TRANSFORMACIÓN DE DATOS AL MODELO DE DATOS OMOP-CDM

Alberto Labarga

2023-09-14

Telecommunications Engineer
Head of Biomedical Data Hub @BSC
More than 20 years teaching

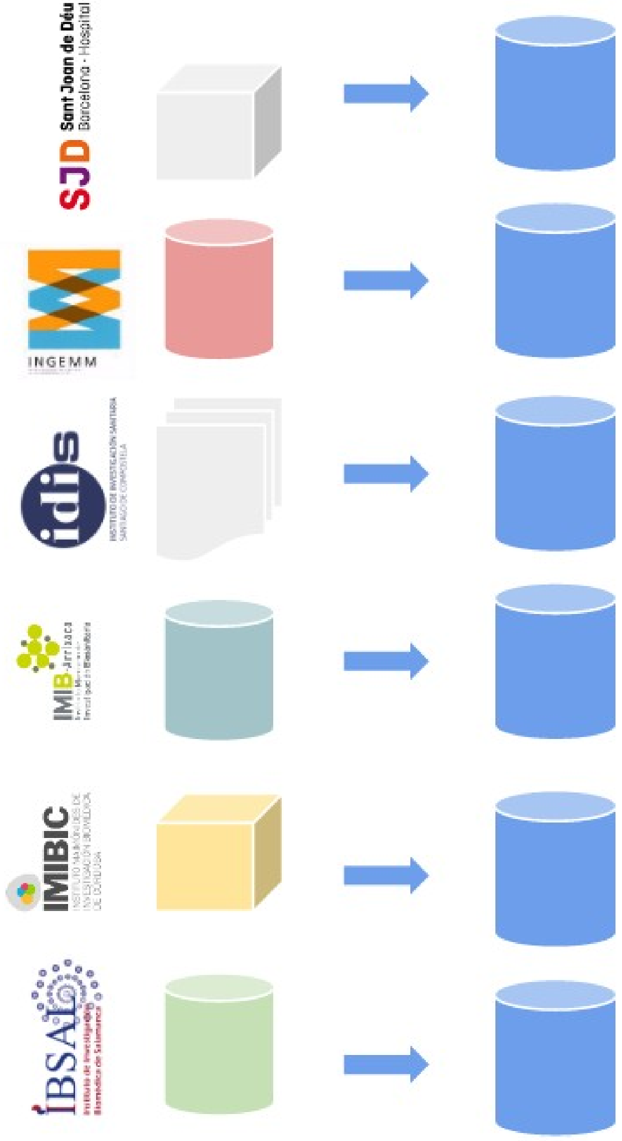
open data - open source – open science

I am Alberto

- alabarga
- alabarga
- /in/albertolabarga

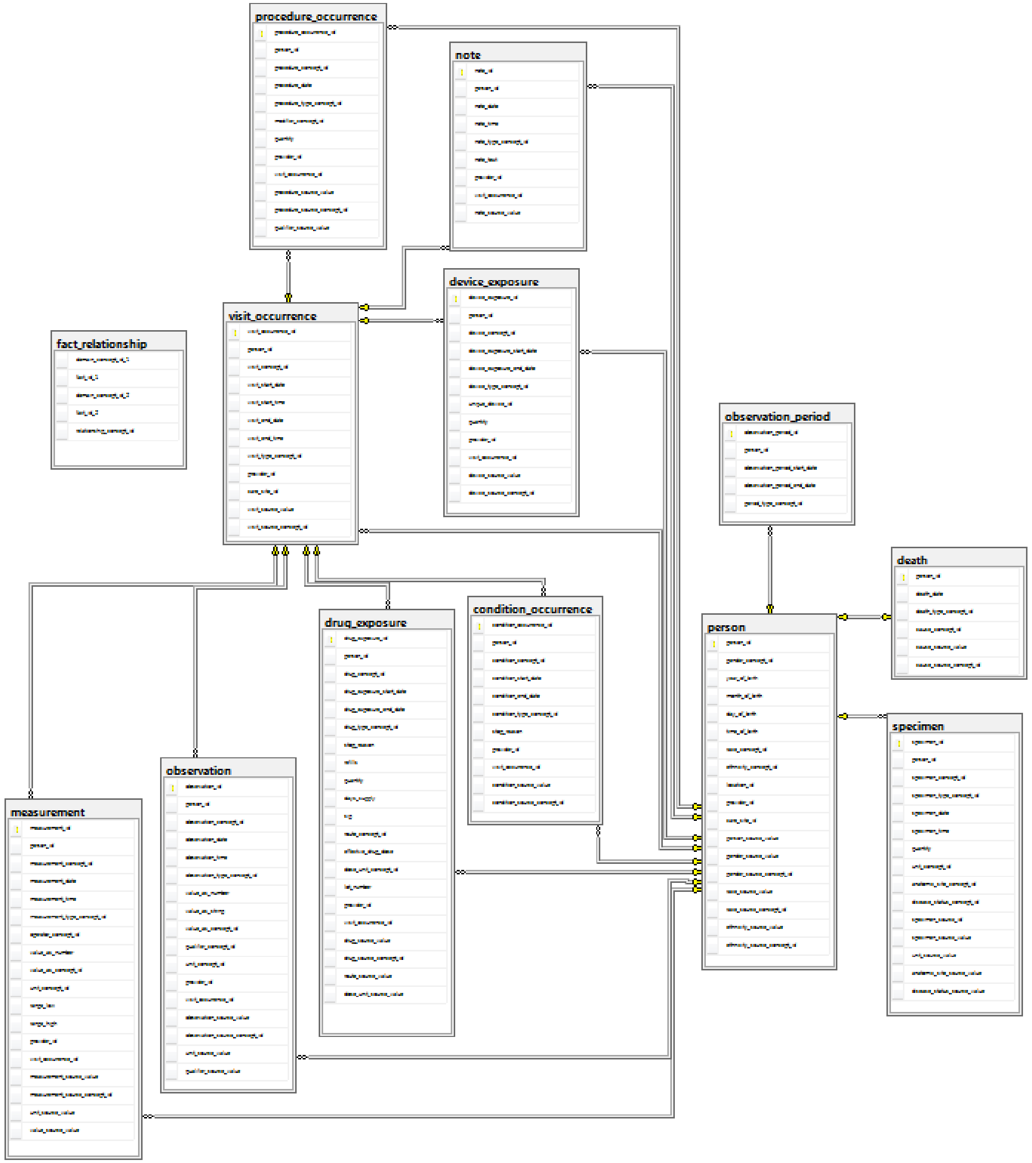
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

SJD Sant Joan de Déu
Barcelona · Hospital
INGEMM
idis
INSTITUTO DE INVESTIGACIÓN SANITARIA
SANTIAGO DE COMPOSTELA
IMIB-Arrixaca
IMIBIC
IBSAL

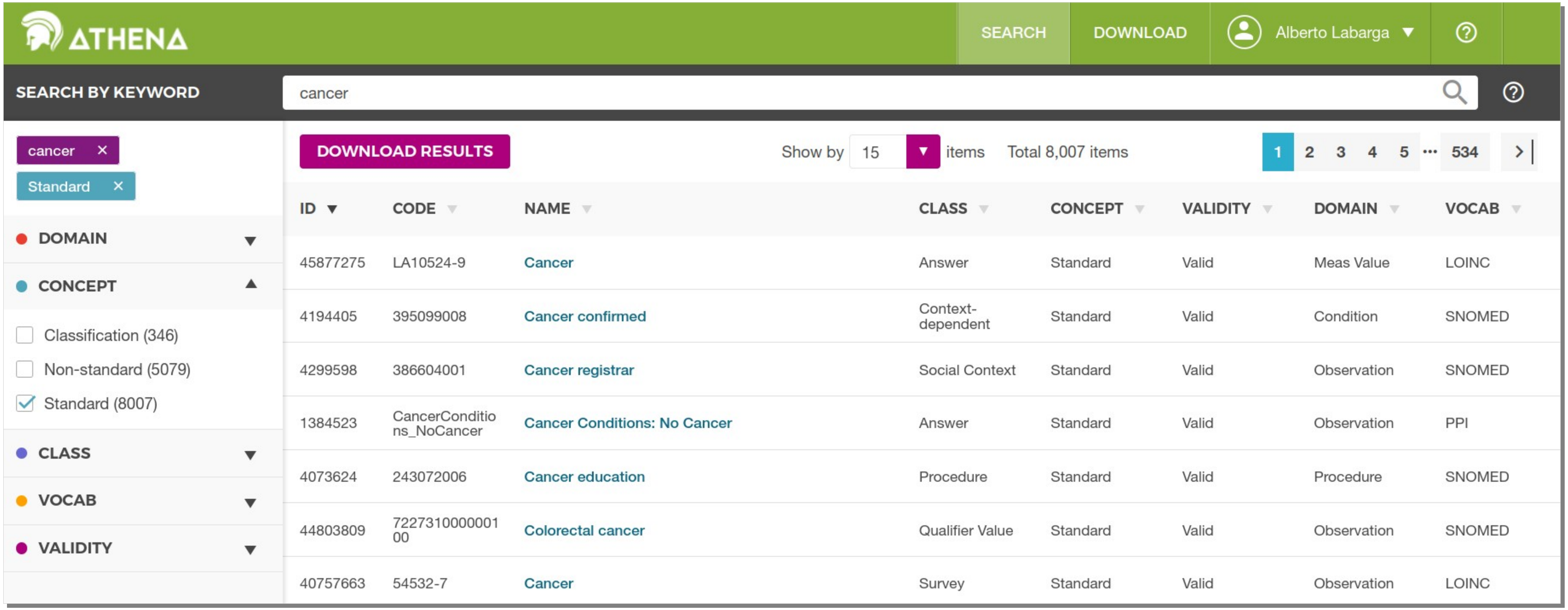

SJD
Sant Joan de Déu
Barcelona · Hospital
INGEMM
idis
INSTITUTO DE INVESTIGACIÓN SANITARIA
SANTIAGO DE COMPOSTELA
IMIB-Arrixaca
IMIBIC
IBSAL

OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS



https://ohdsi.github.io/CommonDataModel/



SEARCH BY KEYWORD: cancer

cancer ✕
Standard ✕

● DOMAIN ▼

● CONCEPT ▲

□ Classification (346)
□ Non-standard (5079)
☑ Standard (8007)

● CLASS ▼

● VOCAB ▼

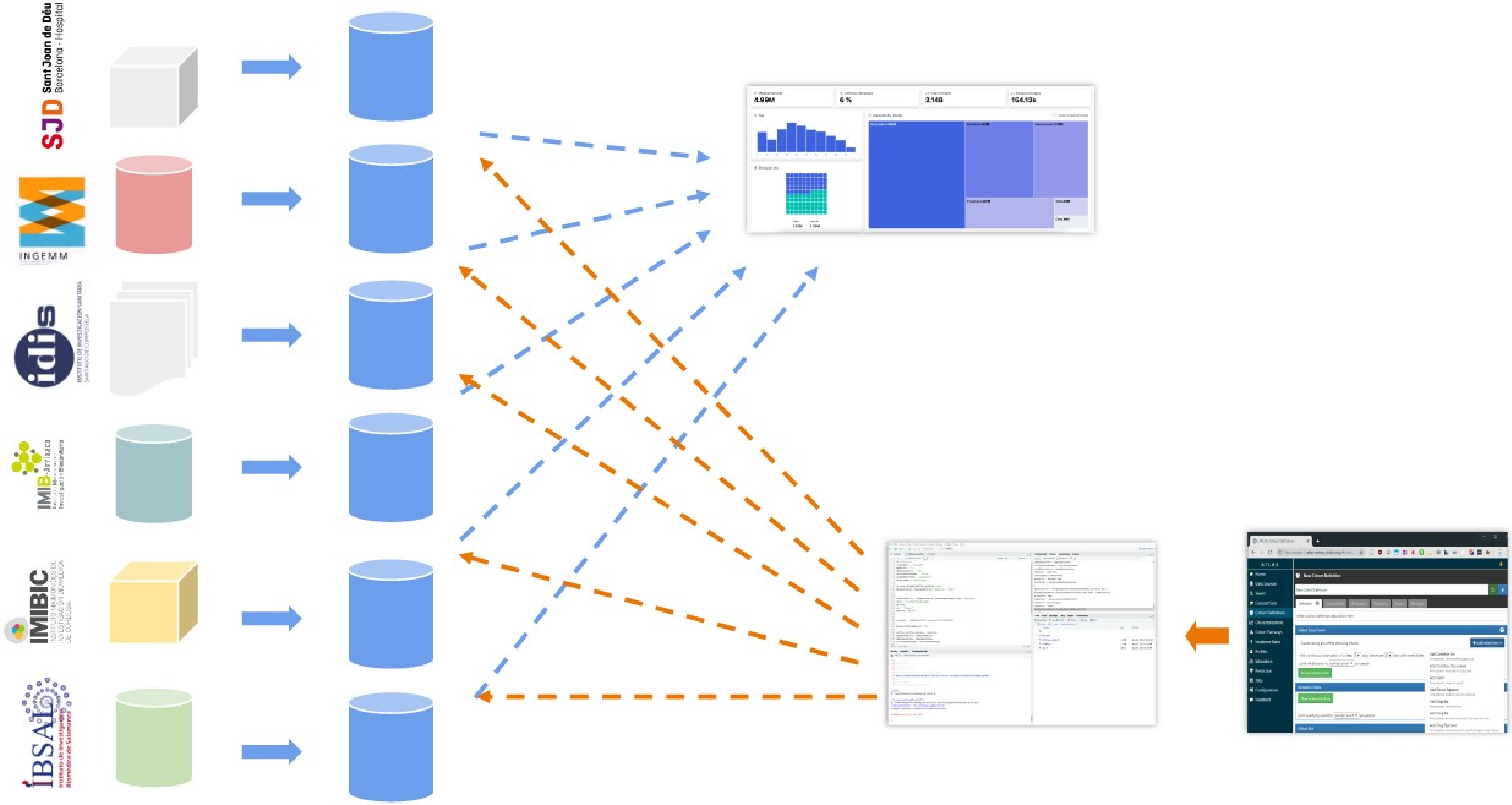
● VALIDITY ▼

Show by 15 items Total 8,007 items

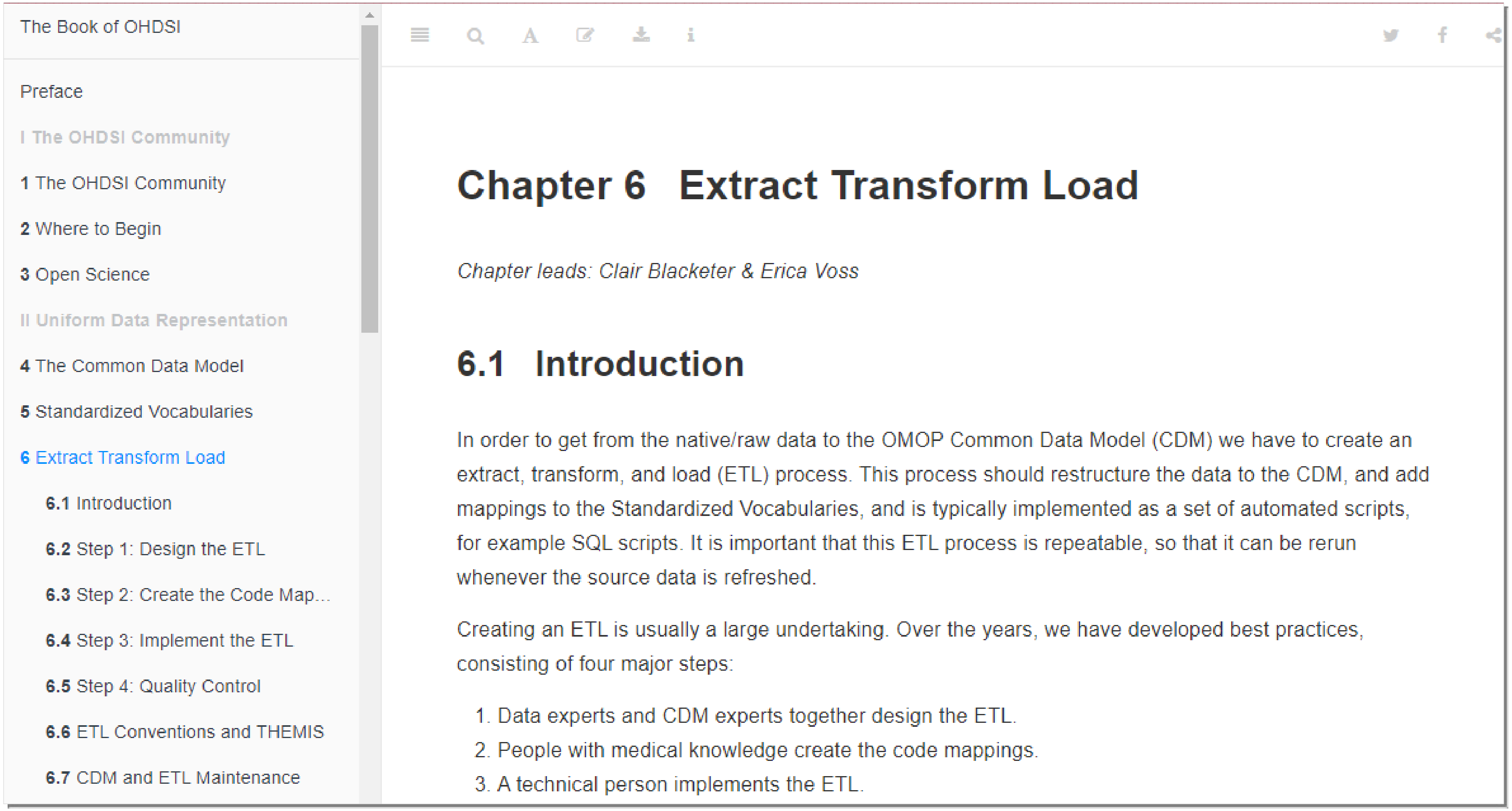
1 2 3 4 5 ... 534 >

ID	CODE	NAME	CLASS	CONCEPT	VALIDITY	DOMAIN	VOCAB
45877275	LA10524-9	Cancer	Answer	Standard	Valid	Meas Value	LOINC
4194405	395099008	Cancer confirmed	Context-dependent	Standard	Valid	Condition	SNOMED
4299598	386604001	Cancer registrar	Social Context	Standard	Valid	Observation	SNOMED
1384523	CancerConditions_NoCancer	Cancer Conditions: No Cancer	Answer	Standard	Valid	Observation	PPI
4073624	243072006	Cancer education	Procedure	Standard	Valid	Procedure	SNOMED
44803809	722731000000100	Colorectal cancer	Qualifier Value	Standard	Valid	Observation	SNOMED
40757663	54532-7	Cancer	Survey	Standard	Valid	Observation	LOINC

https://athena.ohdsi.org



Sant Joan de Déu
Barcelona · Hospital
SJD
INGEMM
idis
INSTITUTO DE INVESTIGACIÓN SANITARIA SANTIAGO DE COMPOSTELA
IMIB-Arrixaca
IMIBIC
INSTITUTO MAIMÓNIDES DE INVESTIGACIÓN BIOMÉDICA DE CÓRDOBA
IBSAL
Instituto de Investigación Biomédica de Salamanca
4.99M
6 %
3.14B
154.13k
ATLAS



The Book of OHDSI

Chapter 6 Extract Transform Load

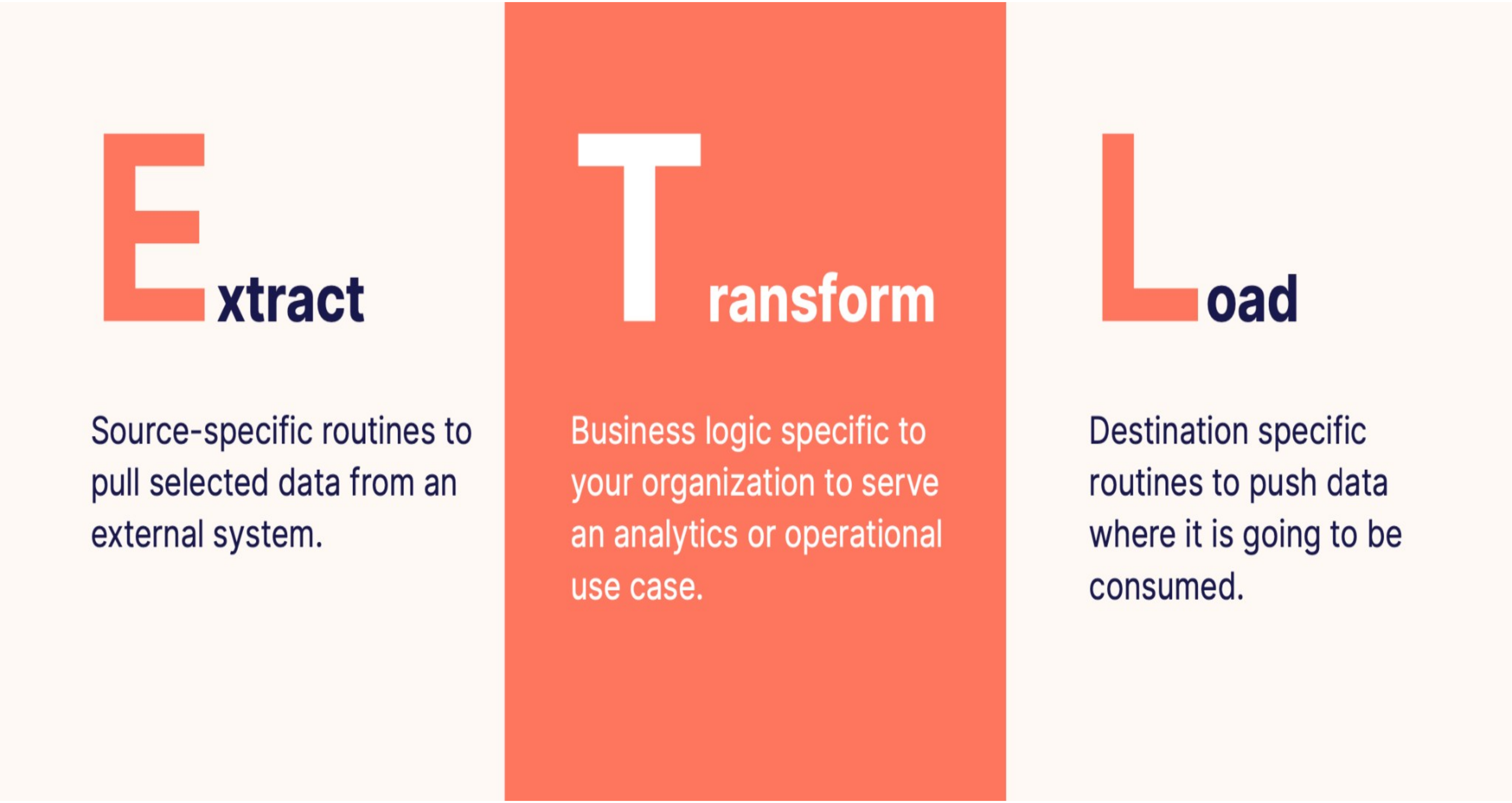
Chapter leads: Clair Blacketer & Erica Voss

6.1 Introduction

In order to get from the native/raw data to the OMOP Common Data Model (CDM) we have to create an extract, transform, and load (ETL) process. This process should restructure the data to the CDM, and add mappings to the Standardized Vocabularies, and is typically implemented as a set of automated scripts, for example SQL scripts. It is important that this ETL process is repeatable, so that it can be rerun whenever the source data is refreshed.

Creating an ETL is usually a large undertaking. Over the years, we have developed best practices, consisting of four major steps:

1. Data experts and CDM experts together design the ETL.
2. People with medical knowledge create the code mappings.
3. A technical person implements the ETL.



Extract

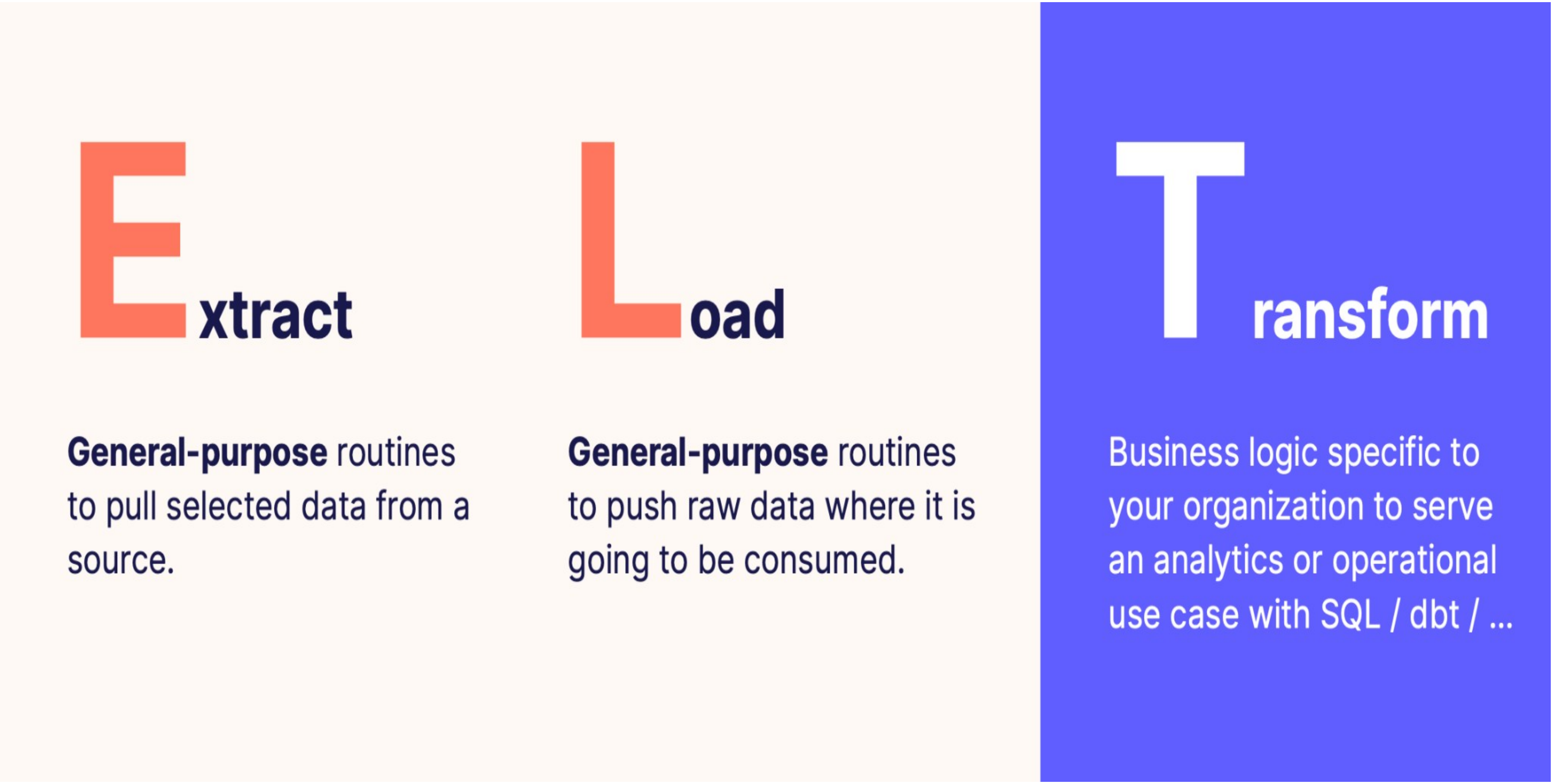
Source-specific routines to pull selected data from an external system.

Transform

Business logic specific to your organization to serve an analytics or operational use case.

Load

Destination specific routines to push data where it is going to be consumed.



Extract

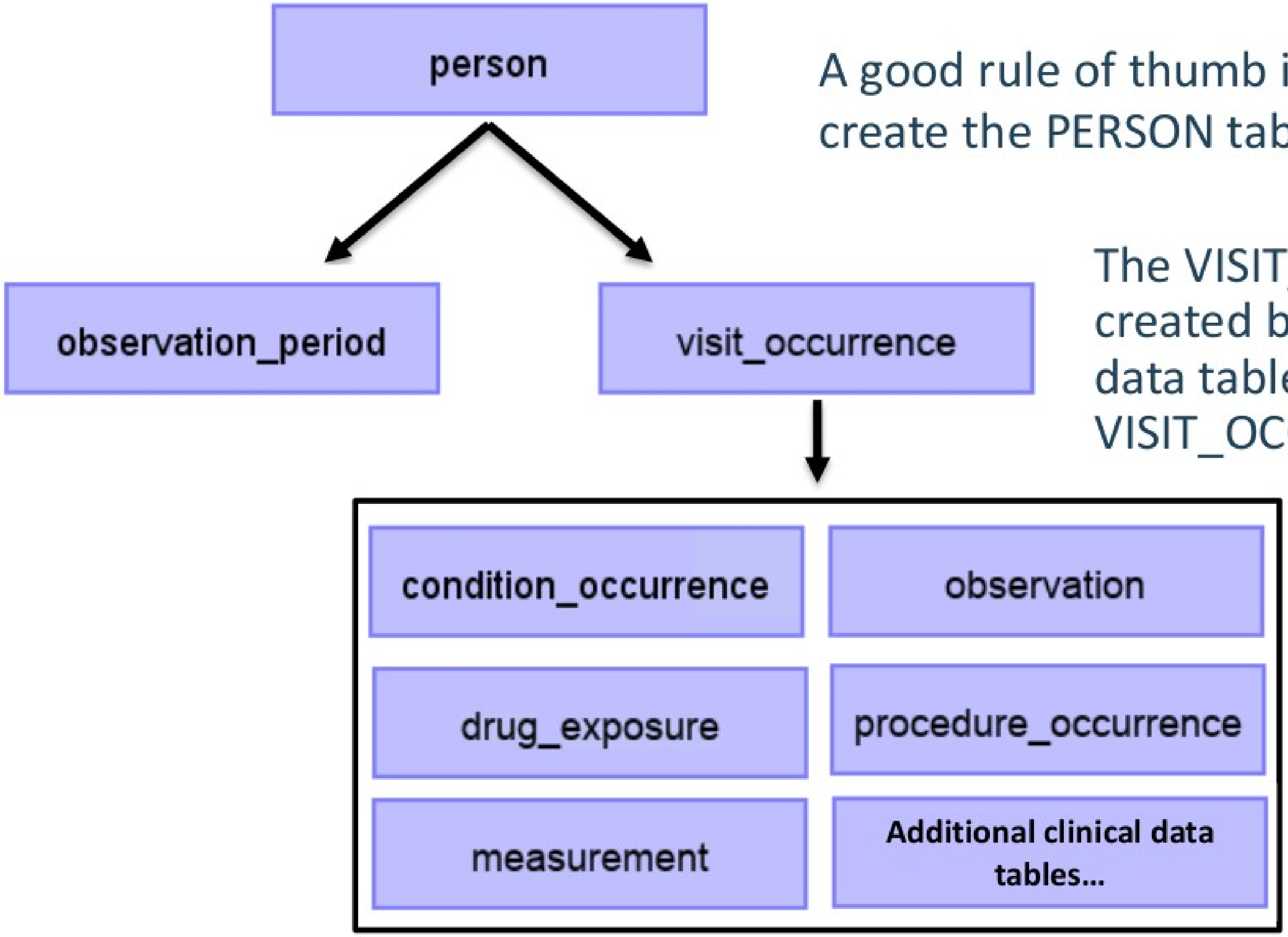
General-purpose routines to pull selected data from a source.

Load

General-purpose routines to push raw data where it is going to be consumed.

Transform

Business logic specific to your organization to serve an analytics or operational use case with SQL / dbt / ...



person

observation_period

visit_occurrence

condition_occurrence

observation

drug_exposure

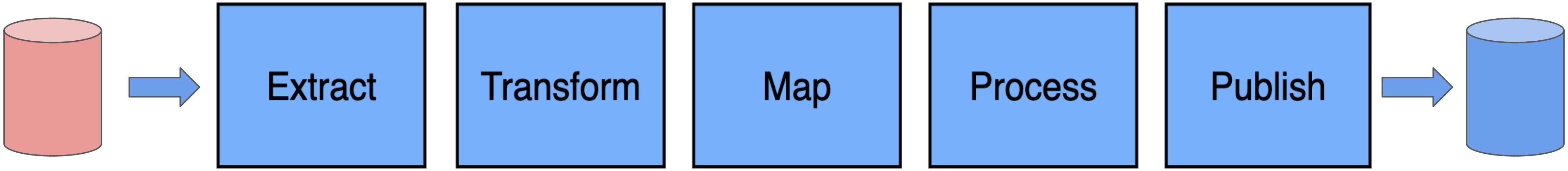
procedure_occurrence

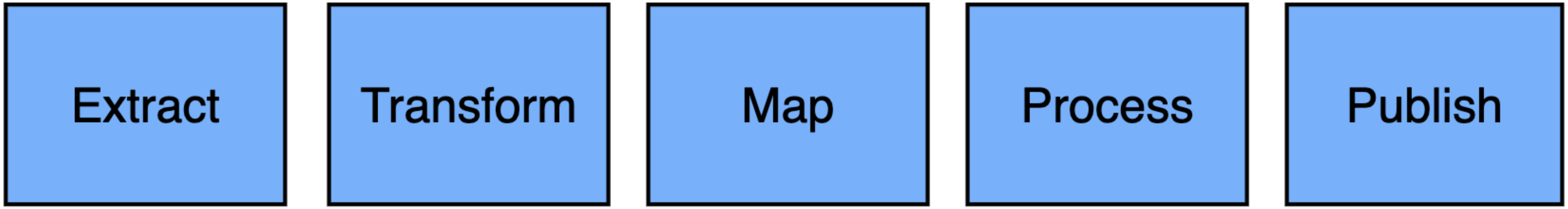
measurement

Additional clinical data tables…

A good rule of thumb is to always create the PERSON table first

The VISIT_OCCURRENCE table must be created before the standardized clinical data tables as they all refer to the VISIT_OCCURRENCE_ID


Extract
Transform
Map
Process
Publish



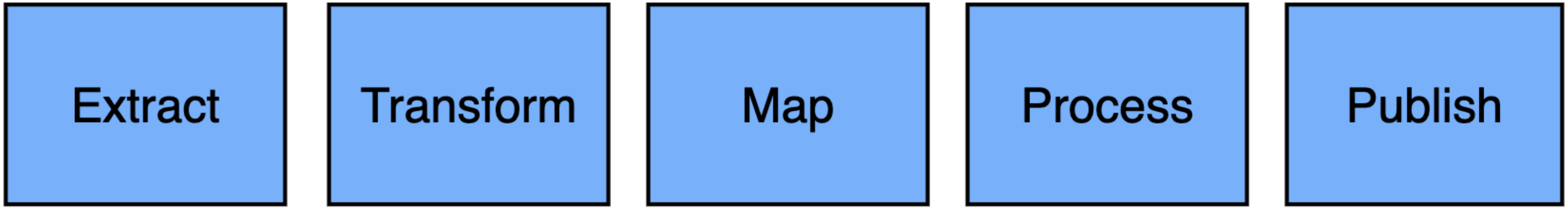
Extract

Transform

Map

Process

Publish



Extract

Transform

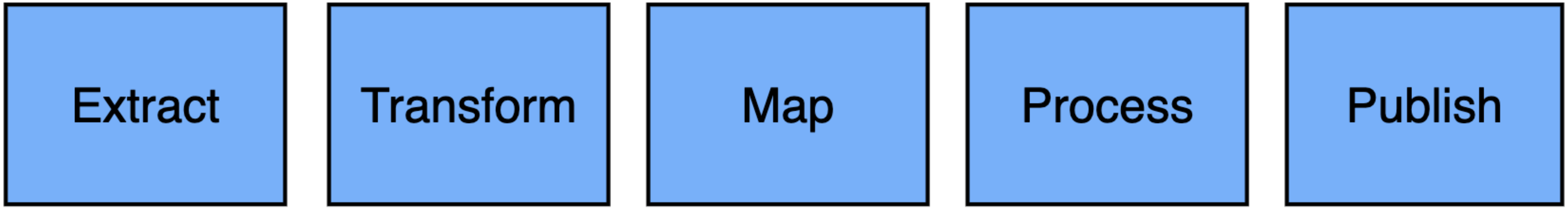
Map

Process

Publish

Meltano

Airbyte



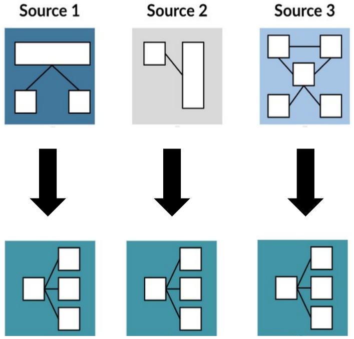
Extract

Transform

Map

Process

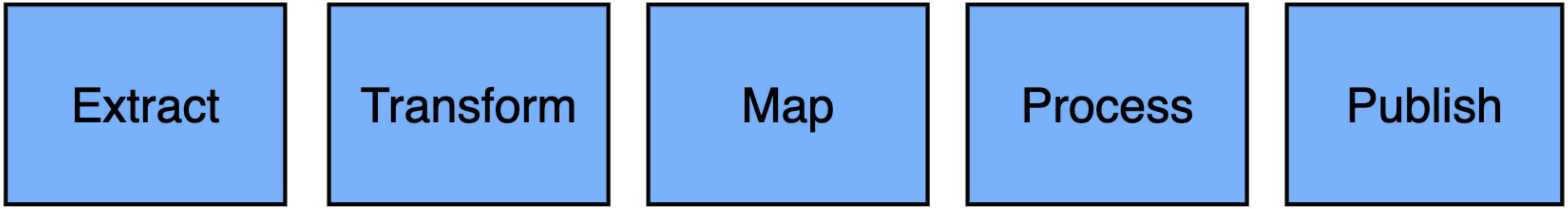
Publish



Esquema de datos común

La misma estructura de base de datos

[Observational Medical Outcomes Partnership (OMOP) CDM]



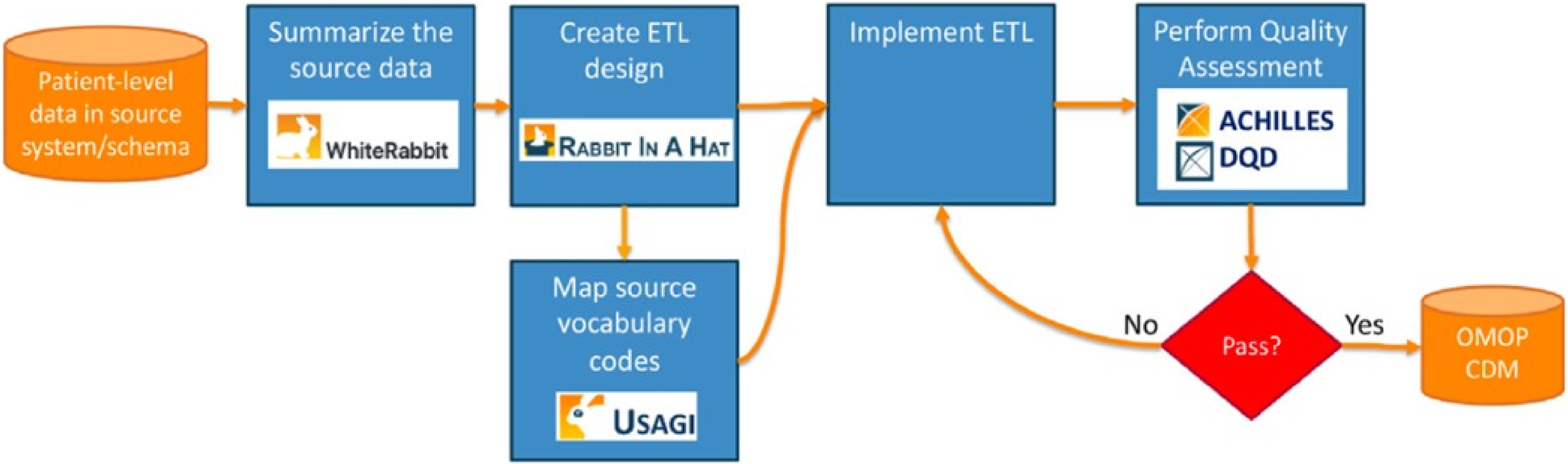
Extract

Transform

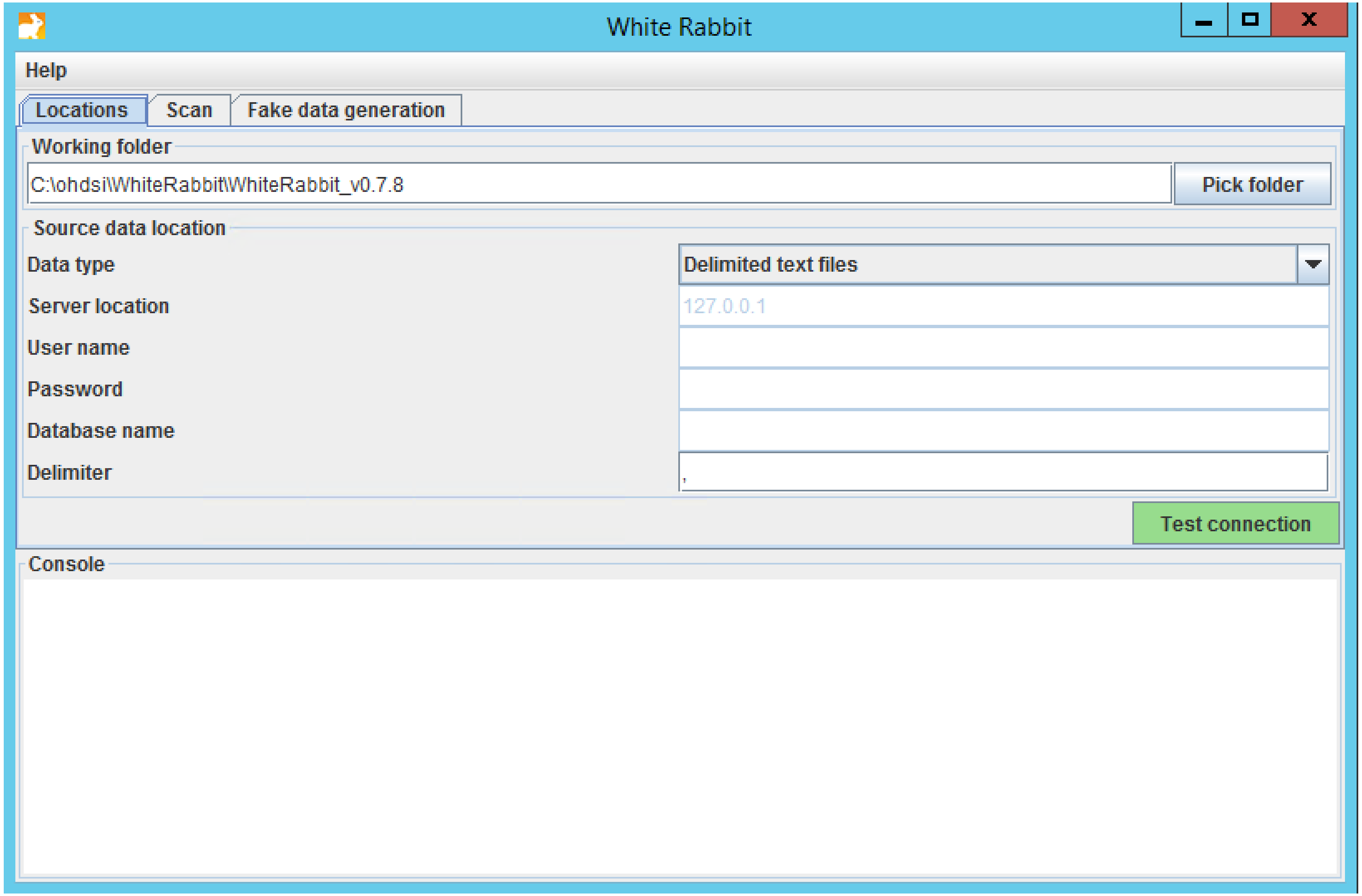
Map

Process

Publish


Patient-level data in source system/schema
Summarize the source data
WhiteRabbit
Create ETL design
Rabbit In A Hat
Map source vocabulary codes
Usagi
Implement ETL
Perform Quality Assessment
ACHILLES
DQD
No
Pass?
Yes
OMOP CDM

WhiteRabbit scans source data & creates a csv report on the source data



White Rabbit

Help

Locations | Scan | Fake data generation

Working folder

C:\ohdsi\WhiteRabbit\WhiteRabbit_v0.7.8

Source data location

Data type: Delimited text files

Server location: 127.0.0.1

User name:

Password:

Database name:

Delimiter: ,

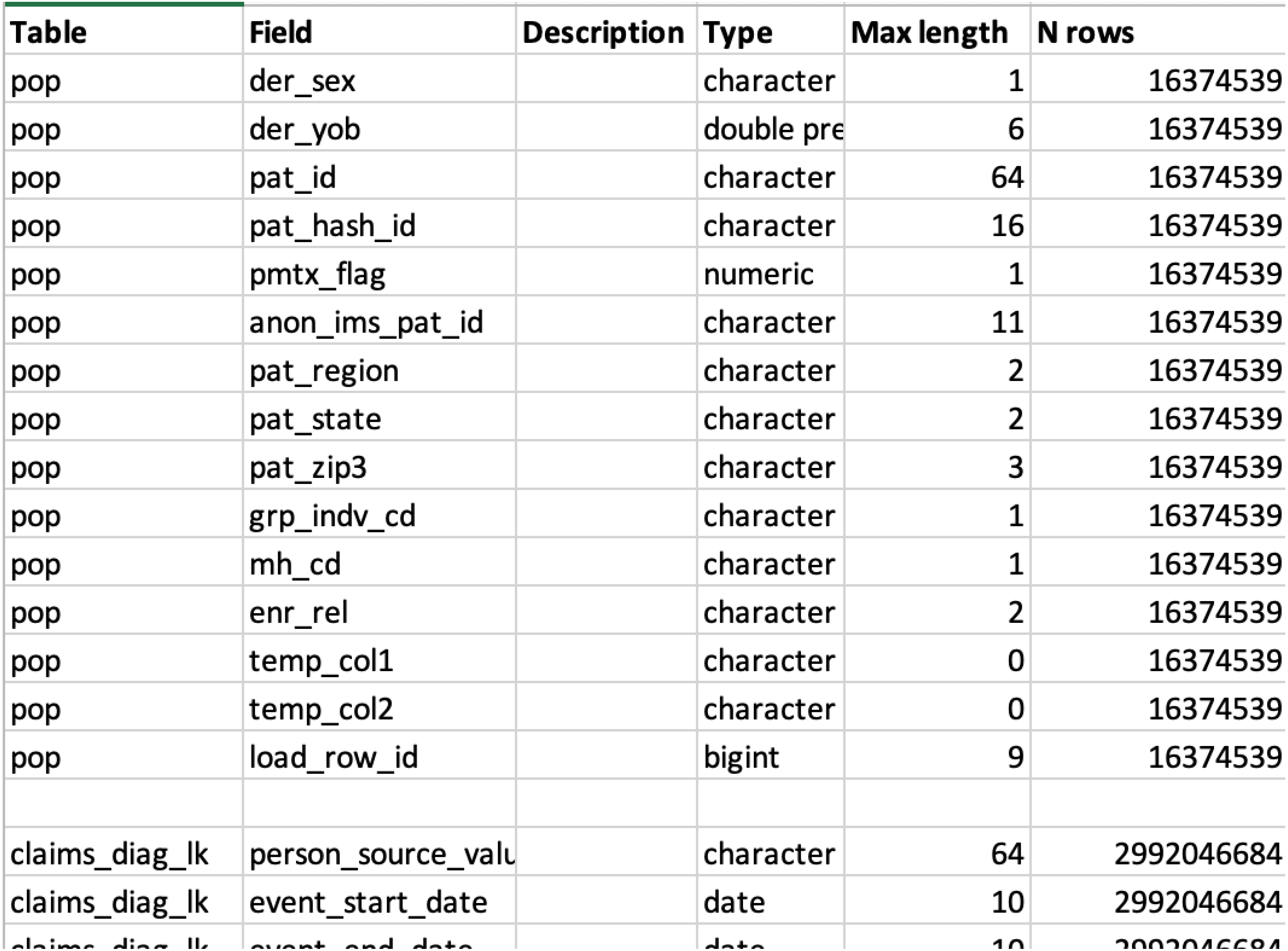
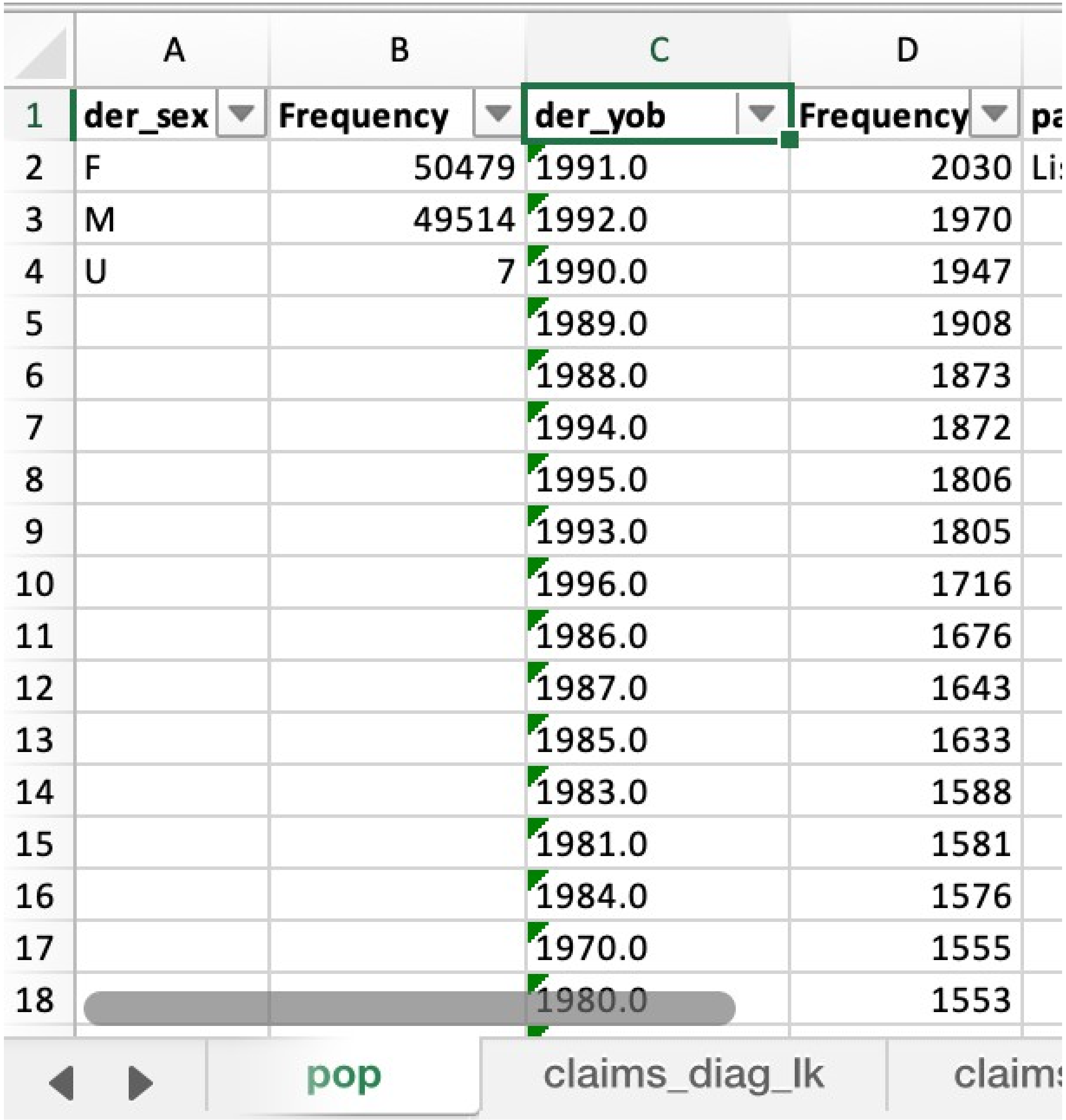


Table	Field	Description	Type	Max length	N rows
pop	der_sex		character	1	16374539
pop	der_yob		double pre	6	16374539
pop	pat_id		character	64	16374539
pop	pat_hash_id		character	16	16374539
pop	pmtx_flag		numeric	1	16374539
pop	anon_ims_pat_id		character	11	16374539
pop	pat_region		character	2	16374539
pop	pat_state		character	2	16374539
pop	pat_zip3		character	3	16374539
pop	grp_indv_cd		character	1	16374539
pop	mh_cd		character	1	16374539
pop	enr_rel		character	2	16374539
pop	temp_col1		character	0	16374539
pop	temp_col2		character	0	16374539
pop	load_row_id		bigint	9	16374539
claims_diag_lk	person_source_valu		character	64	2992046684
claims_diag_lk	event_start_date		date	10	2992046684
claims_diag_lk	event_end_date		date	10	2992046684



	A	B	C	D	E
1	der_sex	Frequency	der_yob	Frequency	pa
2	F	50479	1991.0	2030	Li
3	M	49514	1992.0	1970	
4	U	7	1990.0	1947	
5			1989.0	1908	
6			1988.0	1873	
7			1994.0	1872	
8			1995.0	1806	
9			1993.0	1805	
10			1996.0	1716	
11			1986.0	1676	
12			1987.0	1643	
13			1985.0	1633	
14			1983.0	1588	
15			1981.0	1581	
16			1984.0	1576	
17			1970.0	1555	
18			1980.0	1553	

Rabbit-in-a-Hat Read and display a **WhiteRabbit** scan Document and provides a graphical interface to allow a user to connect source data to CDM tables

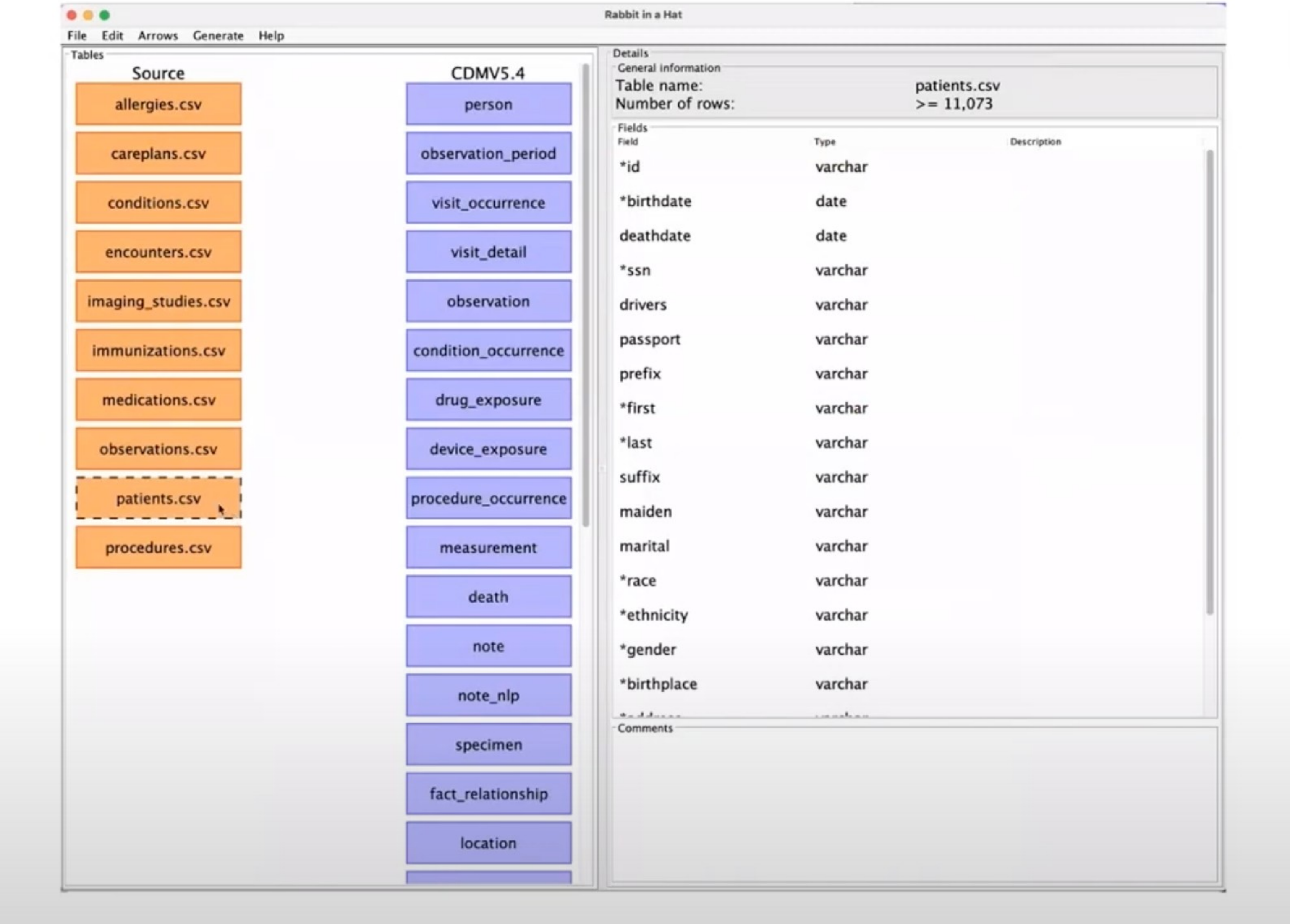
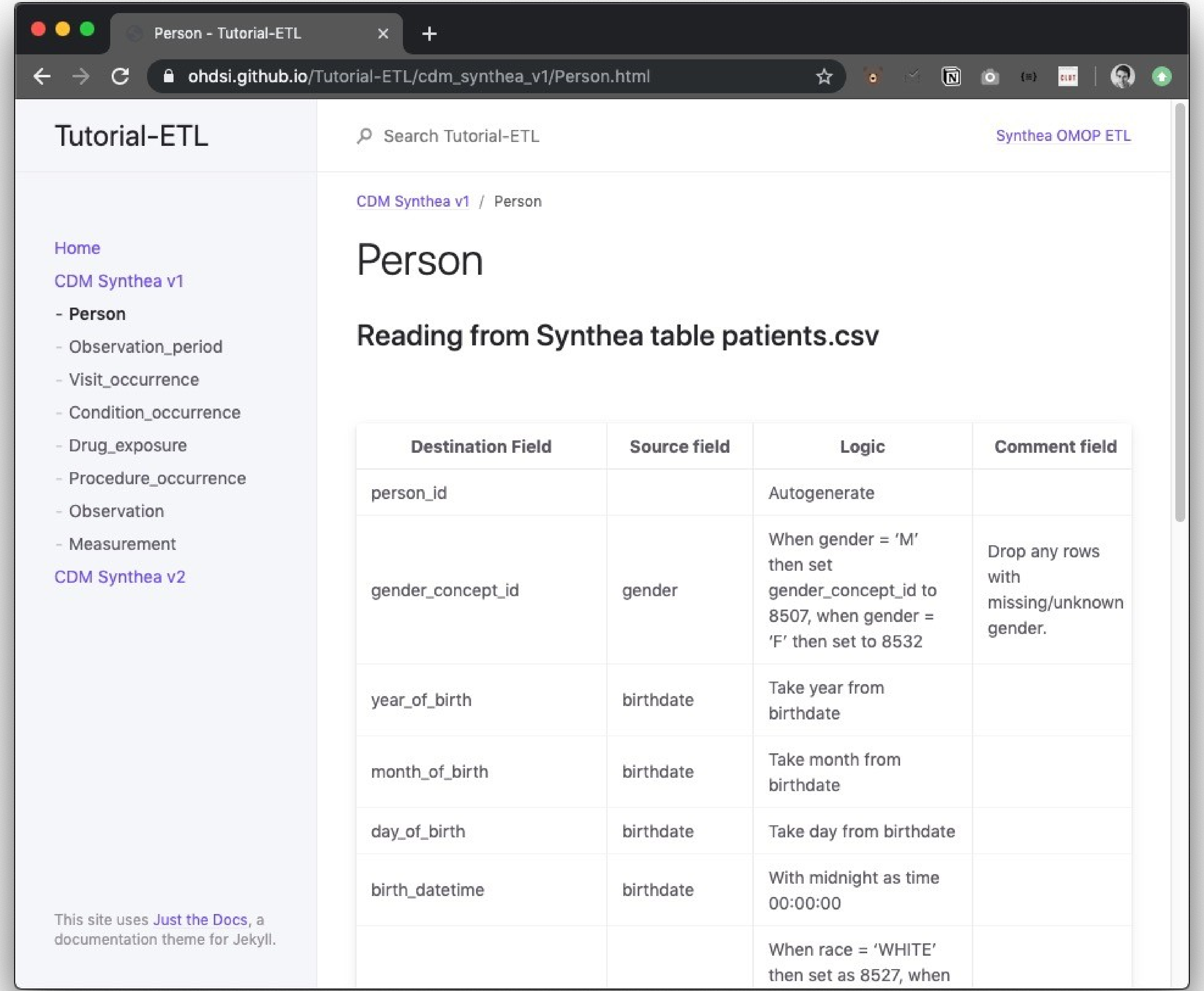


Table name:	patients.csv
Number of rows:	>= 11,073

Field	Type	Description
*id	varchar	
*birthdate	date	
deathdate	date	
*ssn	varchar	
drivers	varchar	
passport	varchar	
prefix	varchar	
*first	varchar	
*last	varchar	
suffix	varchar	
maiden	varchar	
marital	varchar	
*race	varchar	
*ethnicity	varchar	
*gender	varchar	
*birthplace	varchar	

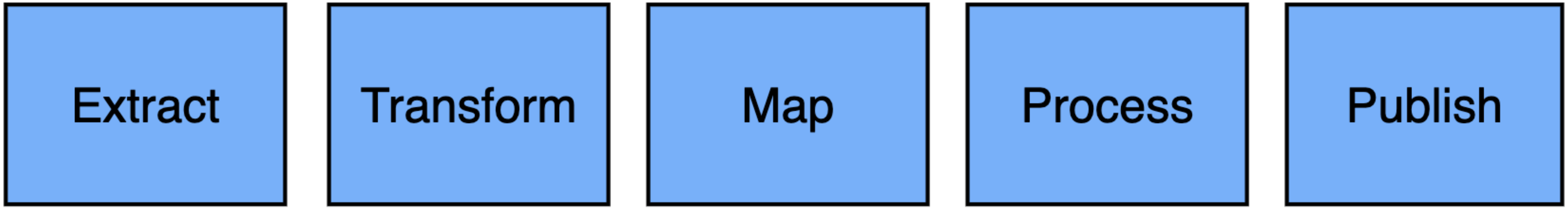


CDM Synthea v1 / Person

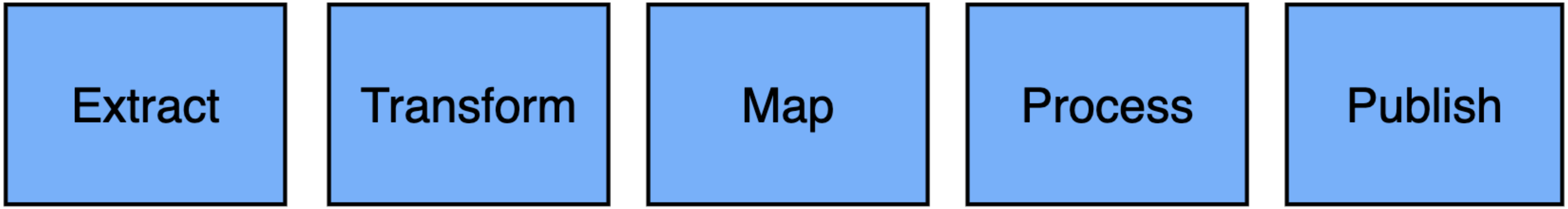
Person

Reading from Synthea table patients.csv

Destination Field	Source field	Logic	Comment field
person_id		Autogenerate	
gender_concept_id	gender	When gender = 'M' then set gender_concept_id to 8507, when gender = 'F' then set to 8532	Drop any rows with missing/unknown gender.
year_of_birth	birthdate	Take year from birthdate	
month_of_birth	birthdate	Take month from birthdate	
day_of_birth	birthdate	Take day from birthdate	
birth_datetime	birthdate	With midnight as time 00:00:00	
		When race = 'WHITE' then set as 8527, when	



dbt



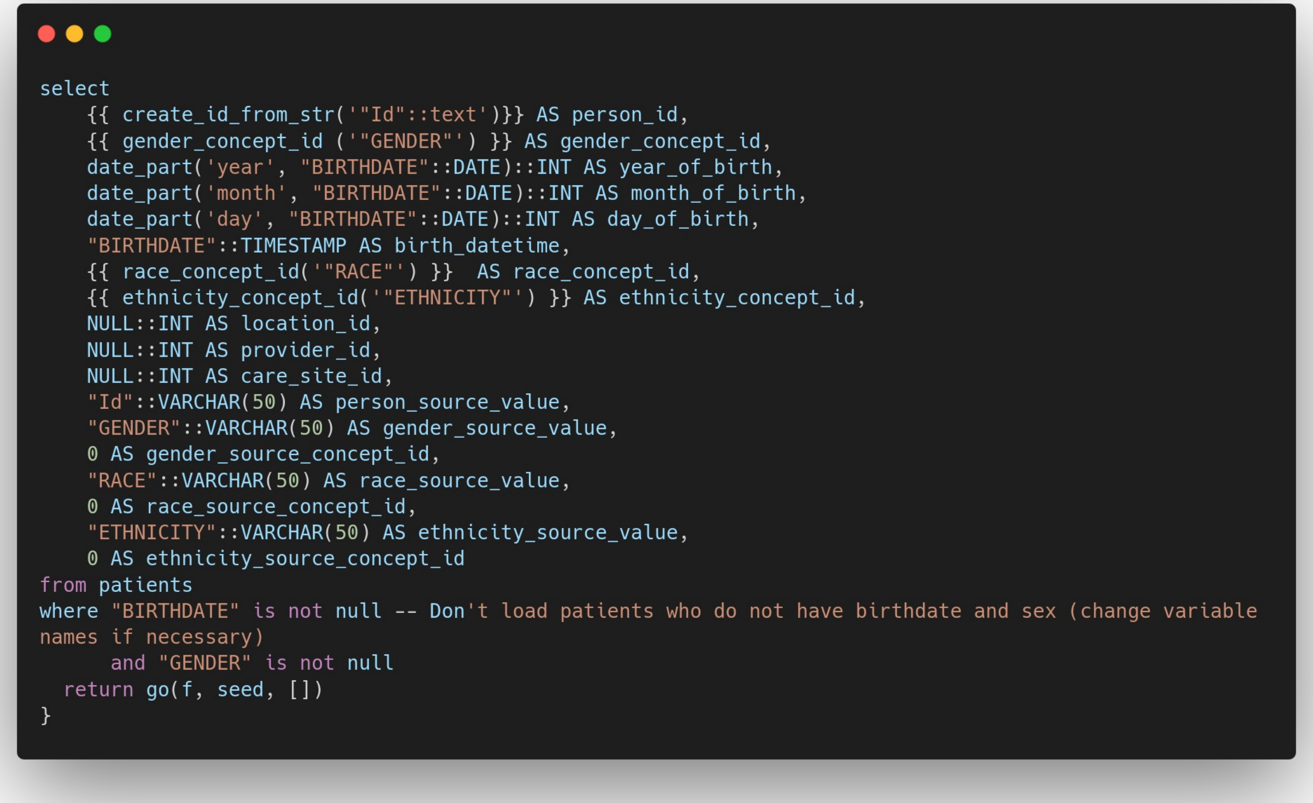
Extract

Transform

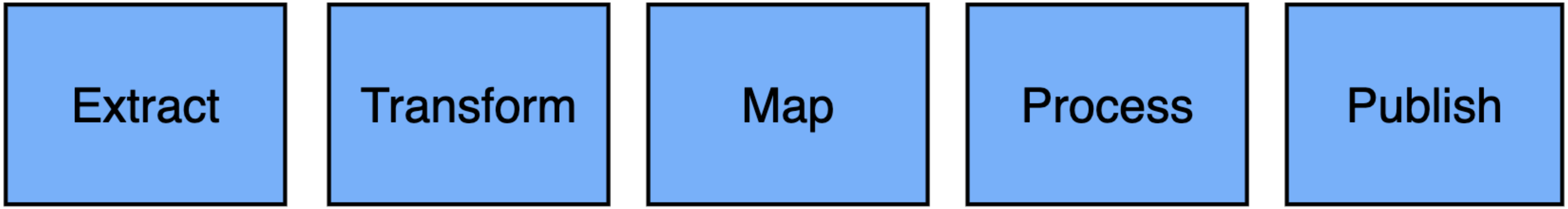
Map

Process

Publish



```
select
    {{ create_id_from_str('"Id"::text')}} AS person_id,
    {{ gender_concept_id ('"GENDER"') }} AS gender_concept_id,
    date_part('year', "BIRTHDATE"::DATE)::INT AS year_of_birth,
    date_part('month', "BIRTHDATE"::DATE)::INT AS month_of_birth,
    date_part('day', "BIRTHDATE"::DATE)::INT AS day_of_birth,
    "BIRTHDATE"::TIMESTAMP AS birth_datetime,
    {{ race_concept_id('"RACE"') }}  AS race_concept_id,
    {{ ethnicity_concept_id('"ETHNICITY"') }} AS ethnicity_concept_id,
    NULL::INT AS location_id,
    NULL::INT AS provider_id,
    NULL::INT AS care_site_id,
    "Id"::VARCHAR(50) AS person_source_value,
    "GENDER"::VARCHAR(50) AS gender_source_value,
    0 AS gender_source_concept_id,
    "RACE"::VARCHAR(50) AS race_source_value,
    0 AS race_source_concept_id,
    "ETHNICITY"::VARCHAR(50) AS ethnicity_source_value,
    0 AS ethnicity_source_concept_id
from patients
where "BIRTHDATE" is not null -- Don't load patients who do not have birthdate and sex (change variable
names if necessary)
      and "GENDER" is not null
  return go(f, seed, [])
}
```



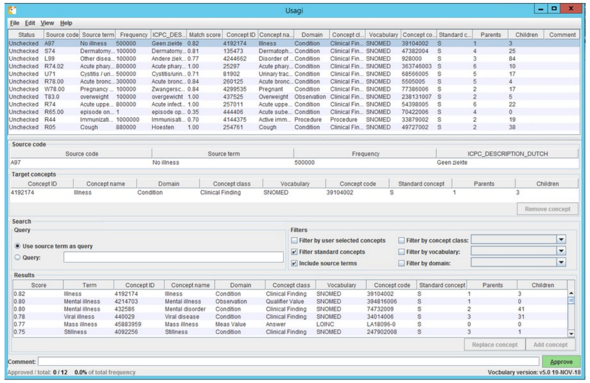
Extract

Transform

Map

Process

Publish



Status	Source code	Source term	Frequency	ICPC_DES...	Match score	Concept ID	Concept na...	Domain	Concept cl...	Vocabulary	Concept co...	Standard c...	Parents	Children	Comment
Unchecked	A97	No illness	500000	Geen ziekte	0.82	4192174	Illness	Condition	Clinical Fin...	SNOMED	39104002	S	1	3	
Unchecked	S74	Dermatomy...	100000	Dermatomy...	0.81	135473	Dermatoph...	Condition	Clinical Fin...	SNOMED	47382004	S	4	25	
Unchecked	L99	Other disea...	100000	Andere ziek...	0.77	4244662	Disorder of ...	Condition	Clinical Fin...	SNOMED	928000	S	3	84	
Unchecked	R74.02	Acute phary...	800000	Acute phary...	1.00	25297	Acute phary...	Condition	Clinical Fin...	SNOMED	363746003	S	6	10	
Unchecked	U71	Cystitis / uri...	500000	Cystitis/urin...	0.71	81902	Urinary trac...	Condition	Clinical Fin...	SNOMED	68566005	S	5	17	
Unchecked	R78.00	Acute bronc...	300000	Acute bronc...	0.84	260125	Acute bronc...	Condition	Clinical Fin...	SNOMED	5505005	S	5	4	
Unchecked	W78.00	Pregnancy ...	100000	Zwangersc...	0.84	4299535	Pregnant	Condition	Clinical Fin...	SNOMED	77386006	S	2	17	
Unchecked	T83.0	overweight	100000	overgewicht	1.00	437525	Overweight	Observation	Clinical Fin...	SNOMED	238131007	S	2	5	
Unchecked	R74	Acute uppe...	800000	Acute infect...	1.00	257011	Acute uppe...	Condition	Clinical Fin...	SNOMED	54398005	S	6	22	
Unchecked	R65.00	episode on...	1	episode op...	0.35	444406	Acute sube...	Condition	Clinical Fin...	SNOMED	70422006	S	4	0	
Unchecked	R44	Immunizati...	1000000	Immunisati...	0.70	4144375	Active imm...	Procedure	Procedure	SNOMED	33879002	S	2	19	
Unchecked	R05	Cough	880000	Hoesten	1.00	254761	Cough	Condition	Clinical Fin...	SNOMED	49727002	S	2	38	

Source code

Source code	Source term	Frequency	ICPC_DESCRIPTION_DUTCH
A97	No illness	500000	Geen ziekte

Target concepts

Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code	Standard concept	Parents	Children
4192174	Illness	Condition	Clinical Finding	SNOMED	39104002	S	1	3

Search

Query

● Use source term as query
○ Query:

Filters

□ Filter by user selected concepts
☑ Filter standard concepts
☑ Include source terms
□ Filter by concept class:
□ Filter by vocabulary:
□ Filter by domain:

Results

Score	Term	Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code	Standard concept	Parents	Children
0.82	Illness	4192174	Illness	Condition	Clinical Finding	SNOMED	39104002	S	1	3
0.80	Mental illness	4214703	Mental illness	Observation	Qualifier Value	SNOMED	394816006	S	1	0
0.80	Mental illness	432586	Mental disorder	Condition	Clinical Finding	SNOMED	74732009	S	2	41
0.78	Viral illness	440029	Viral disease	Condition	Clinical Finding	SNOMED	34014006	S	3	31
0.77	Mass illness	45883959	Mass illness	Meas Value	Answer	LOINC	LA18096-0	S	0	0
0.75	Stillness	4092256	Stillness	Condition	Clinical Finding	SNOMED	247902008	S	3	1

Comment:

Approved / total: **0 / 12** **0.0%** of total frequency

Vocbulary version: v5.0 19-NOV-18

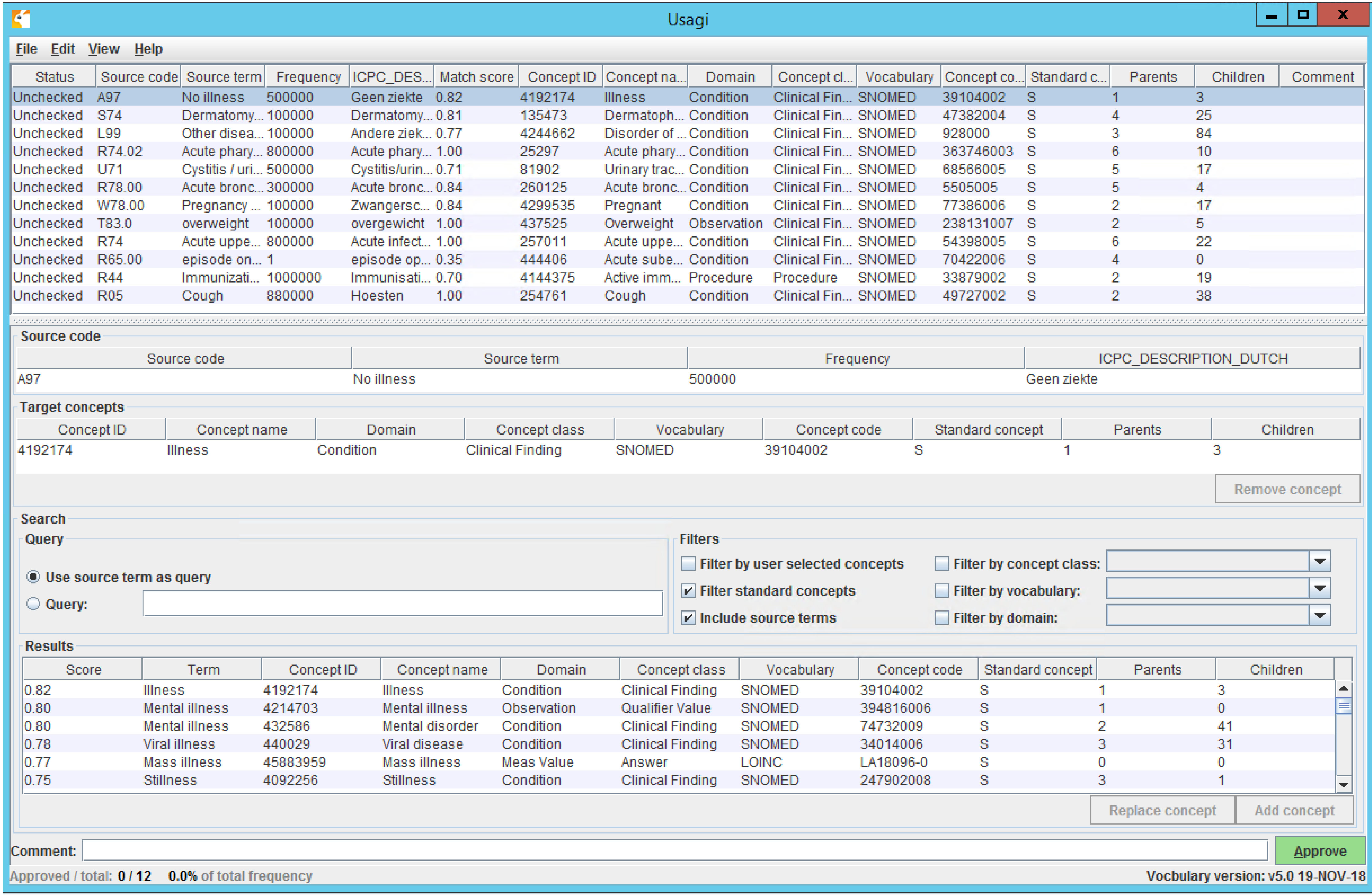
USAGI

Mapeo de vocabularios

- When the Vocabulary does not contain your source terms you will need to create a map to OMOP Vocabulary Concepts
- Usagi helps you to:
  - Find best matches, automatically and/or manually
  - Automatic matching based on text similarities (itf/df)
  - Create ‘source_to_concept_map’ table

Mapeo de vocabularios

- Get a copy of the Vocabulary from ATHENA
- Download Usagi
- Have Usagi build an index on the Vocabulary
- Load your source codes and let Usagi process them
- Review and update suggested mappings with someone who has medical knowledge
- Export codes into SOURCE_TO_CONCEPT_MAP



Status	Source code	Source term	Frequency	ICPC_DES...	Match score	Concept ID	Concept na...	Domain	Concept cl...	Vocabulary	Concept co...	Standard c...	Parents	Children	Comment
Unchecked	A97	No illness	500000	Geen ziekte	0.82	4192174	Illness	Condition	Clinical Fin...	SNOMED	39104002	S	1	3	
Unchecked	S74	Dermatomy...	100000	Dermatomy...	0.81	135473	Dermatoph...	Condition	Clinical Fin...	SNOMED	47382004	S	4	25	
Unchecked	L99	Other disea...	100000	Andere ziek...	0.77	4244662	Disorder of ...	Condition	Clinical Fin...	SNOMED	928000	S	3	84	
Unchecked	R74.02	Acute phary...	800000	Acute phary...	1.00	25297	Acute phary...	Condition	Clinical Fin...	SNOMED	363746003	S	6	10	
Unchecked	U71	Cystitis / uri...	500000	Cystitis/urin...	0.71	81902	Urinary trac...	Condition	Clinical Fin...	SNOMED	68566005	S	5	17	
Unchecked	R78.00	Acute bronc...	300000	Acute bronc...	0.84	260125	Acute bronc...	Condition	Clinical Fin...	SNOMED	5505005	S	5	4	
Unchecked	W78.00	Pregnancy ...	100000	Zwangersc...	0.84	4299535	Pregnant	Condition	Clinical Fin...	SNOMED	77386006	S	2	17	
Unchecked	T83.0	overweight	100000	overgewicht	1.00	437525	Overweight	Observation	Clinical Fin...	SNOMED	238131007	S	2	5	
Unchecked	R74	Acute uppe...	800000	Acute infect...	1.00	257011	Acute uppe...	Condition	Clinical Fin...	SNOMED	54398005	S	6	22	
Unchecked	R65.00	episode on...	1	episode op...	0.35	444406	Acute sube...	Condition	Clinical Fin...	SNOMED	70422006	S	4	0	
Unchecked	R44	Immunizati...	1000000	Immunisati...	0.70	4144375	Active imm...	Procedure	Procedure	SNOMED	33879002	S	2	19	
Unchecked	R05	Cough	880000	Hoesten	1.00	254761	Cough	Condition	Clinical Fin...	SNOMED	49727002	S	2	38	

Source code

Source code	Source term	Frequency	ICPC_DESCRIPTION_DUTCH
A97	No illness	500000	Geen ziekte

Target concepts

Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code	Standard concept	Parents	Children
4192174	Illness	Condition	Clinical Finding	SNOMED	39104002	S	1	3

Search

Query

● Use source term as query

○ Query:

Filters

□ Filter by user selected concepts

☑ Filter standard concepts

☑ Include source terms

□ Filter by concept class:

□ Filter by vocabulary:

□ Filter by domain:

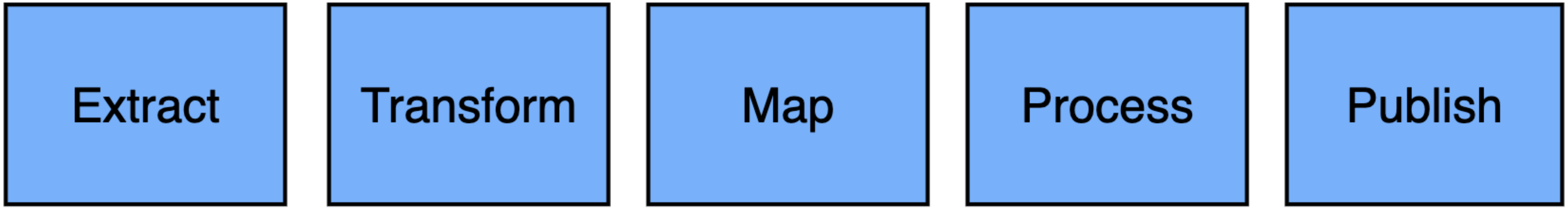
Results

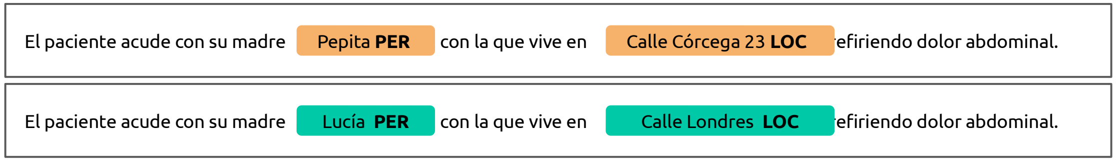
Score	Term	Concept ID	Concept name	Domain	Concept class	Vocabulary	Concept code	Standard concept	Parents	Children
0.82	Illness	4192174	Illness	Condition	Clinical Finding	SNOMED	39104002	S	1	3
0.80	Mental illness	4214703	Mental illness	Observation	Qualifier Value	SNOMED	394816006	S	1	0
0.80	Mental illness	432586	Mental disorder	Condition	Clinical Finding	SNOMED	74732009	S	2	41
0.78	Viral illness	440029	Viral disease	Condition	Clinical Finding	SNOMED	34014006	S	3	31
0.77	Mass illness	45883959	Mass illness	Meas Value	Answer	LOINC	LA18096-0	S	0	0
0.75	Stillness	4092256	Stillness	Condition	Clinical Finding	SNOMED	247902008	S	3	1

Comment:

Approved / total: 0 / 12 0.0% of total frequency

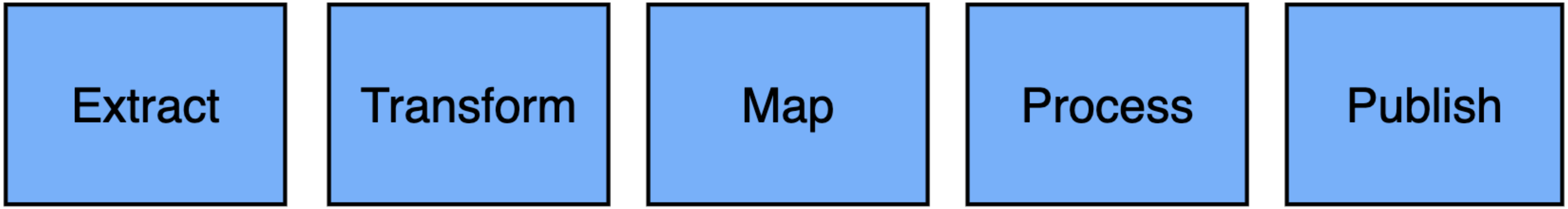
Vocbulary version: v5.0 19-NOV-18





El paciente acude con su madre Pepita **PER** con la que vive en Calle Córcega 23 **LOC** refiriendo dolor abdominal.

El paciente acude con su madre Lucía **PER** con la que vive en Calle Londres **LOC** refiriendo dolor abdominal.



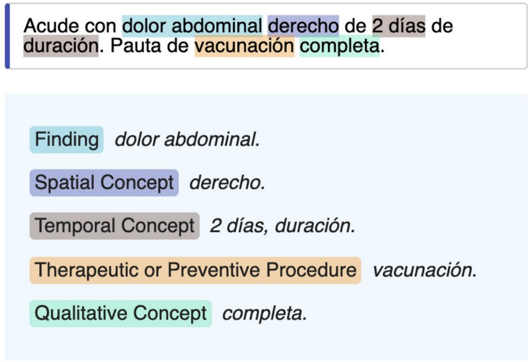
Extract

Transform

Map

Process

Publish



Acude con dolor abdominal derecho de 2 días de duración. Pauta de vacunación completa.

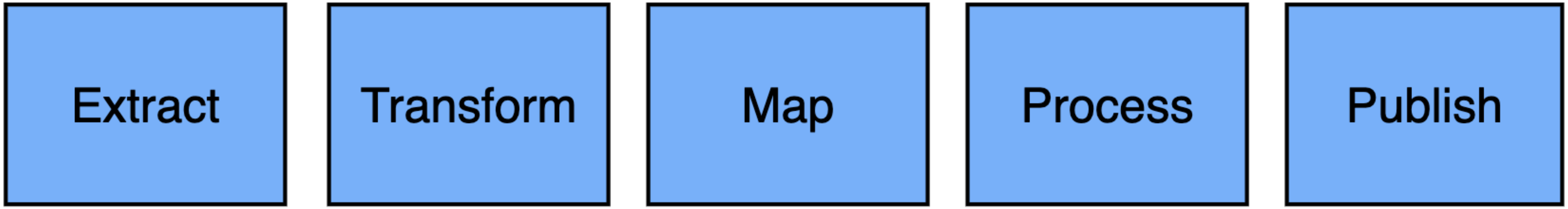
Finding *dolor abdominal.*

Spatial Concept *derecho.*

Temporal Concept *2 días, duración.*

Therapeutic or Preventive Procedure *vacunación.*

Qualitative Concept *completa.*



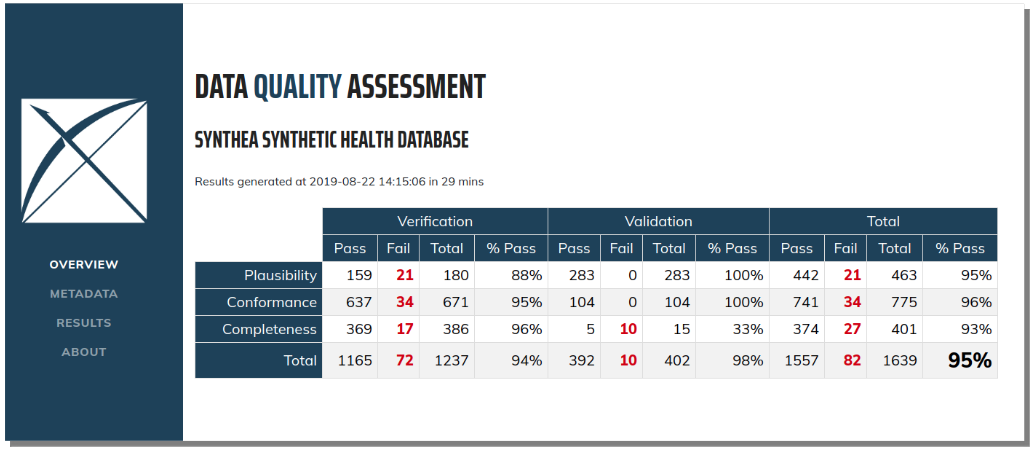
Extract

Transform

Map

Process

Publish

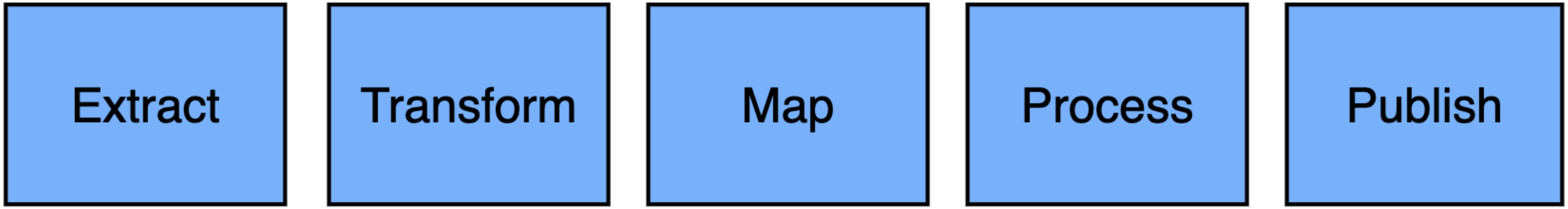


DATA QUALITY ASSESSMENT

SYNTHEA SYNTHETIC HEALTH DATABASE

Results generated at 2019-08-22 14:15:06 in 29 mins

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	159	21	180	88%	283	0	283	100%	442	21	463	95%
Conformance	637	34	671	95%	104	0	104	100%	741	34	775	96%
Completeness	369	17	386	96%	5	10	15	33%	374	27	401	93%
Total	1165	72	1237	94%	392	10	402	98%	1557	82	1639	95%



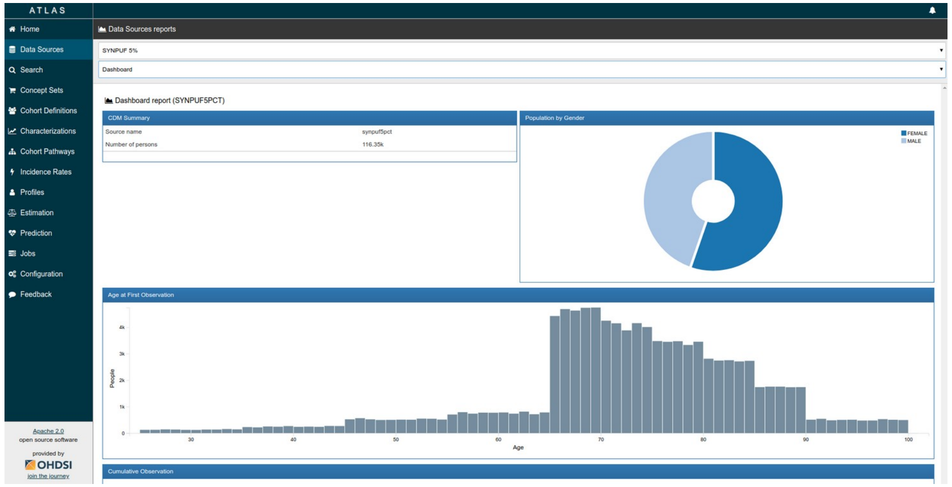
Extract

Transform

Map

Process

Publish



SYNPUF 5%

Dashboard

Dashboard report (SYNPUF5PCT)

CDM Summary	
Source name	synpuf5pct
Number of persons	116.35k