FABIO FIDANZA STUDIO

Feed Basket Schema Gap Analysis Report

Fabio Fidanza, Antonio Lepore

Towards a Unified Reference Schema for Interoperable Feed Basket Data Collection and Research.

Comparative analysis of ILRI, Cornell, and ERA datasets against the Draft Schema

This analysis has been created with support from Claude Sonnet 4.6.

Table of Contents

Table of Contents
1. Executive Summary
2. Context and Objectives
    2.1 The LD4D FeedBasketRegistry Initiative
    2.2 Primary Use Cases
    2.3 Stakeholder Roles
3. Datasets Analysed
    3.1 ILRI Feed Library DB
    3.2 ILRI Livestock DB
    3.3 Cornell SCIOUTPUT
    3.4 ERA Livestock Data Dictionary
4. Draft Schema — Overview and Strengths
5. Gap Analysis
    Priority overview
    5.1 Environmental Output Controlled Vocabularies — HIGH
    5.2 Publication Provenance — HIGH
    5.3 Location Uncertainty — MEDIUM
    5.4 Agroecological Zone — MEDIUM
    5.5 Climate Data at Site Level — MEDIUM
    5.6 Agronomy Context — LOW
    5.7 Digestibility — LOW
    5.8 Soil Profile — LOW
6. Cross-Dataset Coverage Matrix
7. Recommendations
    7.1 For the First Stable Version (v1.0)
    7.2 For Version 1.x (Optional Modules)
8. Proposed Extensions — Entity Summary
9. Conclusions

1. Executive Summary

The LD4D Feed Basket Schema Subgroup is developing a reference database schema intended to serve as a common language for the collection, exchange, and analysis of animal feed basket data across multiple platforms and research programmes. A draft schema (hereinafter referred to as **the Draft**) has already been produced and represents a conceptually sound, highly normalised data model.

This report presents the results of a systematic gap analysis conducted by comparing the Draft schema against four concrete datasets currently available in the working repository: two ILRI operational databases, the Cornell SCIOUTPUT dataset, and the ERA Livestock data dictionary. The analysis identifies dimensions that are absent or insufficiently specified in the Draft, prioritises them by impact on the two primary use cases (empirical studies and model-driven analysis), and proposes targeted, minimally invasive additions.

Key finding: The Draft is structurally solid and does not require redesign. Three additions — explicit environmental output controlled vocabularies, a publication provenance entity, and extended location fields — are sufficient to make the schema mappable to all four analysed datasets and to support both use cases end-to-end.

2. Context and Objectives

2.1 The LD4D FeedBasketRegistry Initiative

LD4D (Linked Data for Development) is a multi-institutional initiative aiming to improve the interoperability of livestock-related data. The FeedBasketRegistry is one of four subgroups within the initiative, and its Schema subgroup is responsible for producing a reference data model.

A **feed basket** is defined as the set of feed items consumed by a defined animal group within a given spatial and temporal window, together with their quantities, nutritional composition, and associated context data.

2.2 Primary Use Cases

The schema must support two distinct analytical workflows:

- **Empirical studies** — correlating recorded feed baskets with production outputs (milk, meat) and environmental outputs (enteric methane, manure emissions, nitrogen excretion), using historical datasets.
- **Model-driven analysis** — applying nutritional formulas and emission factors derived from empirical data to estimate outputs from partial or aggregated feed basket descriptions.

2.3 Stakeholder Roles

The vision document identifies five stakeholder profiles whose needs the schema must address:

Stakeholder	Core need
Data Collector	A structured template for field data collection
Data Aggregator	Combine datasets from multiple sources into consistent indicators
Data Applicator	Query datasets for specific outcome variables (e.g. GHG indicators)
Researcher	Generate and share interoperable research datasets
Application Designer	A canonical model to implement in a data collection app

3. Datasets Analysed

Four datasets were analysed as representative of the current data landscape. They span a range of granularities, contexts, and institutional sources.

Dataset	Source	Type	Geographic scope	Primary unit
ILRI Feed Library DB (feed_library_db)	ILRI / Nepal	Operational app DB	Nepal (province → district → municipality)	Feed record per feed × location × season
ILRI Livestock DB (feed_livestock_db)	ILRI / Ethiopia	Research sample DB	Africa (GPS coordinates)	Sample with full agronomy context
Cornell SCIOUTPUT (ca2006)	Cornell University	Aggregate model dataset	Global (country × production system)	kg DM / TLU / year
ERA Livestock Data Dictionary (2023)	CGIAR / ERA	Literature meta-analysis	Global (research publications)	Experiment × treatment × outcome

3.1 ILRI Feed Library DB

An operational web application deployed in Nepal. The schema is relatively simple, reflecting an app-centric design rather than a research-oriented one.

- Feed catalogue: name, category, scientific name, local name, environmental impact description.
- Feed records: nutritional values (DM, ME, CP, NDF, ADF, ADL, Ca, P, IVOMD) stored twice — estimated values (field) and laboratory-confirmed values (`lab_*` prefix).
- Location: four-level administrative hierarchy (country → province → district → municipality).
- No animal group characterisation; no production or environmental outputs.

3.2 ILRI Livestock DB

A research-oriented database containing sample-level feed composition data with rich agronomic context, used for ILRI field studies primarily in Africa.

- Full GPS coordinates, agroecological zone, elevation.
- Agronomy: cropping system, fertilizer type and N/P/K amounts (kg/ha), soil type, irrigation.
- Plant taxonomy: genus, species, variety, plant part, growth stage (week after planting).
- Nutritional variables: DM, ash, N, CP, NDF, ADF, ADL, GE, OM, in-vitro DMD, yield, ME.
- Provenance: project name, funding agency, sample code, researcher email.
- Workflow: `tbl_feed_request` → `tbl_feed_staging` → `tbl_feed_main` (approval pipeline).

3.3 Cornell SCIOUTPUT

A global aggregate dataset at the level of country × livestock production system (dairy bovine, beef bovine, dairy sheep/goat, meat sheep/goat, pigs, poultry). Each record represents modelled annual averages per Tropical Livestock Unit (TLU).

- Feed basket expressed as four macro-categories: grazing, stover, occasional, grains — with grains disaggregated into 10 crop types (barley, maize, pulses, rice, sorghum/millet, soy, wheat, other cereals, oil seeds, other crops).
- Nutritional summary of the whole diet: CP (g/kg DM) and ME (MJ/kg DM).
- Production outputs: milk yield and meat yield (kg/TLU/year).
- Environmental outputs: enteric CH4, manure CH4, manure N2O (applied, managed, pasture-applied, solids), and N excretion — all in kg CO2eq/TLU/year or kg/TLU/year.
- **This is the only dataset that directly links feed basket → production → GHG emissions.**

3.4 ERA Livestock Data Dictionary

A meta-analysis framework for extracting structured data from scientific publications. Each record traces back to a specific publication, experiment, site, and treatment.

- **Publication provenance:** DOI, journal, author, project, linked publications (up to 4).
- **Site:** GPS with uncertainty fields (lat/lon uncertainty in metres, or bounding radius), elevation, slope, location precision qualifier.
- **Climate:** mean annual rainfall (MAP), mean annual temperature (MAT), rainy season start/end/length (up to 2 seasons), mean seasonal precipitation.
- **Soil:** sand, silt, clay (%), pH, depth upper/lower, texture class (USDA).
- **Herd:** species, breed, sex, production stage, parity, days in milk, age, body weight, number of animals, replicates.
- **Diet:** item type, item name, source, processing (mechanical/chemical/biological/thermal/dehydration), amount, unit (per day per animal), ad libitum flag.
- **Diet composition:** nutritional variable, value, unit, method, DM basis.
- **Digestibility:** per diet item or whole diet, variable, method, DM basis.
- **Outcomes:** variable name, value, unit, error, error type, number of reps.

4. Draft Schema — Overview and Strengths

The Draft schema (Feed Basket Schema DRAFT, 2025) is a conceptually mature, fully normalised relational model. Its main strengths are:

- **Multi-level feed taxonomy** — `feed_term` → `feed_component` → `feed_material` → `feed_composite`, with external linkage fields to Feedipedia, ILRI IID, emission factor databases, water use, and land use registries.
- **Hierarchical processing model** — `feed_process` and `feed_process_profile` capture ingredient-level, group-level, and whole-diet processing (mechanical, chemical, biological, thermal). This is essential for correctly mapping feeds to external nutritional tables where processed and unprocessed variants differ.
- **Separation of delivery and share** — `feed_basket_item_delivery` (actual amounts offered, e.g. kg/d/animal) is kept separate from `feed_basket_item_share` (proportional contribution to diet). This distinction is not present in any of the four analysed datasets but is conceptually important for digestibility studies.
- **Feed target abstraction** — `feed_target` allows grouping basket items into `whole_diet` / `single_element` / `custom_group` for nutritional measurement, enabling analysis at different granularities within the same basket.
- **Full herd characterisation** — `herd_description` with subherd elements (species, breed, age stage, parity, body weight, number of animals) and `herd_target` for analysis groupings.
- **Hierarchical time periods** — `time` table with `parent_time_id` supports nesting (e.g. production phase within annual period).
- **Outcome table** — generic `outcomes` entity with `out_variable` (CV), value, unit, and error fields.

Note: The Draft schema is recommended as the foundational reference model. The additions proposed in Section 5 are targeted extensions, not redesign proposals.

5. Gap Analysis

The following gaps were identified by attempting to map each of the four datasets to the Draft schema and identifying fields that could not be accommodated without schema extensions or loss of information.

Priority overview

Gap	Datasets affected	Use case impact	Priority
Environmental output CVs (CH4, N2O, N excretion)	Cornell, ERA	GHG analysis, model-driven	🔴 HIGH
Publication provenance (DOI, journal, linked pubs)	ERA, Cornell, ILRI Livestock	Research use, data lineage	🔴 HIGH
Location uncertainty (radius / bounding box)	ERA	Spatial analysis, research	🟡 MEDIUM
Agroecological zone	ILRI Livestock, ERA	Contextualisation, modelling	🟡 MEDIUM
Climate data (MAP, MAT, seasons)	ERA	Model-driven analysis	🟡 MEDIUM
Agronomy context (fertilizer, cropping system, soil)	ILRI Livestock, ERA	Research, ILRI mapping	🟢 LOW
Digestibility entity (explicit table)	ERA	Nutritional research	🟢 LOW
Soil profile (texture, pH, depth)	ERA	Soil-feed interaction research	🟢 LOW

5.1 Environmental Output Controlled Vocabularies — HIGH

The Draft schema includes an `outcomes` table with an `out_variable` field designated as a controlled vocabulary (CV). However, neither the CV values nor the unit conventions for greenhouse gas and nitrogen outputs are defined. This is a critical gap for both primary use cases.

The Cornell dataset demonstrates the minimum set of environmental variables required for a complete GHG analysis at the livestock system level:

Proposed CV value	Description	Unit (Cornell)	IPCC source
`enteric_ch4`	Enteric fermentation methane	kg CO2eq / TLU / year	Tier 1/2
`manure_ch4`	Methane from manure management	kg CO2eq / TLU / year	Tier 1/2
`manure_n2o_applied_crops`	N2O from manure applied to cropland	kg CO2eq / TLU / year	Tier 1
`manure_n2o_managed`	N2O from manure management systems	kg CO2eq / TLU / year	Tier 1
`manure_n2o_pasture`	N2O from manure deposited on pastures	kg CO2eq / TLU / year	Tier 1
`n_excretion`	Nitrogen excreted per animal	kg N / TLU / year	Tier 1/2
`milk_yield`	Milk production	kg / TLU / year	—

Proposed CV value	Description	Unit (Cornell)	IPCC source
`meat_yield`	Meat (carcass) production	kg / TLU / year	—
`liveweight_gain`	Average daily liveweight gain	kg / day / animal	—

No structural change to the `outcomes` table is required. The fix is to publish a reference CV list as part of the schema specification, along with standard unit conventions and the IPCC tier that each variable corresponds to.

> The `out_error` field in the outcomes table should also document accepted error types: SD, SE, 95%CI, CV — in line with ERA conventions.

5.2 Publication Provenance — HIGH

The current `project_metadata` entity is a placeholder that does not capture bibliographic information. At least 50% of the datasets analysed (ERA, Cornell, ILRI Livestock) originate from peer-reviewed research or are associated with specific publications. Without structured provenance, it is impossible to:

- Trace a data point back to its primary source.
- Link related studies that report results from the same field experiment.
- Comply with FAIR data principles (Findable, Accessible, Interoperable, Reusable).

Proposed addition — new `publication` table linked to `project_metadata`:

```
publication

  publication_id    PK

  project_id        FK -> project_metadata (nullable)

  doi               varchar  -- strip https://doi.org/ prefix

  url               varchar  -- for publications without DOI

  author_last       varchar  -- first author surname

  year_published    integer  -- YYYY

  journal           varchar

  linked_doi_1      varchar  -- DOI of related publication (same experiment)

  linked_doi_2      varchar

  linked_doi_3      varchar

  linked_doi_4      varchar
```

> The design follows ERA's `Pub.Out` structure. `linked_doi` fields cover the common case where multiple papers report results from the same field trial.

5.3 Location Uncertainty — MEDIUM

ERA systematically records the spatial precision of each site, which is essential for any analysis that uses geographic overlays (e.g. climate zones, soil maps, IPCC region classification). The current `location` entity in the Draft has only `lat/lon` and `geometry`.

Proposed additions to `location` (all nullable, backward-compatible):

```
location  -- additions

  location_precision_type   CV: {GPS_point, facility, city, district, region, country}

  spatial_uncertainty_m     float  -- radius in metres (square/circular region)

  lat_uncertainty_m         float  -- height of bounding box (asymmetric region)

  lon_uncertainty_m         float  -- width of bounding box (asymmetric region)
```

-- Note: spatial_uncertainty_m and lat/lon_uncertainty_m are mutually exclusive

> These fields follow ERA's `Site.Buffer.Manual`, `Site.Lat.Unc`, and `Site.Lon.Unc` conventions. They add no mandatory burden on apps that record precise GPS coordinates.

5.4 Agroecological Zone — MEDIUM

The agroecological zone (AEZ) is a standard CGIAR classification used in both the ILRI Livestock DB and the ERA dataset. It provides a coarser spatial context than GPS coordinates and is often the primary grouping variable in regional analyses.

Proposed addition to `location`:

```
location  -- addition

  agroecological_zone   varchar  -- e.g. "Tepid to cool sub-humid"
```

5.5 Climate Data at Site Level — MEDIUM

ERA records long-term climate variables reported by the study authors at the site level. These differ from values that could be derived algorithmically from GPS coordinates because they reflect conditions as understood and reported in the original study. They are particularly important for model-driven analysis.

Proposed new optional table `location_climate`:

```
location_climate

  location_id             FK -> location

  mean_annual_rain_mm     integer  -- MAP (long-term average)

  mean_annual_temp_c      float    -- MAT (long-term average)

  n_rainy_seasons         integer  -- 1, 2, or year-round

  season_1_start          varchar  -- month name or code

  season_1_end            varchar

  season_1_length_days    integer

  season_2_start          varchar  -- nullable if n_rainy_seasons = 1

  season_2_end            varchar

  season_2_length_days    integer

  mean_seas_precip_s1_mm  integer

  mean_seas_precip_s2_mm  integer
```

5.6 Agronomy Context — LOW

The ILRI Livestock DB records detailed agronomic conditions for each feed sample: the cropping system, fertilizer type and application rates (N, P, K in kg/ha/season), soil type, and irrigation method. The Draft `farming_system` entity uses two generic descriptors that are too vague to accommodate this.

Proposed new optional table `agronomy_context`:

```
agronomy_context

  agronomy_id           PK

  farming_system_id     FK -> farming_system (nullable)

  basket_item_id        FK -> feed_basket_item (nullable)

  cropping_system       CV: {monoculture, intercropping, rotation, ...}

  irrigation_type       CV: {rainfed, irrigated, supplemental, ...}

  soil_type             varchar

  fertilizer_type       varchar

  fertilizer_n_kg_ha    float

  fertilizer_p_kg_ha    float

  fertilizer_k_kg_ha    float
```

5.7 Digestibility — LOW

ERA explicitly tracks digestibility measurements (in-vitro or in-vivo) per diet item or whole diet, with method and basis fields. The ILRI Livestock DB includes in-vitro DMD as a standard nutritional variable.

No new table is strictly necessary. The fix is to define digestibility variables in the `feed_nutrition_measured` CV and to ensure the `method` field documents the technique used (e.g. Tilley-Terry, Daisy Incubator, in-vivo total collection).

Suggested CV values to add: `in_vitro_dmd`, `in_vivo_dmd`, `in_vivo_dom`.

5.8 Soil Profile — LOW

ERA records sand, silt, and clay percentages, pH, and depth horizon for each site. Relevant only for research use cases exploring soil-forage relationships.

Proposed new optional table `soil_profile`:

```
soil_profile

  location_id      FK -> location

  depth_upper_cm   float

  depth_lower_cm   float

  sand_pct         float

  silt_pct         float

  clay_pct         float

  ph               float

  texture_class    CV: {USDA classes}
```

6. Cross-Dataset Coverage Matrix

Legend: ✓ = fully covered, ~ = partially covered, ✗ = absent

Schema dimension	Draft	ILRI Library	ILRI Livestock	Cornell	ERA
Feed taxonomy (name/category)	✓	~	✓	~	✓
Scientific taxonomy (genus/species/variety)	✓	~	✓	✗	✓
Feed processing model	✓	✗	~	✗	✓
Nutritional composition (DM, CP, ME, NDF…)	✓	✓	✓	~	✓
In-vitro / in-vivo digestibility	~	~	~	✗	✓
Feed amounts (kg/d/animal or proportional)	✓	✗	✗	✓	✓
Herd characterisation (species/breed/stage)	✓	✗	✗	~	✓
Production outputs (milk, meat)	✓	✗	✗	✓	✓
Environmental outputs (CH4, N2O, N excretion)	~*	✗	✗	✓	~

Schema dimension	Draft	ILRI Library	ILRI Livestock	Cornell	ERA
Location (lat/lon or admin hierarchy)	✓	✓	✓	~	✓
Location uncertainty	✗	✗	✗	✗	✓
Agroecological zone	✗	✗	✓	✗	✓
Climate (MAP, MAT)	✗	✗	✗	✗	✓
Soil profile	✗	✗	✗	✗	✓
Agronomy (fertilizer, cropping system)	✗	✗	✓	✗	✓
Time period (year, season)	✓	✓	✓	✓	✓
Project metadata	✓	~	~	~	✓
Publication provenance (DOI, journal)	✗	✗	~	✗	✓

** The Draft `outcomes` table structure supports environmental outputs but the controlled vocabulary is not yet defined.*

7. Recommendations

7.1 For the First Stable Version (v1.0)

Three targeted additions with no structural impact on the core model:

Action	Description	Effort
Define outcomes CV	Publish a reference list of standard `out_variable` values for production (`milk_yield`, `meat_yield`, `liveweight_gain`) and environmental (`enteric_ch4`, `manure_ch4`, `manure_n2o_*`, `n_excretion`) outcomes, with standard units and IPCC tier references.	Low — CV spec only, no DDL change
Add publication entity	Add a `publication` table (`doi`, `url`, `author_last`, `year`, `journal`, `linked_doi_1..4`) linked to `project_metadata`. Nullable — apps that do not originate from published research simply leave it unpopulated.	Low — one new table, one FK
Extend location	Add three fields to `location`: `location_precision_type` (CV), `spatial_uncertainty_m` (float), `agroecological_zone` (varchar). All nullable, backward-compatible.	Low — three nullable columns

7.2 For Version 1.x (Optional Modules)

The following extensions can be released as optional modules in subsequent minor versions. They are relevant only to research-oriented datasets and do not affect operational app implementations.

Module	Entities / changes	Triggered by
Climate module	`location_climate` table (MAP, MAT, rainy seasons)	ERA dataset mapping
Agronomy module	`agronomy_context` table (fertilizer NPK, cropping system, soil type, irrigation)	ILRI Livestock DB mapping
Soil module	`soil_profile` table (sand/silt/clay, pH, depth)	ERA soil data
Digestibility CV	Add `in_vitro_dmd`, `in_vivo_dmd` to `feed_nutrition_measured` CV; define accepted method codes	ERA + ILRI Livestock DB

8. Proposed Extensions — Entity Summary

The diagram below shows the proposed additions in relation to existing Draft schema entities. Only entities relevant to the proposed extensions are shown.

```
EXISTING (Draft)                        PROPOSED ADDITIONS

-------------------------------------------------------------------

project_metadata  ------------>  publication  [new table]

  project_id (PK)                   publication_id (PK)

  ...                               project_id (FK, nullable)

                                    doi, url, author_last, year

                                    journal, linked_doi_1..4
```

```
location  -------------------->  location_climate  [new table, optional]
  location_id (PK)                 location_id (FK)
  lat, lon, geometry               map_mm, mat_c, n_seasons
  + location_precision_type [ADD]  season_1/2 start/end/length
  + spatial_uncertainty_m   [ADD]
  + agroecological_zone     [ADD]  soil_profile  [new table, optional]
                                   location_id (FK)
                                   depth_upper/lower
                                   sand/silt/clay_pct, ph
farming_system  ------------>  agronomy_context  [new table, optional]
  farm_sys_id (PK)               agronomy_id (PK)
  farm_sys_descriptor1/2         farming_system_id (FK)
                                 cropping_system, irrigation
                                 fertilizer_n/p/k_kg_ha
outcomes                       (no new table -- CV extension only)
  out_variable (CV) [EXTEND]   add: enteric_ch4, manure_ch4,
  out_value                        manure_n2o_*, n_excretion,
  out_unit                         milk_yield, meat_yield
  out_error
```

9. Conclusions

The Draft Feed Basket Schema is a structurally sound and conceptually mature reference model. Its core design — hierarchical feed taxonomy with processing, separation of delivery and share, herd sub-grouping, and outcome measurement — is well-suited to the diversity of data sources that need to be accommodated.

The gap analysis confirms that no redesign is necessary. The schema can be brought to full coverage of all four analysed datasets by:

1. Publishing a controlled vocabulary for the `outcomes` table that explicitly includes IPCC-aligned GHG variables and standard production variables.
2. Adding a `publication` entity to `project_metadata` for bibliographic provenance.
3. Adding three nullable fields to the `location` entity: precision type, spatial uncertainty, and agroecological zone.

These three changes are backward-compatible, impose no burden on operational apps, and are sufficient to enable the two primary use cases — empirical studies and model-driven analysis — end-to-end.

Climate, soil, and agronomy data are recommended as optional v1.x modules, applicable only to research-oriented datasets.