EXCELERATE Deliverable D1.4

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.	1	
Lead Beneficiary:	38 - DTU	
WP Title	Tools Interoperability and Service Registry	
Contractual delivery date:	31 August 2019	
Actual delivery date:	21 August 2019	
WP leader:	Søren Brunak and Alfonso Valencia	38 - DTU, 12 - BSC
Partner(s) contributing to this deliverable:	12 - BSC, 38 - DTU	

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 Ll. Gelpí (ES), Adam Hospital (ES), Björn Grüning (DE), Peter Løngreen (DK), Søren Brunak (DK)

Reviewers:

ELIXIR-EXCELERATE WP Leaders

Table of contents

Table of contents 2

1. Executive Summary 2

2. Impact 4

3. Project objectives 6

4. Delivery and schedule 6

5. Adjustments made 6

6. Background information 7

7. Appendix 1: Registry release with comprehensive coverage of ELIXIR Node resources, including resource data format curation and analysis 11

7.1 The bio.tools registry of software tools and data resources for the life sciences 11

7.2 Community curation of bioinformatics software and data resources 11

7.3 A Thousand and One Software for Proteomics: Tales of the Toolmakers of Science 14

7.4 M1.1.4 EDAM release with coverage of different resource categories and RIs. Implementation of tooling for sustainable community development. 18

1. Executive Summary

The objective of EXCELERATE Deliverable 1.4 is the development of a discovery portal (bio.tools[1]) 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 (D1.4 at M48) summarises *consolidation* of the registry and portal described in D1.3, into a stable production system. The focus was on improving the registry software robustness and content quality, and on publications around the uptake, utility, scientific application and technical integration of the registry. This deliverable report describes work done with ELIXIR-EXCELERATE resources. Publications include:

- Ison J. et al. (2019). *The bio.tools registry of software tools and data resources for the life*

[1] https://dev.bio.tools (latest development version) or https://bio.tools

Horizon 2020 grant n. 676559

sciences. **Genome Biology** (accepted, doi:10.1186/s13059-019-1772-6).

The article summarises, for a broad audience, the method, progress and challenges in producing a comprehensive, consistent and sustainable tools registry. It is intended to promote the uptake of *bio.tools*.

- Ison J. et al. (2019). *Community curation of bioinformatics software and data resources.* **Briefings in Bioinformatics** (accepted, doi:10.1093/bioinformatics/btt113).

  The article describes the practical means by which scientific communities and individuals can use *bio.tools* to describe and disseminate their software productions.

- Tsiamis V. et al. (2019) *A Thousand and One Software for Proteomics: Tales of the Toolmakers of Science.* **Journal of Proteome Research** (accepted).

  The article presents the curation to a high standard of a collection of 189 software tools for proteomics data analysis. It can serve as an example to other communities.

Design and development aspects include:

- production of a minor new release of the stable data model (biotoolsSchema 3.1.0) which enhances the previous stable version 3.0.0 (described in D1.3)
- revision of the client and server-side software including conformance to biotoolsSchema 3.1.0
- greatly enriched API parameters allow precise queries over tool function and all other fields defined by biotoolsSchema 3.1.0, enabling more sophisticated applications
- software enhancements and usability improvements including enhanced search, interactive annotations, tool tips, richer publication information, links to TeSS *etc.* 71 issues including 21 bugs reported / suggested by users via GitHub[2] were addressed / fixed
- creation of a bio.tools homepage including key information and EDAM-based navigation on highly used operations and topics
- a new utility to text mine software publications and create candidate *bio.tools* entries (once in production, will make content growth and improvement more sustainable)
- technical scoping of a new GitHub-based content management model to improve content maintenance sustainability, encourage contributions and ease integrations: a sandbox of 2000 tool descriptions in JSON format are hosted at GitHub[3]

Operational aspects include:

- registry growth to 12,479 entries (was 11,479 in D1.3) refactored to the new model
- 239,656 aggregated annotations on tools (was 214,771 in D1.3) including 74,748 scientific (EDAM) annotations
- systematic improvement of the tool descriptions as per the Tool Information Information Standard[4]: 88% of entries are annotated to at least "Detailed" level
- consolidation of tool identifiers and entries ensuring usability and uniqueness, enabling easier integration of *bio.tools* with other resources.
- labelling in Tool Cards of official ELIXIR Core Data Resources, ELIXIR Deposition Database and Recommended Interoperability Resources, crediting ELIXIR Nodes and Platforms for tools and databases in ELIXIR Node Service Delivery Plans
- content contributor growth to 1,384 (was 800 in D1.3), following the community build-up process (described in D1.7)

[2] https://github.com/bio-tools/biotoolsregistry/
[3] https://github.com/bio-tools/content/
[4] https://bio-tools.github.io/Tool-Information-Standard/

Horizon 2020 grant n. 676559

- 14,263 users/month (July 2019) with 119,496 users and 184,767 page views in the reporting period (according to Google Analytics).

In addition, 2 new EDAM ontology release (1.22 and 1.23) were developed corresponding to the milestone below (due in M48):

- M1.1.4 "EDAM release with coverage of different resource categories and RIs. Implementation of tooling for sustainable community development"

In D1.3 we summarised our progress towards an envisioned "endgame": to provide a persistent reference to high-quality (curated and verified) "canonical" descriptions of *unique* tools. This vision has now been mostly accomplished, with a stable technical foundation in place for long-term growth of *bio.tools*.
Detailed information on new, or improved technical components is available online. All aspects of the project are interdependent and work will remain ongoing in the long-term, to drive content quality improvement, ensure sustainable growth, develop useful features for end-users and to support emerging integration scenarios and applications.

2. Impact

With sustained development, *bio.tools* can have a profound impact on the efficient comprehension and utilisation of software resources in the life sciences.

1. Substantial growth and upwards trends in **registry content**, **content contributors** and ***bio.tools* users** (see Executive Summary). The site is seeing an upward trend in contributors (Fig. 1), users (Fig. 2) and is used all over the world (Fig. 3)
2. **Publications** in high-impact journals (see Executive Summary).
3. **bio.tools usage** by numerous projects including *e.g.* BioContainers[5], OpenEBench[6], Instruct Toolbox[7], EMBL Australia Tools[8], EMBL-EBI Search[9], and multiple tool utility projects[10].
4. **bio.tools publication**[11] has an AltMetric attention score ranked #1 out of 207 outputs from *Nucleic Acids Research* of similar age, and #72 out of 20,934 outputs from *Nucleic Acids Research*, and is in the 97th percentile of all research outputs ever tracked by Altmetric. It has 49 citations (from Dimensions), 77% in the last 2 years
5. **EDAM publication**[12] has an AltMetric attention score ranked #10 out of 148 outputs from *Bioinformatics* of similar age, and #423 out of 8,466 outputs from *Bioinformatics*, and is in the 94th percentile of all research outputs ever tracked by Altmetric. It has 84 citations (Dimensions), 54% in the last 2 years.

[5] https://biocontainers.pro/#/
[6] https://openebench.bsc.es/
[7] https://instruct-eric.eu/compute/toolbox
[8] https://www.embl-abr.org.au/tools/
[9] https://www.ebi.ac.uk/ebisearch/overview.ebi/about
[10] https://github.com/bio-tools/
[11] Tools and data services registry: a community effort to document bioinformatics resources. Ison, J. et al. (2015). Nucleic Acids Research. doi: 10.1093/nar/gkv1116
[12] EDAM: An ontology of bioinformatics operations, types of data and identifiers, topics, and formats. Ison, J. et al. (2013). Bioinformatics. doi: 10.1093/nar/gkv1116

Horizon 2020 grant n. 676559

6. **EDAM ontology impact** : EDAM is typically in or near the top 10 most visited of 789 ontologies listed at NCBO BioPortal[13] with an average of 1372 monthly views since 2018 in BioPortal alone, and 42,209 daily mentions of "EDAM" in the EBI Ontology Look-up Service (OLS) Tomcat web server log. EDAM is used by numerous applications and communities, including *e.g. bio.tools*, EMBL-EBI Tools[14], Bioschemas[15], Galaxy[16], Debian Med[17], H3Africa[18], TeSS[19], EMBL-EBI training portal[20], CWL[21], NEUBIAS.org[22] and BISE bioimaging tools portal[23].

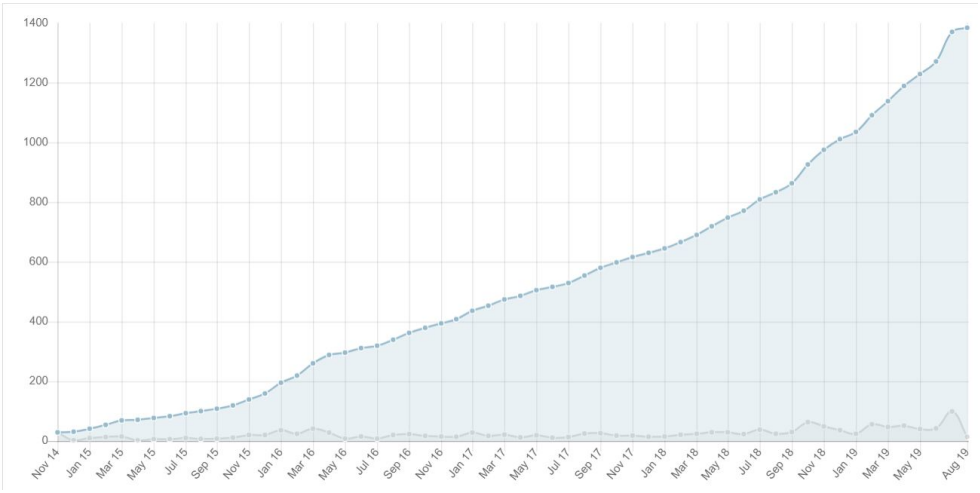

Figure 1. Growth in *bio.tools* contributors (people with bio.tools user accounts)

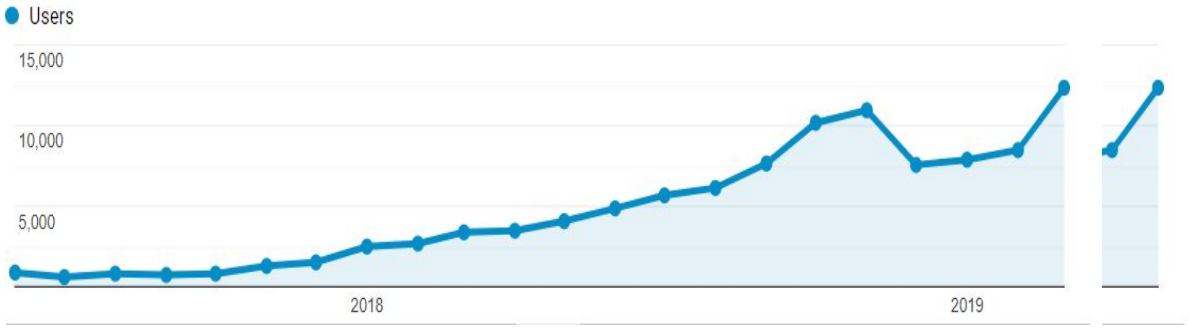

Figure 2. *bio.tools* monthly users since June 2018 (from Google Analytics)

[13] https://bioportal.bioontology.org/visits
[14] https://www.ebi.ac.uk/services
[15] http://bioschemas.org
[16] https://galaxyproject.org
[17] https://salsa.debian.org/med-team
[18] https://h3africa.org/
[19] https://tess.elixir-europe.org/
[20] https://www.ebi.ac.uk/training
[21] https://www.commonwl.org/
[22] http://eubias.org/NEUBIAS/
[23] https://biii.eu/

Horizon 2020 grant n. 676559

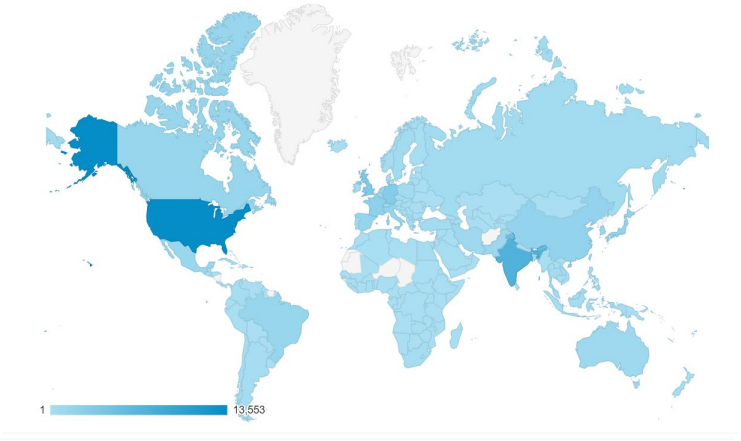

Figure 3. Global distribution of *bio.tools* users in the reporting period (from Google Analytics)

3. Project objectives

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

No.	Objective	Yes	No
1	Establish an ELIXIR discovery portal that provides a transparent route to tools and services for data access and exploitation by users	x	
2	Stimulate innovation by supporting industry uptake of ELIXIR resources, particularly in SMEs	x	

4. Delivery and schedule

The delivery is delayed: Yes • No☑

5. Adjustments made

N/A

Horizon 2020 grant n. 676559

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	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 7.00; 26 - CNRS 9.00; 29 - IP 12.00; 35 - MU 25.80; 38 - DTU 26.00, 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 DataServices Registry is the cornerstone of the WP.

WP 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

Horizon 2020 grant n. 676559

will thus stand as the guiding principle for the entire portal and guard its relevance to the community.

Task 1.1: Federated Registry Curation (100.34PM)

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

Horizon 2020 grant n. 676559

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 a 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

Horizon 2020 grant n. 676559

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 h 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 meetings 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

Horizon 2020 grant n. 676559

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

ELIXIR EXCELERATE D1.4 is delivered by three publications:

- Ison J. et al. (2019). *The bio.tools registry of software tools and data resources for the life sciences.* **Genome Biology** (accepted, doi:10.1186/s13059-019-1772-6). The article summarises, for a broad audience, the method, progress and challenges in producing a comprehensive, consistent and sustainable tools registry. It is intended to promote the uptake of *bio.tools*.
- Ison J. et al. (2019). *Community curation of bioinformatics software and data resources.* **Briefings in Bioinformatics** (accepted, doi:10.1093/bioinformatics/btt113). The article describes the practical means by which scientific communities and individuals can use *bio.tools* to describe and disseminate their software productions.
- Tsiamis V. et al. (2019) *A Thousand and One Software for Proteomics: Tales of the Toolmakers of Science.* **Journal of Proteome Research** (accepted). The article presents the curation to a high standard of a collection of 189 software tools for proteomics data analysis. It can serve as an example to other communities.

Excerpts from the articles are included below to illustrate the work done and impact.

7.1 The bio.tools registry of software tools and data resources for the life sciences

The article (Ison J. et al., doi:10.1186/s13059-019-1772-6) summarises, for a broad audience, the method, progress and challenges in producing a comprehensive, consistent and sustainable tools registry. It is intended to promote the uptake of *bio.tools*.

Abstract: Bioinformaticians and biologists rely increasingly upon workflows for the flexible utilization of the many life science tools that are needed to optimally convert data into knowledge. We outline a pan-European enterprise to provide a catalogue (https://bio.tools) of tools and databases that can be used in these workflows. *bio.tools* not only lists where to find resources, but also provides a wide variety of practical information.

7.2 Community curation of bioinformatics software and data resources

The article (Ison J. et al., doi:10.1093/bioinformatics/btt113) describes the practical means by which scientific communities and individuals can use *bio.tools* to describe and disseminate their software productions.

Abstract: The corpus of bioinformatics resources is huge and expanding rapidly, presenting the life scientist with a growing challenge in selecting tools that fit the desired purpose. To address this, the European Infrastructure for Biological Information, ELIXIR, is supporting a systematic approach towards a comprehensive registry of tools and databases for all domains of bioinformatics, provided under a single portal (https://bio.tools). We describe here the practical means by which scientific communities, including individual developers and projects, through to

major service providers and research infrastructures, can describe their own bioinformatics resources and share these via *bio.tools*.

Community support and engagement: Maintaining a corpus of tool descriptions in the long term depends upon effective community engagement. *bio.tools* offers direct assistance to individual developers of tools or providers of online services, as well as to organisations that foster a community, for example by participating in community-led curation events. ELIXIR is establishing a network of *Thematic Editors*; experts within fields of the life sciences who are motivated to liaise with their respective communities (national or scientific) and provide a bridge to *bio.tools*, supporting developments tailored to that community. In this context, the Danish ELIXIR node ran a studentship scheme, to support early career stage scholars, working under the aegis of a *Thematic Editor*, to contribute to *bio.tools* and gain experience with the ELIXIR infrastructure. This mechanism has proved to be an efficient method for bulk curation work. These initiatives are at an early stage and your involvement is most welcome; more information is available online (see Table 1).

Table 1. Resources for curation of software and database information.

Resource	Description
bio.tools	Registry of life science software and databases bio.tools github.com/bio-tools/biotoolsRegistry/
biotoolsSchema	Formalised XML schema (XSD) for bioinformatics resource information github.com/bio-tools/biotoolsschema
EDAM ontology	Ontology of bioinformatics topics, operations, types of data, data identifiers and data formats github.com/edamontology/edamontology
Tool Information Standard	Standard for bioinformatics resource information requirement at various tiers of description richness bio-tools.github.io/Tool-Information-Standard
Ontology Lookup Service (OLS)	Ontology browser from EMBL-EBI www.ebi.ac.uk/ols/ontologies/edam
BioPortal	Ontology browser from NCBO bioportal.bioontology.org/ontologies/EDAM
EDAM Browser	EDAM browsing and development tool from IFB ifb-elixirfr.github.io/edam-browser github.com/IFB-ElixirFr/edam-browser
EDAMmap	Utility for text mining and mapping to EDAM ontology biit.cs.ut.ee/edammap/ github.com/edamontology/edammap

Horizon 2020 grant n. 676559

The work above is in context of a broader ELIXIR initiative[24] to foster communities and bring together experts to develop standards, services and training within specific life science domains. ELIXIR Communities are international groupings of experts in a particular technical or scientific area, intended to drive the technical evolution of ELIXIR. Communities hold a special place in ELIXIR because they can receive funding from ELIXIR; for activities such as annual workshops and staff exchange, and through ELIXIR Implementation Studies (relatively small projects, funded over two years, that drive the development of the ELIXIR infrastructure). ELIXIR recently announced four new, community-led Implementation Studies that bring together Communities with the ELIXIR Platforms[25]. ELIXIR currently recognises eight Communities: Human Genomics Translational Data, Rare Diseases, Human Copy Number Variation, Crops and Forest Plants, Marine Metagenomics, Proteomics, Metabolomics and Galaxy. A number of other communities have indicated an interest in becoming part of ELIXIR and are in the process of being considered for approval. ELIXIR Communities are ideal resources to draw upon for subject-specific annotation and have been, and will continue to be, drawn upon for this purpose.
bio.tools and the other open projects and initiatives described here welcome your involvement. Information and instructions for new contributors is available online (see Table 2) for *bio.tools* and EDAM and includes details of mailing lists, how to make suggestions and requests, tasks and feature management, forthcoming meetings and events, and so on. Direct assistance with *bio.tools* is available by emailing registry-support@elixir-dk.org. The preferred option for communication, and especially for bug reports and suggestions, is GitHub[26].

Table 2. Links to documentation concerning *bio.tools*.

Documentation	**Description**
bio.tools docs	Documentation for the *bio.tools* registry biotools.readthedocs.io/
Curators Guide	Human-friendly guidelines for writing bioinformatics resource descriptions biotools.readthedocs.io/en/latest/curators_guide.html
Thematic Editors Guide	Emerging guidelines for bio.tools Thematic Editors (see *Community support and engagement*) biotools.readthedocs.io/en/latest/editors_guide.html
API Usage Guide	Usage guidelines with examples for the *bio.tools* API biotools.readthedocs.io/en/latest/user_guide.html
API reference	Comprehensive reference information for the *bio.tools* API biotools.readthedocs.io/en/latest/api_reference.html
bio.tools - getting involved	Overview of ways to get involved with *bio.tools* biotools.readthedocs.io/en/latest/contributors_guide.html
biotoolsSchema docs	Documentation for the biotoolsSchema resource description model biotoolsschema.readthedocs.io
EDAM docs	Documentation for the EDAM ontology edamontologydocs.readthedocs.io

[24] https://www.elixir-europe.org/communities
[25] https://elixir-europe.org/news/new-portfolio-community-led-implementation-studies-selected
[26] http://github.com/bio-tools/biotoolsregistry/

Horizon 2020 grant n. 676559

EDAM - getting involved	How to get involved with EDAM, including guidelines on how to request additions and changes edamontologydocs.readthedocs.io/en/latest/contributors_guide.html#requests
EDAM requests	Request additions and other changes to EDAM via GitHub (using documented issue templates or free-form requests) github.com/edamontology/edamontology/issues

Discussion: We have summarised, as a work in progress, the means by which software and databases can be described and shared via *bio.tools*, putting this in context of the various open projects and community-driven initiatives that are being fostered by the ELIXIR infrastructure. Such efforts provide the best hope for the sustainable provision and maintenance of high quality software information in the long term, required for various contexts and use cases. Beyond merely improving the findability of tools and the dissemination of basic information, the data have exciting applications, for example, in the automated construction and evaluation of alternative bioinformatics pipelines[27,28]. Such applications are only possible if carefully assigned functional annotation is available. To both ends, much work remains to be done, and will include production of "gold standard" tool descriptions for specific communities, provision of the *bio.tools* data in linked open data formats, and integration of *bio.tools* with other products such as Biocontainers.pro[29], Galaxy[30] and EuropePMC[31], to combine the *bio.tools* data with information about where tools can be used or downloaded in an executable form, and put in deeper context of their scientific application. There is also a need to promote better information standards for life science software more generally, such as we have described for EDAM, biotoolsSchema and the Tool Information Standard. All the software described here and the *bio.tools* data itself are made available under open licence. We welcome contributions and collaborations in all areas to improve the corpus of bioinformatics tool descriptions for the benefit of Life Scientists everywhere.

7.3 A Thousand and One Software for Proteomics: Tales of the Toolmakers of Science

The article (Tsiamis V. et al., accepted) presents the curation to a high standard of a collection of 189 software tools for proteomics data analysis. It can serve as an example to other communities.

Abstract: Proteomics is a very active field driven by frequent introduction of new technological approaches, leading to high demand for new software tools and the concurrent development of many methods for data analysis, processing and storage. The rapidly changing landscape of

[27] Lamprecht AL, Naujokat S, Margaria T, et al. "Semantics-Based Composition of EMBOSS Services." Journal of Biomedical Semantics, vol. 2, no. Suppl 1, 2011, doi:10.1186/2041-1480-2-S1-S5
[28] Palmblad M, Lamprecht AL, Ison J, et al. "Automated Workflow Composition in Mass Spectrometry-Based Proteomics." *Bioinformatics*, vol. 35, no. 4, 2018, doi:10.1093/bioinformatics/bty646
[29] da Veiga Leprevost F, Grüning BA, Alves Aflitos S, et al. "BioContainers: an Open-Source and Community-Driven Framework for Software Standardization." *Bioinformatics*, vol. 33, no. 16, 2017, pp. 2580–2582., doi:10.1093/bioinformatics/btx192
[30] Afgan E, Baker D, Batut B, et al. "The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2018 Update." *Nucleic Acids Research*, vol. 46, no. W1, 2018, doi:10.1093/nar/gky379
[31] The Europe PMC Consortium. "Europe PMC: a Full-Text Literature Database for the Life Sciences and Platform for Innovation." *Nucleic Acids Research*, vol. 43, no. D1, 2014, doi:10.1093/nar/gku1061

Horizon 2020 grant n. 676559

proteomics software makes finding a tool fit for a particular purpose a significant challenge. The comparison of software and the selection of tools capable to perform a certain operation on a given type of data relies on their detailed annotation using well-defined descriptors. However, finding accurate information including tool input/output capabilities can be challenging and often heavily depends on manual curation efforts. This is further hampered by a rather low half-life of most of the tools, thus demanding the maintenance of a resource with updated information about the tools. We present here our approach to curate a collection of 189 software tools with detailed information about their functional capabilities. We furthermore describe our efforts to reach out to the proteomics community for their engagement, which further increased the catalogue to >750 tools being about 70% of the estimated number of 1,097 tools existing for proteomics data analysis. Descriptions of all annotated tools are available through https://proteomics.bio.tools.

Results & Discussion: *bio.tools* includes a total of 754 tools with the EDAM Topic annotation of "Proteomics" (EDAM:topic_0121), "Proteomics experiment" (EDAM:topic_3520), or synonyms of these terms. These tools are, for convenience, associated with a *bio.tools* subdomain available for browsing at https://proteomics.bio.tools. The tool descriptions include a total of 30,187 annotations, of which 6,350 are EDAM annotations. The metadata richness of these tools (Table 3) according to the Tool Information Standard shows that 81% of the collection have "Detailed" (or richer) annotation. Of this corpus, this work contributed 189 new tool registrations in *bio.tools*, of which 167 are annotated to at least "Detailed" level. Of these 189 tools, 93% (176) have an input or output defined, with a total of 562 EDAM Data annotations and 963 EDAM Format annotations. The corpus of 754 tools includes command-line tools (276, 31%), web applications (224, 25%), desktop applications (170, 19%), libraries (97, 11%) and a long tail of other tool types defined in biotoolsSchema (note that a single tool can be annotated as of more than one basic type). The ms-utils.org wiki has been structured according to the EDAM ontology, linking most tool categories to EDAM operations. As a single-page wiki, a user can conveniently search the list for keywords appearing in the subheaders or tool descriptions, or look through the tools in a particular category.

Table 3. Metadata richness of curated proteomics tools. The number of *bio.tools* entries compliant to different tiers in the Tool Information Standard is shown.

Tier of Tool Information Standard	**# tools**
Sparse	34
Basic details	105
Detailed	589
Highly detailed	5
Comprehensive	21

At the outset, EDAM only contained a few terms for proteomics data analysis; many additions and changes were needed, and these were made progressively over multiple EDAM releases. Changes included adding new concepts where these were missing, ensuring the preferred label reflected the vernacular, and adding common synonyms of this term. The conceptual hierarchy (concept subsumption relationships) was also extensively revised, to make navigation of EDAM

Horizon 2020 grant n. 676559

and term picking easier in ontology browsers. As an illustration, Figure 4 shows spectral analysis operations (EDAM:operation_3214). A particular effort was focussed on the curation of proteomics data formats as these have high practical value in applications such as workflow composition[32]. 30 new mass spectrometry data formats were added, including proprietary formats created by companies to support specific machines and specific commercial software, such as Thermo RAW format (EDAM:format_3712) supported by Xcalibur, and open source data formats such as mzXML (EDAM:format_3654). EDAM overlaps with a few of the concepts within the PSI-MS ontology[33] which is designed to describe a mass spectrometry experiment. Collaborative efforts between the maintainers of both ontologies will be intensified to ensure the interoperability of these ontologies, for example by cross-referencing equivalent concepts.

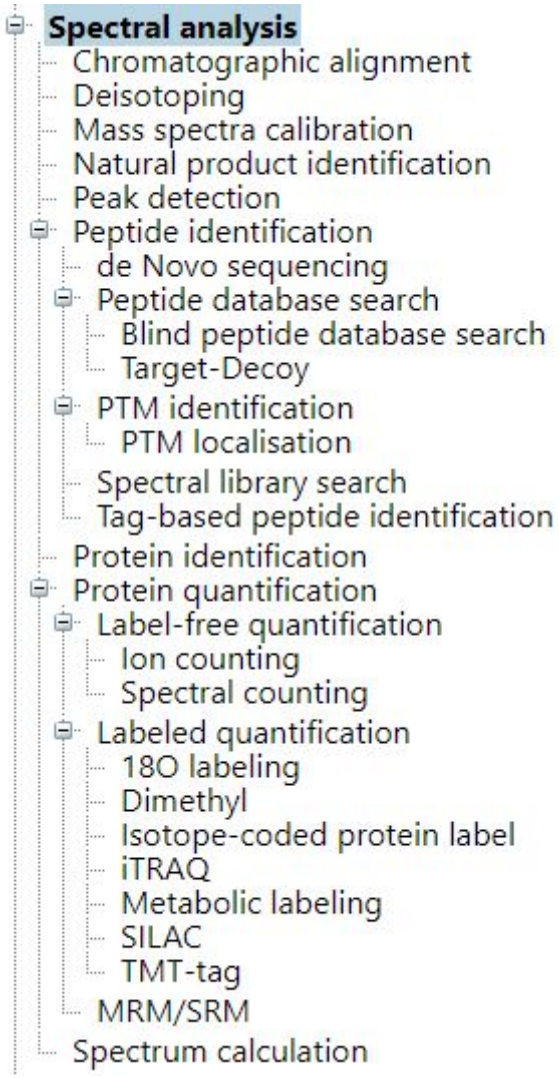

Figure 4. Extract from the EDAM ontology showing operations for spectral analysis (EDAM:operation_3214).

[32] Palmblad, M.; Lamprecht, A.-L.; Ison, J.; Schwämmle, V. Automated Workflow Composition in Mass Spectrometry-Based Proteomics. *Bioinformatics* **2019**, *35* (4), 656–664.
[33] Montecchi-Palazzi, L.; Beavis, R.; Binz, P.-A.; Chalkley, R. J.; Cottrell, J.; Creasy, D.; Shofstahl, J.; Seymour, S. L.; Garavelli, J. S. The PSI-MOD Community Standard for Representation of Protein Modification Data. *Nat. Biotechnol.* **2008**, *26* (8), 864–866.

Horizon 2020 grant n. 676559

Early community engagement helped to prioritise the curation effort and focus on annotations of high practical value to tool interoperability, such as input and output data formats. The emerging Tool Information Standard was helpful to structure the curation effort on a technical level. It was surprising that in many cases, crucial usage information such as programming language, software license and terms of use, were not easy to find. Basic input and output data types and formats were also often not stated in an explicit and clear manner. As a last resort, supported data formats were identified through inspection of the source code or by testing the tools at the command-line. Such challenges in finding information present a barrier to end-users' efficient discovery and use of tools, and underlines the need for resources such as *bio.tools* and ms-utils.org. To provide consistent search and discovery, these resources benefit greatly from the use of controlled vocabularies defined within biotoolsSchema and the EDAM ontology. The numerous EDAM concepts about proteomics are well described and documented, and able to represent almost all the functionality found in any proteomics tool, which in combination with ontology browsers such as OLS, BioPortal and EDAM Browser, greatly facilitates the manual annotation of tools.

The proteomics community benefits from consistent and detailed tool annotations in various ways. *bio.tools* allows researchers to query and find appropriate tools with algorithms that fulfill a certain task, with correctly described input and output file compatibilities, and with valuable references to documentation, tutorials and training. Annotations can furthermore be used to compose workflows comprising multiple operations as previously shown[34]. Collections of single tools and workflows executing identical operation(s) for the analysis of proteomics data can be created and benchmarked by comparing the results using ground-truth data sets. We envision that future efforts will result in a feedback loop where valuable information about tool performance will continuously be updated by monitoring their usage and therefore enhance the annotation, presentation and discovery of optimally performing tools.

Conclusion: Organized catalogues of expert-annotated software, like Pedro's list, that describe software using a standardized vocabulary on one platform, facilitate what can be a daunting search for tools that suit a particular scientific or technical purpose. Providing better and more permanent tool findability should automatically lead to longer half-lives, assuming the software to be functional and maintained. We have summarised a successful curation effort that has enriched *bio.tools,* ms-utils.org and the EDAM ontology, and rendered a significant proportion of all proteomics analysis software more findable, accessible, interoperable and reusable, *i.e.* more FAIR[35]. While detailed annotation of fine-grained details such as data formats are costly, the effort is warranted where it supports valuable scientific applications such as tool interoperability and workflow composition. Detailed tool annotations including input/output data and format will open the door for identifying novel workflows of compatible tools and for implementing alternative workflow components to benchmark their performance. Such efforts can leverage software containers, for example those being registered in Biocontainers[36] in collaboration with *bio.tools*, with the hope to greatly simplify deployment of full data analysis pipelines on local high-performance machines or on the cloud. BioContainers is providing an infrastructure to create, deploy and maintain software containers using Conda and Docker technologies. bio.tools

[34] Palmblad, M.; Lamprecht, A.-L.; Ison, J.; Schwämmle, V. Automated Workflow Composition in Mass Spectrometry-Based Proteomics. *Bioinformatics* **2019**, *35* (4), 656–664.

[35] Wise, J.; de Barron, A. G.; Splendiani, A.; Balali-Mood, B.; Vasant, D.; Little, E.; Mellino, G.; Harrow, I.; Smith, I.; Taubert, J.; et al. Implementation and Relevance of FAIR Data Principles in Biopharmaceutical R&D. *Drug Discov. Today* **2019**.

[36] da Veiga Leprevost, F.; Grüning, B. A.; Alves Aflitos, S.; Röst, H. L.; Uszkoreit, J.; Barsnes, H.; Vaudel, M.; Moreno, P.; Gatto, L.; Weber, J.; et al. BioContainers: An Open-Source and Community-Driven Framework for Software Standardization. *Bioinformatics* **2017**, *33* (16), 2580–2582.

Horizon 2020 grant n. 676559

and BioContainers are coordinated under the ELIXIR Tools Platform[37] . As an example, systematic efforts are ongoing to ensure that all tools which have been containerised are registered in bio.tools, and conversely, bio.tools is being used to provide metadata for exposure in the BioContainers registry.

The difficulties we encountered in finding sometimes even basic information about tools point to a pressing requirement for the promotion of better standards of information for life science software generally. There is a need for upstream provision of richer and more consistent software metadata that can be conveniently reused by efforts such as *bio.tools*. On the cataloguing side, there is no free lunch; high quality content requires an investment of time and manual effort. The curation effort would benefit greatly from more automated ways to harvest trivial annotations such as software license, for example by text mining the literature, and by a closer integration of the ontology construction and tool registration processes, for example harvesting missing terms and synonyms at tool registration time. The quality of annotations would benefit from more powerful and convenient means for term selection, which itself is a significant challenge given the sizeable vocabularies that the typical end-user, pressed for time, may be unfamiliar with.

We approximate that of the potential volume of 1,097 tools existing for proteomics (see *Sources of information*), *bio.tools* captures 68% (754) with 56% (615) annotated to "Detailed" level or better. Thus, there is more curation work to do, and we can expect many new resources to appear in the future, which will also reflect new analytical methods, types of data and data formats. We recently contacted the developers of tools in the proteomics corpus for which contact details were available in *bio.tools*, and hope this will lead to community adoption and maintenance of the corpus in the long-term. We hope the efforts described here for proteomics will stimulate similar efforts in other domains, which are also witness to a large rise in the volume of new tools and data resources[38]. The anchoring of *bio.tools* within the ELIXIR infrastructure will ensure *bio.tools* is maintained in the long term and we encourage the whole proteomics community to collaborate with us on further improving the corpus of tool descriptions.

7.4 M1.1.4 EDAM release with coverage of different resource categories and RIs. Implementation of tooling for sustainable community development.

Two new EDAM releases (1.22 - 1.23) were produced, in summary:

- various new concepts and other changes to support the requirements of the Human Cell Atlas
- extension of Format subontology (24 concepts added)
- simplification of Data subontology (24 concepts deprecated)
- terms for machine learning and statistical methods added as narrow synonyms of *Machine learning*, *Statistics and probability* or *Mathematics* (following alignment of EDAM with the DSEO ontology)
- various clean-ups, and minor bug fixes

A detailed description of changes is available online:
https://github.com/edamontology/edamontology/blob/master/changelog.md (changelog)

https://github.com/edamontology/edamontology/milestone/10?closed=1 (1.22 changes)

https://github.com/edamontology/edamontology/milestone/12 (1.23 changes)

[37] https://elixir-europe.org/platforms/tools
[38] Editorial: The 16th Annual Nucleic Acids Research Web Server Issue 2018. *Nucleic Acids Res.* **2018**, *46* (W1), W1–W4.

Horizon 2020 grant n. 676559