
# Deliverable D4.3

## Release of pilot version of data warehouse

**Date: 10.05.17**

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

**Grant Agreement number:** **654008**

**Project acronym:** **EMBRIC**

**Contract start date:** **01/06/2015**

**Project website address:** **www.embric.eu**

**Due date of deliverable:** **30/05/17 / month 24**

**Dissemination level:** **Public**

## Document properties

| | |
|---|---|
| Partner responsible | EMBL-EBI |
| Author(s)/editor(s) | COCHRANE Guy, RAJAN Jeena and SILVESTER Nicole |
| Version | 1 |

## Abstract

The pilot version of the data warehouse has been developed to link genomic data with data in ELIXIR chemical resources. An SQLite database has been created with tables to link chemical data in the Chemical Entities of Biological Interest (ChEBI) database with taxonomy information, protein data in UniProt, chemical compound and target information in the ChEMBL database, as well as sequence and genome assembly information available in the European Nucleotide Archive (ENA), providing a powerful search function. This SQLite database is available to download from the ENA public FTP site and support is offered for its use.

# Table of Contents

**Table of Contents** ..... 4

**1 Introduction** ..... 5
- **1.1 The Pilot Data Warehouse** ..... 5
- **1.2 Databases** ..... 5

**2 Tables in SQLite database** ..... 6
- **2.1 Description of tables** ..... 6
- **2.2 Example Queries** ..... 7

**3 Limitations** ..... 9

**4 Conclusion and Future** ..... 9

# 1 Introduction

## 1.1 The Pilot Data Warehouse

The main objective of EMBRIC work package 4 is to provide sustainable data management services for the marine science community. One of these tasks is the development of a data warehouse to facilitate better utilisation of data resources across the whole EMBRIC cluster. A pilot version of the data warehouse has been developed to link genomic and chemical data from ELIXIR resources. Following an EMBRIC WP2.1 meeting in Plymouth with participants from WP2 and WP3 in December 2016 and a Chemical Biology workshop in Paris in February 2017 organised by WP4 it became clear there is a need to link data from specific resources. For this pilot we have fulfilled some of these requests, for example it is now possible to search for the source organism of a compound. We are limited on the data that can be mapped by the availability of data in the resources and how it is structured and linked. We have focused on connecting information from the ChEBI, UniPot, ChEMBL and ENA databases. The SQLite database, as well as the README file is available to download here: ftp://ftp.ebi.ac.uk/pub/databases/ena/collaboration/embricDB_v1.tar.gz

ENA already provides cross-references to several external resources both internal to EBI and external. A full list can be viewed here:
http://www.ebi.ac.uk/ena/data/xref/source

## 1.2 Databases

This section describes the databases that information has been extracted from to create tables to populate the SQLite database we have created.

**ChEBI** - **Ch**emical **E**ntities of **B**iological **I**nterest is a freely available dictionary of molecular entities focused on ‘small’ chemical compounds. The term ‘molecular entity’ refers to any constitutionally or isotopically distinct atom, molecule, ion, ion pair, radical, radical ion, complex, conformer, etc., identifiable as a separately distinguishable entity. The molecular entities in question are either products of nature or synthetic products used to intervene in the processes of living organisms.

**UniProt** - The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data.

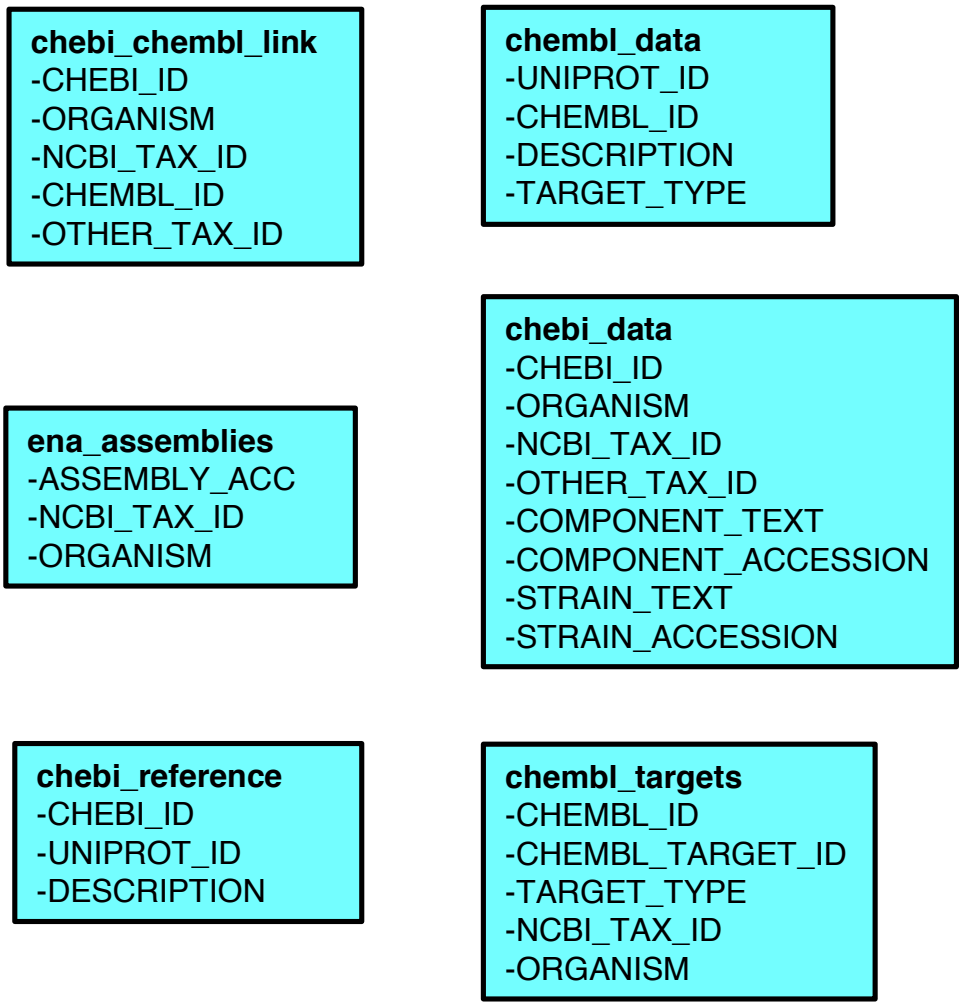
**ChEMBL** - ChEMBL is a database of bioactive drug-like small molecules

**ENA** - the European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation.

# 2 Tables in SQLite database

## 2.1 Description of tables

Below is the schema of the SQLite database. Table names are in bold and column names listed underneath.




The column headers in the tables contain the following information:

CHEBI_ID – ChEBI database identifier
ORGANISM – organism name at species or genus level
NCBI_TAX_ID – NCBI Taxonomy database identifier
CHEMBL_ID – ChEMBL database identifier – (can be compound or target)
OTHER_TAX_ID – non-NCBI taxonomy identifiers
UNIPROT_ID – UniProt database identifier
ASSEMBLY_ACC – ENA assembly accession number
COMPONENT_TEXT – part of the organism the extract originates from
COMPONENT_ACCESSION – accession number of component part
STRAIN_TEXT – strain name

STRAIN_ACCESSION – strain accession in culture collection
TARGET_TYPE – e.g. single protein, protein nucleic-acid complex, protein family
CHEMBL_TARGET_ID – internal CHEMBL target id number

In addition, there are two text files containing the following information:

1) UniProt/SwissProt IDs to INSDC sequence data (from ENA)
Columns: UNIPROT_ID SEQUENCE_ACC PROTEIN_AC

2) UniProt/TrEMBL IDs to INSDC sequence data (from ENA)
Columns: UNIPROT_ID SEQUENCE_ACC PROTEIN_ACC

It was not possible to include the UniProt to ENA mappings contained in the two text files above in the SQLite database due to the sheer volume of data which would result in the database being excessively large.

NCBI taxonomy is the unified taxonomic index used across most ELIXIR resources.

## 2.2 Example Queries

**1) Search for ChEBI ID, CHEMBL ID, NCBI Tax ID and ENA assembly accession for organism "Nocardiopsis dassonvillei"**

Search chebi_chembl_link and ena_assemblies tables:

```
sqlite> select CHEBI_ID,CHEMBL_ID, ASSEMBLY_ACC from chebi_chembl_link
inner join ena_assemblies on ena_assemblies.ORGANISM =
chebi_chembl_link.ORGANISM where chebi_chembl_link.ORGANISM =
'Nocardiopsis dassonvillei';
```

**Returns:**
ChEBI IDs - > CHEBI:69705, CHEBI:69706, CHEBI:69707, CHEBI:69708, CHEBI:69709, CHEBI:69710, CHEBI:69711

NCBI Tax ID -> 2014

ENA assembly accession: GCA_001877055

**2) Retrieve the CHEBI_IDs, CHEMBL_IDs and ENA genome accession IDs for all organisms that occur in both the chebi_chembl_link and ena_accessions tables**

select CHEBI_ID,CHEMBL_ID, ASSEMBLY_ACC from chebi_chembl_link inner join ena_assemblies on ena_assemblies.ORGANISM = chebi_chembl_link.ORGANISM;

**3) Search using CHEBI:82642**

Search chebi_chembl_link table

sqlite> select * from chebi_chembl_link where CHEBI_ID = 'CHEBI:82642';

**Returns:**
Organism -> Neopetrosia exigua
NCBI Tax_ID -> 595373
CHEMBL ID -> CHEMBL371814

**5) Retrieve all targets in ChEMBL that are SINGLE PROTEIN**

select * from chembl_targets where TARGET_TYPE = 'SINGLE PROTEIN';

# 3 Limitations

The current limitations of the pilot data warehouse in connecting data from different resources are due to the availability of the data in the resources and the mapping and structure of this data. For example, within the ChEBI database there are only approximately 7,000 entries with source species information. Also, strain information is not readily available in ChEBI, as either it has not been captured or has not been added, which means it is not currently possible to map data via strain information to the culture collections. There is more species information available than is currently publically accessible and we will liaise with ChEBI to ensure the addition of their species data and strain information where available.

# 4 Conclusion and Future

This report describes the implementation of the pilot Data Warehouse, an SQLite database connecting genomic data with data in ELIXIR chemical resources as described above. The database is available to download from the ENA public ftp site, along with a README file describing the tables and their contents. Queries can be run, connecting small molecules with taxonomy information, compounds, protein, sequence and assembly data.

The choice of SQLite for the database distribution was made on the basis that, while comparatively raw, direct access to a database such as this retains full potential for all kinds of query, while a more intuitive interface would limit potential uses. The interface offered at the moment is not intended for end users, rather these are provided for those involved in moving from prototype to final tool to allow us to understand what kinds of additional content and usage need to be supported. This is simply a prototype and the WP2 deliverable (including integration of the programmatic service interface we provide) is not due until towards the end of the project. The content of the warehouse will be imported into the Marine Metagenomics Portal at University of Tromsø within a couple of months offering an online service. Recognising that not all partners will have sufficient technical skills or support, we offer direct support for project participants in formulating queries against the database; please contact us at datasubs@ebi.ac.uk.
The final version of the Data Warehouse is due towards the end of the EMBRIC project in month 45. For that version we plan to increase the mappings within the current databases that are linked and provide a searchable and more intuitive interface. Design for this eventual interface will be informed by those who use the pilot Data Warehouse and the queries that they construct in this first phase.

We also intend to expand the number of databases that we can map to, including such databases as OBIS (Ocean Biogeographic Information System). We will further explore the use of strain data to link to culture collections and to the literature to facilitate data integration across small molecules, culture collections and genomics databases.