

Data Dictionaries for Humans and Agents

Last modified: 2026-09-13 00:11:34 (PDT)

`data-dict.yaml` is a lightweight, human- and agent-readable data dictionary that records a collection of related tables: their contents, constraints, connections, and the specialized vocabulary needed to understand them. It is described at data-dict.tidyverse.org and developed openly by Posit as a companion to a self-contained CLI, `data-dict`.

The specification is intentionally small. It is a plain-text YAML 1.2 document that assumes data lives in Parquet or database tables, so it does not need to describe CSV wrinkles and it encourages compact, performant storage. It focuses on the most important structures (variable names, types, ranges, uniqueness, and relationships) and leaves the remainder to free-text fields that a human or an LLM can interpret. It does not itself clean data, but it validates that data and dictionary stay consistent, and it is designed to track understanding as it evolves.

A `data-dict` lives with the data it describes. You can use it before any data exists to make expectations concrete, when first encountering a new dataset to record what you learn, retrospectively to get the knowledge out of your head for collaborators, or to guard regularly updated data from unannounced changes. Future tooling will turn a dictionary into a browsable site, simulate data from it, generate a skeleton from Parquet or a database, and index many dictionaries at once.

The project is explicit about why now is the right time. AI lowers the cost of creating a dictionary (an agent can draft boilerplate and surface ambiguities) and raises the benefit (agents need the context that currently lives only in your head). LLMs also change what must be machine-readable: only the most important structures need strict encoding, unusual quirks can stay as prose that an agent reads.

The site shows the same idea through five worked examples (dabstep, elevators, foodbank, loan-application, otters) and notes inspirations that shaped the design (Frictionless Data, Hex semantic models, Snowflake semantic views, Soda, dbt tests, Data Package Standard, BIDS, and DDI), while distinguishing `data-dict.yaml` from a semantic model. For our purposes, it is a practical way to make an AI coding assistant work more accurately: give it the dictionary and it has the context it would otherwise have to ask for.