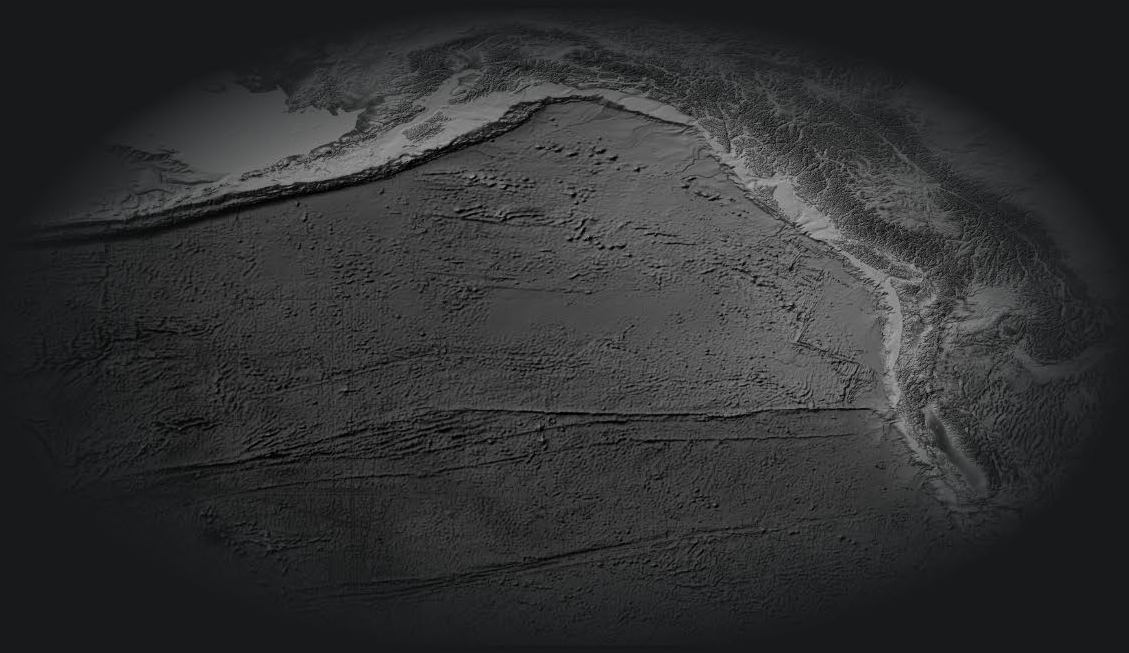
The Salmon Data Integration Toolkit

Removing data friction to scale salmon science

Brett Johnson [1] · Graeme Diack [2] · Bruno Carturan [3] · Gottfried Pestal [4] · Tom Bird [5]

[1] Pacific Salmon Commission
[2] Missing Salmon Alliance / The Atlantic Salmon Trust
[3] Pacific Salmon Foundation
[4] Solv Consulting
[5] Fisheries and Oceans Canada

29 October 2026
PICES 2026
S12
Transboundary Salmon Management

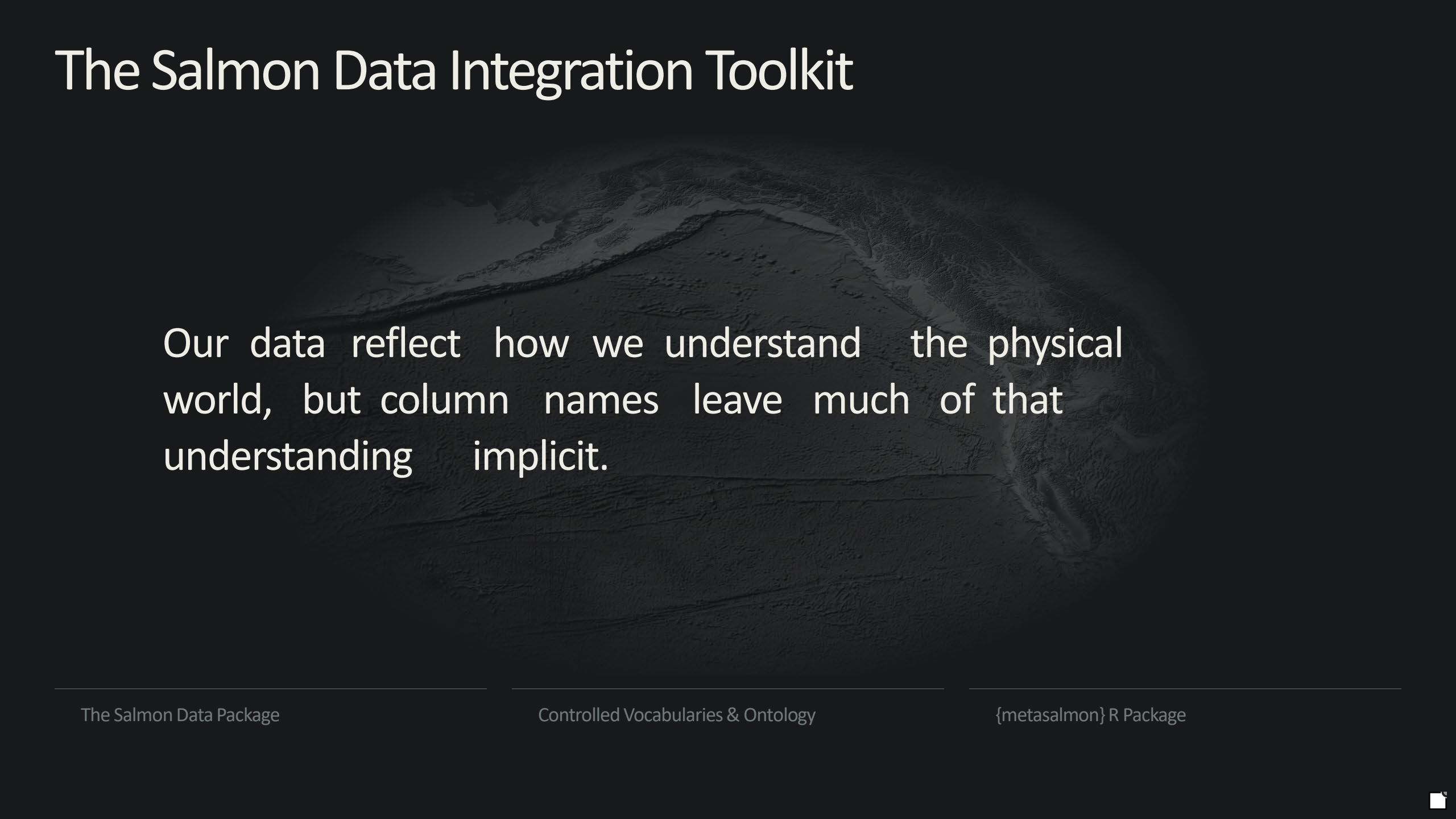


The Salmon Data Package

Controlled Vocabularies & Ontology

{metasalmon} R Package

The Salmon Data Integration Toolkit



Our data reflect how we understand the physical world, but column names leave much of that understanding implicit.

The Salmon Data Package

Controlled Vocabularies & Ontology

{metasalmon} R Package

The Salmon Data Integration Toolkit



What if community-developed tools could help us make the implicit explicit, fill in the missing context, and give everyone a clearer basis for connecting data?

The Salmon Data Package

Controlled Vocabularies & Ontology

{metasalmon} R Package

The Salmon Data Integration Toolkit

The Salmon Data Package

A way to carry precise and explicit meaning with the data

Controlled Vocabularies & Ontology

Reusable definitions and explicit relationships

{metasalmon} R Package

Automated metadata generation and semantic mapping

Salmon Data Integration Tools

The Salmon Data Integration Toolkit

The Salmon Data Package

A way to carry precise and explicit meaning with the data

Controlled Vocabularies & Ontology

Reusable definitions and explicit relationships

{metasalmon} R Package

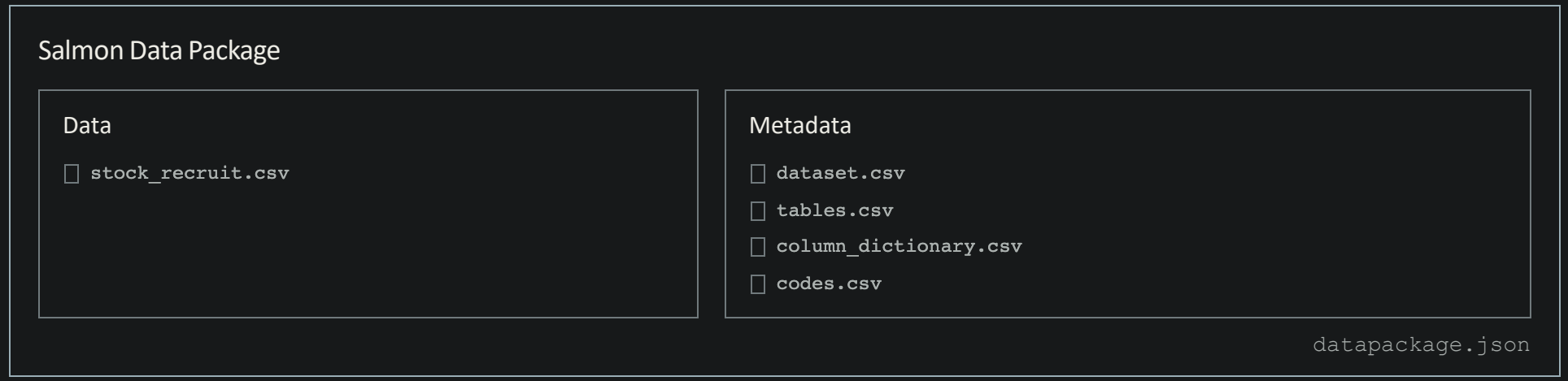
Automated metadata generation and semantic mapping

Salmon Data Integration Tools

The goal is to reduce data friction to scale salmon science and tackle ambitious questions with more joy.

The Salmon Data Integration Toolkit

The Salmon Data Package



The package keeps the values, definitions, units, provenance and context together.

Controlled Vocabularies & Ontology

Reusable definitions and explicit relationships

{metasalmon} R Package

Automated metadata generation and semantic mapping

The Salmon Data Integration Toolkit

The Salmon Data Package

A way to carry precise and explicit meaning with the data

column_dictionary.csv

Controlled Vocabularies & Ontology

column_name	column_label	column_description	term_iri
total_broodyr_EFS	Effective female spawner abundance	Brood-year female spawner abundance adjusted downward for the estimated proportion of eggs not spawned, based on spawning-ground sampling.	https://w3id.org/psc/vocab/concept/PSC-CV-000035

{metasalmon} R Package

Automated metadata generation and semantic mapping

The Salmon Data Integration Toolkit

The Salmon Data Package

A way to carry precise and explicit meaning with the data

Controlled Vocabularies & Ontology

PSC COMMON DATA DEFINITIONS

Provisional · biological review pending

Effective female spawner abundance

`total_broodyr_EFS`

An estimate of female spawner abundance in the brood year (as opposed to the more traditional total of male and female spawners), which is further adjusted downward by the proportion of eggs that were not spawned, as determined by sampling on the spawning grounds.

PSC-CV-000035 ↗

{metasalmon} R Package

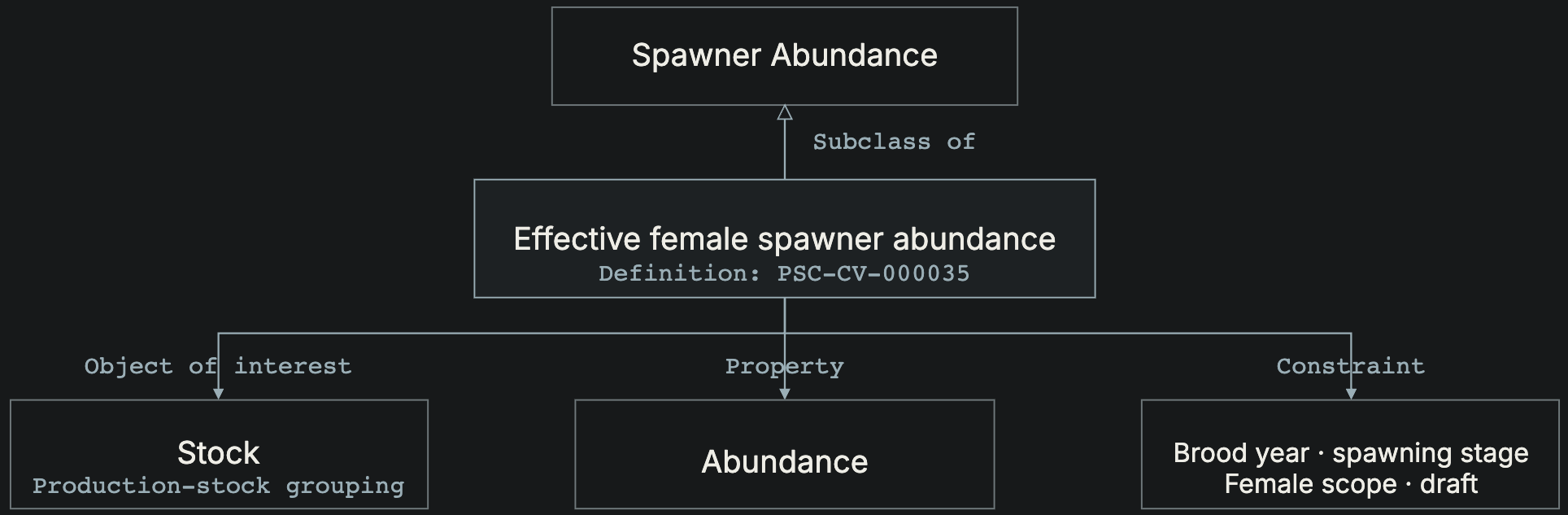
Automated metadata generation and semantic mapping

The Salmon Data Integration Toolkit

The Salmon Data Package

A way to carry precise and explicit meaning with the data

Controlled Vocabularies & Ontology



{metasalmon} R Package

Automated metadata generation and semantic mapping

The Salmon Data Integration Toolkit

The Salmon Data Package	A way to carry precise and explicit meaning with the data
Controlled Vocabularies & Ontology	Reusable definitions and explicit relationships

{metasalmon} R Package



Your R data frame	Start with the data already in R.
Scaffold the package	Re-formats your dataframes as a Salmon Data Package.
Map the terms	Link meanings to published vocabularies.
Review the mappings	You check mappings and scientific fit.
Publish the package	Validate, then review rights and repository access.
Mobilize the contribution	Connect reviewed context to a growing corpus of interoperable data.

Data Archaeology

Clarify the sources so you can integrate them.

Usual preparation

The work still has to be done.

Find scattered files

Data Archaeology

Clarify the sources so you can integrate them.

Usual preparation	The work still has to be done.
Find scattered files	Hunt for documentation

Data Archaeology

Clarify the sources so you can integrate them.

Usual preparation		The work still has to be done.
Find scattered files	→ Hunt for documentation	→ Ask data owners

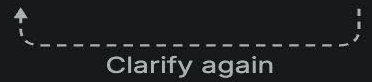
Data Archaeology

Clarify the sources so you can integrate them.

Usual preparation

The work still has to be done.

Find scattered files → Hunt for documentation → Ask data owners → Reconstruct definitions



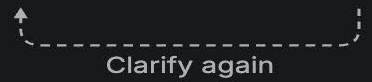
Data Archaeology

Clarify the sources so you can integrate them.

Usual preparation

The work still has to be done.

Find scattered files → Hunt for documentation → Ask data owners → Reconstruct definitions → Rebuild year and population mappings

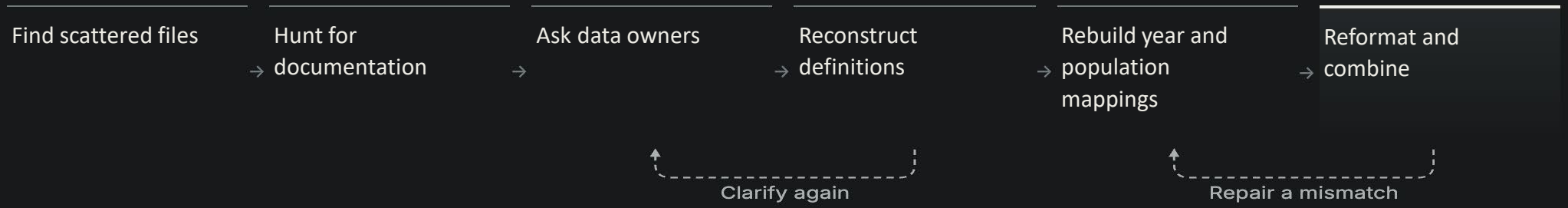


Data Archaeology

Clarify the sources so you can integrate them.

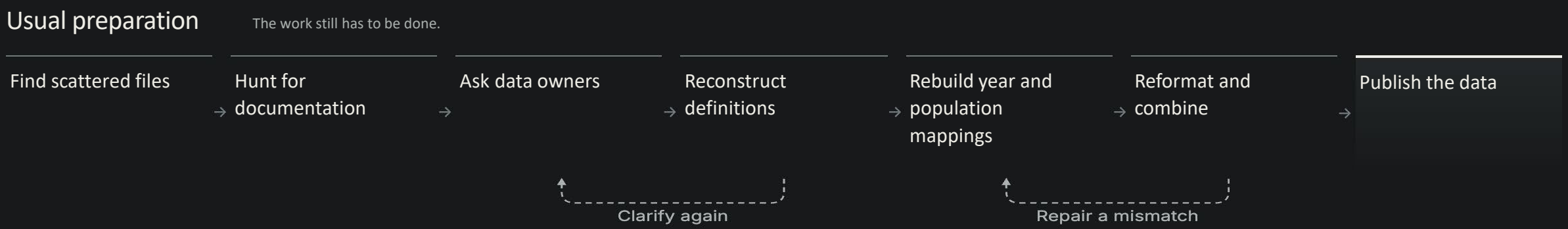
Usual preparation

The work still has to be done.



Data Archaeology

Clarify the sources so you can integrate them.



Data Archaeology

Find scattered files → Hunt for documentation → Ask data owners → Reconstruct definitions → Rebuild year and population mappings → Reformat and combine → Publish the data

Now what do you do with the source data?

- Source dataset A
- Source dataset B
- Source dataset C

→ **Integrate for this question**

Select · map · roll up

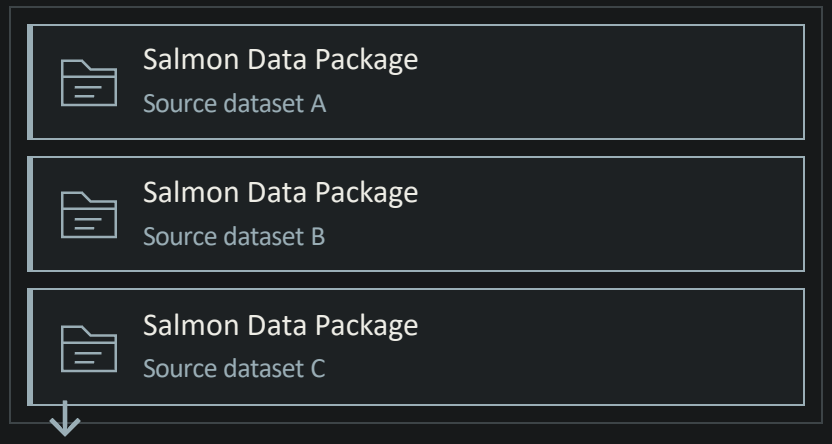
→ **Integrated dataset / database**

A useful result shaped for this question.

Data Archaeology

Find scattered files → Hunt for documentation → Ask data owners → Reconstruct definitions → Rebuild year and population mappings → Reformat and combine → Publish the data

Package each source so its meaning can be reused.



Integrate for this question

Select · map · roll up

Record selections, mappings and aggregation rules.

Integrated dataset / database

Selected records · harmonised fields · chosen roll-ups.

Trace the result back to source versions.

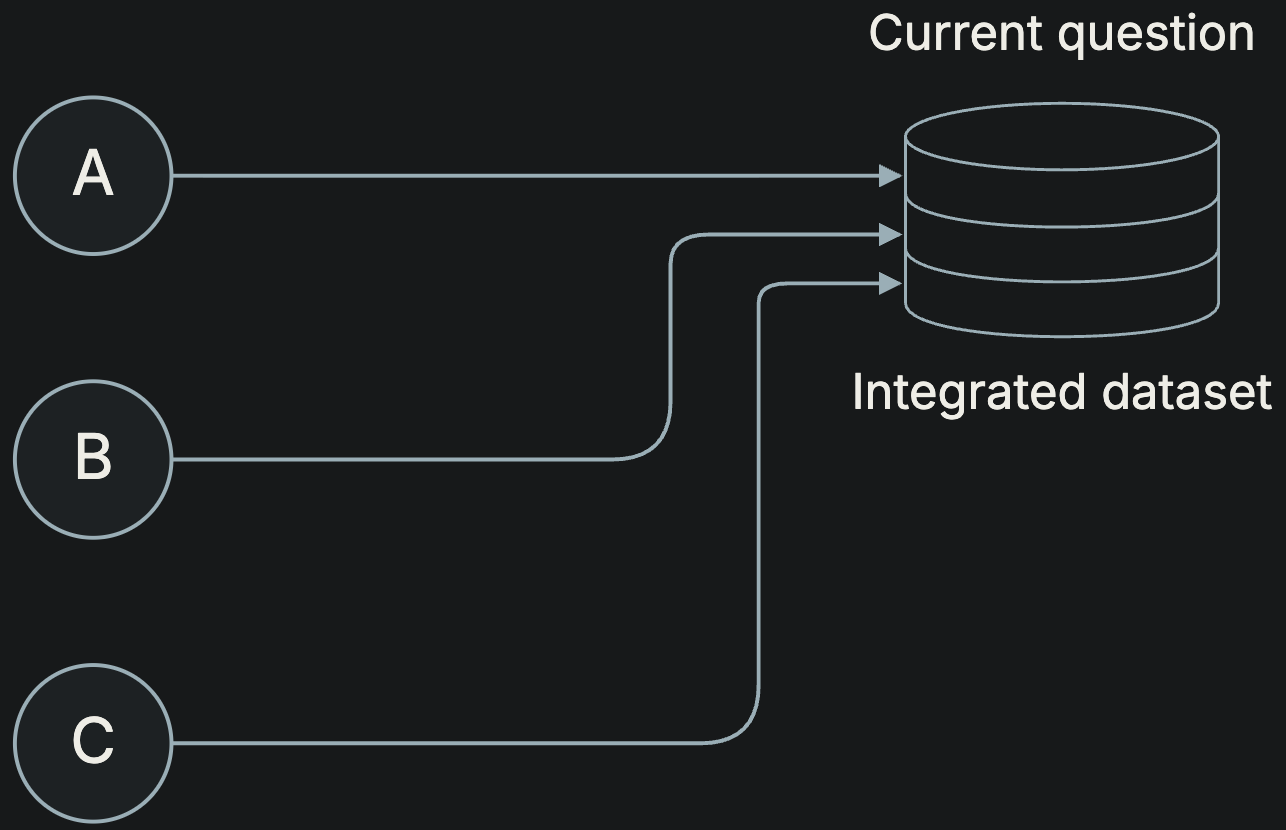
Other integrations

Different questions, groupings and roll-ups.

Data Archaeology

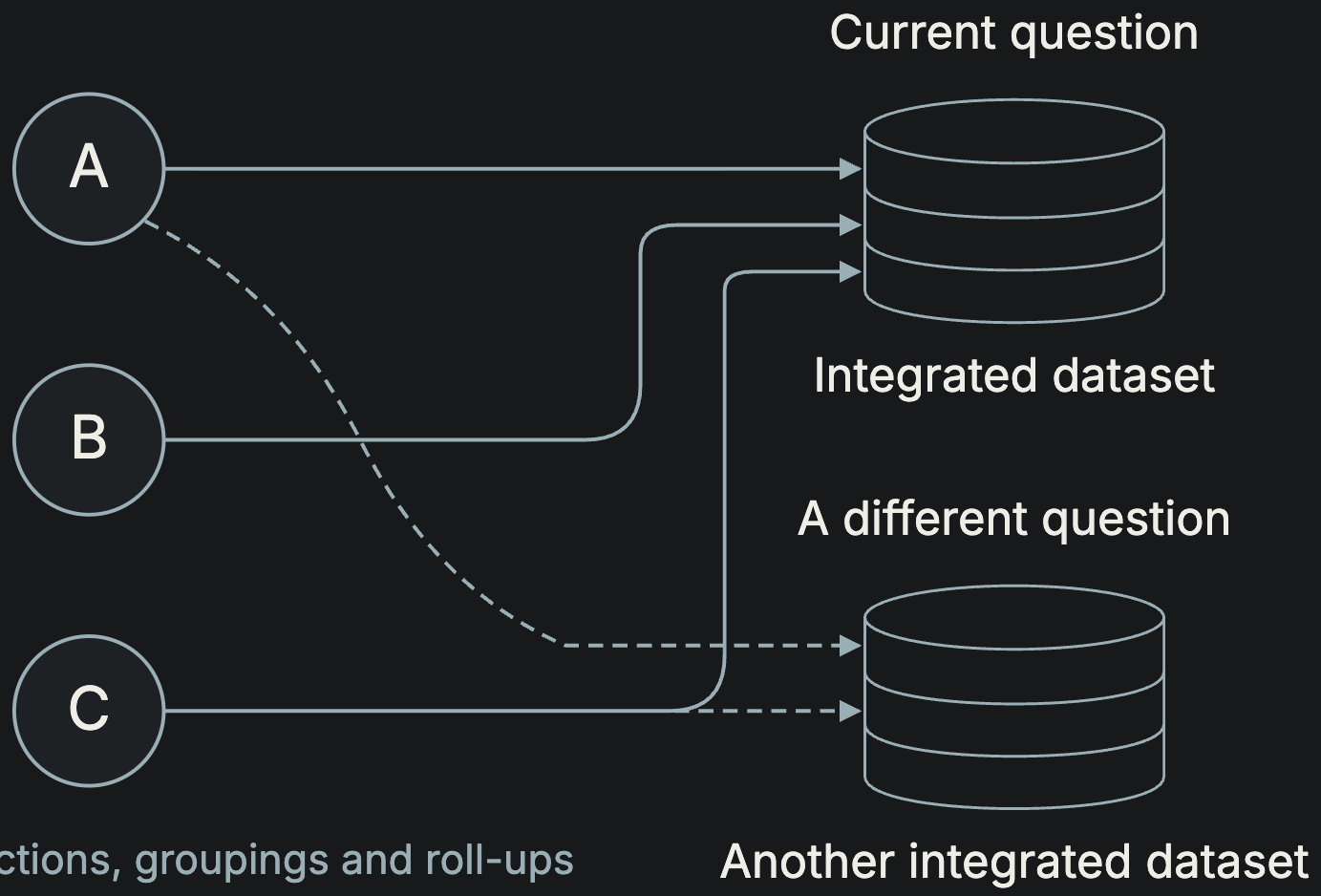
Source packages for this question.

◯ Salmon Data Package



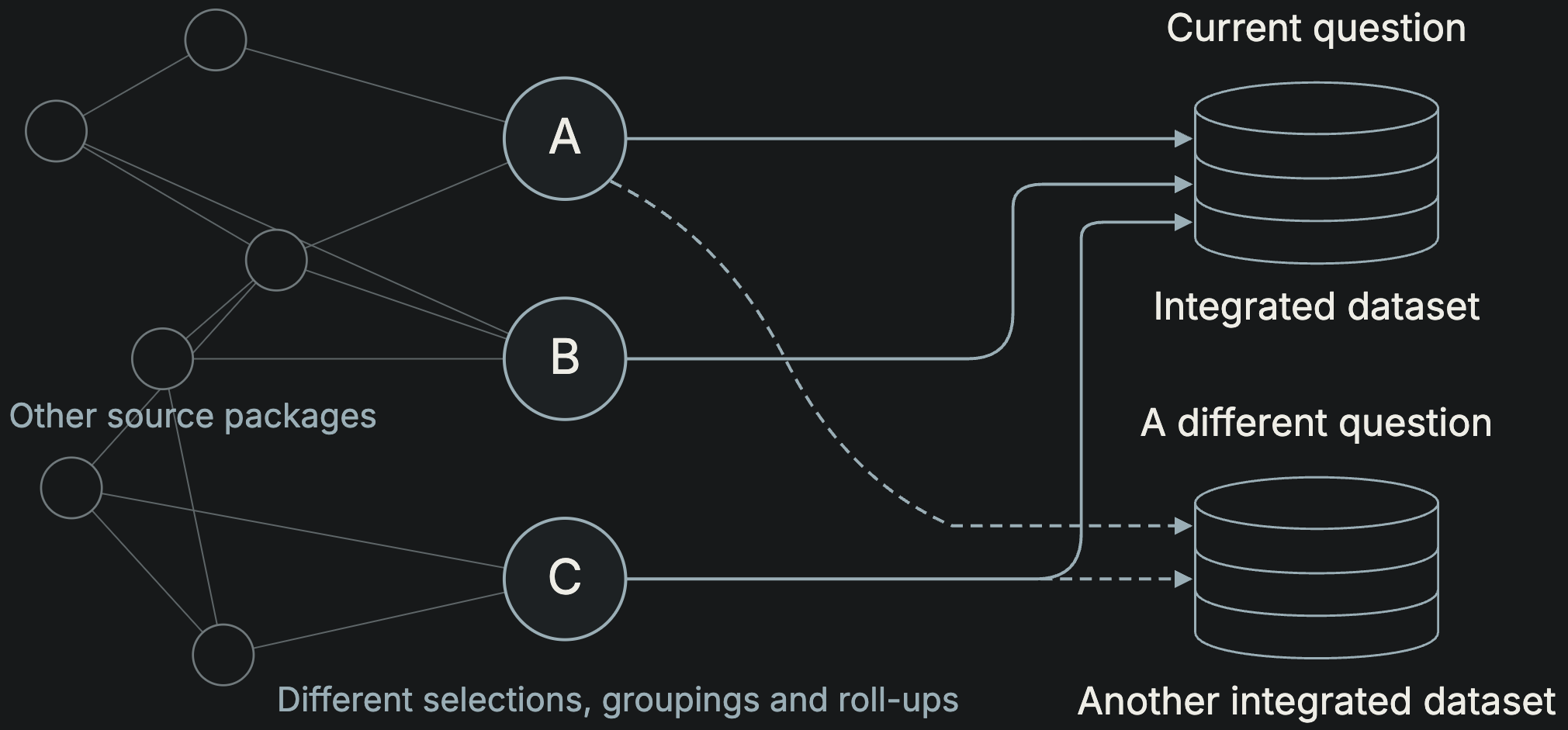
Data Archaeology

Reuse the same sources for another question.



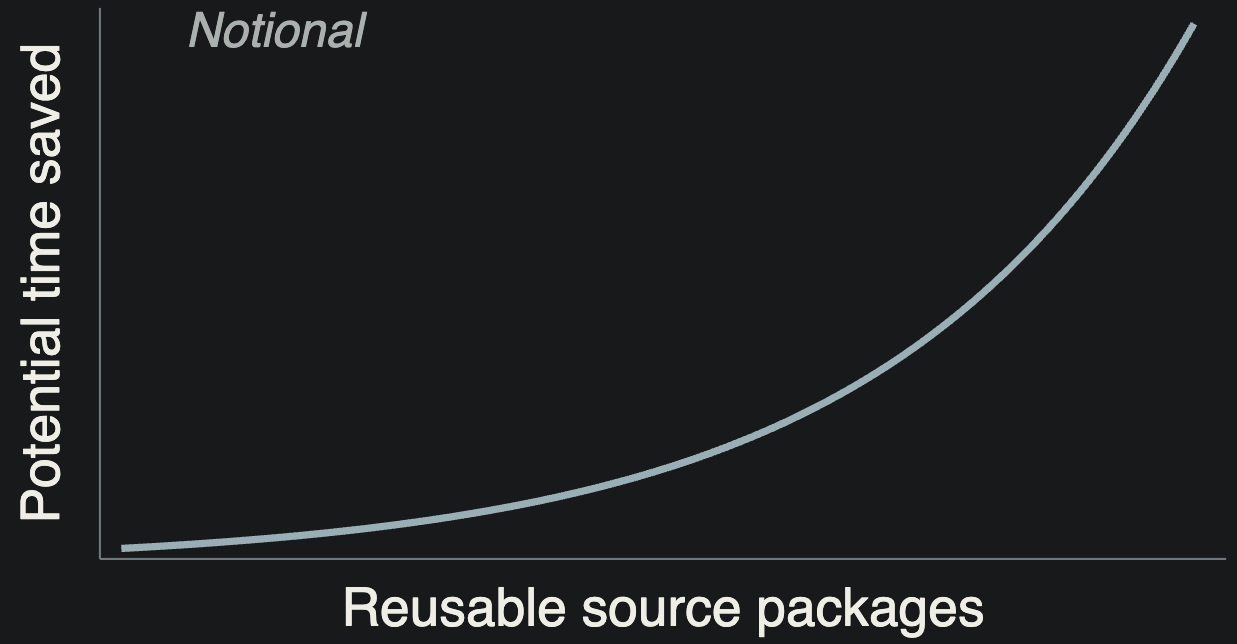
Data Archaeology

The same source packages. Different integrated datasets.



Data Archaeology

Less repeated data archaeology. More capacity for science.



Contribute

One team captures the preparation.

Reuse

Many teams can build on it.

Improve

Return reviewed refinements.

The toolkit makes this easier: scaffold, map, review and share.

A Case Study

What drives Fraser sockeye production?

Spawning grounds

Spawners and juvenile rearing

→

Coast

Migration and local conditions

→

High seas

Growth and survival

→

Returning adults

Return timing and age

Assemble a spawner time series.

Focus on this step ↓

A Case Study

What drives Fraser sockeye production?

Spawning grounds → Coast → High seas → Returning adults

Assemble a spawner time series.

Prepare the data in R ↗

You assembled the data

Your observations and local documentation.

You found a dataset online

Fraser sockeye stock–recruit data.



Read the data as they are

Keep the source detail.

Salmon Data Package

Define · map · review

For integration with other salmon data.

{metasalmon} R pkg




```
# Read and inspect the source.
fraser <- read_csv(source_file,
  col_types = cols(.default = col_character()))
str(fraser, list.len = 3, vec.len = 1,
    give.attr = FALSE)

# Select the Chilko example.
fraser <- fraser |>
  filter(production_stock_name == "Chilko",
         broodyr == "2017",
         age %in% c("32", "42", "52")) |>
  select(production_stock_name, broodyr,
         returnyr, age, total_broodyr_EFS,
         num_recruits)
```

Read the original data

```
spc_tbl_ [25,535 × 22] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ stock_broodyr_age            : chr [1:25535] "Early Stuart_1948_21" ..
 $ production_stock_name        : chr [1:25535] "Early Stuart" ...
 $ production_stock_code        : chr [1:25535] "1" ...
  [list output truncated]
```

Chilko · three selected source rows

production_stock_name	broodyr	returnyr	age	total_broodyr_EFS	num_recruits
Chilko	2017	2020	32	212,956	1,149
Chilko	2017	2021	42	212,956	1,106,873
Chilko	2017	2022	52	212,956	40,083

{metasalmon} R pkg




```
# Scaffold a Salmon Data Package.
pkg <- create_sdp(
  list(detailed_stock_recruit = fraser),
  path = "Fraser-SDP",
  seed_semantics = TRUE,  # Suggest terms.
  semantic_max_per_role = 2
)
```

Salmon Data Package

```
Fraser-SDP/
├─ data/
│  ├─ detailed_stock_recruit.csv
├─ metadata/
│  ├─ codes.csv
│  ├─ column_dictionary.csv
│  ├─ dataset.csv
│  ├─ tables.csv
├─ datapackage.json
├─ semantic_suggestions.csv
```

Read data → Scaffold & review → Publish data

{metasalmon} R pkg




```
# Compare published terms.
review <- review_semantics(
  pkg, columns = "num_recruits",
  max_candidates = 2
)
review |>
  filter(role == "variable") |>
  select(rank, label, source, score)

# Accept and save the matching term.
review <- accept_suggestion(
  review, "num_recruits", "variable",
  rank = 1
)
apply_sdp_semantics(pkg, review)
```

Term candidates for `num_recruits`

```
# A tibble: 2 × 4
   rank label             source score
  <int> <chr>             <chr>  <dbl>
1     1 Recruit abundance smn     19.2
2     2 Abundance         smn     12.3
```

Reviewed mapping

`metadata/column_dictionary.csv`

column_name	term_iri
num_recruits	smn:RecruitAbundance

{metasalmon} R pkg

```
# Preview the catalog deposit.
plan <- publish_sdp_to_knb(
  pkg, eml_path = reviewed_eml,
  public = TRUE, dry_run = TRUE,
  representation = "expanded"
)
```

Publish to a catalog from R

KNB — Knowledge Network for Biocomplexity

A repository in the DataONE network



Ontology annotations support discovery by measurement type.

Publish a citable record and request its DOI.

DFO Enterprise Data Hub

Metadata export

Zenodo

Package upload

Read data → Scaffold & review → Publish data

What have we actually done?

Package data with metadata

Keep definitions, units and provenance with the values.

Use standard terms

Connect meanings through published terms and reviewed mappings.

Publish a citable package

Prepare catalog publication and a DOI request.

Make preparation reusable

Carry reviewed preparation into the next integration.

FAIR salmon data sharing

Findable · Accessible · Interoperable · Reusable

Support organizational mandates and journal requirements.

More joy ;)

Shared tools, different entry points

Salmon Science Plugin

Agentic AI: Codex / Claude

Developing tools

{metasalmon} R package

Scaffold · map · review · validate · publish.

Salmon Domain Ontology

Shared concepts and explicit mappings.

Salmon Data Package

Data, definitions, units and provenance.

Shared tools, different entry points

Salmon Domain Ontology

Salmon Data Package

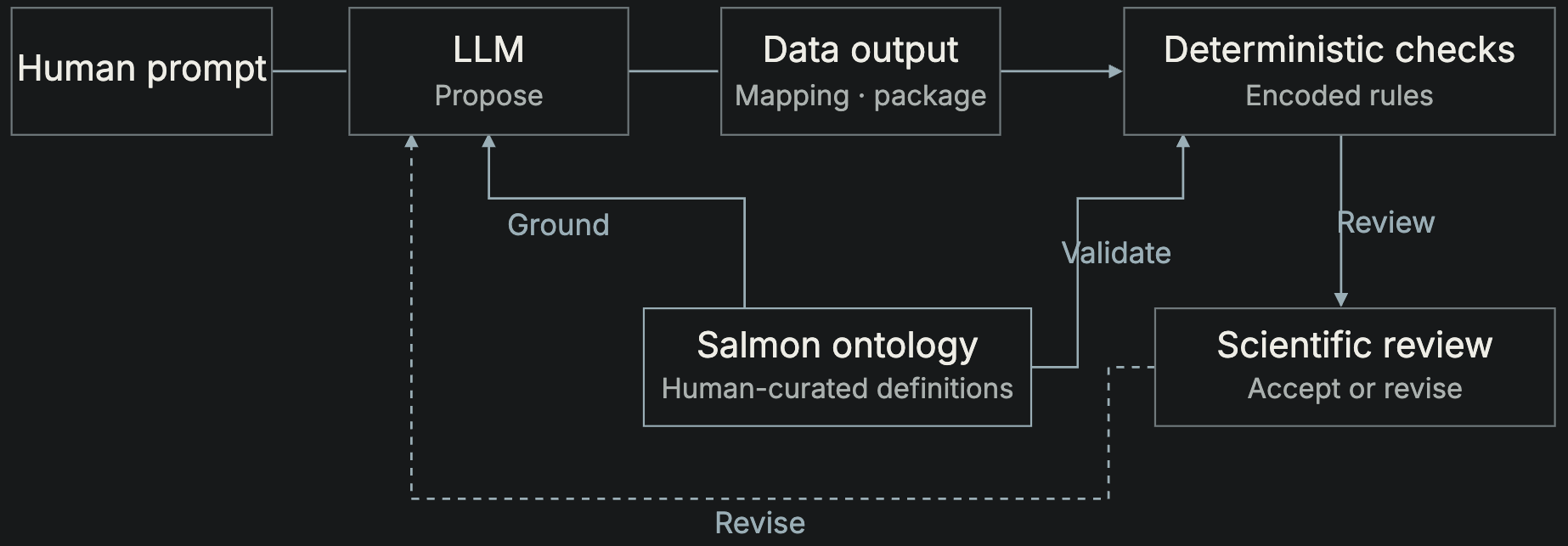
{metasalmon} R package

Salmon Science Plugin

Agentic AI: Codex / Claude

Developing tools

Neurosymbolic AI: ground, check and review.



Ground in salmon expertise. Reduce hallucination risk.

Shared tools, different entry points

Salmon Domain Ontology

Salmon Data Package

{metasalmon} R package

Salmon Science Plugin

Agentic AI: Codex / Claude

Developing tools

A plugin brings skills and bounded tools to the agent.

Skills

Select a workflow and keep its scope explicit.

Tools

Retrieve sources, search terms and prepare packages.

Reviewable work

Return mappings, checks and proposed changes.

Shared tools, different entry points

Salmon Domain Ontology

Salmon Data Package

{metasalmon} R package

Salmon Science Plugin

Agentic AI: Codex / Claude

Developing tools

We can offload tedious labor.
We can't offload our understanding or judgment.

Agents can help

- Recover context
- Draft mappings
- Run checks and document changes

Scientists decide

- Interpret measurements and scope
- Judge population, cohort and exposure
- Accept or reject; decide on publication

Designed for broad interoperability

How do salmon data connect to oceanographic and other data?

Deep within salmon



Salmon Data Package



Salmon Domain Ontology

Keep local meanings precise.



Broad across science

Frictionless Data Package

CSV + metadata. Shared data tools.

Darwin Core

Biodiversity exchange, including OBIS.

Shared observation models

SOSA · I-ADOPT · QUDT · PROV-O

Oceanographic context

CF/NetCDF · NERC vocabularies

metasalmon searches published terms: salmon vocabularies, NERC/NVS, QUDT, OLS, GBIF and WoRMS.

Get involved

Use the tools. Attend the workshop.

The tools are free.

Start by defining your organization's terms in a controlled vocabulary.

- {metasalmon} R package ↗
- Salmon Data Package ↗
- Salmon Domain Ontology ↗
- Salmon Science Plugin ↗
- Salmon Data Integration Workshop ↗

Removing data friction to scale salmon science