Phase 1

📘 Phase 1 Documentation — Canonical Modeling & Aggregation Decisions

Purpose of Phase 1 (in simple terms)

In Phase 1, we took multiple operational clinical reports (each designed for a specific operational purpose) and converted them into **clean, consistent, subject-centric summary datasets**.

Each report:

- has a different level of detail (subject, visit, page, record, drug, event)
- uses different column names
- may contain partial or missing information

Our goal in Phase 1 was **not analytics yet**, but to:

- decide **what entity truly matters** for each report
- choose **canonical keys** (how rows are uniquely identified)
- extract **meaningful signals**
- remove noise and ambiguity
- save clean intermediate datasets for Phase 2

🔑 What are “Canonical Keys”?

Canonical keys define **what one row represents** after aggregation.

Example:

- If the key is `subject_id` → one row = one subject
- If the key is `subject_id + study_id` → one row = one subject within one study

We **intentionally did NOT force the same keys across all datasets**, because each report serves a different operational purpose.

📊 Category-by-Category Decisions

1️⃣ CPID / EDC Metrics

What this report is about

Tracks **data entry and completion metrics** per subject in EDC.

Canonical keys

```
study_id + site_id + subject_id
```

Why

- CPID metrics are **study-specific**
- Site ownership matters
- Metrics are tracked per subject

Descriptive columns kept

- Country
- Source file (traceability)

What we did with the rest

- Aggregated numeric metrics (counts, durations)
- Removed header/footer rows
- Enforced numeric and non-negative checks

2️⃣ Global Missing Pages Report

What this report is about

Shows **missing CRF pages** at visit / form level for subjects.

Canonical keys

```
study_id + site_id + subject_id
```

Why

- Missing pages are **subject-specific**
- Site and study are required for follow-up
- Visit/form details are too granular for Phase 2

Descriptive columns kept

- Country

What we did with the rest

- Dropped rows without subject
- Selected the main metric: **maximum days missing per subject**
- Ignored visit-level noise after aggregation

3️⃣ Visit Projection Tracker

What this report is about

Shows **future / overdue visits** per subject.

Canonical keys

```
site_id + subject_id
```

Why

- Operationally driven by **subject and site**
- Study is not explicitly emphasized in the document
- Subject may exist across studies but site ownership is critical

Descriptive columns kept

- Country
- Study (only as context, not enforced)

What we did with the rest

- Aggregated visit-level rows
- Used **maximum days outstanding** per subject
- Ignored derived duplicate metrics

4️⃣ Missing Lab Name & Missing Ranges

What this report is about

Identifies **lab data quality issues** (missing lab names, ranges, units).

Canonical keys

```
subject_id + site_id
```

Why

- Issues are followed up **per subject**
- Site owns resolution
- Study is contextual, not operational

Descriptive columns kept

- Country
- Lab category
- Study (context only)

What we did with the rest

- Converted issue text into **countable issue types**
- Created metrics:
- total lab issues
- missing lab name count
- missing range/unit count

5️⃣ SAE Dashboard

What this report is about

Tracks **Serious Adverse Event (SAE) discrepancies** and their resolution status.

Canonical keys

```
subject_id + study_id + site_id
```

Why

- SAEs are **patient-centric**
- Study context is regulatory-critical
- Site is responsible for action

Descriptive columns kept

- Country
- Report update required flag

What we did with the rest

- Translated multiple status columns into **open vs resolved signals**
- Created metrics:
- total SAE discrepancies
- open SAE count
- review pending / action pending / case open counts

6️⃣ Inactivated Forms & Loglines

What this report is about

Audit-style log of **inactivated data pages / records**.

Canonical keys

```
subject_id
```

Why

- This is a **log dataset**
- Site and study are often missing
- Subject is the only consistently meaningful anchor

Descriptive columns kept

- Country
- Study (if present)
- Site (only if available)

What we did with the rest

- Strictly used only columns documented in the guidance
- Detected inactivation via `audit_action`
- Counted **number of inactivated records per subject**
- Did NOT drop rows due to missing site/study

7️⃣ Compiled EDRR

What this report is about

Pre-compiled **open issue count per subject**.

Canonical keys

```
subject_id + study_id
```

Why

- Already aggregated at subject level
- Study context matters
- No site information provided

Descriptive columns kept

- None (minimal dataset)

What we did with the rest

- Treated the metric as authoritative
- Used defensive aggregation (`max`) in case of duplicates

8️⃣ Coding Reports — MedDRA (Adverse Events)

What this report is about

Tracks **AE term coding completeness** using MedDRA.

Canonical keys

```
subject_id + study_id
```

Why

- Coding completeness is assessed per subject
- Study determines dictionary and regulatory context

Descriptive columns kept

- Dictionary
- Dictionary version

What we did with the rest

- Converted term-level rows into subject-level counts:
- terms requiring coding
- coded terms
- uncoded required terms

9️⃣ Coding Reports — WHO-DRA / WHODrug (Medications)

What this report is about

Tracks **medication and therapy coding completeness**.

Canonical keys

```
subject_id + study_id
```

Why

- Medication safety is subject-centric
- Study context is required
- Mirrors MedDRA logic but for drugs

Descriptive columns kept

- Dictionary
- Dictionary version

What we did with the rest

- Same pattern as MedDRA
- Created drug-specific coding backlog metrics

Phase 2 goals (Preview)

- Join all 9 aggregated datasets
- Create a **single master subject table**
- Add cross-category data quality checks
- Enable EDA and AI/ML automation in Phase 3 & 4
