



[www.emif.eu](http://www.emif.eu)

## European Medical Information Framework


*Grant Agreement n°115372*

# **D13.1 Evaluation of technologies and tools available for data analysis and visualisation**

**WP13–Analysis, processing & visualisation methods and tools**


**V1.9  
[Final]**

Lead beneficiary: EMBL  
Date: 20/10/2014  
Nature: R  
Dissemination level: PU

 IMI - 115372	<b>D13.1</b> - Catalogue of technologies and tools available for data analysis and visualization		
	<b>WP13.</b> Analysis, processing & visualization methods and tools		<b>Version:</b> v1.9 - Final
	<b>Authors:</b> Natalja Kurbatova, Rudi Verbeek		<b>Security:</b> PU

## TABLE OF CONTENTS

<b>DOCUMENT INFORMATION .....</b>	<b>3</b>
<b>DOCUMENT HISTORY .....</b>	<b>3</b>
<b>KEY WORDS (WORDLE STYLE) .....</b>	<b>6</b>
<b>1. INTRODUCTION.....</b>	<b>7</b>
<b>2. SOFTWARE ACCEPTED AND USED BY EMIF PLATFORM.....</b>	<b>8</b>
2.1. TRANSMART .....	8
2.2. EMIF CATALOGUE.....	9
2.3. JERBOA RELOADED .....	9
<b>3. SOFTWARE POTENTIALLY USEFUL FOR THE EMIF PARTNERS .....</b>	<b>11</b>
3.1. GENESTACK .....	11
3.2. GALAXY .....	12
3.3. R CLOUD WORKBENCH .....	12
3.4. AUTOMATIC BIO CURATOR (ABC) .....	13
3.5. E-LABS .....	14
3.6. DISGENET.....	15
3.7. ADVERSE REACTION SUBSTANTIATION WORKFLOW .....	16
3.8. SPOTFIRE.....	16
3.9. D3.JS AND NVD3 .....	17
3.10. PIPELINE PILOT .....	18
<b>4. SOFTWARE TO DEVELOP .....</b>	<b>19</b>
4.1. SAMPLEDAS .....	19
4.2. TREATMENT PATHWAYS .....	20
<b>ANNEXES.....</b>	<b>21</b>
ANNEX I. ANNEX_D13.1 – TABLE OF TOOLS .....	22

 IMI - 115372	<b>D13.1</b> - Catalogue of technologies and tools available for data analysis and visualization		
	<b>WP13.</b> Analysis, processing & visualization methods and tools		<b>Version:</b> v1.9 - Final
	<b>Authors:</b> Natalja Kurbatova, Rudi Verbeek		<b>Security:</b> PU

## DOCUMENT INFORMATION

<b>Grant Agreement Number</b>	115372	<b>Acronym</b>	EMIF
<b>Full title</b>	European Medical Information Framework		
<b>Project URL</b>	<a href="http://www.emif.eu">http://www.emif.eu</a>		
<b>IMI Project officer</b>	Ann Martin ( <a href="mailto:Ann.Martin@imi.europa.eu">Ann.Martin@imi.europa.eu</a> )		


<b>Deliverable</b>	<b>Number</b>	13.1	<b>Title</b>	Evaluation of technologies and tools available for data analysis and visualisation
<b>Work package</b>	<b>Number</b>	13	<b>Title</b>	Analysis, processing & visualization methods and tools

<b>Delivery date</b>	<b>Contractual</b>	Month 18	<b>Actual</b>	20/10/2014
<b>Status</b>	Current version / V1.9		Draft <input type="checkbox"/> Final <input checked="" type="checkbox"/>	
<b>Nature</b>	Report <input checked="" type="checkbox"/> Prototype <input type="checkbox"/> Other <input type="checkbox"/>			
<b>Dissemination Level</b>	Public <input checked="" type="checkbox"/> Restricted <input type="checkbox"/> Confidential <input type="checkbox"/>			

<b>Authors (Partner)</b>	Natalja Kurbatova (EMBL), Rudi Verbeek (JANSSEN)			
<b>Responsible Author</b>	Natalja Kurbatova		<b>Email</b>	<a href="mailto:natalja@ebi.ac.uk">natalja@ebi.ac.uk</a>
	<b>Partner</b>	EMBL	<b>Phone</b>	+44 (0) 1223 492 597


## DOCUMENT HISTORY

NAME	DATE	VERSION	DESCRIPTION
Natalja Kurbatova	13/05/14	1.0	First draft
Rudi Verbeek	17/06/14	1.2	Review and changes
Natalja Kurbatova	24/06/14	1.3	Review and changes
Graham Foster	07/07/14	1.4	Review and comments
Platform ExCom	22/07/14	1.4	Review of scope and comments
Natalja Kurbatova	06/08/14	1.5	Review and changes
Rudi Verbeek	11/08/14	1.6	Review and changes
Peter Rijnbeek	15/08/14	1.7	Additions to the Jerboa section
Graham Foster, Anthony Brookes, José Luis Oliveira, Natalja Kurbatova, Rudi Verbeek, Ugis Sarkans	23/09/14	1.8	Internal review and changes
Jennifer Waiting, Mark Gordon, Paul Avillach, Natalja Kurbatova	14/10/14	1.9	Consortium review and production of final version

 IMI - 115372	<b>D13.1</b> - Catalogue of technologies and tools available for data analysis and visualization		
	<b>WP13.</b> Analysis, processing & visualization methods and tools		<b>Version:</b> v1.9 - Final
	<b>Authors:</b> Natalja Kurbatova, Rudi Verbeek		<b>Security:</b> PU

## DEFINITIONS

- **Aggregated data.** It is the aggregation of several data, of a specific population or study.
- **Analysis (Broad definition).** Any “analytical method” used to get insights from data, based on descriptive or predictive statistics, modelling, simulation, graphs and other visualisation methods.
- **API.** Stands for application programming interface. In computing, API specifies how some software components should interact with each other.
- **Framework (Broad definition).** It is a real or conceptual structure intended to serve as a support or guide for the building of something that expands the structure into something useful.
- **Helpdesk.** It is a resource providing the customer or end user with information and support related to a company’s or institution’s products and services.
- **Home Page.** It is the main page of a website and the screen from which all other screens on the website can be linked.
- **HTTP.** It stands for Hypertext Transfer Protocol – a protocol for distributed, collaborative, hypermedia information systems. HTTP is the foundation of data communication for the World Wide Web (Web).
- **Interface (In computing).** It is a device or program enabling a user to communicate with a computer.
- **Ontology (In computing).** Ontology defines a set of representational primitives with which to model a domain of knowledge or discourse.
- **Pipeline (In the scope of this document).** Pipeline is a chain of data analysis components. Terms “pipeline” and “workflow” are interchangeable.
- **Platform (In the scope of this document).** The meaning of the term “platform” is very similar to the term “framework” – any base of technologies on which other technologies or processes are built. Platform in most of the cases has tools for developers and may provide computational power.
- **REST API.** REST stands for representational state transfer protocol – the way to create, read, update or delete information on a server using simple HTTP calls. A REST call is simply an HTTP request to the server. REST API means that software supports REST calls and allows communication through REST calls.
- **Web UI.** It stands for Web User Interface and refers to the interaction between a user and software on a Web server. The user interface in such a case is the Web browser and the Web page it has downloaded and rendered.
- **Workflow (In the scope of this document).** A series of computational steps usually programmed to run at once. Terms “pipeline” and “workflow” are interchangeable.
- **Workflow management system (In the scope of this document).** Software system that provides infrastructure to setup, execute and monitor scientific workflows.

 IMI - 115372	<b>D13.1</b> - Catalogue of technologies and tools available for data analysis and visualization		
	<b>WP13.</b> Analysis, processing & visualization methods and tools	<b>Version:</b> v1.9 - Final	
	<b>Authors:</b> Natalja Kurbatova, Rudi Verbeek	<b>Security:</b> PU	5/22

## EXECUTIVE SUMMARY


In EMIF, data analysis and visualization are the components of the platform used by the vertical tracks to generate their results. They form the layer between the platform databases and the user who wants to access and analyse these data. As such, the required functionality will be highly dependent on the types of analysis or scientific questions proposed by these projects.

Deliverable D13.1 is based on a software collection created by EMBL-EBI for the description of tools and frameworks that EMIF partners develop, use or would like to use for data analysis and visualization. The collection is organized as a Google spreadsheet that is available online to the EMIF partners. It is a “living” document that allows for the addition of tools and that will evolve during the project. The intention and scope of the deliverable D13.1 is to provide an overview of the current state of the collection.

The software collection consists of tools and services that are divided into the following conceptual groups:

- software already accepted and used by the EMIF Platform,
- software developed by EMIF partners or by external organisations and potentially useful for the consortium according to the EMIF partners’ current thinking,
- software that is not developed yet but is needed by EMIF partners for analysis and visualization tasks. In other words this section is a ‘gap’ analysis between the two preceding sections: it specifies areas or uses for which it is believed that no software exists currently.



 IMI - 115372	<b>D13.1</b> - Evaluation of technologies and tools available for data analysis and visualization		
	<b>WP13.</b> Analysis, processing & visualization methods and tools		<b>Version:</b> v1.9 - Final
	<b>Authors:</b> Natalja Kurbatova, Rudi Verbeek		<b>Security:</b> PU

## 1. INTRODUCTION

Deliverable D13.1 is the document describing the collection of available technologies used, developed or needed by EMIF partners for EHR and omics data analysis and visualization.

Each software system has been evaluated according to the following criteria: necessity level (3 – high, 2 – medium, 1 - low), commercial or open source license (2 – open source, 1 – commercial or other), integration potential with software systems that are already accepted by EMIF Platform, namely EMIF Catalogue, Jerboa reloaded and tranSMART (3 – high, 2 – medium, 1 - low). The sum of scores represents our evaluation: the higher score corresponds to the better evaluation of the software system. The maximum value for the overall score is 8, the minimum value is 3. Please note that evaluation criteria can be changed and/or new criteria can be added at further stages of the EMIF project.

The collection of tools is not exhaustive. As the EMIF project evolves new scientific questions may surface that require functionality not covered by the current selection. In addition, tools will need to be integrated with the EMIF architecture that is being developed by WP14.


In order to collect data from EMIF partners EMBL-EBI has adopted the BioMedBridges ([www.biomedbridges.eu](http://www.biomedbridges.eu)) approach to the creation of a software registry (<http://bioregistry.cbs.dtu.dk/>). More specifically, we have used the BioMedBridges template for the description of software. To label the tools in a consistent way, we selected the EDAM ontology (<http://edamontology.org/page>) to control values for drop down menus. We created a data collection form that was successfully used by partners and can be found at: <http://goo.gl/1uyXoo>. The responses are stored in a Google spreadsheet that represents the EMIF software collection (<http://goo.gl/DkdEgm>).

The intention and scope of deliverable D13.1 is to provide an overview of the current state of the software collection. This document consists of technologies, frameworks, tools, platforms and services that are divided into the following conceptual groups:

- software already accepted and used by the EMIF Platform,
- software developed by EMIF partners or by external organisations and potentially useful for the consortium according to the EMIF partners ideas,
- software that is not developed yet but is needed by EMIF partners for analysis and visualization tasks.

The first group represents the core EMIF Platform's components already supported and accepted by the EMIF Platform: EMIF Catalogue, tranSMART and Jerboa Reloaded software.

The second group consists of the tools, platforms and services developed by EMIF partners or that can be leveraged from other projects. We would like to emphasize that this group represents software that potentially can be useful for data analysis and visualization. We evaluate the available tools, platforms and services, but decisions about EMIF Platform's support will be made later by all the partners according to their expertise, the fit to the EMIF architecture and the needs that are derived from the scientific questions formulated by the verticals.

 IMI - 115372	<b>D13.1</b> - Evaluation of technologies and tools available for data analysis and visualization		
	<b>WP13.</b> Analysis, processing & visualization methods and tools	<b>Version:</b> v1.9 - Final	
	<b>Authors:</b> Natalja Kurbatova, Rudi Verbeek	<b>Security:</b> PU	8/22

The third group represents the initial specification of tools and services that are potentially needed for the analysis and visualization tasks. These tools do not exist yet but might be developed within EMIF Platform to help to solve interesting clinical research or data visualization problems.

## 2. SOFTWARE ACCEPTED AND USED BY EMIF PLATFORM

### 2.1. tranSMART

**Home Page:** <http://transmartfoundation.org>

**Type:** Framework.

**Description:** The open source tranSMART platform provides researchers with a single self-service web portal with access to phenotypic, 'omics, and unstructured text-based data from multiple sources, combined with search and analysis capabilities. All of the data integration is hypothesis free. The tranSMART platform helps scientists develop and refine research hypotheses by investigating correlations within genomic and phenotypic data. The data model behind tranSMART platform is that of the i2b2 clinical data warehouse.

**Functions:** Data handling, genomics, ontologies, nomenclature and classification, literature and reference, transcriptomics.

**Publications:** PMID 20376001, PMID 20642836, PMID 24303286, PMID 24608524

**License type:** Open source.

**Input types:** Core data, identifier, parameter, plotting and rendering, report, search and retrieval.

**Input formats:** Textual format.

**Output types:** Core data, identifier, parameter, plotting and rendering, report, search and retrieval.

**Contact name:** Rudi Verbeek, Paul Avillach.

**Contact:** [rverbeec@its.inj.com](mailto:rverbeec@its.inj.com), [paul\\_avillach@hms.harvard.edu](mailto:paul_avillach@hms.harvard.edu)

**Help desk:** <https://groups.google.com/forum/?fromgroups#!forum/transmart-discuss>

**Interfaces:** Web UI, REST API, Other.


**Why and how tool can be useful for EMIF platform:** tranSMART is a data integration and analysis platform for genotypic and phenotypic data. It is used in EMIF to bring together patient level data for the vertical tracks. End users from the verticals have direct access to the data and can perform advanced analyses and visualization for biomarker discovery.

Functions include (in addition to those listed below): Anovar Annotations of WES, Bio repository information.

**Integration potential:**

- 1) Groovy API is under development: <https://github.com/thehyve/transmart-core-api>;
- 2) Any query tool linked to an Oracle or Postgress database;



 IMI - 115372	<b>D13.1</b> - Evaluation of technologies and tools available for data analysis and visualization		
	<b>WP13.</b> Analysis, processing & visualization methods and tools	<b>Version:</b> v1.9 - Final	
	<b>Authors:</b> Natalja Kurbatova, Rudi Verbeek	<b>Security:</b> PU	9/22

3) R tranSMART interface:

<https://github.com/transmart/RInterface/blob/master/R/transmart.getClinicalData.r>

The tranSMART source code is available from [transmartfoundation.org](http://transmartfoundation.org). Interactions with or between developers use the [transmart-discuss](#) Google group or the tranSMART Community LinkedIn group. A first version of the tranSMART API is available, but further developments are ongoing. A plugin is available to integrate with R scripts.

We would score its integration potential as high.

**Evaluation:** necessity level – 3, license – 2, integration potential – 3, overall score - 8.

## 2.2. EMIF Catalogue

**HomePage:** <http://bioinformatics.ua.pt/emif-dev/>

**Type:** Service.

**Description:** Fingerprint data collected from EHR databases. Code written in python.

**Functions:** Data handling

**Publications:** NA.

**License type:** Other (Currently it is "Undefined yet". The authors have no objection to deliver it as open source. However, this decision needs to be discussed in EMIF project).

**Input types:** Report.

**Input formats:** HTML.

**Output types:** Report.

**Contact name:** José Luis Oliveira.

**Contact:** [jlo@ua.pt](mailto:jlo@ua.pt)

**Help desk:** NA.

**Interfaces:** Web UI.

**Why and how tool can be useful for EMIF platform:** It's a core software resource of the EMIF Platform. For more details see D14.3.


**Integration potential:** EMIF Catalogue is already prepared to integrate with Jerboa and tranSMART output formats. Web services are also available for data update. The programmatic access for data download is under discussion with EMIF partners. We would score its integration potential as high.

**Evaluation:** necessity level – 3, license – 1, integration potential – 3, overall score - 7.

## 2.3. Jerboa Reloaded

**HomePage:** None. <http://vaesco.net/vaesco/results/Tools.html>

**Type:** Tool.

 IMI - 115372	<b>D13.1</b> - Evaluation of technologies and tools available for data analysis and visualization		
	<b>WP13.</b> Analysis, processing & visualization methods and tools		<b>Version:</b> v1.9 - Final
	<b>Authors:</b> Natalja Kurbatova, Rudi Verbeek		<b>Security:</b> PU

**Description:** The initial version of Jerboa extraction software was developed within the EU-ADR project (FP7-ICT-2007-215847) and has since then been used successfully in many projects.

In the EMIF project a new version of Jerboa (Jerboa Reloaded) is being developed to address feature requests and suggestions of the early adopters. Additionally, custom-built modules will be developed tailored to research questions that will arise in the EMIF project.

The Jerboa software is used in a so-called distributed network design, i.e. it runs de-identification, harmonisation and aggregation locally at each data source site. Jerboa runs a script that contains all parameters of a specific study design. This has the advantage that local data processing is performed in a common way and not subject to small differences in implementation by local statisticians. Jerboa aggregates data to comply with the privacy regulations that are imposed on the member states by the European and local legislation (see also D10.2 Federation Procedures). This ensures that no identifiable patient data is present in the output generated by Jerboa, e.g. no dates, patient identifiers, etc.

To allow for easy adaptation and customization Jerboa employs a modular design approach. The JAVA Jerboa Core contains classes for loading, transforming and extracting data (ETL), logging, script handling, and several utilities like graphical tools. The Jerboa Modules are responsible for the specific study designs, e.g. primary data extraction (PDE) or filter operations, e.g., population definition or cohort definition.

**Functions:** Data handling.

**Publications:** PMID 21182150, D12.1 Data extraction software v1.

**License type:** Other (Comment: There is an intention to make it open source. The decision should be made how to support the open source version and how to define the open source license. This will be decided during the EMIF project).

**Input types:** Core data.

**Input formats:** Textual format.

**Output types:** Plotting and rendering, Report.

**Contact name:** Peter Rijnbeek.


**Contact:** [p.rijnbeek@erasmusmc.nl](mailto:p.rijnbeek@erasmusmc.nl)

**Help desk:** [rre@erasmusmc.nl](mailto:rre@erasmusmc.nl)

**Interfaces:** Desktop GUI.

**Why and how tool can be useful for EMIF platform:** Jerboa Reloaded facilitates the federation of EHR data sources. It provides graphical feedback and intermediate results to allow the data custodian to check and approve the generated results before sharing. Its flexible and modular design allows for the addition of more study specific modules based on requests of the current vertical projects.

**Integration potential:** Jerboa runs on study-specific input files created by the data custodian. An SQL module is being developed to connect to a database, for example, an instance of the OMOP Common Data Module. The output of Jerboa could be integrated with

 IMI - 115372	<b>D13.1</b> - Evaluation of technologies and tools available for data analysis and visualization		
	<b>WP13.</b> Analysis, processing & visualization methods and tools		<b>Version:</b> v1.9 - Final
	<b>Authors:</b> Natalja Kurbatova, Rudi Verbeeck		<b>Security:</b> PU

the EMIF Catalogue. For example, the output of the primary data extraction module can already be uploaded or pushed to the EMIF Catalogue. The output of Jerboa can be shared and post-processed in a remote research environment. Command line interface is under development. We would score its integration potential as high.

**Evaluation:** necessity level – 3, license – 1, integration potential – 3, overall score - 7.

### 3. SOFTWARE POTENTIALLY USEFUL FOR THE EMIF PARTNERS

#### 3.1. Genestack

**HomePage:** [www.genestack.com](http://www.genestack.com)

**Type:** Framework.

**Description:** Genestack is a universal collaborative platform for bioinformatics research and development. It allows users to store and share large data sets securely within and across organizations, with free access to public data from major databases. The platform includes open-source and proprietary genomics applications, working together independent of file formats. For developers an SDK, APIs and a marketplace are provided.

More details: There are slides from last year's Bio-IT World 2013:

<http://www.slideshare.net/genestack> and a long blog post

<http://www.genestack.com/blog/2013/03/18/platform-alpha/> with an explanation of key concepts.

**Functions:** Biological data resources, Data handling, Transcriptomics, Workflow management system.

**Publications:** NA.

**License type:** Commercial.

**Input types:** Core data.

**Input formats:** NA.

**Output types:** Core data.

**Contact name:** Misha Kapushesky.

**Contact:** [misha@genestack.com](mailto:misha@genestack.com)


**Help desk:** [info@genestack.com](mailto:info@genestack.com)

**Interfaces:** Web UI, API.

**Why and how tool can be useful for EMIF platform:** An environment is needed to support the creation of new analysis pipelines by experts, documentation of the pipelines and analytical tools integration - especially for new technologies and methods.

**Integration potential:** It has an open code framework and a large community, and so we would score its integration potential as high.

**Evaluation:** necessity level – 3, license – 1, integration potential – 3, overall score - 7.

 IMI - 115372	<b>D13.1</b> - Evaluation of technologies and tools available for data analysis and visualization		
	<b>WP13.</b> Analysis, processing & visualization methods and tools	<b>Version:</b> v1.9 - Final	
	<b>Authors:</b> Natalja Kurbatova, Rudi Verbeeck	<b>Security:</b> PU	12/22

## 3.2. Galaxy

**HomePage:** [www.galaxyproject.org](http://www.galaxyproject.org)

**Type:** Framework.

**Description:** Galaxy is an open, web-based platform for accessible, reproducible, and transparent computational biomedical research. In addition to using the public Galaxy server, one can also install a local instance of Galaxy, or create an instance of Galaxy on the cloud using CloudMan. Another option would be to use one of the increasing number of Public Galaxy Servers hosted by other organizations. The Galaxy Framework at the highest level is a set of reusable software components. Galaxy is a general bioinformatics workflow management system that helps to integrate data, tools and scripts for analysis and publishing. Its aim is to make computational biology accessible to research scientists that do not have computer programming experience.

**Functions:** Biological data resources, Data handling, Workflow management system.

**Publications:** PMID 20738864, PMID 20069535.

**License type:** Open source.

**Input types:** Core data.

**Input formats:** NA.

**Output types:** Core data.

**Contact name:** NA.

**Contact:** NA.

**Help desk:** <http://wiki.galaxyproject.org/Support>

**Interfaces:** Web UI, REST API.

**Why and how tool can be useful for EMIF platform:** Workflow management system is needed to assure the creation of new analysis pipelines by experts, documentation of the pipelines and analytical tools integration - especially for new technologies and methods.

**Integration potential:** It has an open code framework and a large community, and so we would score its integration potential as high.

**Evaluation:** necessity level – 3, license – 2, integration potential – 3, overall score - 8.


## 3.3. R Cloud Workbench

**HomePage:** [www.ebi.ac.uk/Tools/rcloud/](http://www.ebi.ac.uk/Tools/rcloud/)

**Type:** Framework.

**Description:** The ArrayExpress R Cloud Workbench is an R visualization framework, scalable and distributed for exposure of R/Bioconductor packages to Java applications. It is also a general resource pooling framework suitable for dispatching compute-intensive tasks to the server farm infrastructure at the EBI or other institutions.

**Functions:** Biological data resources.

 IMI - 115372	<b>D13.1</b> - Evaluation of technologies and tools available for data analysis and visualization		
	<b>WP13.</b> Analysis, processing & visualization methods and tools		<b>Version:</b> v1.9 - Final
	<b>Authors:</b> Natalja Kurbatova, Rudi Verbeek		<b>Security:</b> PU

**Publications:** NA.

**License type:** Open source.

**Input types:** Core data.

**Input formats:** Textual format.

**Output types:** Core data.

**Contact name:** Andrew Tikhonov.

**Contact:** [andrew@ebi.ac.uk](mailto:andrew@ebi.ac.uk)

**Help desk:** NA.

**Interfaces:** Web UI, Java API.

**Why and how tool can be useful for EMIF platform:** R development platform can help to combine together biostatisticians' and bioinformaticians' efforts in analysis of clinical data in R. R scripts aggregation into R packages would help to ensure "good practise" examples and approaches are spread amongst the community.

**Integration potential:** It is an open code framework with a Java API that developers could use for new solutions and integration activities.

**Evaluation:** necessity level – 3, license – 2, integration potential – 3, overall score - 8.

### 3.4. Automatic Bio Curator (ABC)

**HomePage:** NA.

**Type:** Tool.

**Description:** ABC is a partially developed, standalone tool designed to allow users to validate their data based on a set of rules. The program will check the data against the rules chosen by the user and report which aspects of the data fail. Options for data conversion and automatic data fixing are currently being considered.

**Functions:** Data handling.

**Publications:** NA.

**License type:** Open source.

**Input types:** Core data.

**Input formats:** Binary format, Textual format.


**Output types:** Core data.

**Contact name:** Anthony Brookes.

**Contact:** [ajb97@le.ac.uk](mailto:ajb97@le.ac.uk)

**Help desk:** NA.

**Interfaces:** Desktop GUI, Command-line.

 IMI - 115372	<b>D13.1</b> - Evaluation of technologies and tools available for data analysis and visualization		
	<b>WP13.</b> Analysis, processing & visualization methods and tools	<b>Version:</b> v1.9 - Final	
	<b>Authors:</b> Natalja Kurbatova, Rudi Verbeek	<b>Security:</b> PU	14/22

**Why and how tool can be useful for EMIF platform:** The tool could be used to check data of different types and formats for consistency, errors and problems. The ABC tool is designed as a standalone GUI tool.

**Integration potential:** The tool is written in python and the rule engine used has an API that could be queried via different programs if needed. We would score its integration potential as high.

**Evaluation:** necessity level – 2, license – 2, integration potential – 3, overall score - 7.

### 3.5. E-Labs

**HomePage:** NA.

**Type:** Framework.

**Description:** Platform for Social Collaboration on the Analysis and Visualization of Healthcare Data:

- secure, web-based environment that brings together multiple sources of data, tools and expertise to produce information, insights and intelligence efficiently.
- collaboration on methods.
- encapsulation of the methods/data/pipeline of research that are shareable and reusable.

**Functions:** Data handling.

**Publications:** PMID: 22941986.

**License type:** Other (Comment: 'case-by-case license'. E-Labs are the output of university research mainly, so the license is held by the university, but if they were to be reused they would be licensed on a case-by-case basis depending on the use. In essence, if EMIF wanted to use them we would license the use of the technology to EMIF).

**Input types:** Core data.

**Input formats:** NA.

**Output types:** Search and retrieval.

**Contact name:** James Cunningham.

**Contact:** [j.a.cunningham@gmail.com](mailto:j.a.cunningham@gmail.com)


**Help desk:** NA.

**Interfaces:** Web UI.

**Why and how tool can be useful for EMIF platform:** EHR data based framework - mappings and nice data visualisation.

**Integration potential:** Additional development work will be required for the E-Labs integration with other software components. So called 'case-by-case license' requires discussions with authors, and so we would score its integration potential as medium.

**Evaluation:** necessity level – 1, license – 1, integration potential – 2, overall score - 4.

 IMI - 115372	<b>D13.1</b> - Evaluation of technologies and tools available for data analysis and visualization		
	<b>WP13.</b> Analysis, processing & visualization methods and tools		<b>Version:</b> v1.9 - Final
	<b>Authors:</b> Natalja Kurbatova, Rudi Verbeek		<b>Security:</b> PU

## 3.6. DisGeNET

**HomePage:** <http://ibi.imim.es/tools/disgenet/>

**Type:** Other.

**Description:** DisGeNET is a database that integrates gene-disease associations from several public data sources and the literature that supports research on the molecular mechanisms of human diseases. DisGeNET covers the full spectrum of human genetic diseases (Mendelian, complex, environmental) and considers different types of associations between genes and diseases, as described in the DisGeNET association ontology. DisGeNET database can be queried through a web interface, which supports Search and Browse functionalities, or through a Cytoscape plugin, which allows users to query and analyse a network representation of DisGeNET data. A Mysql version of the DisGeNET database is also available for download. Additionally, an RDF (Resource Description Framework) representation of DisGeNET database has been created that can be queried using a matching SPARQL endpoint.

**Functions:** Biological data resources, Data handling, Genetics, Literature and reference, Ontologies, nomenclature and classification.

**Publications:** DOI: 10.1371/journal.pone.0020284, DOI: 10.1093/bioinformatics/btq538.

**License type:** Open source [GNU GPL 3.0 licence].

**Input types:** Identifier, Search and retrieval.

**Input formats:** Textual format.

**Output types:** Identifier, Report, Search and retrieval.

**Contact name:** Laura I. Furlong.

**Contact:** [lfurlong@imim.es](mailto:lfurlong@imim.es), [jpinero@imim.es](mailto:jpinero@imim.es)

**Help desk:** NA.


**Interfaces:** Web UI, Command-line, Other.

**Why and how tool can be useful for EMIF platform:** DisGeNET can be used for exploration of the complex relationship between genotype and phenotype that underlies human diseases. Examples of questions that can be answered with DisGeNET are:

1. what are the genes associated with a particular disease?
2. what are the diseases associated with a particular gene or protein?
3. what are the genes and proteins that are shared by a group of unrelated diseases?

**Integration potential:** DisGeNET information is standardized and therefore is prepared to be integrated with other sources of information (e.g. disease codes and genomic or proteomic data). The database is available in a variety of formats, thus there is also the possibility to integrate it and query it in different ways. For example there is a Cytoscape plugin to query and analyse three network representations of DisGeNET data. We also have web services to query the data that have been integrated in other tools (e.g. EU-ADR ADR-Substantiation



 IMI - 115372	<b>D13.1</b> - Evaluation of technologies and tools available for data analysis and visualization		
	<b>WP13.</b> Analysis, processing & visualization methods and tools		<b>Version:</b> v1.9 - Final
	<b>Authors:</b> Natalja Kurbatova, Rudi Verbeek		<b>Security:</b> PU

workflow implemented in Taverna consuming the DisGeNET web service URL). We would score its integration potential as high.

**Evaluation:** necessity level – 2, license – 2, integration potential – 3, overall score - 7.

### 3.7. Adverse Reaction Substantiation workflow

**HomePage:** <http://ibi.imim.es/tools/adr-substantiation/>

**Type:** Framework.

**Description:** ADR-Substantiation is a computational framework to aid in the collection and exploration of evidence about the causal inference of ADRs detected by mining clinical records. This framework was implemented as publicly available tools integrating state-of-the-art bioinformatics methods for the analysis of drugs, targets, biological processes and clinical events.

**Functions:** Biological data resources, Data handling, Chemo-informatics, Literature and reference, Systems biology.

**Publications:** DOI: 10.1371/journal.pcbi.1002457.

**License type:** Open source.

**Input types:** Identifier.

**Input formats:** Textual format.

**Output types:** Identifier, Plotting and rendering, Report.

**Contact name:** Laura I. Furlong.

**Contact:** [lfurlong@imim.es](mailto:lfurlong@imim.es), [jpinero@imim.es](mailto:jpinero@imim.es)

**Help desk:** NA.

**Interfaces:** Other.

**Why and how tool can be useful for EMIF platform:** ADR-Substantiation can be used to explore the biological plausibility of adverse drug reactions identified from clinical health records.

**Integration potential:** A very high integration potential. It is an open source framework that could be connected easily to the output data from other web services.

**Evaluation:** necessity level – 2, license – 2, integration potential – 3, overall score - 7.

### 3.8. Spotfire

**HomePage:** <http://spotfire.tibco.com/>


**Type:** Framework.

**Description:**

Advertised as follows:

Visualize and interact with data. Instantly spot and act on insights. Monitor key operational metrics. Mashup all your data and explore freely. Share and collaborate with teammates.



 IMI - 115372	<b>D13.1</b> - Evaluation of technologies and tools available for data analysis and visualization		
	<b>WP13.</b> Analysis, processing & visualization methods and tools		<b>Version:</b> v1.9 - Final
	<b>Authors:</b> Natalja Kurbatova, Rudi Verbeeck		<b>Security:</b> PU

Predict future direction. Explore real-time and historical data side-by-side. Use advanced statistics to discover unexpected opportunities and risks. Analytics at your desk or on-the-go. On-premise or in the Cloud.

From users reviews:

Spotfire uses data structure in a flexible and robust way. Spotfire allows you to directly connect with database and lets the developers to develop powerful reports intuitively.

Spotfire has an ability to mash together multiple data sources creating real-time dashboards that are both visual pleasing and decision driving.

Spotfire is user friendly, and has great visualizations available: the filters, marking, and user interface.

**Functions:** Data handling, Chemo-informatics, Eco-informatics.

**Publications:** NA.

**License type:** Commercial.

**Input types:** Core data.

**Input formats:** Structured data from client or server side RDBMS connections; File based data from CSV; XLS; SAS data formats; proprietary binary formats; selected instrument data interfaces.

**Output types:** Core data, Plotting and rendering, Report.

**Contact name:** NA.

**Contact:** [spotfireconsulting@tibco.com](mailto:spotfireconsulting@tibco.com)

**Help desk:** <https://support.tibco.com>

**Interfaces:** Web UI, Desktop GUI, Other.

**Why and how tool can be useful for EMIF platform:** Mentioned by WP14 as a potential tool for Fingerprint statistics and Population Characteristics graph creation. Spotfire is also a powerful tool for guided analytics: a data analysis process that follows a guided path for decision making, where the user can interact with and interrogate that data along the way. In EMIF it could be used as a tool for data analysts to distribute standard analyses to the verticals and collaborate on scientific questions.


**Integration potential:** Spotfire can be used as a stand-alone tool, on top of the data in a private remote research environment. Spotfire has a fully documented C# and ASP.NET API and can be integrated with existing tools or portals. Spotfire also has a direct connector to R, Splunk, SAS or Matlab code that can be used for advanced statistical calculations. We would score its integration potential as high.

**Evaluation:** necessity level – 2, license – 1, integration potential – 3, overall score - 6.

### 3.9. D3.JS and NVD3

**HomePage:** <http://d3js.org/>; <http://nvd3.org/>; <https://github.com/mbostock/d3/releases>

**Type:** Tool.

 IMI - 115372	<b>D13.1</b> - Evaluation of technologies and tools available for data analysis and visualization		
	<b>WP13.</b> Analysis, processing & visualization methods and tools		<b>Version:</b> v1.9 - Final
	<b>Authors:</b> Natalja Kurbatova, Rudi Verbeeck		<b>Security:</b> PU

**Description:** D3.js is a JavaScript library for manipulating documents based on data. D3 helps you bring data to life using HTML, SVG and CSS. D3's emphasis on web standards gives you the full capabilities of modern browsers without tying yourself to a proprietary framework, combining powerful visualization components and a data-driven approach to DOM manipulation. The NVD3 project seeks to build re-usable charts and chart components for d3.js without taking away the power that d3.js gives you. This is a very young collection of components, with the goal of keeping these components very customizable, staying away from your standard cookie cutter solutions.

**Functions:** Data handling.

**Publications:** NA.

**License type:** Open source.

**Input types:** NA.

**Input formats:** NA.

**Output types:** NA.

**Contact name:** Mike Bostock.

**Contact:** [mike@ocks.org](mailto:mike@ocks.org)

**Help desk:** NA.

**Interfaces:** Web UI.

**Why and how tool can be useful for EMIF platform:** EMIF Catalogue needs to show graphs with fingerprint statistics, population characteristics graphs (Jerboa results).

**Integration potential:** A lot of development work is therefore needed, but D3.JS library and NVD3 look to be very useful for the needs of the EMIF Catalogue. We would score its integration potential as medium.

**Evaluation:** necessity level – 3, license – 2, integration potential – 2, overall score - 7.

### 3.10. Pipeline Pilot

**HomePage:** <http://accelrys.com/products/pipeline-pilot/>

**Type:** Framework.


**Description:** Built on the Accelrys Enterprise Platform, Pipeline Pilot enables scientists to rapidly create, test and publish scientific services that automate the process of accessing, analyzing and reporting scientific data, either for the scientist's personal use or for sharing across the scientific community.

**Functions:** Data handling. Data Sharing, Genomic, Bioinformatics, Chemistry awareness, Database integration, Workflow management system.

**Publications:** <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3414708/>

**License type:** Commercial.

**Input types:** Core data (contextual data, chemical structures).

 IMI - 115372	<b>D13.1</b> - Evaluation of technologies and tools available for data analysis and visualization		
	<b>WP13.</b> Analysis, processing & visualization methods and tools		<b>Version:</b> v1.9 - Final
	<b>Authors:</b> Natalja Kurbatova, Rudi Verbeek		<b>Security:</b> PU

**Input formats:** Structured data from client or server side RDBMS connections; File based data from CSV; XLS; SAS data formats; proprietary binary formats; selected instrument data interfaces.

**Output types:** Core data, Plotting and rendering, Report.

**Contact name:** Accelrys.

**Contact:** <http://accelrys.com/about/contact/>

**Help desk:** <http://accelrys.com/services/support.html>

**Interfaces:** Web UI, Desktop GUI.

**Why and how tool can be useful for EMIF platform:** Provides a commercial, supported, extensible, stable and generic platform for doing all forms of data pipeline analyses. It has separate components for industry specific solutions / libraries.

**Integration potential:** High – fully scriptable; programmable; API extensible and reusable with integration into other commercial and open source tools (such as R, SAS). We would score its integration potential as high.

**Evaluation:** necessity level – 3, license – 1, integration potential – 3, overall score - 7.

## 4. SOFTWARE TO DEVELOP

### 4.1. SampleDAS

**HomePage:** NA.

**Type:** Framework.

**Description:** The overall purpose of the proposed SampleDAS system is to provide a simple mechanism that would enable federation of biological and biomedical data on the biological sample dimension. A system similar to DAS (Distributed Annotation System) to enable distributed management, integration and visualization of data mapped to one or more related ontologies is envisaged.

**Functions:** Data handling, Ontologies, nomenclature and classification.

**Publications:** NA.

**License type:** Open source.

**Input types:** Core data, Search and retrieval.

**Input formats:** XML.


**Output types:** Plotting and rendering, Report.

**Contact name:** Ugis Sarkans.

**Contact:** [ugis@ebi.ac.uk](mailto:ugis@ebi.ac.uk)

**Help desk:** NA.

**Interfaces:** Web UI, REST API.

 IMI - 115372	<b>D13.1</b> - Evaluation of technologies and tools available for data analysis and visualization		
	<b>WP13.</b> Analysis, processing & visualization methods and tools	<b>Version:</b> v1.9 - Final	
	<b>Authors:</b> Natalja Kurbatova, Rudi Verbeeck	<b>Security:</b> PU	20/22

**Why and how tool can be useful for EMIF platform:** To enable integrated exploration of patient data from all EMIF data sources, mapped to a common set of controlled vocabularies/ontologies. Based on simple open protocols, so should be easy to create both information sources ("annotation servers") and information clients (visualization tools).

**Integration potential:** We would score its integration potential as high since the required integration components will be added during development process.

**Evaluation:** necessity level – 1, license – 2, integration potential – 3, overall score 6.

## 4.2. Treatment Pathways

**HomePage:** NA.

**Type:** Tool.

**Description:** The overall purpose of the proposed treatment pathways searching tool is to provide a simple algorithm that would help to find patterns in the EHR data related to treatment processes and to create pathways based on such kind of patterns.

**Functions:** Data handling.

**Publications:** NA.

**License type:** Open source.

**Input types:** Core data, Search and retrieval.

**Input formats:** NA.

**Output types:** Report.

**Contact name:** Nada Boudiaf

**Contact:** [nada.x.boudiaf@gsk.com](mailto:nada.x.boudiaf@gsk.com)


**Help desk:** NA.

**Interfaces:** Open source.

**Why and how tool can be useful for EMIF platform:** Treatment pathway searching is one of the possible analysis uses of EHR data. If this algorithm and tool were developed it might provide a customizable part of EMIF Platform for the partners.


**Integration potential:** We would score its integration potential as high since the required integration components will be added during development process.

**Evaluation:** necessity level – 1, license – 2, integration potential – 3, overall score – 6.

 IMI - 115372	<b>D13.1</b> - Evaluation of technologies and tools available for data analysis and visualization		
	<b>WP13.</b> Analysis, processing & visualization methods and tools	<b>Version:</b> v1.9 - Final	
	<b>Authors:</b> Natalja Kurbatova, Rudi Verbeeck	<b>Security:</b> PU	21/22

---

## ANNEXES

 IMI - 115372	<b>D13.1</b> - Evaluation of technologies and tools available for data analysis and visualization		
	<b>WP13.</b> Analysis, processing & visualization methods and tools	<b>Version:</b> v1.9 - Final	
	<b>Authors:</b> Natalja Kurbatova, Rudi Verbeeck	<b>Security:</b> PU	22/22

## Annex I. Annex\_D13.1 – table of tools

**EMIF\_D13.1\_Tools\_for\_data\_analysis\_and\_visualisation\_v1\_Annex.xlsx** 