

Coral Reef - A FAIR metadata layer to provide clinical study data at your fingertips

Javier D. Fernández

Principal Data Scientist at Hoffmann-La Roche

Nelia Lasierra Beamonte

Senior Principal Data Scientist at Hoffmann-La Roche



The topics for today

1. OCEAN
2. The Clinical Data Repository
3. Lighthouse
4. Coral Reef

OCEAN

Data Sciences under one roof

OCEAN - One Centralised Analytics Environment

Next Generation Clinical Trials Analysis platform where our Data Scientists (Clinical Data Managers, Statistical Programmers, Biostatisticians, and others) do their work.

We are designing OCEAN to allow us to be more productive, realize efficiencies, take advantage of new ways of working and ensure FAIR data assets generation

Data Science under one roof

Breaking the silos

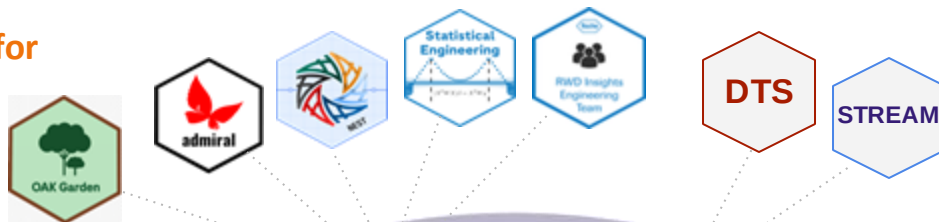
Enabling new ways of working



OCEAN supports new ways of working with embedded NextGen tools and Legacy tools

NextGen Tools/Toolkits for R workflow

Additional tools will follow over time



Legacy Tools for SAS workflow

OCEAN

The backbone system that hosts the NextGen Tools under one roof for both *regulatory* and *exploratory* work

Access to Rstudio/SAS Studio/Jupyter containers via OCEAN Workbench

Code & Collaboration Management: GitLab

Clinical Data Repository (CDR):
Data Ingestion (DCH & other sources)
Data Storage (Harbour)
Connected Metadata (Coral Reef)

Production Runs : The Orchestration Engine

Package Management & Publishing (RStudio
Package Manager (RSPM), RSCONNECT, RIM)

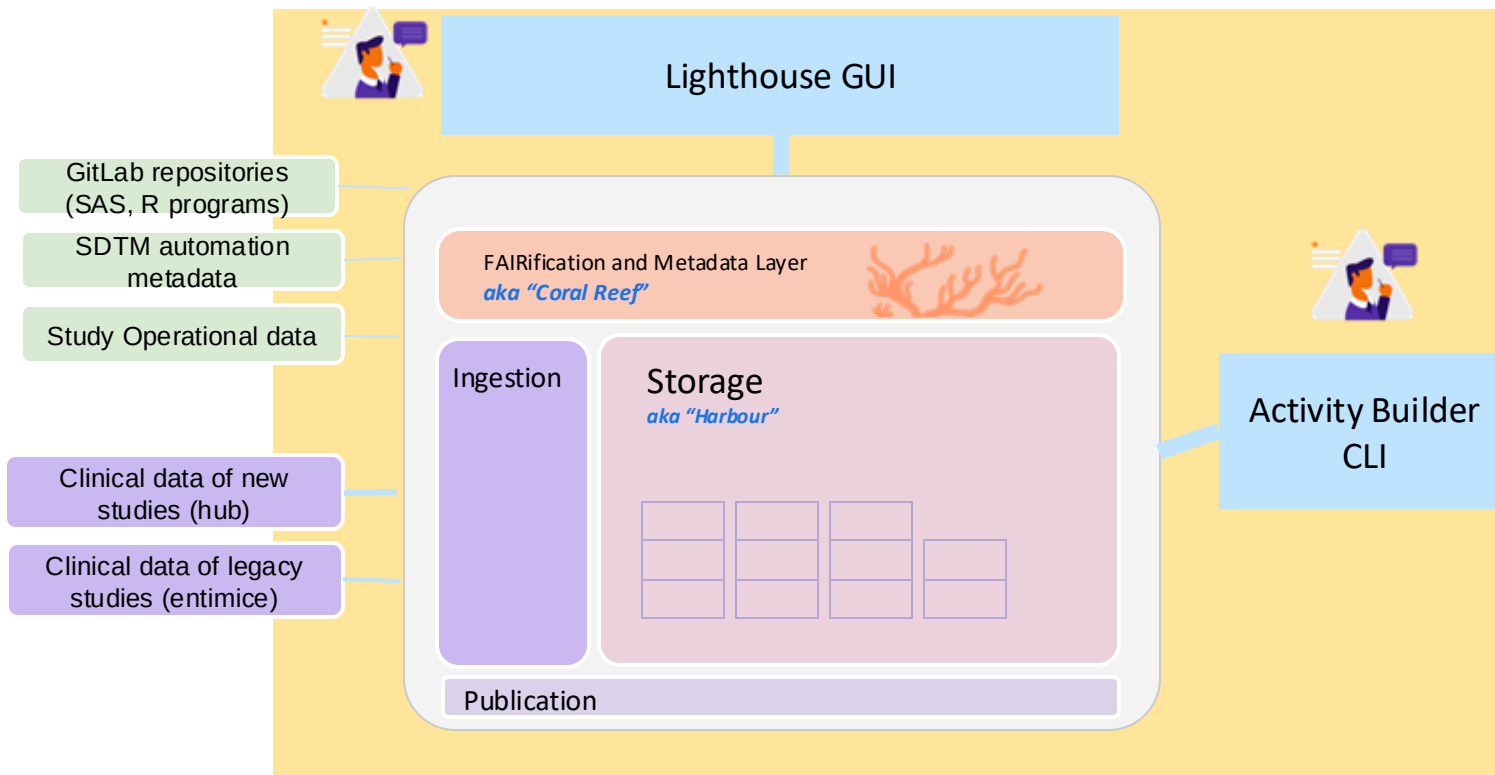
Why the CDR - Clinical Data Repository?

Our goal, our mission

Ensure clinical data can be safely ingested,
stored, analysed and transferred outside of
OCEAN to support the needs of Data
Scientist working on datasets development
and analysis

What is the Clinical Data Repository?

Repository to ingest, store and analyse clinical data



Enabling Findability in OCEAN - Lighthouse and Coral Reef



Why Coral Reef and Lighthouse?

Our goal, our mission

*We want to help Data Scientists to find, understand, discover,
and reproduce any Study, Activity, Dataset, Program,
Workflows and - more generally, any work performed in OCEAN
in an efficient manner*

Overcome existing challenges and bring new discovery opportunities

*Focus on **data scientists** workflows and usability*

*Focus on **data centricity** - automation & reuse*

For Whom? When?

Available module and planned modules

Use cases

- 16 analysed scenarios
- More than 200 questions!

Data Scientist

Activity Understanding (Study)

I would like to be able to understand the different Activities conducted for a Study, its original purpose, the snapshot used and the products created (including specs) so that I have a good overview and understanding of the Study and its deliverables

Traceability and Reproducibility

I would like to be able to point to the source data and code used for the creation of an output, demonstrate the audit history of deliverables being able to reproduce them.

Project Lead

Activity and Deliverable Management - overview

I would like to have a good overview of the progress done for each activity across all the Studies that belong to a Molecule project, their similarities and understand what activities contribute towards a submission.

Data Publishing & Sharing

I would like to publish datasets, the associated specifications and reviewer's guide to other publication sites seamlessly and in a way others can understand the content and whether they can re-use it or not.

Discovery - data and activity

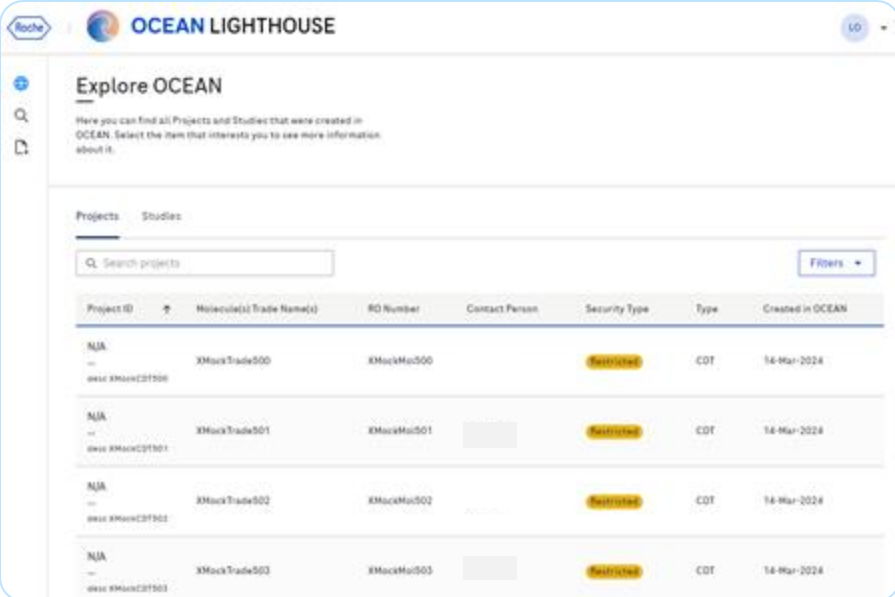
I would like to find studies for a particular molecule or TA/indication to understand both the availability and accessibility of data for secondary data use

Explore OCEAN

Allows navigation over entities registered in OCEAN

Overview - Project & Studies

It offers an overview of registered Projects and Studies. Data Scientists can be filtered based on the Molecule Name, Study Number, Security Type, Project Type or Creation date in Ocean



Project ID	Molecule(s) Trade Name(s)	RD Number	Contact Person	Security Type	Type	Created in OCEAN
NJA --- desc XMockCDT500	XMockTrade500	XMockMac500		Restricted	CDT	14-Mar-2024
NJA --- desc XMockCDT501	XMockTrade501	XMockMac501		Restricted	CDT	14-Mar-2024
NJA --- desc XMockCDT502	XMockTrade502	XMockMac502		Restricted	CDT	14-Mar-2024
NJA --- desc XMockCDT503	XMockTrade503	XMockMac503		Restricted	CDT	14-Mar-2024

Explore OCEAN

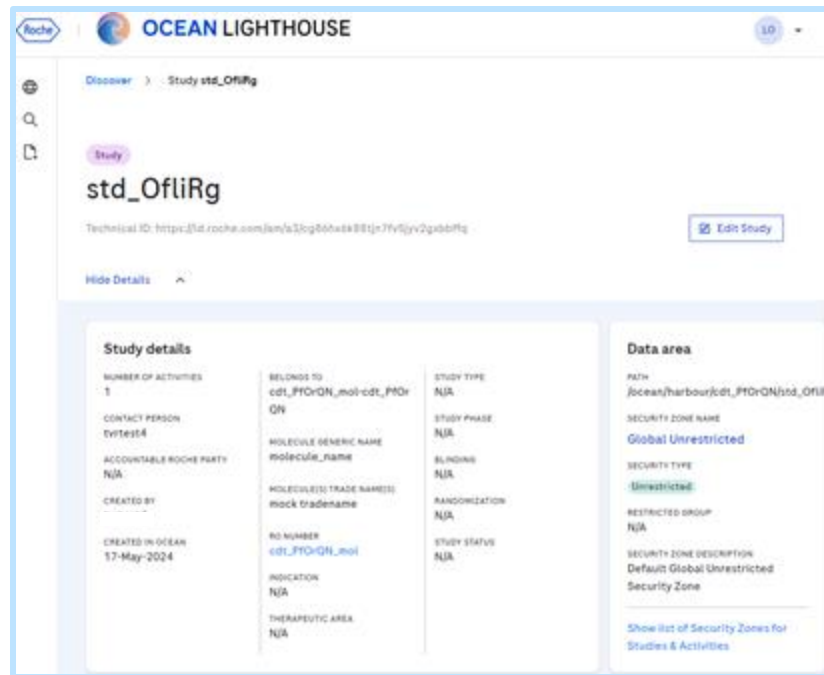
Allows navigation over entities registered in OCEAN

Info Pages

Offers detailed and dedicated information of each registered Project, Study, Activity and related datasets and workflows created and executed in OCEAN. It allows navigation across InfoPages according to the connectivity of the entities

Available infoPages:

- CDT project
- Generic project
- Molecule
- Study
- Activity
- Release
- GitLab Repository
- IO Job (PROD run)
- Security Zone



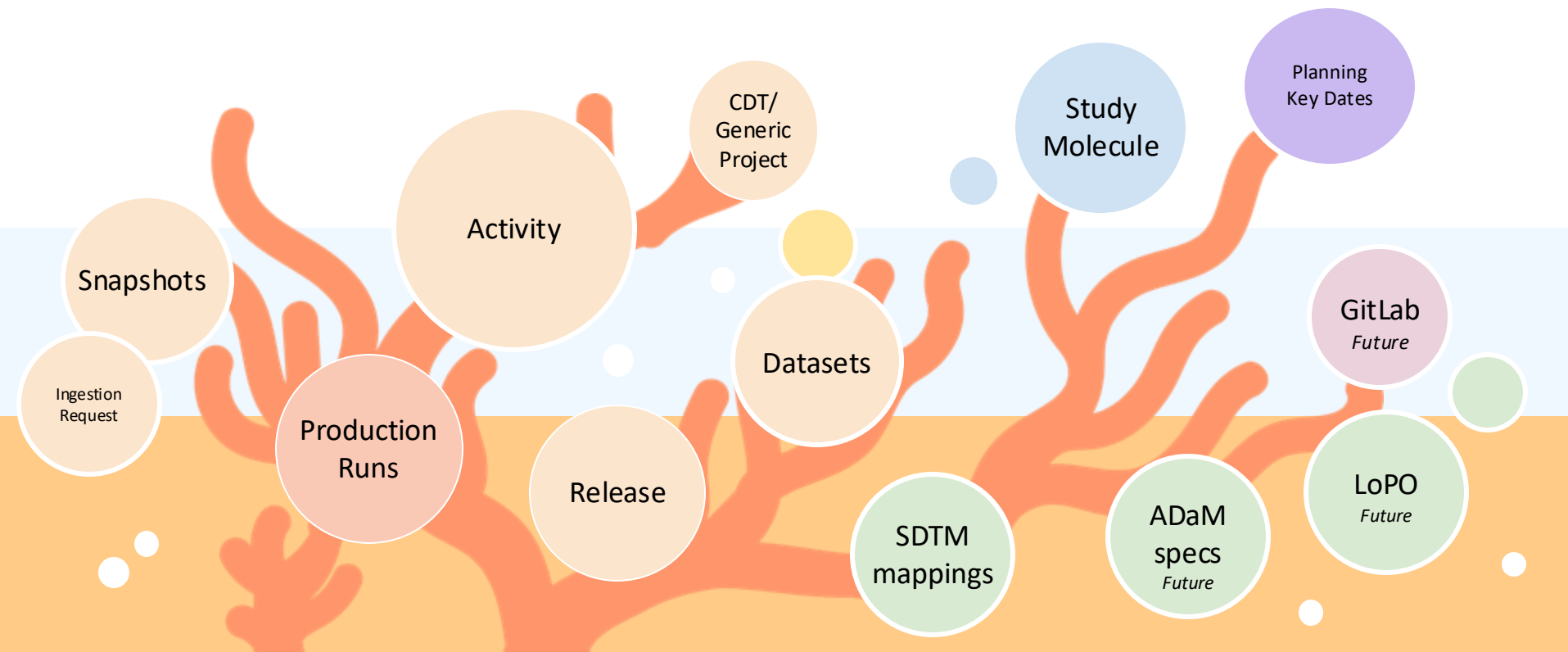
The screenshot displays the 'std_OfliRg' study details page within the OCEAN LIGHTHOUSE application. The page is divided into several sections:

- Header:** Includes the Roche logo, 'OCEAN LIGHTHOUSE' title, and a user profile icon.
- Breadcrumbs:** Shows the navigation path: Discover > Study std_OfliRg.
- Study Card:** Features a 'Study' label, the study name 'std_OfliRg', a technical ID, and an 'Edit Study' button.
- Study details:** A table-like section providing key information:

NUMBER OF ACTIVITIES	BELONGS TO	STUDY TYPE
1	cdt_PROD_mol-cdt_PROD	N/A
CONTACT PERSON	MOLECULE GENERIC NAME	STUDY PHASE
tyvtest4	molecule_name	N/A
ACCOUNTABLE ROCHE PARTY	MOLECULE(S) TRADE NAME(S)	BLINDING
N/A	rock tradename	N/A
CREATED BY	INDICATION	RANDOMIZATION
	N/A	N/A
CREATED IN OCEAN	THERAPEUTIC AREA	STUDY STATUS
17-May-2024	N/A	N/A
- Data area:** Contains metadata such as PATH, SECURITY ZONE NAME (Global Unrestricted), SECURITY TYPE (Unrestricted), RESTRICTED GROUP, and SECURITY ZONE DESCRIPTION.

What? Connected Knowledge Graph

Integration of metadata from multiple sources



How do we enable this?



User centric

*We enable
FAIR by DESIGN*



Data centric

Impact

Supports knowledge management in OCEAN and beyond

■ *Increased Productivity*

- Search less to find relevant information
- Gathers information needed for reproducibility and full E2E traceability

■ *Informed Data Scientists*

- Knowledge about anything happening in OCEAN at your fingertips
- Clear understanding of what has been done in a Project by anyone working on it
- Facilitates decision making

■ *Automation*

- Knowledge connected with the wider Roche ecosystem
- Machine readable content available to be used by other digital applications

FAIR by Design

Annotating Data Assets



The FAIR principles

Designed to enable data findability and re-use

- Designed to enhance the reusability of data assets
- FAIR principles define characteristics for data assets to be discovered and reused
 - Both machines and humans can understand: what is this object?
How can I process it? Can i get it? Is what I need?

The FAIR principles

Designed to enable data findability and re-use

F

Findable

Finding data and associated context

A

Accessible

Accessing data from different systems

I

Interoperable

Data interoperability and sharing culture – “our” data

R

Reusable

Reusing data/identifiers, re-linking (patient/sample) and properly tracking data

FAIR by Design

Principles applied to the (meta)data



Data centric

- Identifiers
 - We use unique and persistent identifiers
- Data Standards
 - Our assets are described according to a formal data model (ontology)
- Rich metadata how, what, why, by whom
 - Enables connectivity of rich machine readable metadata
 - Knowledge graph at the center
- Access and re-use
 - Access to the Coral Reef is enabled through standard interfaces

URI Generation - id.roche.com

We use unique and persistent identifiers

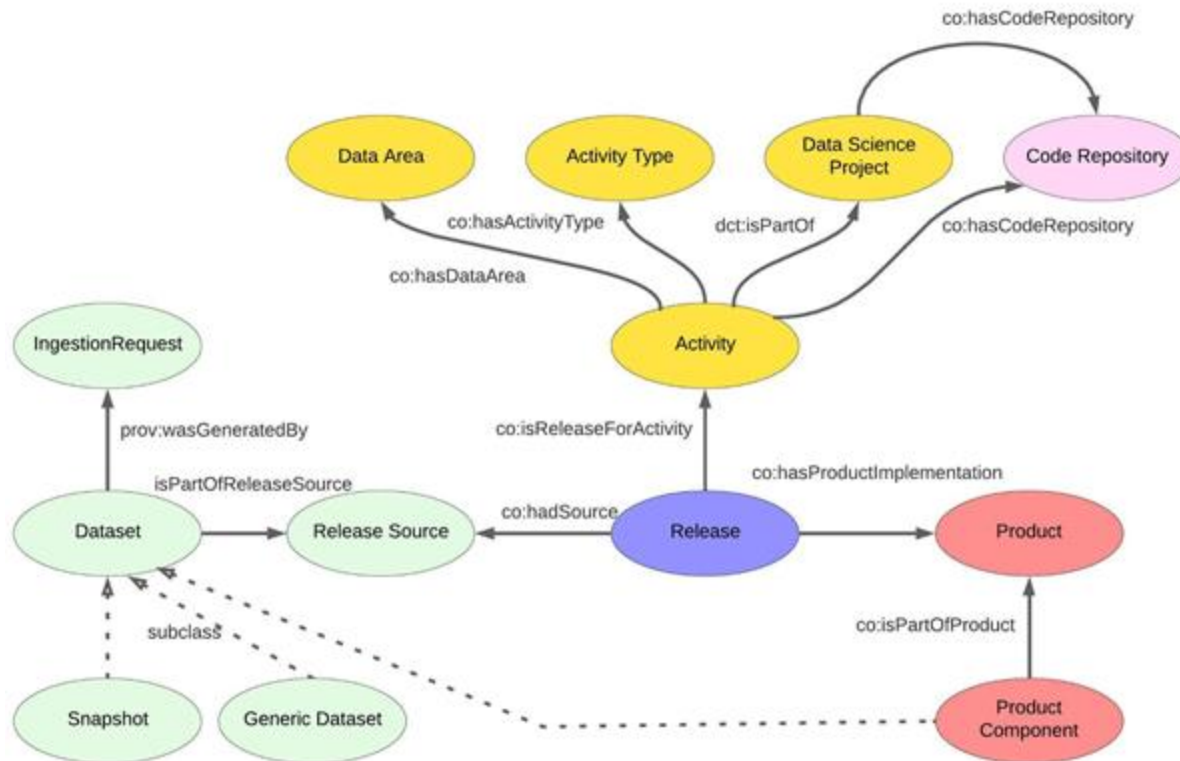
- Resolvable, unique, persistent identifier
- 2 stateless algorithms (hash+random)
- **Hash-based URI:** md5 (class + set of features) + erdi8 *
- **Random URI:** xid + erdi8 *
- API endpoint and Javascript for GraphDB
- **id.roche.com** namespaces

* erdi8:

- <https://github.com/athalhammer/erdi8-py> Thanks to Andreas Thalhammer

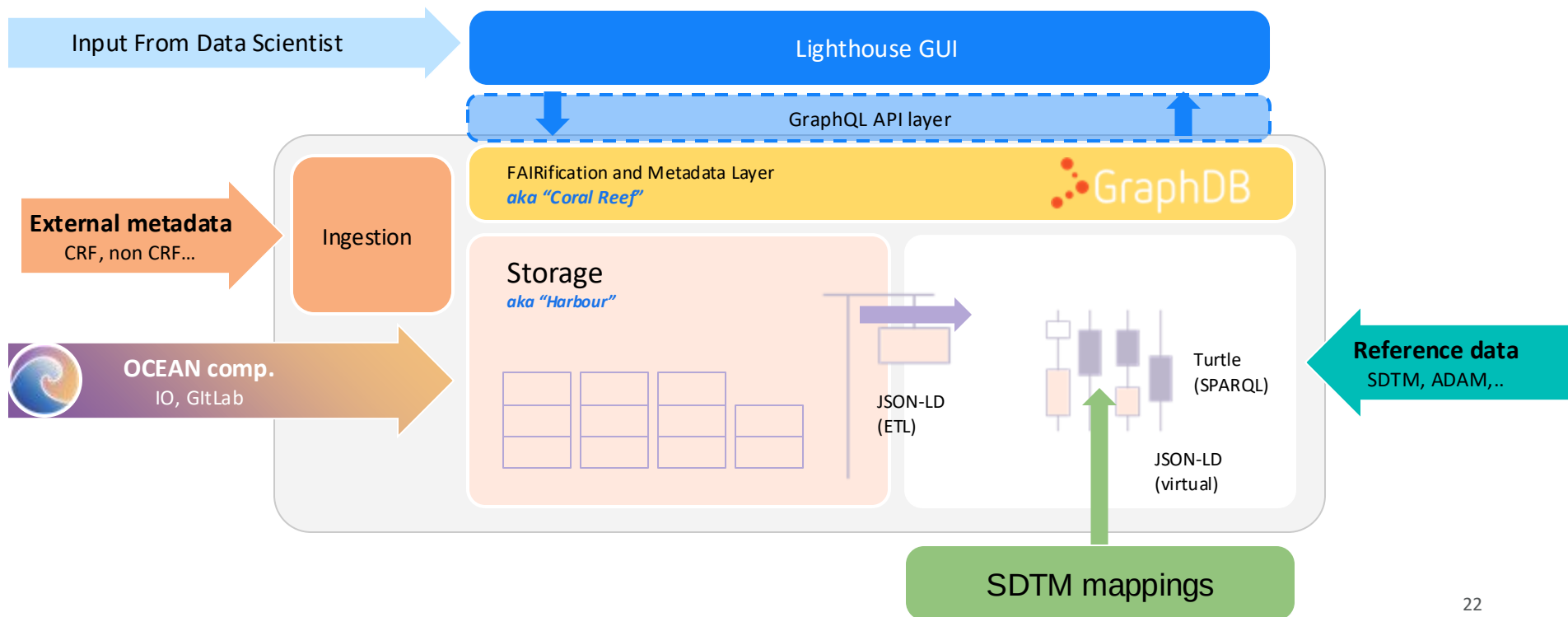
Ontology - Activity Map

Data model - how all the items are connected



Knowledge Graph

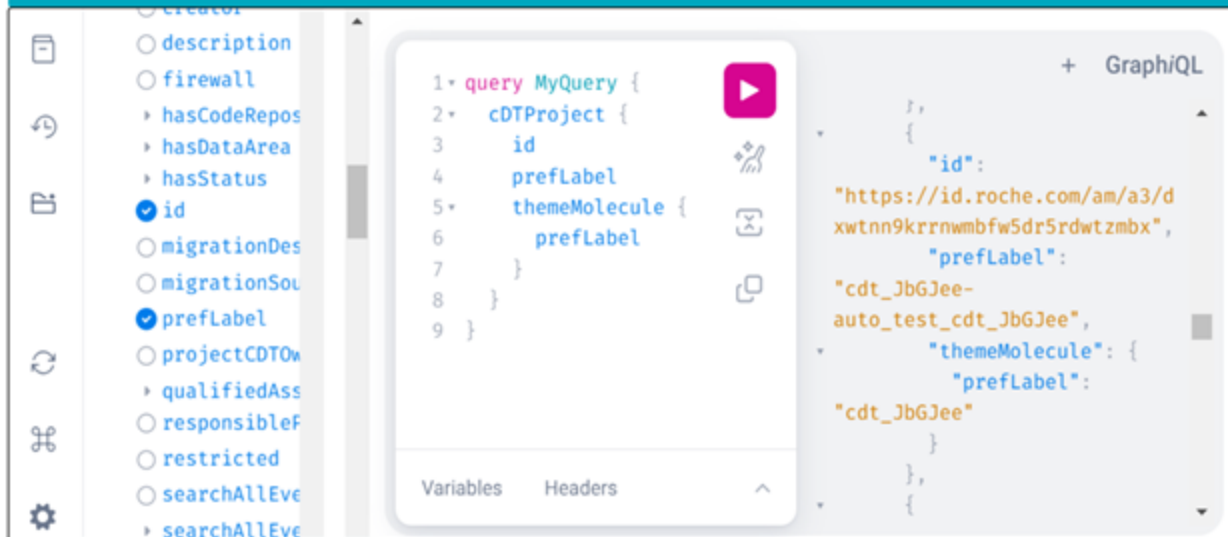
Population via ETL (persisted) + virtualization



GraphQL

API abstraction for reading and writing

Enhanced Developer Experience: A single, well-documented API fosters efficient development and simplifies maintenance.



- Open source standard developed by Facebook
- GraphQL Foundation
- GraphQL schema
- Introspection
- Queries, mutations, subscriptions
- Many tools, e.g. GraphiQL, Apollo GraphQL
- Many libraries for Python, JS, Java...
- Used by
 - Facebook
 - Linked In
 - AWS
 - Gitlab
 - And many more

**One single, consistent interface for client applications (web app, CLI):
A GraphQL layer successfully unified disparate APIs**

FAIR by Design

Maximizing User Experience



FAIR by Design

Making the FAIR way the easy way for our Data Scientist



User centric

- Business oriented questions
 - We focus on data scientist needs and workflows
- User Centric Design
 - Easy to use and embed as part of business workflows
 - Usability and Controlled terminologies for enabling easy FAIR
- Rich metadata and Controlled terminologies
 - The majority of metadata is collected automatically. Minimal metadata requested to the user and collected only once
- Prospective FAIRification
 - Avoid retrospective tagging

Scenario - Understanding Activities

Example of key questions

As a recently onboarded team member to Study GO1234, Jack needs to understand what Activities were completed and are in progress for the Study. Jack goes to OCEAN to better understand the following

- Which Activities are Completed?
- Which Activities are still Ongoing?
 - What Products still need to be created?
 - Who last worked on the Activity and the Products? Who can he go to to ask questions?
- Where are all of the Products and programs located?
- What was the purpose of the Activity?
- What Source Data exists for the Study?
- When was the Source Data extracted?
- What other Data Products were generated from the specific Source Data version? What Activity Type(s) are they associated with?
- Which Activities influence or are influenced by another? Which Activity(s)?
- Where is the Git Repo that was used?
- What Dictionaries were used?

User Interviews and Usability Testing

Example of key questions

**User Interviews - discovery
of needs**

**Usability testing - evaluation
of designs**

Key Outcomes and feedback for implementation

- We simplified and improve language we use in the interface.
- We discover improvements and navigation gaps and also highlight
- We could identify key priorities as the users were highlighting the most important features for them

Controlled Terminologies, Definitions and Synonyms

Key Components

- **Repositories for Standards and Controlled Vocabularies at Roche**
 - **GDSR** - Global Data Standards Repository
 - Provides with CDISC aligned standards and controlled vocabularies list
 - **RTS** - Roche Terminology System
 - Provides terminologies to more than 80 applications across all sites
 - **Lynx** - Reference Data Connectivity:
 - An application that provides a point of access to connected reference data

Activity Types - 62

E.g. Primary CSR, Monitoring Committee, Simulations

Product Components and Model conformance Types - 8

E.g. Displays, SDTM, ADaM,

Code Repositories Types - 12

E.g. R Sharing, SAS Sharing, IDD type

Activity Type - 12 main types

Overview - Controlled terminologies, definitions & synonyms

Type	Subtype					
Clinical Study Report	Primary CSR	Update CSR	Interim CSR	Final CSR	Addendum CSR	Other
Other Regulatory Document	Data Memo	Briefing Package	4MSU / 120DSU	RMP		
	Investigator's Brochure	Licence Renewal	DSUR / PBRER	IND Annual Report		
	Core Data Sheet	SCS / SCE	ISS / ISE	SmPC / USPI	Other	
Data Monitoring & Review	Monitoring Committee	Advisory Board	Event Tracking	Visualisation Data	RBQM	Other
Strategic Decision Making	Interim Analysis	Dose Decision	PivGo	Other		
Study Design	Sample Size	Randomisation	Simulations	PTS	Other	
TAI Data Handling	PK	Immunogenicity	Biomarkers	Other		
Health Authority Request	FDA	EMA	Health Canada	SwissMedic	MHRA	
	NMPA	PMDA	TGA	Other		
Insights Generation	Data Mart	Other				
Publications						
Exploratory Analysis						
Data Sharing Request						
Molecule Enabling Project						
Other						

Capturing metadata through Lighthouse

Creation workflows for : Project, Study and Activity

The screenshot displays the Roche OCEAN LIGHTHOUSE web application. On the left is a sidebar with navigation icons and a menu containing: Context* (green checkmark), Additional information* (green checkmark), Data Area and Access Configuration* (green checkmark), Code Repository* (green checkmark), and Summary (yellow circle with 'i'). The main content area is titled 'Summary' and shows a configuration form for a 'Data Area'. The form has four sections, each with a green checkmark icon and an expand/collapse arrow:

- Context:** Shows 'Relation: Project' and 'Project: mo_perft_test_06' with a dropdown menu showing 'dummy_molecule_mo_perft_test_06', 'dummy_molecule_name: dummy_molecule_trade', and 'Mock CDT mo_perft_test_06'.
- Additional information:** Shows 'Type: Insights Generation', 'Sub-type: Curation', and 'ddgdyfj dfgf'. There is an unchecked checkbox 'Use Activity to create Pooling' and 'Contact Person (Roche ID): Javier D. Fernandez (ferja29)'.
- Data Area in Harbour and Access Configuration:** Shows 'for Data Area path: /ocean/harbour/mo_perft_test_06/InsightsCuration_test'. Below this is a green success message: 'Data Area was created successfully! This Data Area is **Unrestricted** (info icon) Everyone in OCEAN has access to it.'
- Code Repository:** Shows 'InsightsCuration_test_Clinical_Reporting_Activity-specific_6305126', 'Type: Clinical Reporting Activity-specific', and 'Description: test desc'.

An 'Activate' button is located at the bottom right of the form.

Lighthouse allows the registration of CDTs, Studies and Activities and also capturing metadata using the **Controlled Terminologies** list .

Look Ahead

Next Steps!

Planned in our roadmap for 2025

Continue improving findability and enable access to more operational Insights

- Exploration of LLMs to query the Knowledge Graph
- Additional info Pages planned for Lighthouse
 - Visibility of Transfers to other systems
 - Information about ingestion of datasets

Help Data Scientist with Project management capabilities

- Capture metadata about Key Dates and Programs per Activity
- Enable visualizations across Studies and Projects efficiently manage resources

*Thank you very much to the
Coral Reef, Lighthouse and
OCEAN teams*

Doing now what patients need next