bayesian computation algorithm.	<ul style="list-style-type: none"> Indel detection Phylogenetic tree generation (maximum likelihood and Bayesian methods) Phylogenetic tree reconstruction 	<ul style="list-style-type: none"> Phylogeny Gene transcripts 	<ul style="list-style-type: none"> General General 	
ARTS	Antibiotic Resistant Target Seeker. Specific and efficient genome mining for antibiotics with interesting and novel targets. Automate the screening of large amounts of sequence data and to focus on the most promising strains that produce antibiotics with new Read More	<ul style="list-style-type: none"> Query and retrieval Antimicrobial resistance prediction 	<ul style="list-style-type: none"> Phylogeny Plant biology 	<ul style="list-style-type: none"> General General 	
I-TASSER-MR	Automated molecular replacement for distant-homology proteins using iterative fragment assembly and progressive sequence truncation	<ul style="list-style-type: none"> Molecular replacement Phasing Protein modelling 	<ul style="list-style-type: none"> Sequence assembly X-ray diffraction 	<ul style="list-style-type: none"> Training material General 	

Figure 8. Grid view with new look and features.

Add new tool

Validate
Save

Summary *
Function
Labels
Links
Download
Documentation
Publications
Credits
JSON

Permissions

Basic information about the software.

You need to specify at least the name, homepage and a short description of the tool. See the [Curation Guidelines](#).

Name *
Name

Description *
Description

Homepage URL *
Homepage URL

Software version(s)
Add version

Noticed any problems or have some suggestions for improvement? Please add your feedback [here](#) or email us

Figure 9. Registration interface with embedded usage guidelines and crosslinks to Curation Guide.

The major revision to the look and feel of the bio.tools UI, described in mock-ups or prototypes (from M1.7.1 or M1.1.2) is still pending:

- **bio.tools landing page** (D1.2, Appendix G.1)
- **topics view** (D1.2, Appendix G.4) for scientific topic-based navigation based on the EDAM ontology, and incorporating a cleaner summary view (“mini-cards”)
- **curators interface** (D1.2, Appendix G.7) for validating bio.tools entries as per the Tool Information Standard
- **tool annotator** (D1.2, Appendix G.6) for greatly improved tool registration

6.2 Back-end developments

These include:

- **back-end architecture** (Section 3.6.2) now includes plug-in framework to support multiple serialisation (input/output) formats of biotoolSchema 3.0.0-compatible data.
- **search architecture**. Further improvements to the parameterisation of the Elastic search engine were made, following the critique of the search (from D1.2), achieving satisfactory search performance
- **tool IDs and persistent URLs** (Section 3.6.2). The simplified “cool URI” scheme (persistently resolvable tool Card URIs that are intuitive and human-readable) - described in D1.2 - has been implemented, alongside a major refactoring of tool names for this purpose (Section 3.5.3).

In D1.2 we summarised the partial implementation of a Quality Control / Quality Assurance (QC/QA) mechanism, performing and reporting on various QC checks. Further development of this mechanism is planned, to allow reporting of errors on a per tool basis, by collection, or for all tools via a public API serving error information in JSON format. It will be based on the Curation Guidelines and will (in due course) power a curator interface for validating and improving bio.tools entries as per the Tool Information Standard.

Future improvements will also include:

- improved SEO including marking-up Tool Cards with schema.org JSON-LD (compatible with BioSchemas³⁷)
- further rounds of critique of the search bar performance and optimisation of the search architecture

7. bio.tools integration with other resources

In D1.2 we listed numerous resources that (are used to) independently maintain software metadata, which complement and should be more fully integrated with bio.tools. We explored in detail a strategy that supports these resources, underpins the sustainable growth of bio.tools and helps to establish bio.tools as the primary archive of basic tool metadata for unique tools. Mindful of feedback from the ELIXIR EXCELERATE midterm review to the effect that leveraging the tool developer community is vital to improve the sustainability of the effort, and of the limited resources available to EXCELERATE, the strategy is being rolled-out initially for BioContainers³⁸ and Galaxy³⁹:

³⁷ <http://bioschemas.org/>

³⁸ <https://github.com/BioContainers>

³⁹ <https://galaxyproject.org/public-galaxy-servers/#deepTools>

- mapping entries in the resources and bio.tools, which in turn implies a mapping of package, container, and service names to tool names stored in bio.tools
- efficient mechanism to create new bio.tools entries as required to provide - and maintain - coverage of resources, conformant to at least a minimum acceptable quality as per the Tool Information Standard (Section 3.4)
- an efficient mechanism to allow resource managers to update bio.tools, specifying certain attributes across, in principle, all entries whilst respecting the notion of entry ownership and edit rights. Such attributes include links to an online service for a tool, links out e.g. to BioContainers, download links, installation commands etc.

The other technical requirements from D1.2 are now mostly complete:

- unique identifiers and thus Tool Card URLs (Section 3.6.2) that provide a persistent reference to unique tools and resolve to a “canonical” tool description
- systematic refactoring of existing bio.tools content (Section 3.5.3) to ensure
 - duplicates entries are consolidated
 - supplied tool names and thus toolIDs are sensible
 - entries for canonical tools are included for tools underlying online services registered thus far⁴⁰
- common or at least compatible metadata exchange formats with shims (transformer utilities) as needed

Technical developments to fulfill the requirements are ongoing on all fronts^{41,42} and will be advanced in due also at the forthcoming Paris BioHackathon⁴³ (Oct 2018, France).

8. Utilities

8.1 biotoolsSum (collections view)

bio.toolsSum, described as a prototype in D1.2 has been further developed and documented. biotoolSum is a client-side web application for rendering views and reports of tool descriptions from bio.tools. It provides a quick overview of tools (Fig. 10) within a particular collection divided into meaningful categories by topics covered. Clicking on an icon in the overview brings up a grid view (Fig. 11) for comparing tools. Clicking on an arrow in the “More” column in the grid view brings up details (Fig. 12) for that tool, including a plot of citations over time of the tool publication, and fine-grained EDAM and biotoolsSchema annotations.

A working example⁴⁴ customised for tools and services from the ELIXIR Czech node⁴⁵ renders several domains (DNA, RNA, protein and drugs) divided by the type of data (1D,

⁴⁰ <https://biotools.sifterapp.com/issues/40>

⁴¹ <https://biotools.sifterapp.com/issues/100>

⁴² <https://biotools.sifterapp.com/issues/206>

⁴³ <https://www.elixir-europe.org/events/biohackathon-2018-paris>

⁴⁴ <https://biotools-sum.firebaseio.com/>

⁴⁵ <https://www.elixir-czech.cz/services>

2D, 3D and xD). Development of this utility can be tracked on GitHub⁴⁶ and is available for customisation and use by all.



Figure 10. bio.tools Sum overview. An overview of ELIXIR-CZ tools is shown.

⁴⁶ <https://github.com/bio-tools/biotoolssum>

elixir

Services Matrix

All Services

DNA

RNA

Protein

Drugs and other small molecules







Report mode

All elixir-cz Services for studies on protein sequences.
There is a total number of 6 tools available.

Previous

Page 1 of 1

Next

Name (Sortable, filterable)	Institute	Description (Filterable)	Publications info (Sortable)	More
ProteinCutter v.1  <div>Web application</div>	<ul style="list-style-type: none"> Palacký University Olomouc, Czech Republic ELIXIR-CZ 	Protein mass spectroscopy cleavage prediction.	Publications: no Total Citations: -	▼
Biospean v.1  <div>Web application</div>	<ul style="list-style-type: none"> Palacký University Olomouc, Czech Republic ELIXIR-CZ 	Comparison of mass spectrometry spectra.	Publications: [1] Total Citations: -	▼
PredictSNP v.1.0  <div>Command-line tool</div> <div>Web application</div>	<ul style="list-style-type: none"> Brno University of Technology, Brno, Czech Republic Masaryk University, Brno, Czech Republic International Centre for Clinical Research, Brno, Czech Republic Mayo Clinic, Rochester, New York, United States of America Loschmidt Laboratories 	A consensus classifier that combines six of the top performing tools for the prediction of the effects of mutation on protein function. The obtained results are provided together with annotations extracted from the Protein Mutant Database and the UniProt database.	Publications: [1]? Total Citations: 65	▼
HERVd  <div>Web application</div> <div>Database portal</div>	<ul style="list-style-type: none"> Institute of Molecular Genetics, AS CR ELIXIR-CZ 	Human endogenous retroviruses database.	Publications: [1]? Total Citations: 27	▼
Repeat Explorer v.0.9.7.8  <div>Command-line tool</div> <div>Web application</div>	<ul style="list-style-type: none"> Biology Centre, CAS, Czech Republic ELIXIR-CZ 	Includes utilities for Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data and tools for the detection of transposable element protein coding domains.	Publications: [1]? Total Citations: 106	▼
HotSpot Wizard v.2.0  <div>Web application</div>	<ul style="list-style-type: none"> Brno University of Technology, Brno, Czech Republic Masaryk University, Brno, Czech Republic International Centre for Clinical Research, Brno, Czech Republic Loschmidt Laboratories 	An automated design of mutations and smart libraries for engineering of protein function and stability and annotation of protein structures.	Publications: [1]? Total Citations: 11	▼

Previous

Page 1 of 1

Next

Figure 11. bio.tools Sum grid view. The grid view allows for easy comparison of tools.

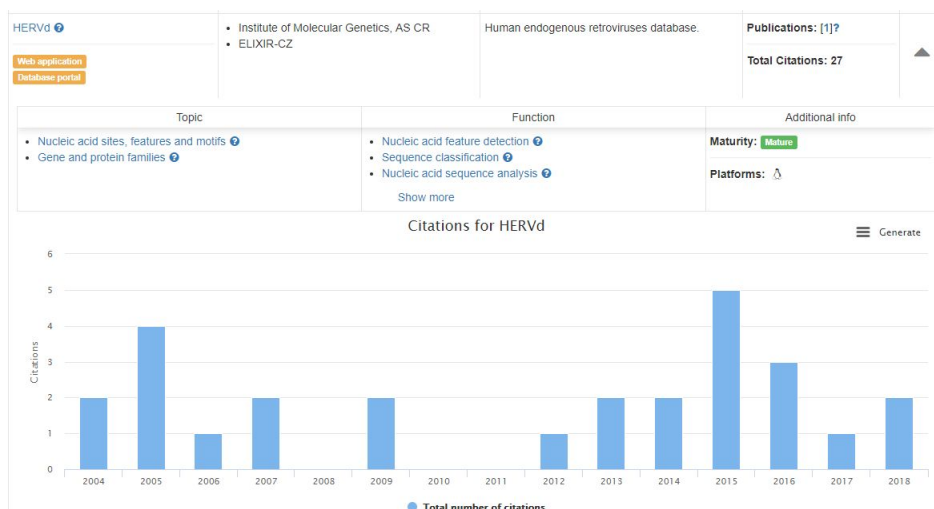


Figure 12. bio.tools Sum details view. More information including citation history is shown when a down-arrow is clicked.

8.2 edamMap

In D1.2, we introduced the **edamMap utility**: a Java based command-line software to assist curators in the creation of bio.tools entries. This utility has been significantly developed, including new inputs sources, improvement in light of its application to bio.tools content imports *en masse*, and its provision as a Web application (Figs. 13-14)

and Web API. edamMap takes as input 1) the EDAM ontology, and 2) a variety of sources of tool description, including tool name, description, keywords, links to relevant webpages and, and Pubmed identifiers of relevant publications. Based on the supplied descriptors, edamMap automatically (using text mining algorithms) fetches relevant information and propose annotation terms for the curator to use when annotating entries in bio.tools. As a future work, the text mining algorithms will be further optimised.

The edamMap software⁴⁷ and web application⁴⁸ are freely available under open license. Extensive documentation including for the API⁴⁹ are available.

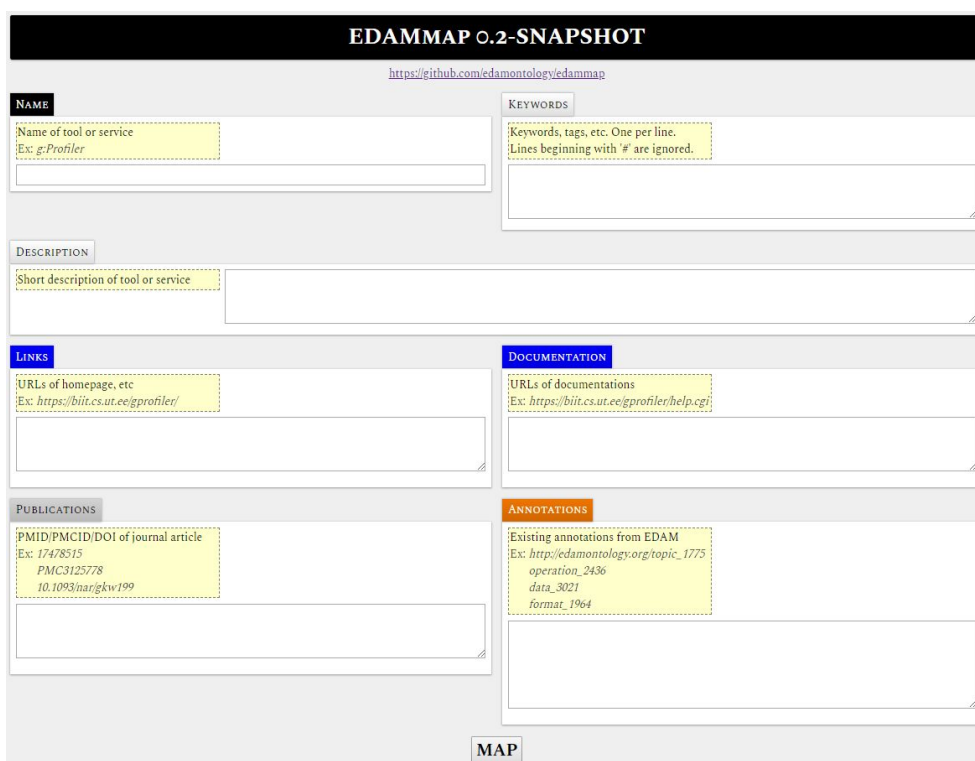


Figure 13. edamMap web application. The input fields are shown where inputs for text mining are specified.

⁴⁷ <https://github.com/edamontology/edammap>

⁴⁸ <https://biit.cs.ut.ee/edammap/>

⁴⁹ <https://github.com/edamontology/edammap/wiki/API>

PARAMETERS

MAIN	
PROCESSING	
PREPROCESSING	
FETCHING	
MAPPING	<div style="display: flex; align-items: flex-start;"> <div style="margin-right: 10px;">Branches:</div> <div style="border: 1px solid #ccc; padding: 2px;"> <div style="background-color: #f2f2f2; padding: 2px;">topic</div> <div style="padding: 2px;">operation</div> <div style="padding: 2px;">data</div> <div style="padding: 2px;">format</div> </div> </div>
MAPPING ALGORITHM	Top matches per branch: <input style="width: 100px;" type="text" value="3"/>
IDF	Obsolete concepts: <input checked="" type="radio"/> false
CONCEPT MULTIPLIERS	Done annotations: <input checked="" type="radio"/> true
QUERY NORMALISERS	Inferior parents & children: <input checked="" type="radio"/> false
QUERY WEIGHTS	
SCORE LIMITS	
COUNTS	

Figure 14. edeamMap configuration. The text mining algorithms are configurable via the web user interface.

8.3 edamBrowser

edamBrowser is a web application that provides a novel graphical user interface for browsing resources (in bio.tools, TeSS *etc.*) based on their scientific annotation. It is optimised specifically for visualisation of EDAM (Fig. 15) and of EDAM usage / annotations. It also allows for community suggestions and additions (Fig. 16) to the ontology, via GitHub integration. The edamBrowser software⁵⁰ and web application⁵¹ are freely available under open license. An article⁵² describing the work has been submitted for publication in *The Journal of Open Source Software*.

⁵⁰ <https://github.com/IFB-ElixirFr/edam-browser>

⁵¹ <https://ifb-elixirfr.github.io/edam-browser/>

⁵² <https://github.com/IFB-ElixirFr/edam-browser/blob/master/paper.md>

EDAM ontology

EDAM ▾ Custom

Type at least 2 letters

EDAM is a simple ontology of well established, familiar concepts that are prevalent within bioinformatics [edamontology.org]

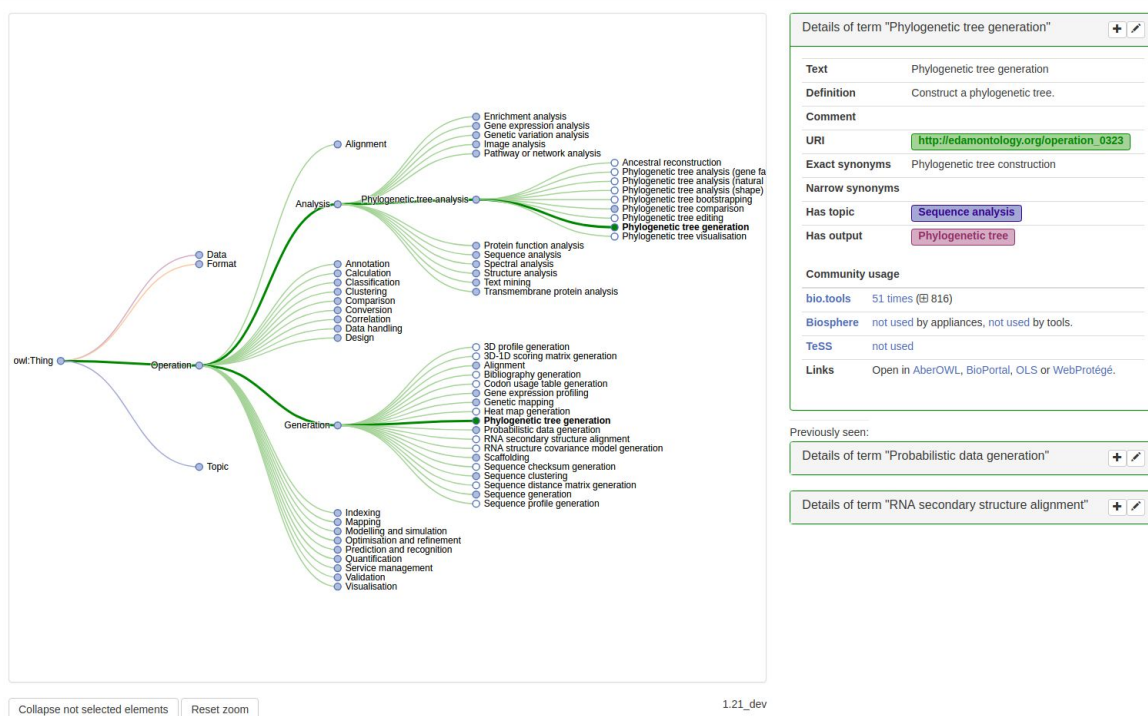


Figure 15. edamBrowser EDAM browsing. The utility offers a tree-based view of EDAM with information panel for the concept and its applications (annotations in various resources).

Figure 16. edamBrowser EDAM suggestions. Various types of changes or additions to the EDAM ontology can be suggestions can be made, which are formatted and forwarded to the EDAM developers via GitHub integration.

9. Future Work

In D1.2 we described future work, mostly directed towards delivering the desired “endgame” for the registry (Section 1), including:

- targeted actions to improve metadata quality, including content review and clean-up
- more accessible curation mechanisms
- clear information standards that motivate and support quality improvements
- better quality assurance and control process
- better engagement with and leverage of tool developers, service providers and other contributors
- curation guidelines and other documentation to facilitate contributions

Progress on specific actions (see D1.2, Table 5) is monitored mostly in GitHub and the “bio.tools features” and “bio.tools content” strands of sifterapp. This includes actions that address the positive suggestions raised from the EXCELERATE midterm review process.

10. Licensing

The registry source code is freely available under the GNU General Public License v3.0 only (GPL-3.0), guaranteeing the freedom to share and change the software and ensure it remains free software for all its users. The registry content is freely available to all under the Creative Commons Attribution Licence (CC BY 4.0), accordingly, members of the community may re-use the content for their own purposes and create their own interfaces tailored to their specific needs. EDAM and biotoolsSchema are licensed under a "Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0), which afford the same freedoms, but require copies or adaptations of the work to be released under the same or similar licence as the original.