

Dealing with historical data and making it accessible



Strobbe F, Lagring R, Scory S, Stojanov Y, Vandenberghe T

Royal Belgian Institute of Natural Sciences, Operational Directorate Natural Environment, Belgian Marine Data Centre (BMDC), Brussels, Belgium (bmdc@naturalsciences.be)

Introduction

Long-term datasets are important to study global trends in the environment and to define a reference status of an ecosystem. Results of analyses on such datasets allow policy makers to define proper strategies to reach regulatory objectives. Access to these data is also a crucial issue for the scientific community.

Belgium has a long oceanographic data collection history. In that context, the Belgian Marine Data Center (BMDC), with its continuously growing experience in data rescuing and management, worked out a tool to identify the marine datasets and retrieve as much data as possible.

The **Data Inventory and Tracking System (DITS)** (dits.bmdc.be) was designed by BMDC in the frame of the project '4 Decades of Belgian Marine Monitoring: uplifting historical data to today's needs' (www.4DEMON.be).

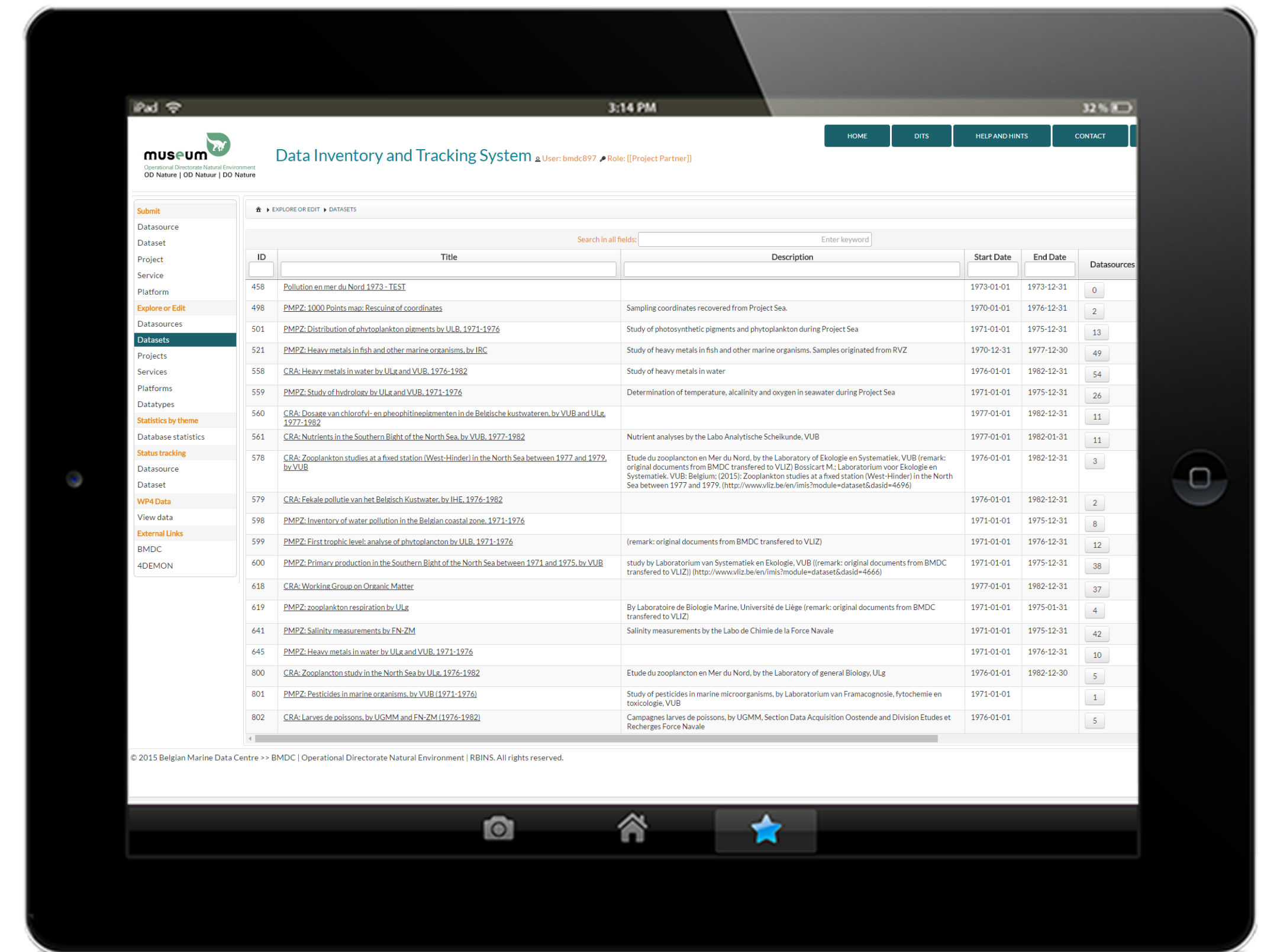
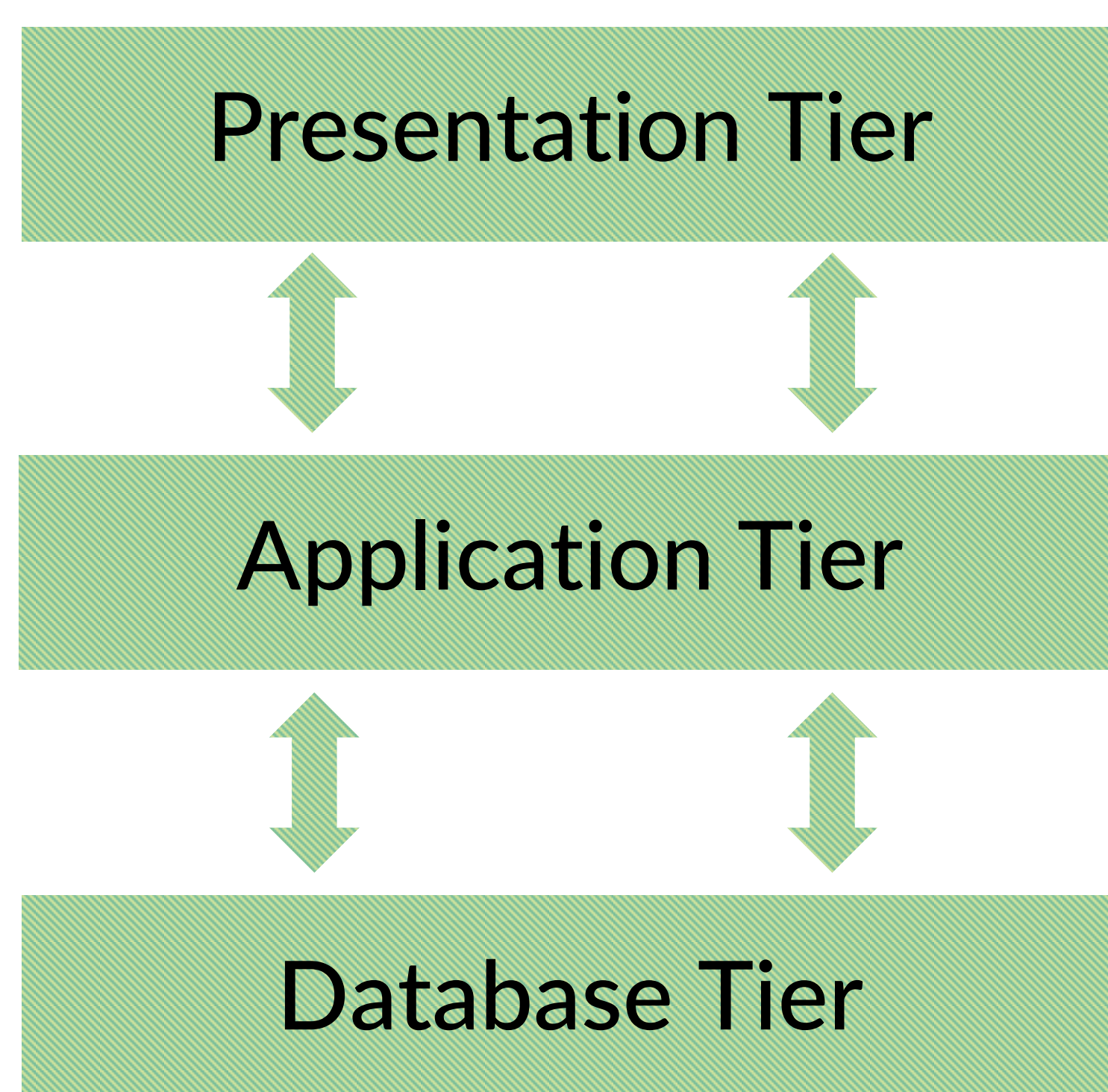


Figure 1: Screenshot of DITS with overview page of available datasets.

Architecture



"A highly convertible 3-tier database architecture. All its components are independent and can be modified accordingly."

The users operate on this tier and need no knowledge of the underlying layers. To make full use of the web application, each user needs to register and receives a user role with specific privileges.

JavaServer Faces - Primefaces JSF Framework - XHTML

This is the mediator between the user and the database. All views are also generated here.

Java with Java EE - Java Persistence API

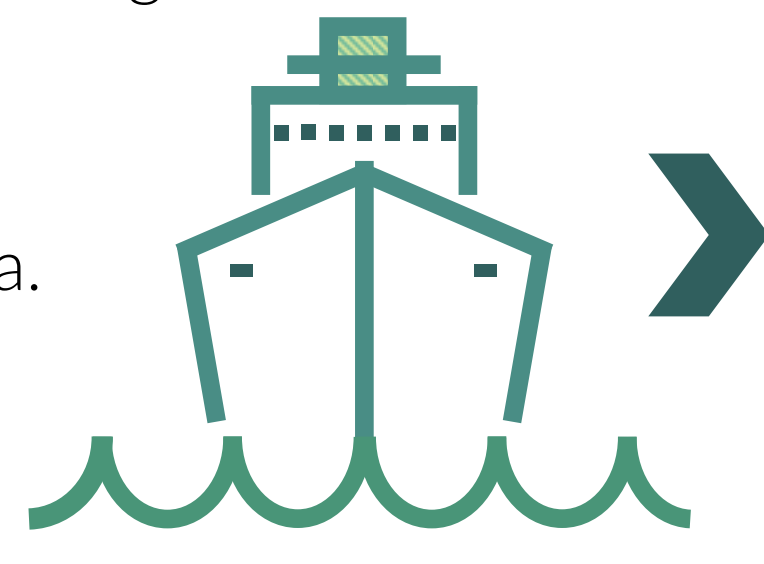
The database resides here, with its query processing languages
Oracle (or any other database engine (e.g. MySQL))

Figure 2: Schematic representation of DITS three tier architecture.

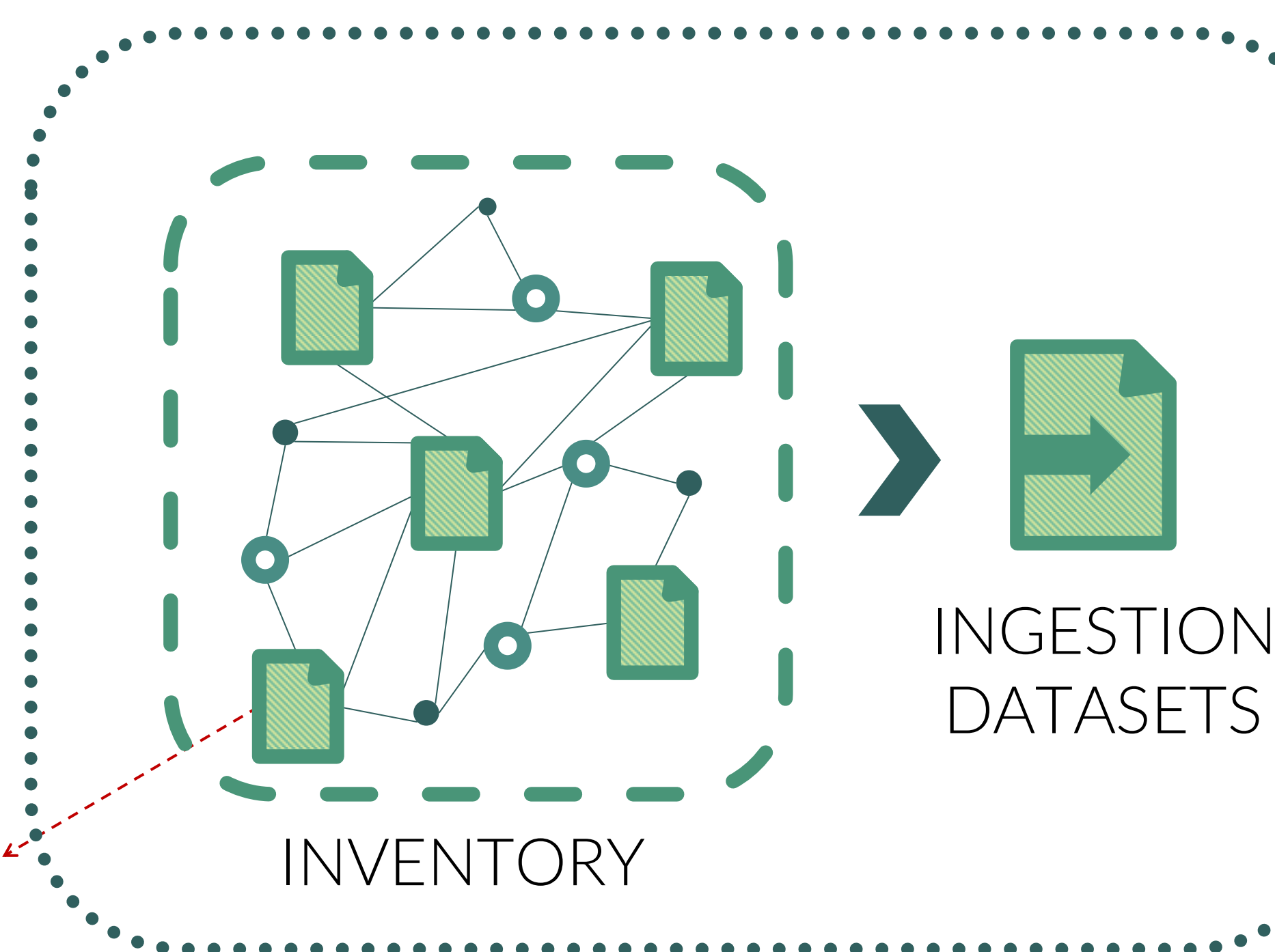
Data Flow

Requirements

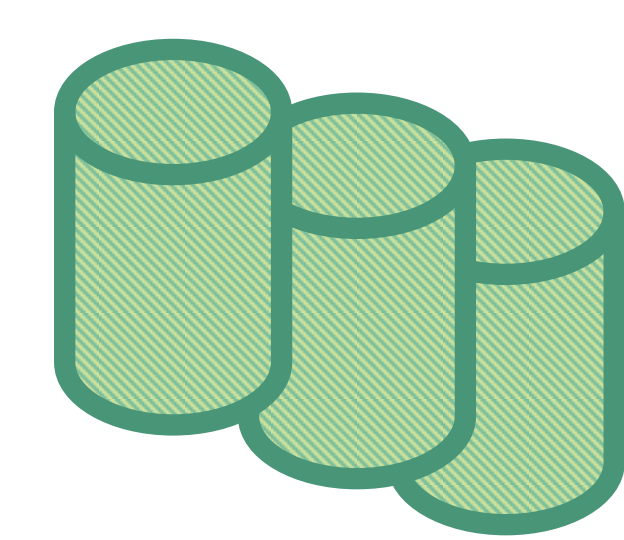
- **Inventory of sources** (e.g. publications, databases, files...) Submitting new sources, uploading or linking files with its metadata, editing if required and browsing the inventory.
- **Reporting data to BMDC** Uploading files and defining its metadata.
- **Archival of files** Storage on a separate secure server with versioning of files.
- **Identification of datasets** Combining datasources in a comprehensive ingestion dataset with upload of processed files.
- **Identification of missing data** Evaluating the completeness of a dataset (e.g. missing sampling gear, analysis method, units) before import into a central database.
- **Follow-up: status tracking of files and datasets** Tracking the status of datasources and datasets from reporting till import, with upload of the processing files throughout the process.
- **Exploration of database and export of data** Browsing specific tables (services, projects, platforms), editing if required and access to database statistics.
- **Accessibility** Import via Presentation Tier by all users and via Database Tier by the data managers depending on user role at login.



Status



ARCHIVAL OF FILES



DATABASIS



Conclusions

- DITS provides us with a **comprehensive** and easy to browse inventory of data sources and datasets with its import status accessible for both data originators and data managers.
- The import of data into central databases is facilitated thanks to the **standardization of metadata** already provided during the data submission.
- As it encourages **digitization of (historic) documents**, it reduces the bulk of physical documents in the office making the information accessible to a wider public.
- The flexible architecture of DITS enables us to **implement add-ons** if desired.