EXCELERATE Deliverable D1.8

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:	Matchmaking service: implementation & evaluation of impact	
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	

Authors and Contributors:
Magnus Palmblad, Anna-Lena Lamprecht, Veit Schwämmle and Jon Ison, DTU

Reviewers:

N/A

Table of contents

1. Executive Summary

The objective of EXCELERATE Deliverable 1.8, as stated in the original proposal, was to analyse the data format-usage landscape of tools registered the ELIXIR Tools and Data Services Registry (bio.tools) to provide a basis for targeted software developments to improve interoperability of registered tools. We originally foresaw those developments to be facilitated via a "Matchmaking Service mechanism" possibly including conversion of tools to use common formats, and development of format converter software where needed. Whilst such specific developments are laudable, it is now obvious that significant progress depends on the bioinformatics community at large, and requires vastly greater capacity than WP1 has at its disposal.

Instead, we took a much more targeted approach, focussing - as an exemplar - upon the systematic curation of tools for proteomics data analysis, and subsequent exploitation of the annotations for automated workflow composition. The work addresses the challenge of tool interoperability and is a major stride in delivering a practical "Matchmaking Service mechanism". The work is described in two articles:

- Tsiamis, V., Ienasescu, H., Palmblad, M. Schwämmle, V. and Ison J. *Community curation of software tools as illustrated for mass spectrometry-based proteomics.* **In preparation,** see https://tinyurl.com/proteomics-tools.

Horizon 2020 grant n. 676559

- Palmblad, M., Lamprecht, A., Ison, J. and Schwämmle, V. (2018). *Automated workflow composition in mass spectrometry based proteomics.* Accepted for publication in **Bioinformatics**[1]

The first article (Tsiamis *et al.*) describes a systematic approach towards the comprehensive coverage in bio.tools of the prevalent tools for proteomics data analysis, including expert curation of many tools to a high standard including consistent annotation of data formats and operation using the EDAM ontology. The second (Palmblad *et al.*) explores automated workflow composition from the tool semantic annotation, and provides a toolkit to support researchers in identifying, comparing and benchmarking multiple workflows from individual bioinformatics tools.

For this deliverable report, we link to and summarise the publications in press. The project files and workflows are freely available:

- https://github.com/bio-tools/biotoolsCompose/tree/master/Automatic-Workflow-Composition

2. Impact

The work addresses a problem of fundamental importance in bioinformatics, namely, the selection of practical and effective data analysis pipelines for a specific experimental design. This was done by providing a toolkit to support researchers in identifying, comparing and benchmarking multiple workflows from individual bioinformatics tools. Automated workflow composition is enabled by the tools' semantic annotation in terms of the EDAM ontology. We demonstrated that the specification of operations, data types and formats enables the identification of compatible tools and composition of a set of tentatively viable workflows as permutations of a data analysis plan. This is a small but important step towards plug-and-play workflow composition, where researchers only need to specify the overall setup of their bioinformatics pipeline to automatically receive directly executable and benchmarked software solutions.

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	

[1] https://doi.org/10.1093/bioinformatics/bty646

Horizon 2020 grant n. 676559

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.

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.			

Horizon 2020 grant n. 676559

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,

Horizon 2020 grant n. 676559

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 resource 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.

Horizon 2020 grant n. 676559

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

Horizon 2020 grant n. 676559

5. Appendix 1: Matchmaking service: implementation & evaluation of impact

ELIXIR EXCELERATE D1.8 is delivered by two publications:

- Tsiamis, V., Ienasescu, H., Palmblad, M. Schwämmle, V. and Ison J. *Community curation of software tools as illustrated for mass spectrometry-based proteomics.* **In preparation,** see https://tinyurl.com/proteomics-tools and Section 3.1
- Palmblad, M., Lamprecht, A., Ison, J. and Schwämmle, V. (2018). *Automated workflow composition in mass spectrometry based proteomics.* Accepted for publication in **Bioinformatics.** See Section 3.2.

1. Publications

1.1 Community curation of software tools for mass spectrometry-based proteomics

The article (Tsiamis *et al*.) describes a systematic approach towards the comprehensive coverage in bio.tools of the prevalent tools for proteomics data analysis, including expert curation of many tools to a high standard including consistent annotation of data formats and operation using the EDAM ontology.

- Tsiamis, V., Ienasescu, H., Palmblad, M. Schwämmle, V. and Ison J. *Community curation of software tools as illustrated for mass spectrometry-based proteomics.*

The article is being prepared for publication in *Briefings in Bioinformatics*, see https://tinyurl.com/proteomics-tools**.**

1.2 Automated workflow composition in mass spectrometry based proteomics

The article (Palmblad *et al.*) explores automated workflow composition from the tool semantic annotation, and provides a toolkit to support researchers in identifying, comparing and benchmarking multiple workflows from individual bioinformatics tools.

- Palmblad, M., Lamprecht, A., Ison, J. and Schwämmle, V. (2018). *Automated workflow composition in mass spectrometry based proteomics.* Accepted for publication in **Bioinformatics**
  https://doi.org/10.1093/bioinformatics/bty646
  https://www.overleaf.com/12337446kyqdktwyctny#/61745978/

2. Summary

2.1 Introduction

Numerous software utilities operating on mass spectrometry (MS) data are described in the literature and provide specific operations as building blocks for the assembly of on-purpose workflows. Working out which tools and combinations are applicable or optimal in practice is often hard. Thus researchers face difficulties in selecting practical and effective data analysis pipelines for a specific experimental design.

Horizon 2020 grant n. 676559

In our article[2] we explore the value of formalized semantic tool descriptions for guided construction of practical workflows for mass spectrometry-based proteomics. Workflow creation follows a minimal framework, *i.e.* definition of operations, input files and output files of the complete workflow. We use PROPHETS to synthesize workflows based on a library of EDAM-annotated tools registered within bio.tools. We implement different workflows for four typical use cases in the analysis of mass spectrometry data: peptide retention time prediction, protein identification and enrichment analysis, localization of phosphorylation and protein quantitation using isotopic labeling. Finally, we test and compare the workflows on public data and thereby demonstrate both the feasibility and potential of automatic workflow composition in bioinformatics exemplified for mass spectrometry-based proteomics.

2.2 Summary of results

We provide a toolkit to support researchers in identifying, comparing and benchmarking multiple workflows from individual bioinformatics tools. Automated workflow composition is enabled by the tools' semantic annotation in terms of the EDAM ontology. To demonstrate the practical use of our framework, we created and evaluated a number of logically and semantically equivalent workflows for four use cases representing frequent tasks in MS-based proteomics. Indeed we found that the results computed by the workflows could vary considerably, emphasizing the benefits of a framework that facilitates their systematic exploration.

2.3 Availability and implementation

The project files and workflows are available from https://github.com/bio-tools/biotoolsCompose/tree/master/Automatic-Workflow-Composition

2.4 Future work

We have demonstrated a proof of concept. Much further work is needed to provide an implementation suitable for the non-expert bench user, and many challenges stand in the way of this vision. In summary, this includes the of modelling complex tools with multiple inputs and outputs, flexible ways to rank and filter prospective workflows, the algorithmic complexity of synthesis algorithms, specification and synthesis of non-linear workflows (i.e., workflows with parallelism, loops or conditional branchings), provision of reliable benchmarking datasets for different analyses, and the automated implementation of workflow solutions, to name a few. We envision to work together with the developer of workflow management systems, to integrate our method there to provide additional workflow composition support to their users. For example, if running the synthesis from within, say, a Galaxy server, then it could be made aware of what software tools are available on that server. Workflow generation could then be restricted to already installed software, or components already available on the server could be prioritized over those that are not. Furthermore, the synthesized workflows could be exported in exchange formats like the Common Workflow Lanugage (CWL) and provided along with containerized tool packages to facilitate their execution on various platforms.

[2] https://doi.org/10.1093/bioinformatics/bty646

Horizon 2020 grant n. 676559

Horizon 2020 grant n. 676559