Paper AD06

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

Javier D. Fernández, Hoffmann-La Roche, Basel, Switzerland
Nelia Lasierra, Hoffmann-La Roche, Basel, Switzerland

ABSTRACT

OCEAN (One CEntralised ANalytics platform) is the next generation data and analytics platform for delivering studies for PD Data Sciences at Roche. At the core of OCEAN, the Coral Reef FAIR metadata layer offers a new, rich approach to capture, store and use metadata, integrating multiple sources in a common model (Knowledge Graph). Coral reef empowers our Data Scientists to easily search, find, explore, reproduce and understand any work performed in OCEAN (e.g. Projects, Studies, Activities, Repositories, Datasets) in an efficient manner. This connected metadata layer impacts productivity, enabling our scientists to invest less time searching, as well as knowledge management, facilitating better informed decisions. Additionally, the FAIR foundation enables downstream automation using machine-readable content. Access to Coral Reef metadata is enabled via a graphical user interface (the Lighthouse) and APIs, enabling further knowledge connection across other digital applications in the Roche ecosystem.

INTRODUCTION

OCEAN (One Centralized Analytics Platform) represents the cutting-edge data and analytics solution designed to facilitate the delivery of studies for PD Data Sciences at Roche. Using this platform, Data Scientists (Clinical Data Managers, Statistical Programmers, Biostatisticians, and others) develop datasets (SDTM, ADaM) and outputs (tables, listings, graphs) created as part of the study conduct, enabling its analysis and insights generation. OCEAN brings *Data Science under one roof,* allowing not only regulatory and study related activities but also exploratory work, breaking in this way current silos and enabling further collaboration and alignment activities among Data Scientists.

We are designing OCEAN to allow our Data Scientists to be more productive, achieve greater efficiencies, leverage modern technologies and enable new ways of working. Roche Data Scientists can execute both SAS and R workflows and make use of tools like OAK [1] for boosting automation of SDTM datasets creation, apply Admiral templates [2] for ADaM creation, and use the NEST [3] suite for output development. Furthermore, OCEAN promotes the management of code via GitLab and offers a core component, the Clinical Data Repository (CDR), for safely ingesting and storing data.

CDR is a cloud-based and file-based storage solution that allows storing the datasets created by Data Scientists organized in Activities as part of Studies or Projects. An Activity has been defined generally as, “any type of work performed by the Data Scientist”. In the context of a Study, Activity refers to “the description and reason for why a set of analysis is being performed with a defined set of source data”. The CDR allows Data Scientist to ingest raw data (CRF, non CRF, protocol deviations, etc.), but also other types of datasets (e.g. SDTM or ADaM from legacy Studies), in each Activity, and to further transfer datasets to external systems, such as P21 for conformance checking, or to other systems in Roche involved in the submission process.

A key requirement within the design of OCEAN, and CDR, was to ensure findability and reproducibility of datasets in an efficient manner. This led us to the design and development of the Coral Reef, a FAIR By Design [4] metadata layer within the CDR. This metadata layer allows us to capture and to connect metadata within each Activity created in the system. This metadata is further exploited by the Coral Reef Lighthouse to enable an easy navigation, search and consumption of knowledge by our Data Scientists. The Lighthouse is a graphical web interface connected to the Coral Reef that serves as an entry point to find and understand the work performed in OCEAN. Figure 1 shows the main components of the CDR, including Coral Reef and Lighthouse.

The FAIR principles, key to increase productivity and impact knowledge management, have been embedded since the initial design of OCEAN. In the following, we describe the methods, principles and technology choices that we

selected to enable FAIR By Design in Coral Reef and Lighthouse, and how it is essential to enable finability and reproducibility of the data assets generated in OCEAN, and to maximize user experience and adoption.

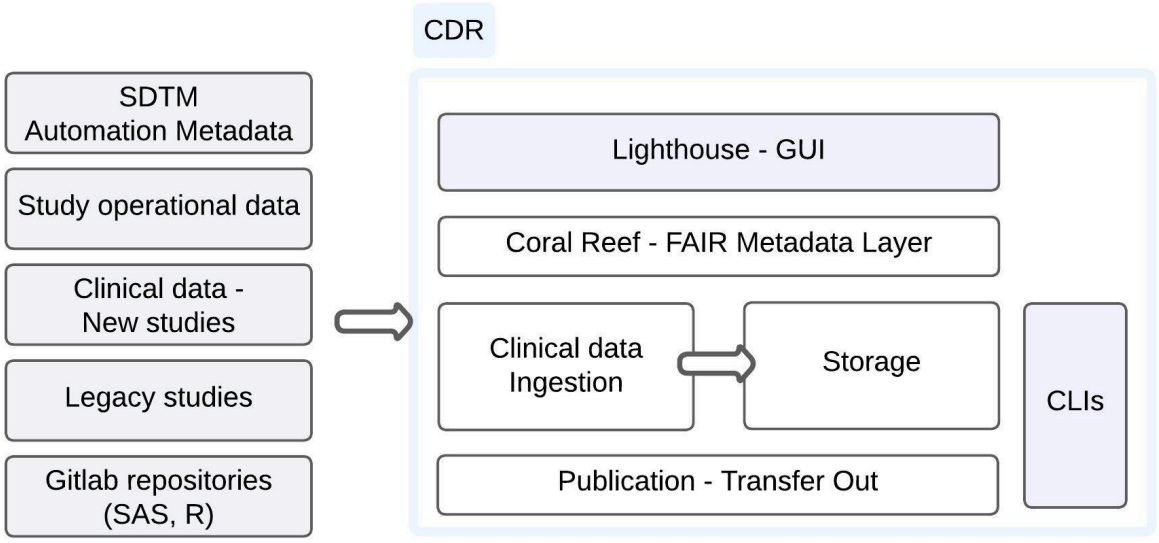

Figure 1. Clinical data repository architecture

The need - enabling findability in OCEAN

Executing and managing clinical trials involves the generation of a diverse myriad of metadata that needs to be meticulously collected, organized and made findable: Not only the clinical study data *per se* (CRF, non CRF, protocol deviations, etc.) need to be properly annotated, but also how the data is processed, when, by whom, which with methods, etc., so one can ensure data integrity, provenance, reproducibility, regulatory compliance and future reuse.

Coral Reef is designed as the comprehensive metadata layer of the Clinical Data Repository of OCEAN with the twofold objective of being data centric and keeping the focus on Data Scientist needs. On the one hand, data centricity is the overarching framework to effectively manage such a diverse array of metadata, making it interoperable across different components and processes, from the ingestion of the data, to the analysis and transfer/publication of results. In this regard, Coral Reef can be seen as “metadata contracts” between parties to enable interoperability and foster automation. On the other hand, helping Data Scientists to find, understand, discover and reproduce any Study, Activity, Dataset, Program, Workflow and, more generally, any work performed in OCEAN in an efficient manner has been our goal and main driver to build the Coral Reef layer.

In order to make sure that we were meeting these needs, our first step towards the development of this metadata layer was to understand the different scenarios, stakeholders and questions that they had with regards to the Activities performed in OCEAN. This activity led us to identify 2 main different personas:

- **Data Scientists**: Coral Reef aims to help Data Scientists working in a particular Study or Project by providing the right metadata at their fingertips so that they can understand the different Activities conducted, the original purpose, the snapshot used and the products created (including specs), reaching a good overview and understanding of the Study/Project and its deliverables. In the particular case of reproducibility, Data Scientists need to point to the source data and code used for the creation of a deliverable, and demonstrate the audit history of deliverables, being able to reproduce them. Finally, they need 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.
-
- **Project leads**: Coral reef aims to provide project leads with activity and deliverable overview and management capabilities. Project leads need 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 to understand which activities contribute towards a submission.

Note that, in both cases, users might need basic findability features to find studies and activities, e.g. for a particular molecule or TA/indication, to understand both the availability and accessibility of data for secondary data use. During the initial development months we conducted a series of workshops that led us to identify 18 different scenarios and to derive more than 200 questions that we wanted to answer with the Coral Reef metadata.

FAIR BY DESIGN

The need to enable findability in OCEAN led us to the adoption of FAIR data principles [4]. The FAIR Data principles, published for the first time in 2016, were designed to enhance findability and reusability of digital assets and are widely recognized and adopted today across academia and industry. These principles describe a set of characteristics for digital assets to be discovered and re-used with a special emphasis on machine actionability so that machines can help with the findability of these assets.

The recognition and popularity of the FAIR principles to enable findability led, in recent years, to retrospectively act on the data assets already developed and stored in our systems in order to enable its FAIRification. In many situations, and due to the lack of initial consideration of these principles in the design of the systems, this led to involving the Data Scientists in the process of collecting and enriching metadata being an afterthought complex to follow, and undermining the success of the FAIRification activities. This was an important aspect to consider within the design of Coral Reef where we adopted FAIR principles “by design”. This meant not only to ensure that the metadata was born FAIR, but also to ensure that, whenever we needed the involvement of the Data Scientist in the FAIRification process, this was going to be only if strictly necessary (prioritizing automatic metadata collection), in a prospective manner and integrated within their business workflows.

Applying FAIR Data Principles

To ensure our metadata was described according to the FAIR principles, we implemented the following key functionalities within our system:

- **Identifiers:** Ensuring metadata is identified with globally unique and persistent identifiers is the first of the FAIR principles, that we further develop to achieve Global Unique Persistent Resolvable Identifiers (GUPRI). To ensure global and unique identifiers, we developed 2 algorithms to mint new URI identifiers, either random and non-speakable, or hash-based on certain properties, so that the identifier is repeatable. Persistency is obtained by making use of an internal redirection service (id.roche.com) that guarantees persistent and stable identifiers. Finally, we also obtain resolvable identifiers redirecting each entity to a central metadata hub (Lynx [5]) that describes the metadata behind each identifier and points to the system where the data is.
- **Ontology**: To facilitate interoperability and ensure a common and formal understanding of our metadata, we developed a Coral Reef OWL ontology. The design of this ontology is particularly tailored to help answering the competency questions identified during our workshops with the Data Scientist. To date, this ontology has more than 130 classes, 90 object type properties and 150 data type properties. In addition, we reused well-known existing vocabularies such as dublin core, foaf, prov and dcat or schema.org.The ontology is being extended with every new release of the system as we see the need of covering more use cases (due to new stakeholders joining OCEAN or new data sources being connected to the CDR). To facilitate its understanding, we also developed documentation using WIDOCO [6]. The main classes of our ontology are displayed in Figure 2.
- **Machine readable metadata**: Providing rich metadata is at the core of the FAIR principles. In Coral Reef, our metadata is materialized in the form of an RDF knowledge graph, stored in GraphDB [7].
- **Accessibility:** Access to the Coral Reef metadata is possible across Roche users through the aforementioned Lighthouse Graphical User Interface, but also programmatically via GraphQL.

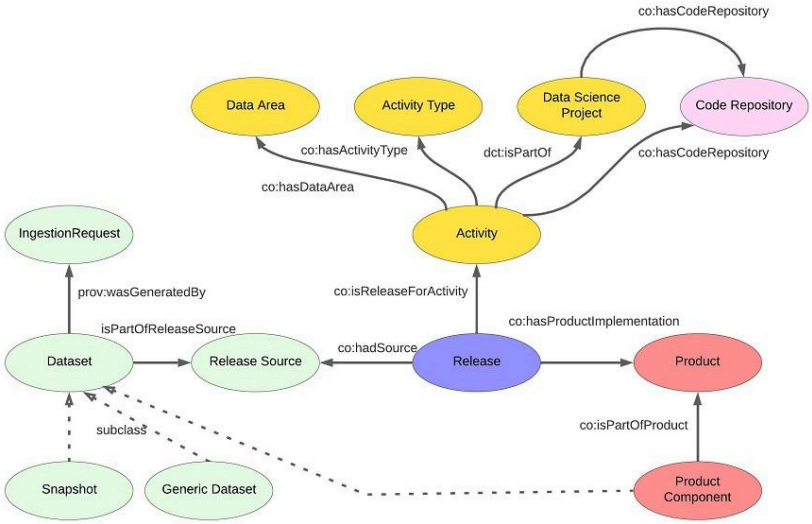

FIGURE 2. CORE ENTITIES OF THE CORAL REEF ONTOLOGY

MAKING THE FAIR WAY THE EASY WAY FOR OUR SCIENTISTS

Similarly, to ensure that the FAIRification activities were integrated within the business workflows and user-friendly for our Data Scientists, we implemented the following key functionalities within the design of our system:

- **Business scenarios**: As explained above, we initially conducted a set of workshops to identify the different scenarios and requirements when our Data Scientists need to understand the work performed in OCEAN. During these sessions, we presented the Data Scientist with a scenario and discussed their needs to successfully implement their job in that scenario, with particular attention to their needs when understanding the work performed. An example of a scenario is when someone new joins the Study. In that scenario, some of the questions that we identified and for which Data Scientists can find an answer with Coral Reef metadata are: (1) Which Activities are Completed? (2) Which Activities are still Ongoing? (3) What Products have been created in each Release? (4) Who last worked on the Activity and the Products? To whom we can go to ask questions? (5) Where are all of the Products and programs located? (6) What was the purpose of the Activity? (7) What Source Data exists for the Study? When was the Source Data extracted?

- **User centric Design:** User experience plays a key role in the adoption of a new system. To ensure user acceptance and integration within the business workflows, a key aspect within the development of Lighthouse was user centricity. As part of the User experience (UX) research activities we conducted: (1) User Interviews to discover needs, and (2) Usability testing sessions of Mock-ups to validate designs and identify improvements. These activities were complementary to our business analysis workshops as they were focused on the usability of the system and led to co-create and iterate in the design of Lighthouse. User Interviews allowed us to learn from the Data Scientist about their needs to access information learning from the challenges they were experiencing with the former system. Based on the outcomes of these sessions we simplified and improved the language that we used in the interface and discovered improvements and navigation gaps.

- **Rich metadata and Controlled Terminologies**: To ensure consistency, common understanding and to facilitate capturing information through Lighthouse, we developed lists of controlled terminologies that are shared across different CDR components. When possible, we reused existing terminologies from Roche reference metadata repositories like the Roche Terminology Service (RTS [8]) or the Global Data Standards Repository (GDSR [9]). An essential list of controlled vocabularies is the Activity Type. We developed 52 different terms including description and synonyms used to describe the different types of Activities performed in OCEAN. When creating a new Activity, the Data Scientist needs to select one of these types, so that it eventually helps to understand the purpose and the datasets created in the Activity. Other examples of controlled vocabularies are the list of product models to which the products created conform to (e.g. SDTM) or the list of systems from which the data can be ingested in the CDR.

- **Prospective FAIRification:** The vast majority of our metadata is collected in an automatic manner and early in the process e.g. at the ingestion or creation time. Only in a few situations do we require intervention from the Data Scientist to provide information that aids in understanding. This is done only once in the system and always in a prospective manner.

CONCLUSION

The implementation of OCEAN, Roche's next-generation data and analytics platform, marks a significant advancement to bring Data Science (regulatory and exploratory work) under one roof, to enable new and more productive ways of working and to execute R workflows. In this paper, we provide details on Coral Reef, its central keystone and metadata layer for capturing, storing, and utilizing metadata through its integrated Knowledge Graph model. This layer is essential to allow findability of Data Assets in OCEAN and provides Data Scientist with rich knowledge to understand and reproduce the work performed in OCEAN. We show the foundational FAIR principles of Coral Reef, and how they are implemented in practice. The application of these principles allowed us not only to deliver our goal of finability but also to enable automation and further metadata integration opportunities in the wider Roche ecosystem.

REFERENCES

[1] Franklin, A., Krochmal, M., Ganapathy, R. (2023). Oak Garden - Metadata-driven SDTM automation in R. In PHUSE EU Connect 2023.

[2] Farrugia, R. (2024). {admiral} - from Open Source to Company Implementation. In PHUSE US Connect 2024

[3] Shen, V., Black, J. An introduction to NEST, an open-sourced R toolkit for TFL delivery and interactive data displays used in all new Roche trials NEST. In PHUSE US Connect 2023

[4] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3(1), 1-9.

[5] Fernández, JD., Lasierra, N. (2021). Lynx: A FAIR Knowledge Graph Engine for Reference Data Integration & Look up Service. In Industry track of Semantics. Available at https://2021-eu.semantics.cc/lynx-fair-knowledge-graph-engine-reference-data-integration-look-service.

[6] Garijo, D. (2017). WIDOCO: a wizard for documenting ontologies. In The Semantic Web–ISWC 2017, Proceedings, Part II 16 (pp. 94-102). Springer International Publishing.

[7] https://graphdb.ontotext.com/

[8] Thalhammer, A., Romacker, M., & Rupp, J. (2017). Semantic Terminology Management for Applications: Contextualized SKOS-XL. In ISWC (Posters, Demos & Industry Tracks).

[9] Mason, H. (2019). FAIR Data Standards in the GDSR. PHUSE EU Connect 2019.

ACKNOWLEDGMENTS

We would like to thank all the members of Coral Reef, CDR and the overall OCEAN team, as well as Roche users involved in this effort.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name:Javier D. Fernandez, Nelia Lasierra
Company: F. Hoffmann La Roche
Email:javier_d.fernandez@roche.com, nelia.lasierra@roche.com