A **comprehensive data dictionary** goes beyond just column names and descriptions. Here’s what’s typically included, structured for clarity and utility:

**1. Basic Column Metadata**

| **Field** | **Description** | **Example** |
| --- | --- | --- |
| **Column Name** | Exact name in the dataset | num\_comments |
| **Description** | Purpose of the column | "Total number of comments on the post." |
| **Data Type** | Integer, string, timestamp, etc. | INTEGER |
| **Format** | How values are structured | Unix epoch (UTC) |

**2. Data Characteristics**

| **Field** | **Description** | **Example** |
| --- | --- | --- |
| **Allowed Values** | Valid ranges/enumerations | 0 to ∞ |
| **Default Value** | Fallback if empty | NULL or 0 |
| **Missing Data** | How nulls are represented | NaN, NULL, "" |
| **Uniqueness** | Whether values are unique | id is primary key |

**3. Relationships & Dependencies**

| **Field** | **Description** | **Example** |
| --- | --- | --- |
| **Parent Table** | Linked tables (if relational) | posts → comments |
| **Foreign Keys** | Columns linking to other tables | subreddit\_id → subreddits.id |
| **Derived From** | Source of calculated fields | score = ups - downs |

**4. Business Context**

| **Field** | **Description** | **Example** |
| --- | --- | --- |
| **Owner** | Team/person responsible | "Data Engineering Team" |
| **Sensitivity** | PII/confidentiality level | "Public" / "Internal Only" |
| **ETL Process** | How data is populated | "Daily API pull from Reddit" |

**5. Usage Notes**

| **Field** | **Description** | **Example** |
| --- | --- | --- |
| **Analytic Use** | Common queries/filters | "Filter over\_18=True for NSFW content" |
| **Known Issues** | Data quality caveats | "downs is always 0 due to Reddit API" |
| **Example Value** | Sample data | num\_comments: 42 |

**6. Versioning & Governance**

| **Field** | **Description** | **Example** |
| --- | --- | --- |
| **Last Updated** | Date of latest refresh | 2024-03-15 |
| **Change Log** | History of modifications | "Added treatment\_tags on 2024-01-10" |

**Why This Matters**

* **Reproducibility**: Documents assumptions (e.g., "score excludes downvotes").
* **Onboarding**: Helps new team members understand the data.
* **Quality Control**: Flags known issues (e.g., "thumbnail URLs expire after 30 days").

**Tools to Automate**

* **Python**: Use pandas-profiling or D-Tale to auto-generate summaries.
* **SQL**: Comment tables/columns directly in database schemas.