Deliverable D4.2

Configurator service available for the EMBRIC community

Date: November 2016

HORIZON 2020 - INFRADEV
Implementation and operation of cross-cutting services and solutions for clusters of ESFRI

Grant agreement no.:	654008
Project acronym:	EMBRIC
Project website:	www.embric.eu
Project full title:	European Marine Biological Research Infrastructure cluster to promote the Bioeconomy
Project start date:	June 2015 (48 months)
Submission due date :	November 2016
Actual submission date:	November 2016
Work Package:	WP 4 – Data services and reporting standards
Lead Beneficiary:	EMBL-EBI
Version:	2.0
Authors:	COCHRANE Guy, RAJAN Jeena TEN HOOPEN Petra

Project funded by the European Union's Horizon 2020 research and innovation programme (2015-2019)		
Dissemination Level		
PU	Public	**X**
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services	

Abstract

The EMBRIC Configurator has been made available to the marine biotechnology community. The service provides an entry point into the complex landscape of existing informatics data resources, especially around molecular biology. Targeting those embarking on the design of new marine projects, clients of the service provide a scientific description into an online form and, following several rounds of structured discussion and review involving input from informatics and scientific experts, are provided with a project-specific 'configuration'. This configuration includes a description of the elements of infrastructure (such as databases, standards, formats, curation groups, analysis methods and cloud compute capacity), advice on accessing and setting these elements up for the project and data management guidelines.

Table of Contents

1 Introduction

1.1 The EMBRIC Configurator concept

The EMBRIC Configurator was conceived to assist marine biotechnology professionals in making the most effective use of existing informatics resources around biological, especially molecular, data. The service provides an entry point for marine scientists in the design stages of new projects and programmes into an extensive and powerful, but complex, landscape of resources. While we have established formal process and will increasingly lean upon registries and knowledgebases of informatics tools, services and other resources, the EMBRIC Configurator service is fundamentally provided by experts in informatics infrastructure with an understanding of the requirements of marine science. We describe the service, therefore, as one of 'structured consultancy'.

The core function of the Configurator is translation; in the most simple terms, the 'client' of the service provides a description in scientific terms of the planned work (such as the number of samples, the molecular methods to be used and endpoints for interpretation of data) and the service returns a description of the required informatics system that will serve the client's needs (such as data management systems, analysis methods, curation capacity and relevant data standards). We consider that for a given marine project, many, or at least most, of the required infrastructure elements will largely already exist. However, we also expect that the elements must be selected expertly for specific cases and 'configured' into a working infrastructure that serves the needs of the project.

The **EMBRIC Configurator**, therefore, provides specific **configurations** that serve to describe how **infrastructure elements** can be combined and utilised for the purposes of a given marine biotechnology project or programme.

1.2 Scope of deliverable

In this deliverable, we announce the first broadly available iteration of the EMBRIC Configurator service to the EMBRIC community, including its trans-national access partners. We invite scientists to take advantage of the service and work with us to improve its utility. This deliverable marks the start of work on the second iteration, in which we will steadily tune and otherwise improve the client-facing interface and increase automation, where possible, and make other improvements within internal workflows.

1.3 Organisation of this report

This deliverable report comprises two sections. In the first, 'Client Perspective', section, we describe core concepts and present the client workflow. In the second, 'Internal Processes', section, we detail the current internal workflow.

1.4 Definitions

Element: A component of informatics infrastructure including the machine-bound, such as databases, tools and data standards, and the human, such as curation capacity and data analysis expertise.

Configuration: A description of the set of elements organised to interoperate to serve as a specific 'instance' of infrastructure to support a marine science project or programme.

EMBRIC Configurator: The service to return configurations of elements of informatics for client-defined marine science projects and programmes.

Case officer: An individual charged with coordinating the start-to-finish requirements gathering, design and presentation of an infrastructure configuration.

Case consultant: A community expert, called from the pool of EMBRIC partners that contribute to the EMBRIC Configurator service (see pool of community experts document[1]), that assists in design and decision making in the drawing up of configurations.

[1] The pool of community experts comprises those directly involved in the EMBRIC Configurator (WP4 task 4.1) and those with appropriate expertise who volunteer to join this group; the document is available from https://3.basecamp.com/3372748/buckets/974183/google_documents/282474910.

2 The client perspective

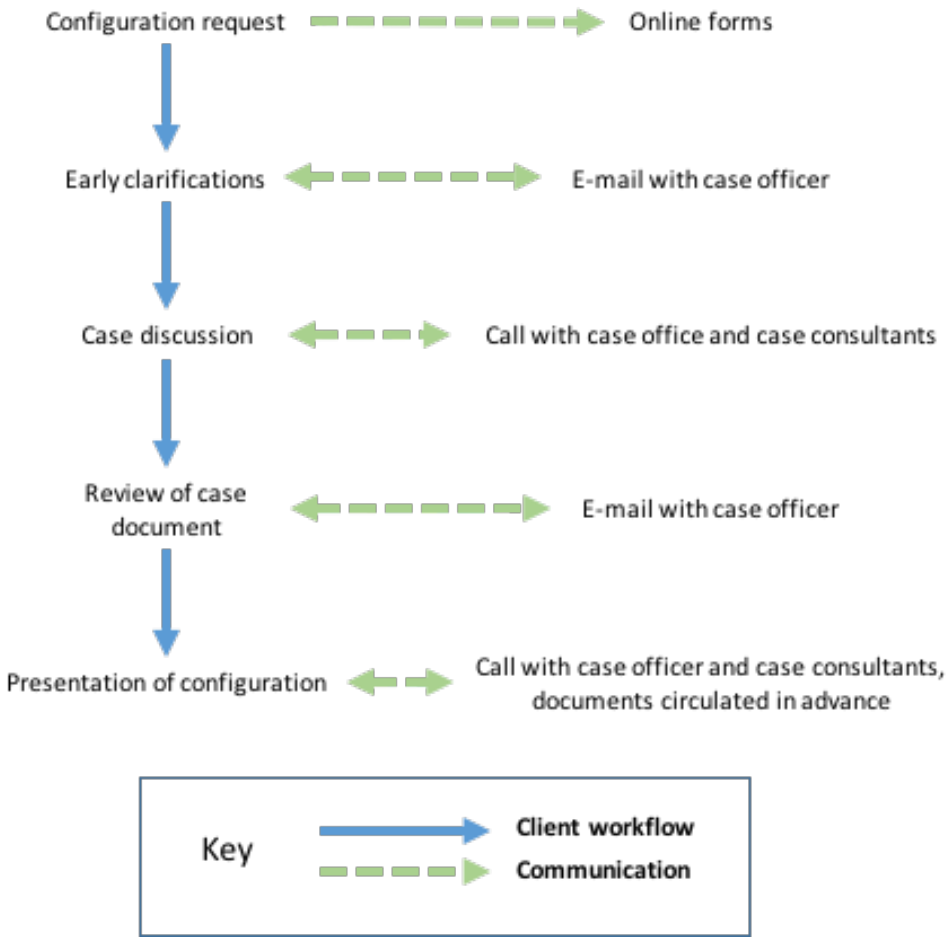


FIGURE I: CLIENT WORKFLOW SHOWING MAJOR STEPS AND POINTS OF COMMUNICATION WITH EMBRIC CONFIGURATOR INTERFACES AND STAFF

The overall client workflow is shown in Figure I. The first approach is taken by the client by providing outline details of the project or programme to be undertaken through the online form at http://tinyurl.com/h6b9led (see Figure II). While there will be a notification of receipt of the request and may be some simple questions for clarification, the intention at this stage is solely to communicate enough detail to portray the essence and scale of the study sufficient to allow the case officer to plan a more detailed information capture process. All e-mail communication will be handled through a central mail tracking system, using address embric-configurator@ebi.ac.uk. The case discussion provides the opportunity to communicate detailed information about the project or programme and takes the form of a video- or audio- conference, or in-person meeting. The case officer leads the client through a series of questions targeting key details to be captured. Based on the case discussion, a case document is prepared and returned to the client for review.

Further e-mail communications serve to refine the case document. Finally, a second call (video, audio or in-person) allows the case officer to present and explain the configuration to the client.

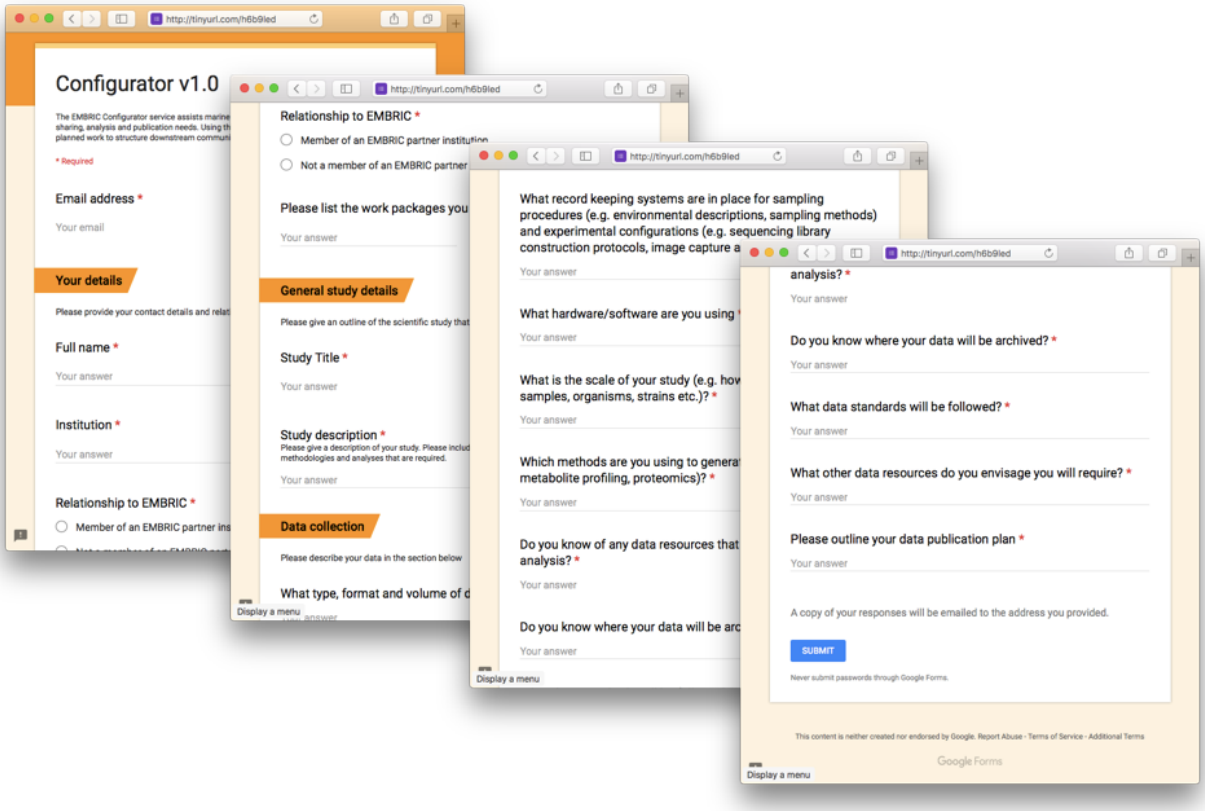


FIGURE II: SCREENSHOTS OF ONLINE FORM FOR INITIAL DESCRIPTION OF PROJECT OR PROGRAMME.

3 Internal processes

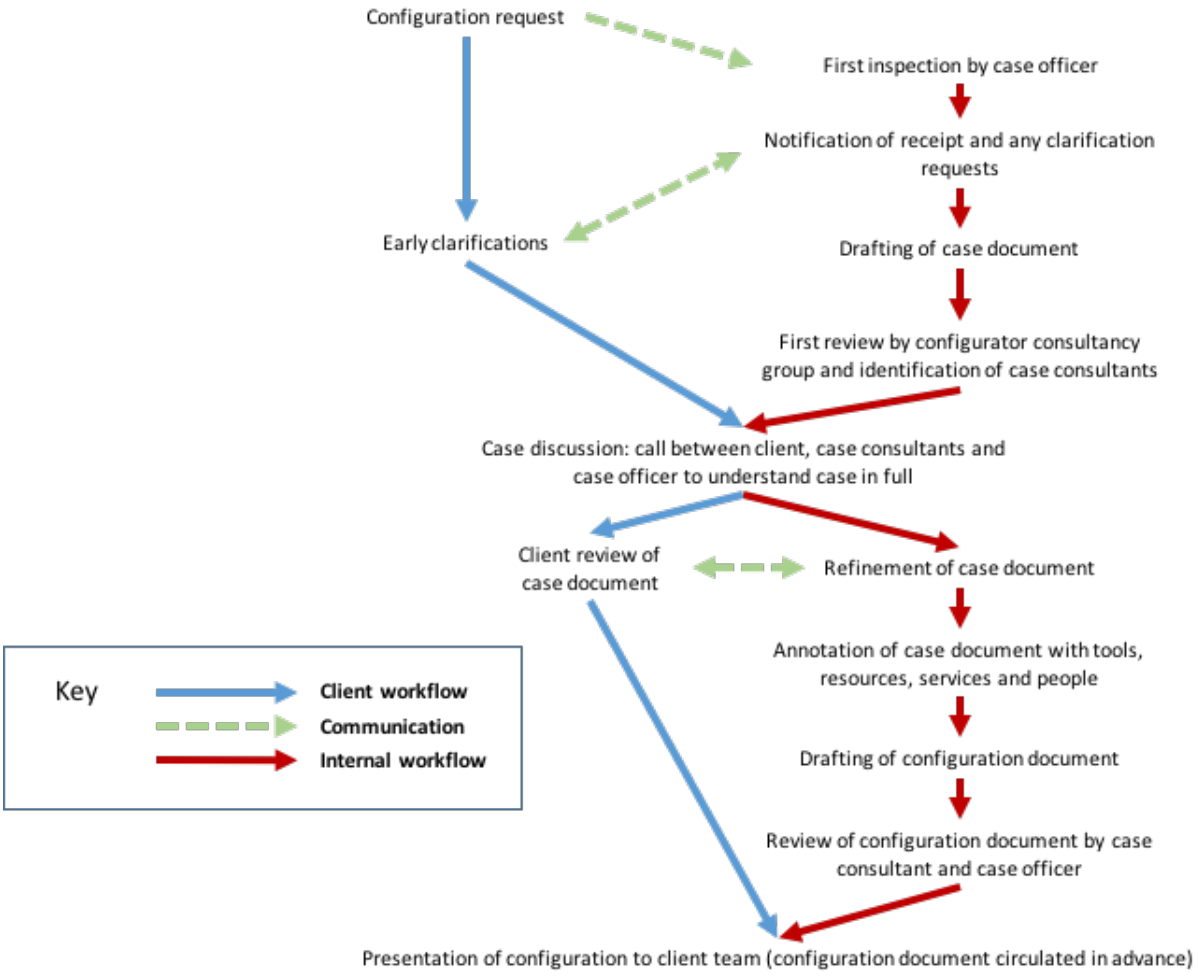
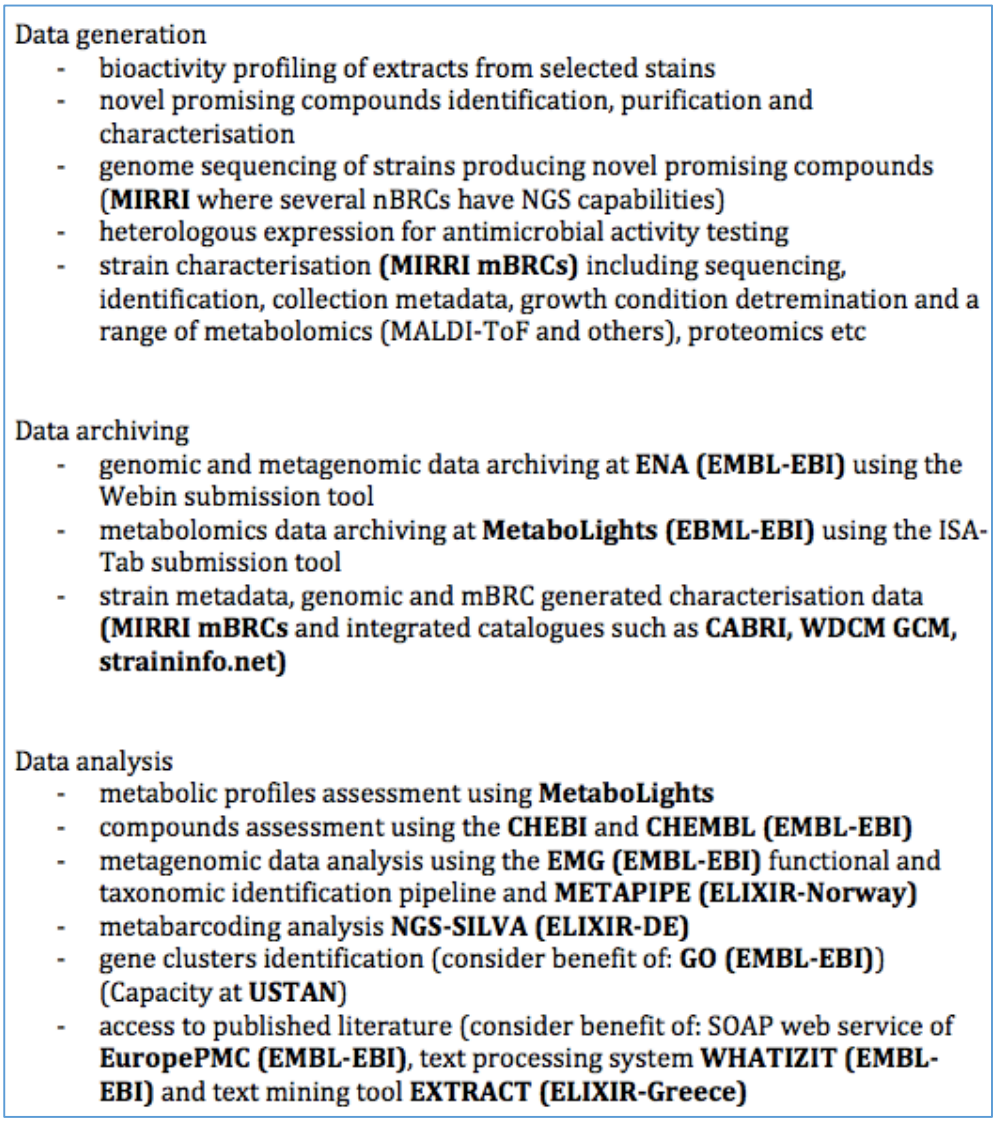


FIGURE III: INTERNAL WORKFLOW SHOWING MAJOR STEPS AND POINTS OF COMMUNICATION; THE CLIENT WORKFLOW (SEE ALSO FIGURE I) IS ALSO INDICATED.

Helpdesk staff at EMBL-EBI charged with EMBRIC responsibilities are first alerted to a request for a new configuration through an automatic notification from the online form into the mail tracking system. Corresponding through the central e-mail address, embric-configurator@ebi.ac.uk, a case officer is appointed and sends a notification of receipt including any immediate requests for clarification. In this e-mail, instructions on later stages of the process are also explained to set client expectations. The case officer prepares a case document with assistance of further team members at EMBL-EBI. Invitations are sent to appropriate community experts on the pool (see pool of community experts document*) to contribute to the case document and planning. The case officer sets up the case discussion call (video, audio or, if convenient, in-person) with themselves, the client and case consultants. Having refined understanding and captured more fully the details of the project, the case document is refined and,

ultimately, returned to the client for further review and e-mail communication. The case officer and case consultants then annotate the case document with infrastructure elements of relevance (see Figure IV). The configuration document is then drafted by the case officer and reviewed by case consultants. Finally, the case officer sets up a second call with the client and case consultant for the presentation of the final configuration document, having circulated documents in advance.



Data generation
- bioactivity profiling of extracts from selected stains
- novel promising compounds identification, purification and characterisation
- genome sequencing of strains producing novel promising compounds (**MIRRI** where several nBRCs have NGS capabilities)
- heterologous expression for antimicrobial activity testing
- strain characterisation **(MIRRI mBRCs)** including sequencing, identification, collection metadata, growth condition detremination and a range of metabolomics (MALDI-ToF and others), proteomics etc

Data archiving
- genomic and metagenomic data archiving at **ENA (EMBL-EBI)** using the Webin submission tool
- metabolomics data archiving at **MetaboLights (EBML-EBI)** using the ISA-Tab submission tool
- strain metadata, genomic and mBRC generated characterisation data **(MIRRI mBRCs** and integrated catalogues such as **CABRI, WDCM GCM, straininfo.net)**

Data analysis
- metabolic profiles assessment using **MetaboLights**
- compounds assessment using the **CHEBI** and **CHEMBL (EMBL-EBI)**
- metagenomic data analysis using the **EMG (EMBL-EBI)** functional and taxonomic identification pipeline and **METAPIPE (ELIXIR-Norway)**
- metabarcoding analysis **NGS-SILVA (ELIXIR-DE)**
- gene clusters identification (consider benefit of: **GO (EMBL-EBI)**) (Capacity at **USTAN**)
- access to published literature (consider benefit of: SOAP web service of **EuropePMC (EMBL-EBI)**, text processing system **WHATIZIT (EMBL-EBI)** and text mining tool **EXTRACT (ELIXIR-Greece)**

FIGURE IV: EXCERPT OF EXAMPLE ANNOTATED CASE DOCUMENT (FROM EMBRIC WP6 MICROBIAL WORKFLOWS)

In the first iteration of the EMBRIC Configurator, we have chosen not to predefine the format of the configuration document; rather, we expect that different projects will require different presentations. However, certain components will be provided in all cases, including a list of infrastructure elements, guidelines on accessing and/or setting up these elements and comments on data management requirements. With usage, we expect a number of common components to emerge (e.g. data flow diagrams, data resource descriptions, metadata checklists, etc.) and we will introduce more fixed structure around such components over time as appropriate.

4 Conclusion and future

This report describes the first implementation of the EMBRIC Configurator, a service now introduced into operation. Over time, the internal processes required to run the Configurator will be refined, especially with regard to knowledgebases and registries to make the process of defining elements of infrastructure more efficient. Further automation of the information gathering (currently a Google form) is not planned in the short-term, although not ruled out. It is expected that the Configurator will always require a human expert for its operation. Beyond the funded term for EMBRIC, resource for the case officer must be sought – appropriate cost models will be explored and reported - but by this time the level of resource required, given the accumulated knowledge and its organisation, will be reduced. It is expected that contributions from the case consultants will remaining minimal and supported through good will and mutual interest in the success of the projects under configuration.