EXCELERATE Deliverable 1.6

Project Title:	ELIXIR-EXCELERATE: Fast-track ELIXIR implementation and drive early user exploitation across the life sciences	
Project Acronym:	ELIXIR-EXCELERATE	
Grant agreement no.:	676559	
	H2020-INFRADEV-2014-2015/H2020-INFRADEV-1-2015-1	
Deliverable title:	REPORT: Workbench integration enabler: implementation & evaluation of impact	
WP No.	1	
Lead Beneficiary:	38: DTU	
WP Title	Tools Interoperability and Service Registry	
Contractual delivery date:	31 August 2017	
Actual delivery date:	31 August 2017	
WP leader:	ren Brunak (DK) and Alfonso Valencia (ES)	DTU, BSC

Authors and contributors:

Hervé Ménager (FR), Jon Ison (DK), Kenzo-Hugo Hillion (FR), Ivan Kuzmin (EE), Anton Khodak (UA), Eric Rasche (DE), Hedi Peterson (EE)

Table of content

1. Executive Summary

The objective of EXCELERATE Deliverable 1.6 is to integrate the ELIXIR Tools and Data Services Registry (bio.tools) with workbench environments such as Galaxy, respecting the 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. The ELIXIR EXCELERATE proposal envisioned integration in two ways:

a Workbench Integration Enabler service to develop the vision “register your software once - enable its support everywhere”. The idea, technically, is to translate the description of any tool or service that is registered in bio.tools, into the metadata format required by the existing major workbench environments. In so, doing, the cost of maintaining quality tool description for us in workbench environments is lowered, giving a practical boost to tool interoperability.

a utility for en masse registration of services from Galaxy instances. In the context of a generic registry for bioinformatics software, such as bio.tools, Galaxy instances constitute a major source of valuable content (online services and underlying tools). The idea is to provide a sustainable means to harvest this content.

We have made major strides in delivering this functionality, having published two papers with a third in preparation:

Ménager, H., Kalaš, M., Rapacki, K. and Ison, J. (2016). Using registries to integrate bioinformatics tools and services into workbench environments. Int J Softw Tools Technol Transfer, doi:10.1007/s10009-015-0392-z

Hillion KH, Ison J and Ménager H. (2017). ToolDog – generating tool descriptors from the ELIXIR tool registry. [version 1; not peer reviewed]. F1000Research 2017, 6:767 (poster). doi:10.7490/f1000research.1114125.1

Doppelt-Azeroual, O., Mareuil, F., Deveaud, Kalaš, M., Soranzo, N., van den Beek, M., Grüning, B., Ison, J. and Ménager, H. (2017). ReGaTE: Registration of Galaxy Tools in Elixir. GigaScience, doi:10.1093/gigascience/gix022

The first article (Ménager et al.) outlines the general strategy behind the Workbench Integration Enabler service, scopes the problem, and provides the technical groundwork. The second (Hillion KH et al.) is a poster summarising progress thus far in implementing this service and the basis for a publication in preparation. The third (Doppelt-Azeroual et al.) describes a utility for en masse registration of services from Galaxy instances, which has been applied to the registration of services from the Galaxy instance hosted by Institut Pasteur, as described in D1.2 report.

For this deliverable report, we attach and summarise the publications in press. The software is available under open source license:

ToolDog (MIT license)

https://github.com/bio-tools/ToolDog

ReGaTE (GPL license)

https://github.com/C3BI-pasteur-fr/ReGaTE

2. 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	

3. Delivery and schedule

The delivery is delayed: Yes ☑ No

4. Adjustments made

N/A

5. Background information

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

Work package number	WP1	**Lead beneficiary**	38 - DTU
Work package title	Tools Interoperability and Service Registry		
Start Month	1	**End Month**	48
Work Package Lead	Søren Brunak (DK) and Alfonso Valencia (ES)		
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.			

Description of work and role of partners

WP1 - Tools Interoperability and Service Registry [Months: 1-48]

DTU, EMBL, UOXF, UTARTU, CNIO, CRG, IRB, 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 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.orgcompatible 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

Partner number, short name and effort: 1 - EMBL 12.00; 2 - UOXF 6.00; 5 - UTARTU 43.00; 7 - CNIO 2.00; 8 - CRG 11.00; 10 - IRB 8.00; 17 - INESC-ID 4.00; 21 - UiB 18.00; 25 - SIB 9.50; 26 - CNRS 9.00; 29 - IP 12.00; 35 - MU 18.20; 38 - DTU 76.00

6. REPORT: Workbench integration enabler: implementation & evaluation of impact

Summary

The ELIXIR EXCELERATE proposal envisioned integration of bio.tools with workbench environments in two ways (Figure 1).

- a *Workbench Integration Enabler* service translates the description of any tool or service that is registered in bio.tools, into the metadata format required by the existing major workbench environments. The strategy, scoping and technical groundwork for this have been published (Section 6.1). At the core of the service is a utility called ToolDog which is described in a publication in preparation (Section 6.2).
- a utility for *en masse* registration of services from Galaxy instances. This utility (ReGaTE) has been published and applied to the registration of services from the Galaxy instance hosted by Institut Pasteur, as described in D1.2 report.

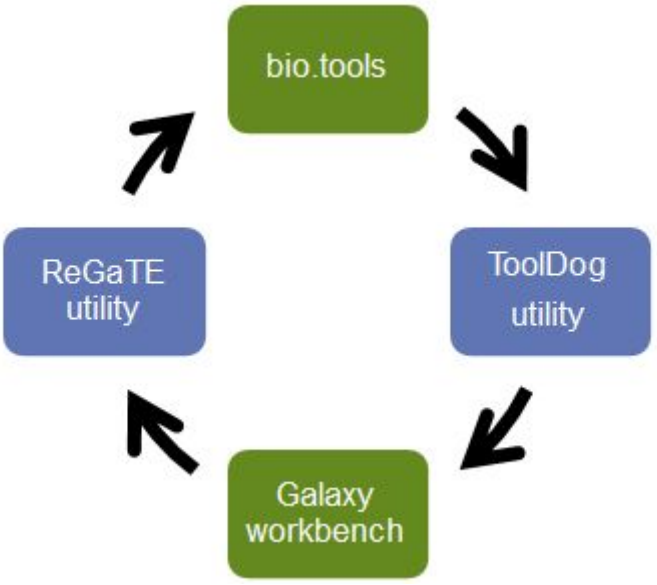

Figure 1. Overview of tooling for bio.tools workbench integration

6.1 Using registries to integrate bioinformatics tools and services into workbench environments

The strategy, scoping and technical groundwork of the workbench integration enabler service are published (see Abstract below), with specific implementation *via* ToolDog utility (Section 6.3).

- Ménager, H., Kalaš, M., Rapacki, K. and Ison, J. (2016). *Using registries to integrate bioinformatics tools and services into workbench environments.* **Int J Softw Tools Technol Transfer**, doi:10.1007/s10009-015-0392-z

Abstract: The diversity and complexity of bioinformatics resources presents significant challenges to their localisation, deployment and use, creating a need for reliable systems that address these issues. Meanwhile, users demand increasingly usable and integrated ways to access and analyse data, especially within convenient, integrated "workbench" environments. Resource descriptions are the core element of registry and workbench systems, which are used to both help the user find and comprehend available software tools, data resources, and Web Services, and to localise, execute and combine them. The descriptions are, however, hard and expensive to create and maintain, because they are volatile and require an exhaustive knowledge of the described resource, its applicability to biological research, and the data model and syntax used to describe it. We present here the Workbench Integration Enabler, a software component that will ease the integration of bioinformatics resources in a workbench environment, using their description provided by the existing ELIXIR Tools and Data Services Registry.

6.3 ToolDog – generating tool descriptors from the ELIXIR tool registry

ToolDog (Figure 2) is the utility at the core of workbench integration enabler service. It helps a system administrator generate high quality tool descriptions for use in Galaxy or generically via the Common Workflow Language[1] (CWL). It integrates general tool information from bio.tools with command-line specifications from various sources.

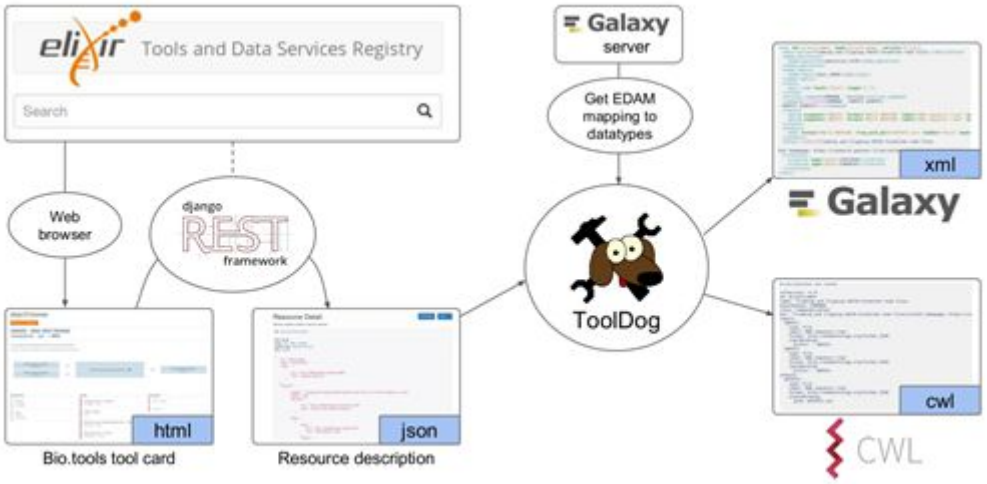

Figure 2. Overview of ToolDog utility

The ToolDog utility is described in a publication in preparation (see Abstract below):

- Hillion KH, Ison J and Ménager H. (2017). *ToolDog – generating tool descriptors from the ELIXIR tool registry.* [version 1; not peer reviewed]. **F1000Research** 2017, 6:767 (poster). doi:10.7490/f1000research.1114125.1

ToolDog is a prerequisite to providing, as a future work, the functionality *via* an online service, invocable from the bio.tools user interface and thus fully realising the *Workbench Integration Enabler* idea. The impact of this service will be evaluated in a future deliverable report.

The source code for ToolDog is freely available under MIT license:
https://github.com/bio-tools/ToolDog

Abstract: Bioinformatics workbench and workflow systems such as Galaxy, Taverna, or Common Workflow Language (CWL)-based frameworks, facilitate the access to bioinformatics tools in a user-friendly, scalable and reproducible way. Still, the integration of tools in such environments remains a cumbersome, time consuming and error-prone process. A major consequence is the incomplete or outdated description of tools that are often missing information such as some parameters, a description or metadata.

[1] https://github.com/common-workflow-language/common-workflow-language

ToolDog (Tool Description Generator) is the main component of the Workbench Integration Enabler service of the ELIXIR bio.tools registry. The goal of this tool is to guide the integration of tools into workbench environments. ToolDog is divided in two modules: the first analyses the source code of the bioinformatics software with language dedicated tools and generates a Galaxy XML or CWL tool description. The second is dedicated to the enrichment of the generated tool description using metadata provided by bio.tools. This last module can also be used on its own to complete or correct existing tool descriptions with missing metadata.

6.3 ReGaTE: Registration of Galaxy Tools in Elixir

ReGaTE (Figure 3) is a utility that automates the process of *en masse* registration in bio.tools of services from Galaxy instances. It extracts service metadata from a Galaxy server, enhances the metadata with the scientific information required by bio.tools, and pushes it to the registry.

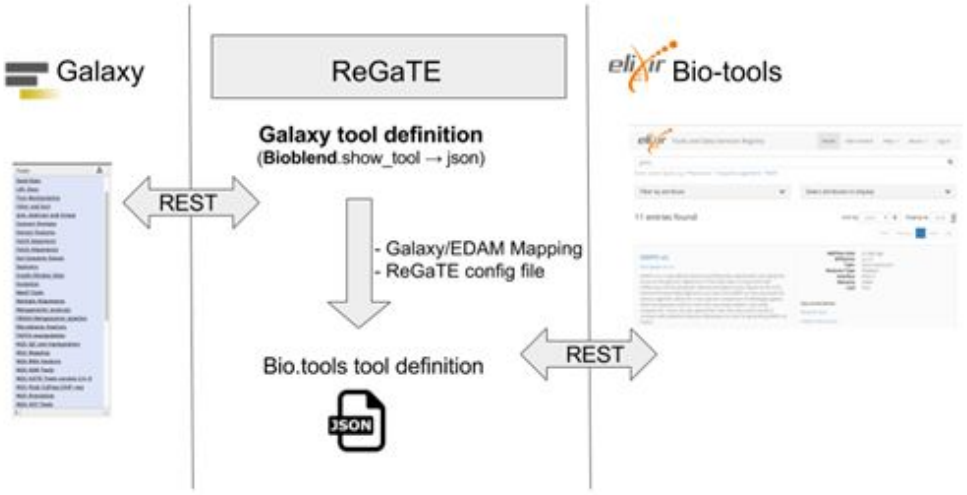

Figure 3. Overview of ReGaTE utility

The work has been published (see Abstract below):

- Doppelt-Azeroual, O., Mareuil, F., Deveaud, Kalaš, M., Soranzo, N., van den Beek, M., Grüning, B., Ison, J. and Ménager, H. (2017). *ReGaTE: Registration of Galaxy Tools in Elixir.* **GigaScience**, doi:10.1093/gigascience/gix022

ReGaTE has been applied to the registration of services from the Galaxy instance hosted by Institut Pasteur, as described in D1.2 report. As a future work, ReGaTE will be extended to process metadata not only for online services within Galaxy instances, but also the underlying tools, ensuring coverage of these within bio.tools, in preparation for the ongoing refactoring of bio.tools which is providing persistent references to canonical descriptions of unique tools (see D1.2 report).

The source code for ReGaTE is available under GPL license:

- https://github.com/C3BI-pasteur-fr/ReGaTE

Abstract: Bioinformaticians routinely use multiple software tools and data sources in their day-to-day work and have been guided in their choices by a number of cataloguing initiatives. The ELIXIR Tools and Data Services Registry (bio.tools) aims to provide a central information point, independent of any specific scientific scope within bioinformatics or technological implementation. Meanwhile, efforts to integrate bioinformatics software in workbench and workflow environments have accelerated to enable the design, automation, and reproducibility of bioinformatics experiments. One such popular environment is the Galaxy framework, with currently more than 80 publicly available Galaxy servers around the world. In the context of a generic registry for bioinformatics software, such as bio.tools, Galaxy instances constitute a major source of valuable content. Yet there has been, to date, no convenient mechanism to register such services *en masse*. We present ReGaTE (Registration of Galaxy Tools in Elixir), a software utility that automates the process of registering the services available in a Galaxy instance. This utility uses the BioBlend application program interface to extract service metadata from a Galaxy server, enhance the metadata with the scientific information required by bio.tools, and push it to the registry. ReGaTE provides a fast and convenient way to publish Galaxy services in bio.tools. By doing so, service providers may increase the visibility of their services while enriching the software discovery function that bio.tools provides for its users. The source code of ReGaTE is freely available on Github at https://github.com/C3BI-pasteur-fr/ReGaTE.