Coordination Centre

Data quality metrics in the context of DARWIN EU®

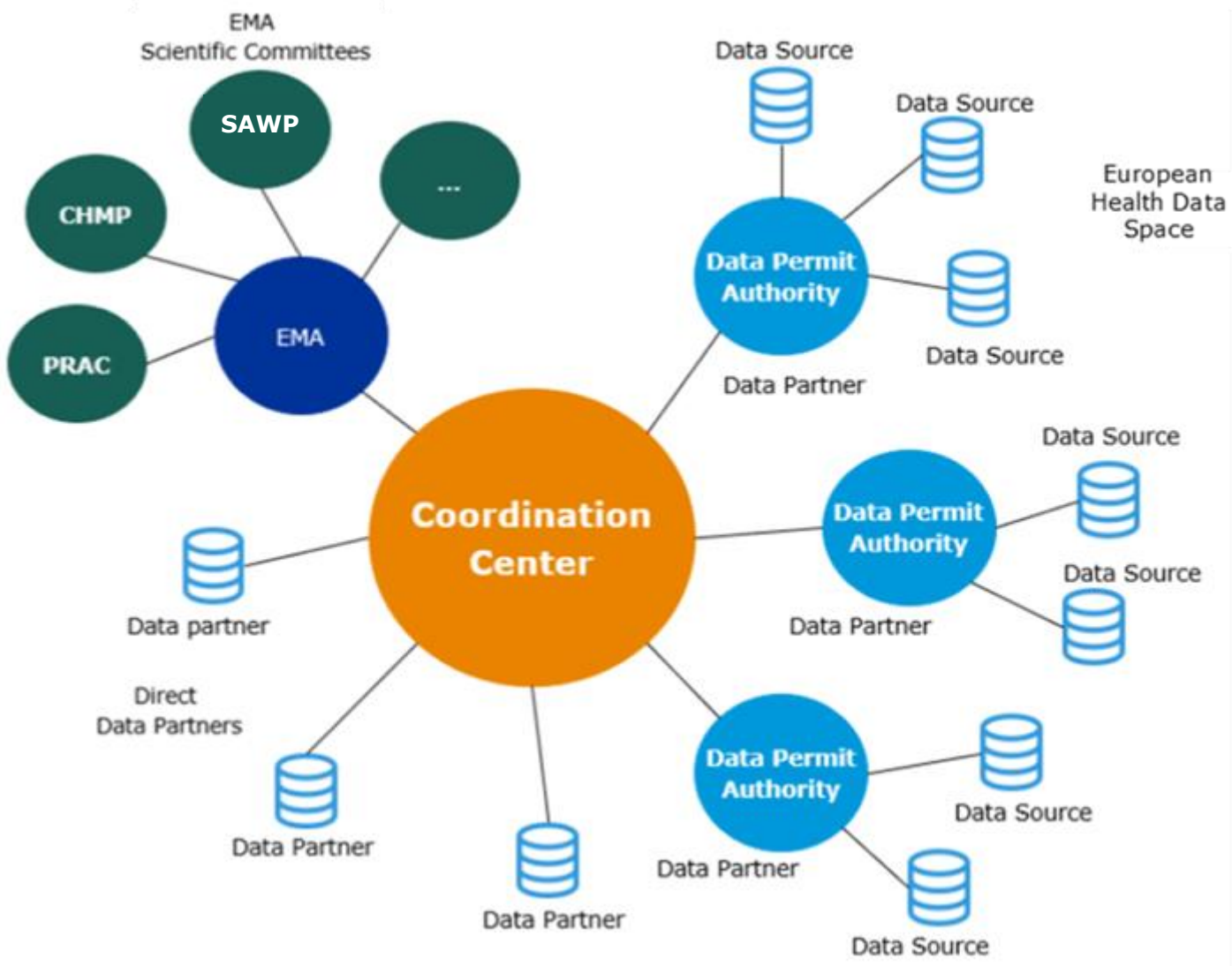
Maxim Moinat, DARWIN-EU® Coordinating Centre

2023-06-26

Disclosure

This presentation represents the views of the DARWIN EU® Coordination Centre only and cannot be interpreted as reflecting those of the European Medicines Agency or the European Medicines Regulatory Network.

DARWIN EU® is a federated **network** of **data**, **expertise** and **services** that supports better decision-making throughout the product lifecycle by generating reliable **evidence from real world healthcare data**



FEDERATED NETWORK PRINCIPLES

- Data stays **local**
- **Use of the OMOP CDM** to perform studies in a timely manner and increase consistency of results

Kahn Framework for DQA

eGEMs
Generating Evidence & Methods to improve patient outcomes

A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data

Michael G. Kahn, MD, PhD;[i] Tiffany J. Callahan, MPH;[ii] Juliana Barnard, MA;[ii] Alan E. Bauck;[ii] Jeff Brown, PhD;[iii] Bruce N. Davidson, PhD;[iv] Hossein Estiri, PhD;[v] Carsten Goerg, PhD;[ii] Erin Holve, PhD, MPH, MPP;[vi] Steven G. Johnson, MS;[vii] Siaw-Teng Liaw, MBBS, PhD, FRACGP, FACHI;[viii] Marianne Hamilton-Lopez, PhD, MPA;[ix] Daniella Meeker, PhD;[x] Toan C. Ong, PhD;[xi] Patrick Ryan, PhD;[xii] Ning Shang, PhD;[xiii] Nicole G. Weiskopf, PhD;[xiv] Chunhua Weng, PhD, FACMI;[xiii] Meredith N. Zozus, PhD;[xv] Lisa Schilling, MD[xi]

ABSTRACT

Objective: Harmonized data quality (DQ) assessment terms, methods, and reporting practices can establish a common understanding of the strengths and limitations of electronic health record (EHR) data for operational analytics, quality improvement, and research. Existing published DQ terms were harmonized to a comprehensive unified terminology with definitions and examples and organized into a conceptual framework to support a common approach to defining whether EHR data is 'fit' for specific uses.

Materials and Methods: DQ publications, informatics and analytics experts, managers of established DQ programs, and operational manuals from several mature EHR-based research networks were reviewed to identify potential DQ terms and categories. Two face-to-face stakeholder meetings were used to vet an initial set of DQ terms and definitions that were grouped into an overall conceptual framework. Feedback received from data producers and users was used to construct a draft set of harmonized DQ terms and categories. Multiple rounds of iterative refinement resulted in a set of terms and organizing framework consisting of DQ categories, subcategories, terms, definitions, and examples. The harmonized terminology and logical framework's inclusiveness was evaluated against ten published DQ terminologies.

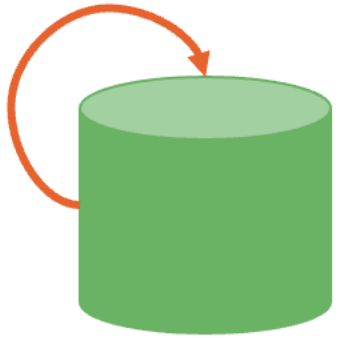
Table 1. Harmonized DQ Terms, Definitions, and Examples: Organized by Verification and Validation Contexts Within Categories and Subcategories

VERIFICATION		VALIDATION	
DEFINITION	EXAMPLE	DEFINITION	EXAMPLE
CONFORMANCE: DO DATA VALUES ADHERE TO SPECIFIED STANDARDS AND FORMATS?			
VALUE CONFORMANCE			
a. Data values conform to internal formatting constraints. b. Data values conform to allowable values or ranges.	a. Sex is only one ASCII character. b. Sex only has values "M," "F," or "U."	a. Data values conform to representational constraints based on external standards.	a. Values for primary language conform to ISO standards.
RELATIONAL CONFORMANCE			
a. Data values conform to relational constraints. b. Unique (key) data values are not duplicated. c. Changes to the data model or data model versioning.	a. Patient medical record number links to other tables as required. b. A medical record number is assigned to a single patient. c. Version 1 data does not include medical discharge hour.	a. Data values conform to relational constraints based on external standards.	a. Data values conform to all not-NULL requirements in a common multi-institutional data exchange format.
COMPUTATIONAL CONFORMANCE			
a. Computed values conform to computational or programming specifications.	a. Database- and hard-calculated Body Mass Index (BMI) values are identical.	a. Computed results based on published algorithms yield values that match validation values provided by external source.	a. Computed BMI percentiles yield identical values compared to test results and values provided by the CDC.
COMPLETENESS: ARE DATA VALUES PRESENT?			
a. The absence of data values at a single moment in time agrees with local or common expectations. b. The absence of data values measured over time agrees with local or common expectations.	a. The encounter ID variable has missing values. b. Gender should not be null. c. Medical discharge time is missing for three consecutive days.	a. The absence of data values at a single moment in time agrees with trusted reference standards or external knowledge. b. The absence of data values measured over time agrees with trusted reference standards or external knowledge.	a. The current encounter ID variable is missing twice as many values as the institutionally validated database. b. A drop in ICD-9CM codes matches implementation of ICD-10CM
PLAUSIBILITY: ARE DATA VALUES BELIEVABLE?			
UNIQUENESS PLAUSIBILITY			
a. Data values that identify a single object are not duplicated.	a. Patients from a single institution do not have multiple medical record numbers.	a. Data values that identify a single object in an external source are not duplicated.	a. An institution's CMS facility identifier does not refer to a multiple institutions.

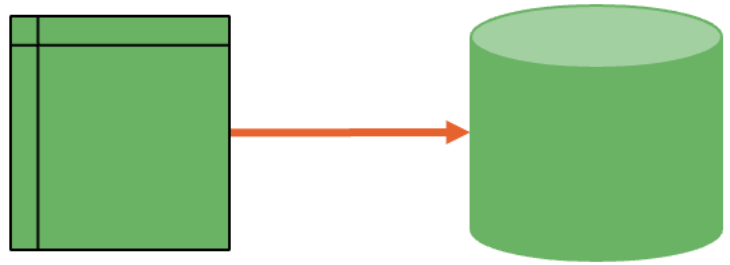
Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, Estiri H, Goerg C, Holve E, Johnson SG, Liaw ST, Hamilton-Lopez M, Meeker D, Ong TC, Ryan P, Shang N, Weiskopf NG, Weng C, Zozus MN, Schilling L. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC). 2016 Sep 11;4(1):1244. doi: 10.13063/2327-9214.1244. PMID: 27713905; PMCID: PMC5051581.

Kahn Framework – Two Contexts

Verification
*"focuses on how data values match expectations with respect to metadata **constraints**, **system assumptions**, and **local knowledge**"*



Validation
*"focuses on the alignment of data values with respect to relevant **external benchmarks**"*



Kahn et al., eGEMS 2016

Kahn Framework – Three Categories

Conformance

*"focuses on DQ features that describe the **compliance** of the representation of data against internal or **external formatting, relational, or computational** definitions", e.g.:*

- ✓ Diagnosis is linked to person table via valid foreign key
- ✓ All persons have a birthdate
- ✓ Given BMI is equal to calculated BMI value

Completeness

*"focuses on features that describe the **frequencies of data attributes present** in a data set without reference to data values", e.g.:*

- ✓ At least one diagnosis code given for a healthcare visit
- ✓ Fewer hospitalisations during the COVID lockdown

Plausibility

*"focuses on features that describe the **believability or truthfulness** of **data values**", e.g.:*

- ✓ Height and weight values are positive
- ✓ Medication administrated after birthdate

Kahn et al., eGEMS 2016

Where to begin with Data Quality?

CONTEXTS

CATEGORIES

	Verification	Validation
Plausibility	?	?
Conformance	?	?
Completeness	?	?

Clair Blacketer, Frank J Defalco, Patrick B Ryan, Peter R Rijnbeek, Increasing trust in real-world evidence through evaluation of observational data quality, *Journal of the American Medical Informatics Association*, Volume 28, Issue 10, October 2021, Pages 2251–2257, https://doi.org/10.1093/jamia/ocab132

An example plausibility check...

	Verification	Validation
Plausibility	?	?
Conformance	?	?
Completeness	?	?

The number and percent of records with a value in the *year_of_birth* field of the *person* table less than 1850.

...which we can make more generic

	Verification	Validation
Plausibility	?	?
Conformance	?	?
Completeness	?	?

The number and percent of records with a value in the *CDM field* of the *CDM table* less than *a low value*.

…which we can make more generic

	Verification	Validation
Plausibility	?	?
Conformance	?	?
Completeness	?	?

The number and percent of records with a value in the *quantity* field of the *drug_exposure* table less than 0.

An example completeness check...

	Verification	Validation
Plausibility	?	?
Conformance	?	?
Completeness	?	?

The number and percent of values which are mapped to a standard concept in the CONDITION_CONCEPT_ID field of the CONDITION_OCCURRENCE table.

…which we can make more generic

	Verification	Validation
Plausibility	?	?
Conformance	?	?
Completeness	?	?

The number and percent of values which are not mapped to a standard concept in the *CDM field* of the *CDM table*.

Resulting in a set of Check Types

Conformance

- Primary Key
- Foreign Key
- Data type
- CDM Field Exists
- Required Field

Database constraints

- Standard Valid Concept
- Domain
- Class

Completeness

- Records per person
- Field value
- Standard Concept
- Source Concept
- Source Value

Plausibility

- Gender
- Low Value
- High Value
- Unit concept

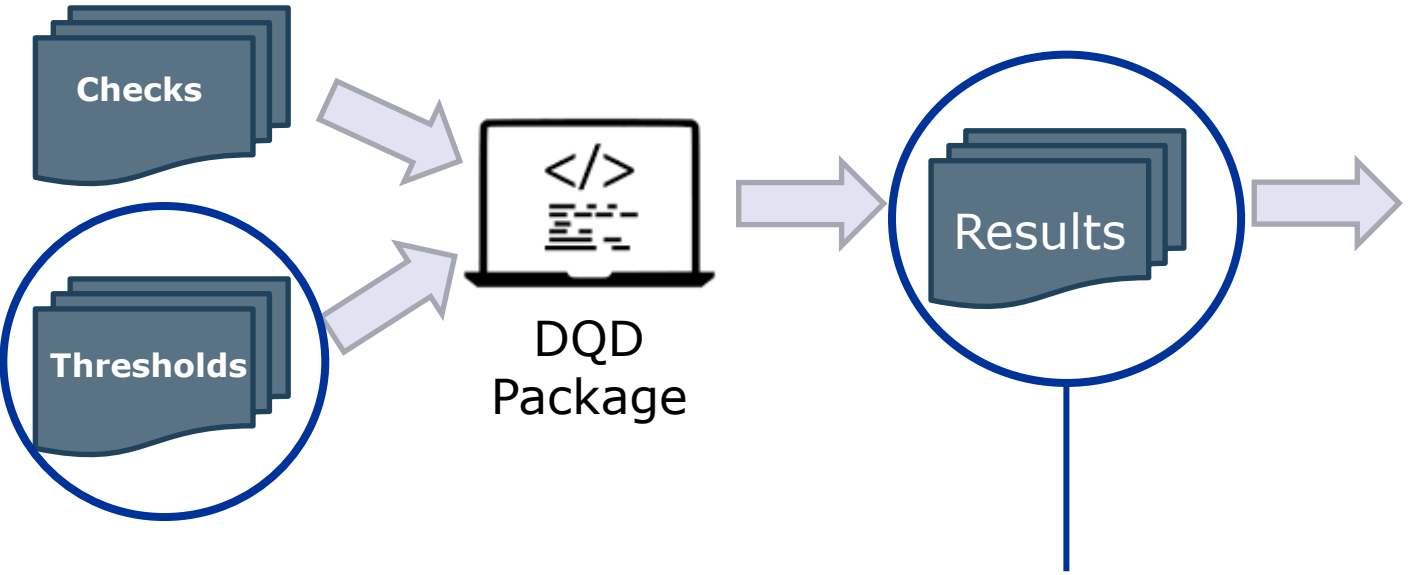
Examples of validation context

- During Life
- Temporal after

A total of >3000 checks

	Verification	Validation
Plausibility	1878	287
Conformance	681	104
Completeness	386	15

Data Quality Dashboard



To be shared as part of the onboarding process

DATA QUALITY ASSESSMENT

MEDICARE CLAIMS SYNTHETIC PUBLIC USE FILES (SYNPUFS)

DataQualityDashboard Version: 1.0.0
Results generated at 2021-11-25 05:41:37 in 14 hours

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	1859	1	1860	100%	89	198	287	31%	1948	199	2147	91%
Conformance	588	3	591	99%	97	0	97	100%	685	3	688	100%
Completeness	332	6	338	98%	12	2	14	86%	344	8	352	98%
Total	2779	10	2789	100%	198	200	398	50%	2977	210	3187	**93%**

https://data.ohdsi.org/DataQualityDashboardSynpuf/

How to decide if result are good enough?

% of records that violate check

0% = any violation → FAIL
100% = always PASS

Check Category	Check Description	Check Result	Decision Threshold	Pass / Fail	
Verification - Plausibility	The number and percent of records with a value in the YEAR_OF_BIRTH field of the PERSON table less than 1850.	0%	0%	PASS	
Verification - Plausibility	The number and percent of records with a value in the DAYS_SUPPLY field of the DRUG_EXPOSURE table less than 0.	0%	1%	PASS	
Verification - Plausibility	For Hemoglobin A1c percent, the number and percent of records with a value in the VALUE_AS_NUMBER field of the MEASUREMENT table less than 4.	0.01%	5%	PASS	
Verification - Completeness	The number and percent of records with a value of 0 in the standard concept field CONDITION_CONCEPT_ID in the CONDITION_OCCURRENCE table.	0.02%	5%	PASS	
Verification - Completeness	The number and percent of records with a value of 0 in the standard concept field UNIT_CONCEPT_ID in the MEASUREMENT table.	93.66%	5%	FAIL	

OHDSI Data Quality Dashboard (DQD)

OHDSI / DataQualityDashboard

Unwatch 7 | Star 5 | Fork 6

Code | Issues 16 | Pull requests 0 | Actions | Projects 1 | Wiki | Security | Insights | Settings

A tool to help improve data quality standards in observational data science. https://ohdsi.github.io/DataQualityDa... Edit

data-quality Manage topics

150 commits | 4 branches | 0 releases | 1 environment | 6 contributors | Apache-2.0

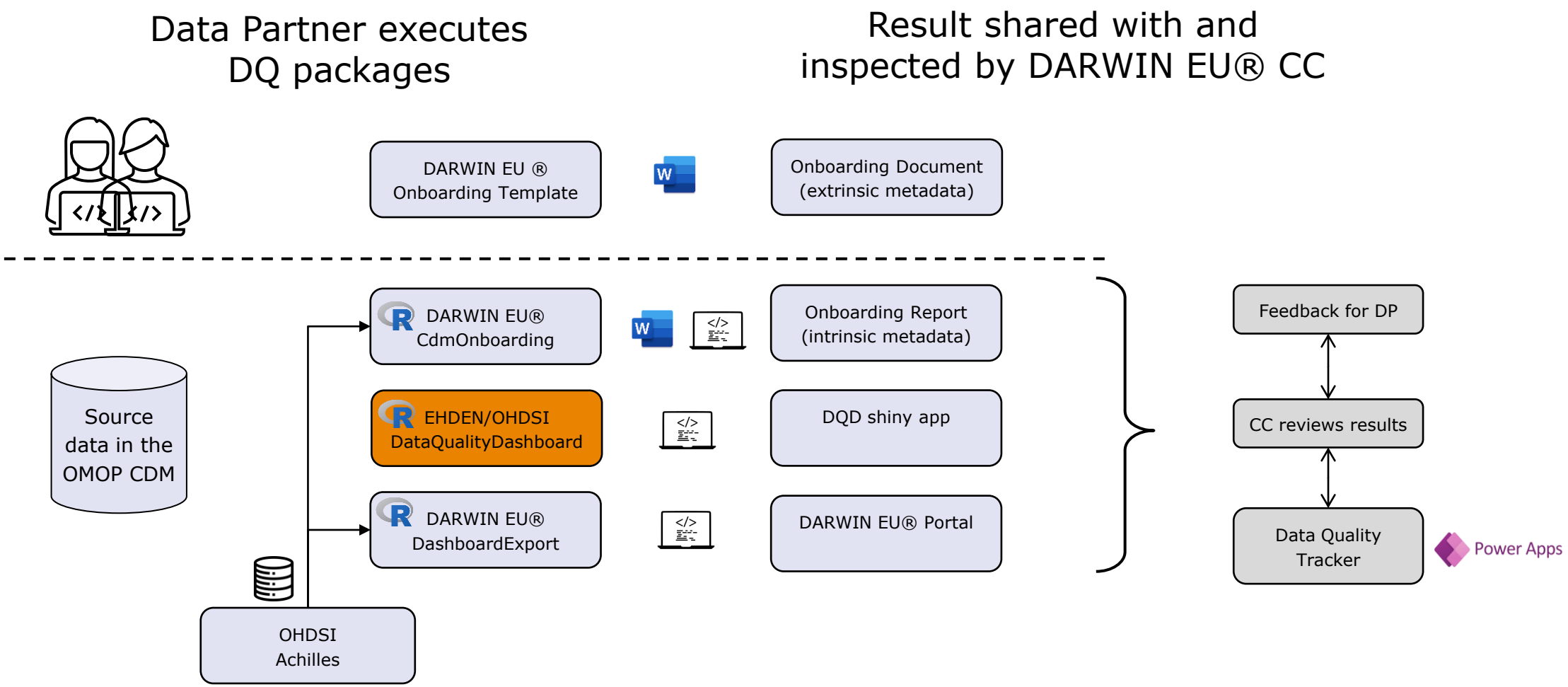
Branch: master | New pull request | Create new file | Upload files | Find File | Clone or download

alondhe Disable scientific notation in execution — Latest commit 4676653 12 hours ago

R	Disable scientific notation in execution	12 hours ago
docs	Added tablesToExclude parameter to allow skipping of tables if known ...	13 days ago
extras	Added screenshot	18 days ago
inst	Remove duplicates from concept level	4 days ago
man	Added tablesToExclude parameter to allow skipping of tables if known ...	13 days ago
tests	Added missing dbType in test	18 days ago
.Rbuildignore	updated R package wrappers	2 months ago
.gitignore	Fixes #30, making autocommit setting specific to connectionDetails db...	18 days ago
.travis.yml	Added devtools to travis. Added travis and codecov statuses to readme...	18 days ago
DESCRIPTION	Added pkgdown documentation. Added one testthat test.	19 days ago
DataQualityDashboard.Rproj	updated R package wrappers	2 months ago
LICENSE	Initial commit	3 months ago
NAMESPACE	Added pkgdown documentation. Added one testthat test.	19 days ago

https://github.com/OHDSI/DataQualityDashboard

DARWIN EU® Data Quality Assessment



More Information

DQ Packages and documentation:

https://github.com/ohdsi/DataQualityDashboard

https://github.com/darwin-eu/CdmOnboarding

https://github.com/darwin-eu/DashboardExport

Coordination Centre website

https://darwin-eu.org/

For questions to the Coordination Centre, please contact: enquiries@darwin-eu.org