

The background is a dark blue gradient with various abstract shapes. On the left, there are several vertical bars of varying heights and widths, some with a blue-to-purple gradient. Scattered throughout are circles, triangles, and a grid of small dots. In the bottom right, there is a large white circle with a smaller solid blue circle inside it.

Governance and Cataloguing of Analytics Assets

Paul McCaffery
Clement Benamara

November 14th, 2022

Table of Contents



 Topic	 Description
Context Sharing	Growth of Data and Data as an asset
Data Cataloging	Our solution for data cataloging The Roche Data Marketplace and Collibra Data Governance principles
Extend our cataloging capabilities	Solution Description Use case : Cataloging a R Packages
Q&A	

Context sharing



About the project

objective: to design and develop an approach to catalogue the analytics used across our organisation to gain new insights, enable faster decision making from data assets, the cataloguing, and re-use of analytics assets is fundamental.



Business Problem

The growth of data has lead to an exponential growth of our analytics and machine learning models. While cataloging the data is part of our normal practices, cataloging the analytics and linking them back to the data has become even more essential.



Solution

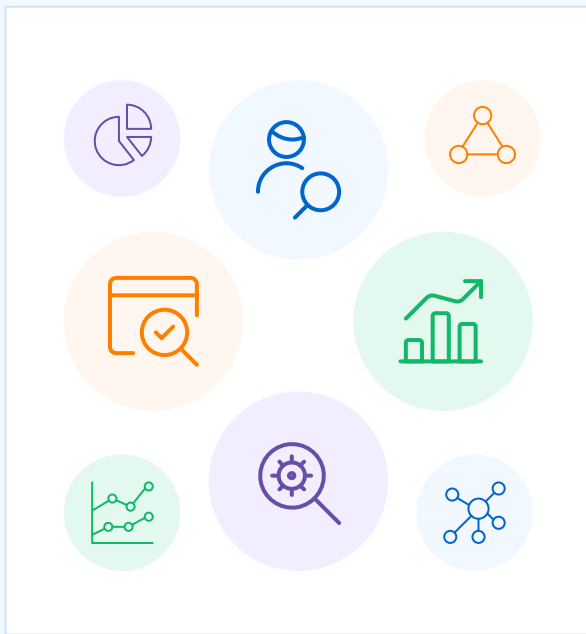
Extend our cataloging capabilities beyond data and provide a data lineage between data and analytics such as machine learning models, R packages and Python libraries.



Value

- Provide key information about our data and related analytics
- Make analytics and data searchable next to each other
- Allow reusability and avoid duplication of efforts

Data Governance Definition



- Solution for **Consistent** and **Reliable** Data
- Details rules, data quality and lineage
- Clear data standards adopting **FAIR principles**
- Ensures data ownership and compliance



Findable
"I know where
data is"



Accessible
"I can access
data I need"

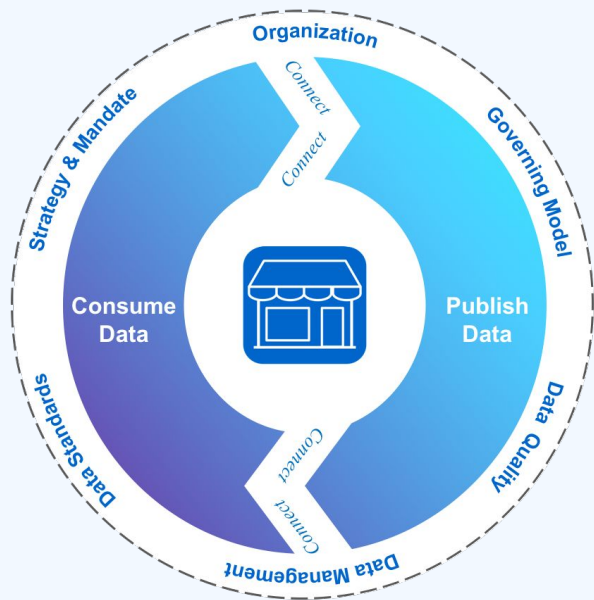


Interoperable
"I use consistent
standards & terms"



Reusable
"I use existing data to
answer new questions"

The Roche Data Marketplace (RDM)



The **Roche Data Marketplace** is the enterprise-wide solution that connects data producers and data consumers.

A Single place to ...

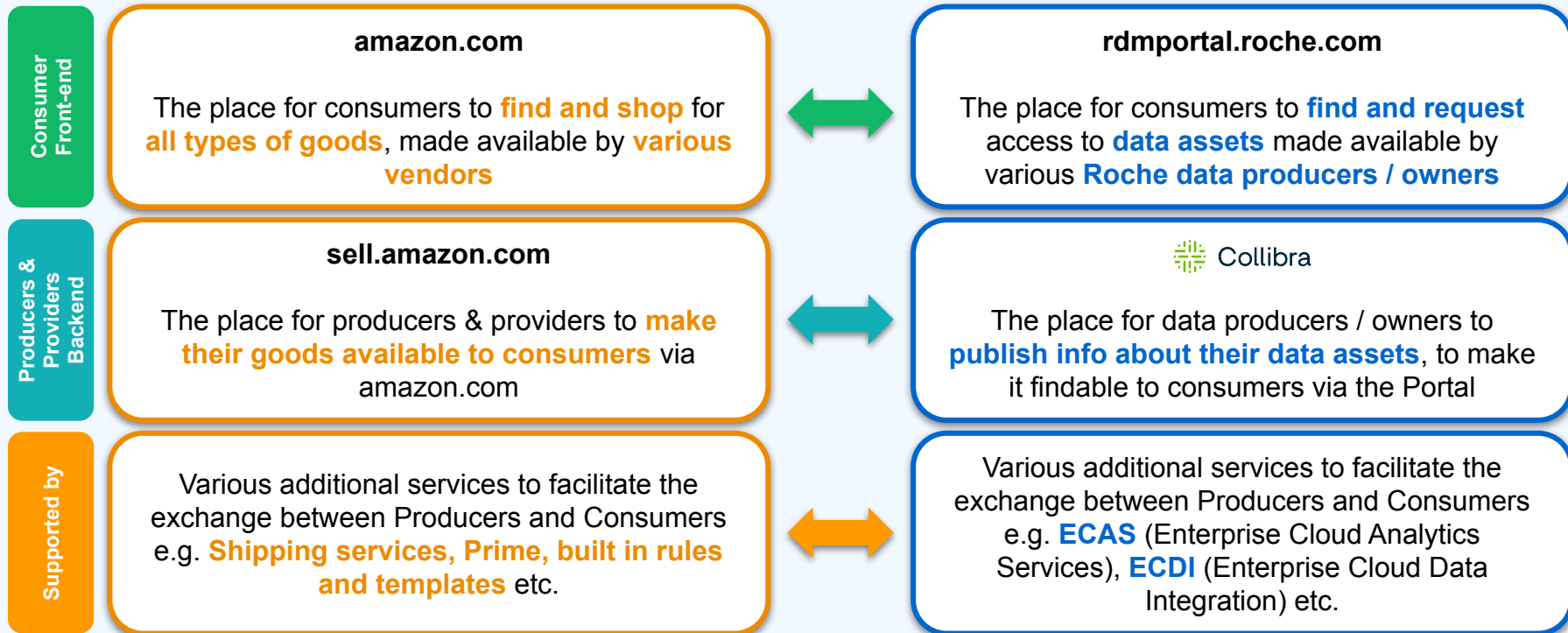
- Document
- Publish data
- Govern data
- Find data
- Request data

Which allow us to ...

- Support Data Governance
- Embodies FAIR principles

What is the Roche Data Marketplace?

Like Amazon, but for data: a comparison



The Roche Data Marketplace (RDM)



Find



Document and govern:

- **Published** data assets
- Data asset **lineage and structure**



Know what data is available, where it comes from, and how to access it

Understand



- **Business terms** and their **relationship** to our **data assets**



Know what the data means

Trust



- **Rules** applying to our data assets
- **Governance structure**
- **Roles & responsibilities**

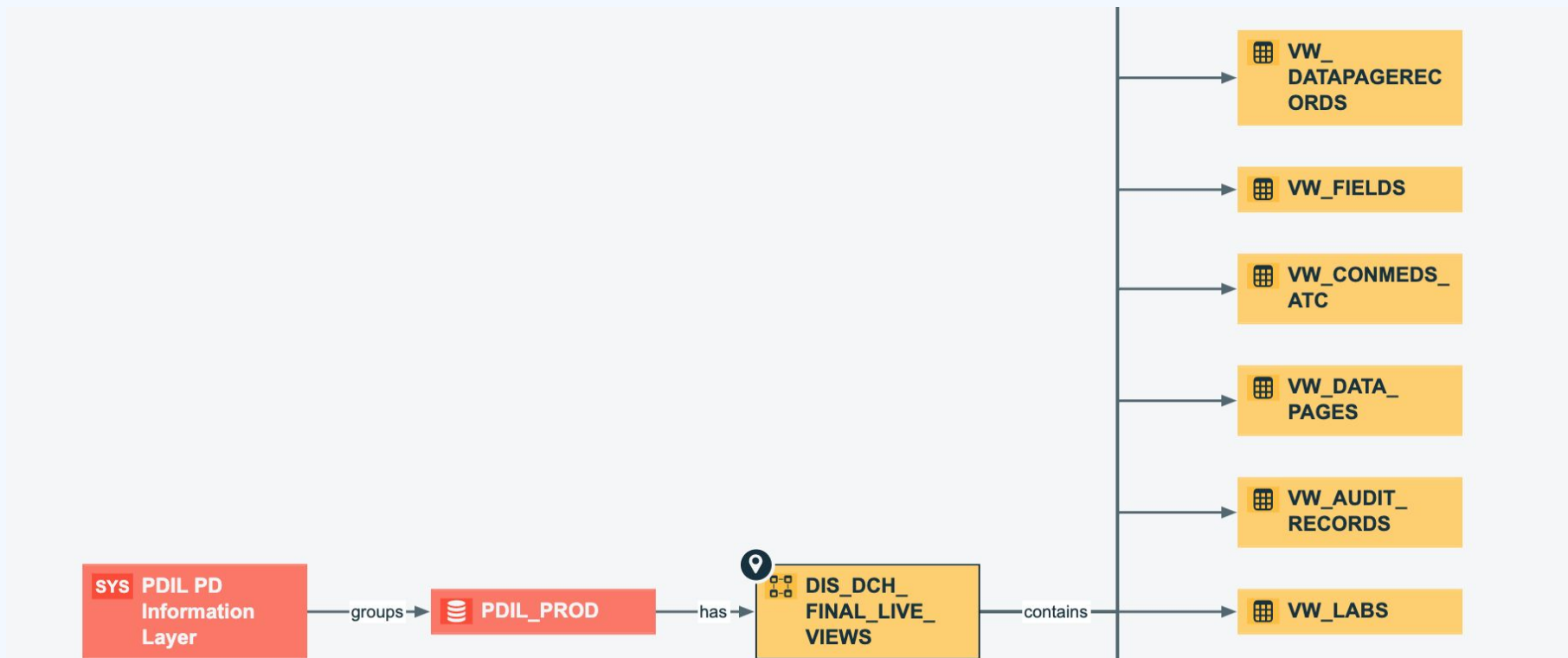


Know how complete and accurate the data is
Know who to contact if issues with the data

type what data
you're looking
for here

this tells you
that the data is
already
available for
querying

The Roche Data Marketplace (RDM)



Extending our Cataloging Capabilities



Leverage existing tools

- GitLab/GitHub is already well established within Roche to ensure version control of our code
- Use existing practices for code documentation (Roxygen / Doxygen)
- Roche Data Marketplace as our single cataloging solution



Ensure compatibility

Ensure such solution is compatible with the programming languages we use for Analytics and Machine Learning Models :

- R
- Python



Avoid additional Work

The solution should not create any extra work and should be well integrated within our existing ecosystem. Automation is a must and it should come as a plug and play solution for developers or analytics and machine learning models.



A full Data Lineage

- A full data catalogue
- Which includes data sources, data location
- An connected analytics and machine learning models to all the data items
- All condensed in one single place

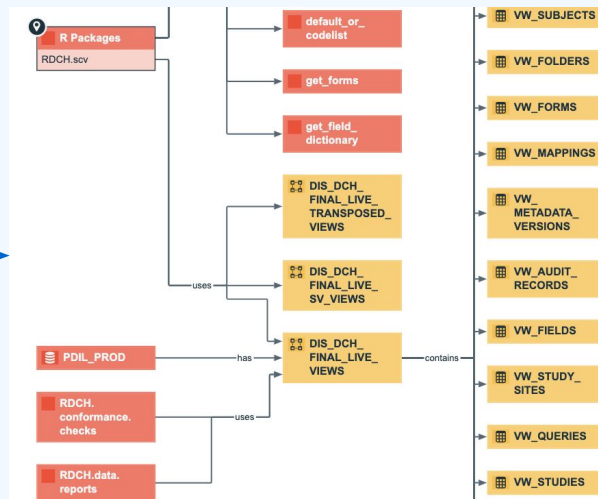
Our Solution



Use case : Cataloging R packages



```
#' @title SCV metadata get_folders function
#'
#' @description Fetches Rave FOLDERS metadata through DCH.
#'
#' @param study string
#' @param con (optional) RDCH.snowflake connection object
#' @param metadata_version_id (optional) string
#'
#' @return data.frame
#' @ctable DIS_DCH_FINAL_LIVE_VIEWS.VW_FOLDERS
#' @export
#'
#' @examples
```



RDCH.scv

Draft

Created on Jun 14, 2022, 5:11:17 AM

Last modified on Jun 22, 2022, 12:21:02 PM

Description

RDCH.scv helps you compute and extract data from DCH Study Clinical Views (SCV). RDCH is a key component to help you play and access DCH SCV data. RDCH.scv allows you to transpose SCV data automatically in a similar format to the classic Rave extracts. Additionally, all of the existing Rave study builds are available in SCV such as (Forms, Fields, Dictionaries, Visits) SCV functions. <https://rdch.pages.roche.com/rdch.packages/RDCH.scv/>

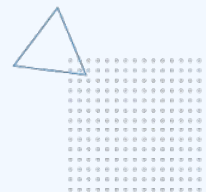
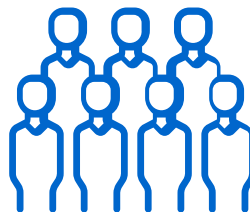
External Link

<https://code.roche.com/rdch/rdch.packages/RDCH.scv>

Since then



- Published 139 assets in the R Packages and R Functions catalog
- Published assets for Machine Learning models written in R and Python
- Our Data Scientists can now find existing analytics next to the data they are looking with a direct link to the source code



The text "Q&A" is centered in the image in a large, white, elegant serif font. The background is a solid blue color with a faint, out-of-focus image of two women's faces, one on the left and one on the right, looking towards the center.

Q&A

Doing now what patients need next