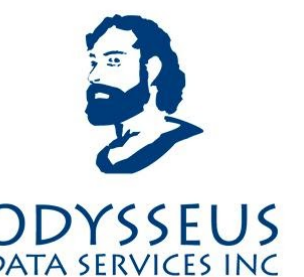# Source vocabulary mapping, typical pitfalls, solutions and quality assurance

Oleg Zhuk, 28-apr-2020

e-mail: oleg.zhuk@odysseusinc.com
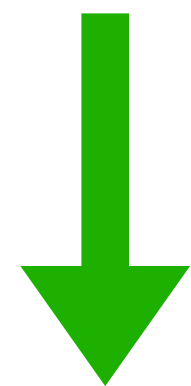
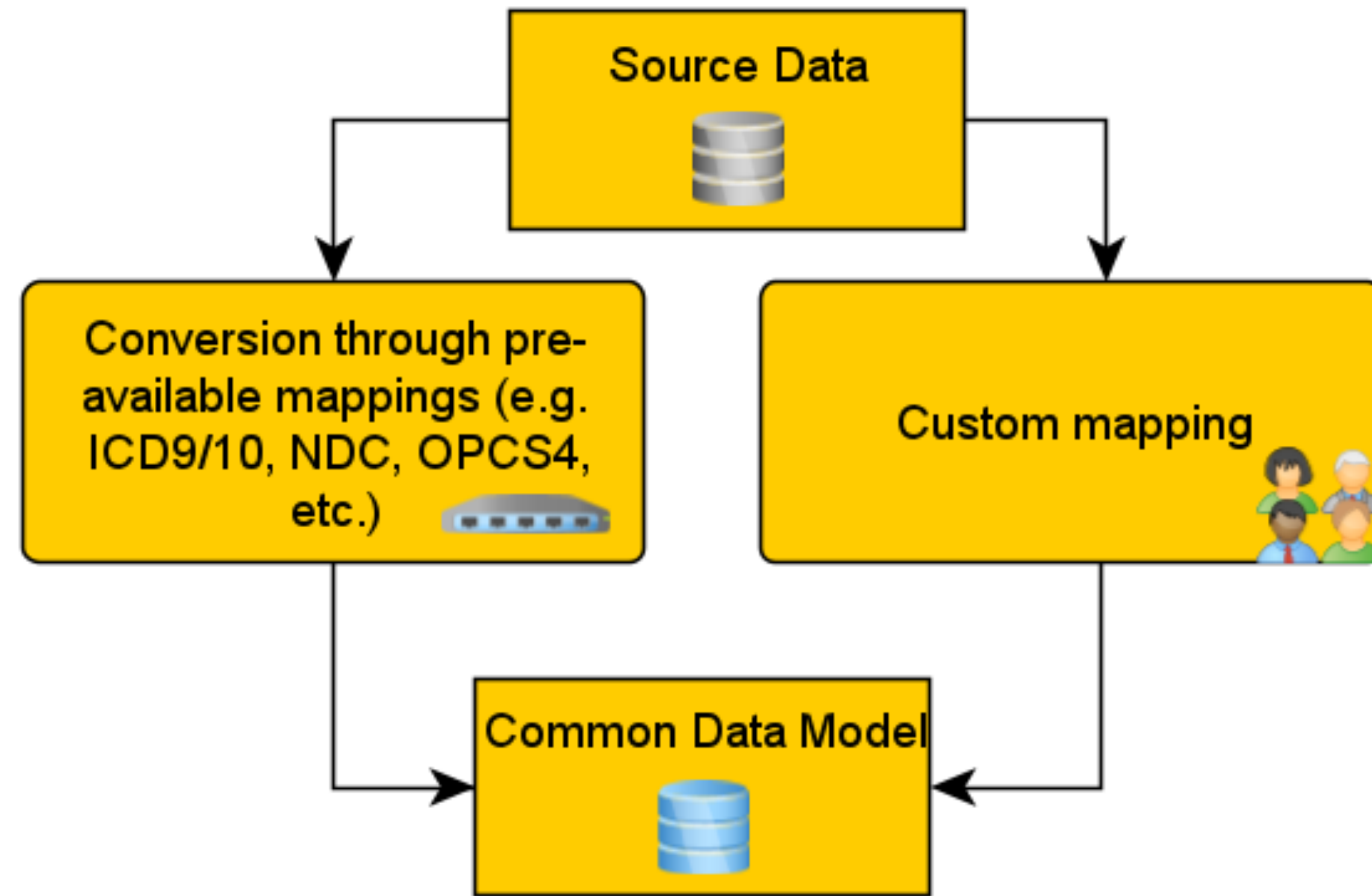OHDSI Community forum user link

ETL process

Source data:

Code	Vocabulary	Name
I48	ICD10CM	Atrial fibrillation and flutter

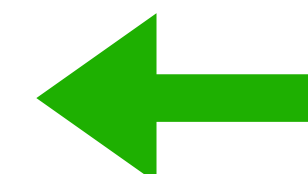**Maps_to** relationship *(source vocabulary ICD10CM has already been OMOPed and mapped to standard)*

Target concept:

concept_code	concept_id	vocabulary_id	concept_name
195080001	4108832	SNOMED	Atrial fibrillation and flutter

Source Data

Conversion through pre-available mappings (e.g. ICD9/10, NDC, OPCS4, etc.)

Custom mapping

Common Data Model

Source data:

Code	Vocabulary	Name
A22	source_specific	Atrial fibrillation and flutter

Custom creation of new **Maps_to** relationship *(source specific vocabulary hasn't been OMOPed and mapped to standard)* - ***custom mapping process***

OHDSI

ODYSSEUS
DATA SERVICES INC

Custom mapping process and Q&A steps

3 major steps

Working with source data: define **source_code** and **source_code_description**	Creation of new Maps_to relationships	Integration into existing CDM

Source testing:

- source_code uniqueness
- length of the fields

Unit testing + semantic testing:

- To check general mapping consistency and accordance with CDM specification/constraints

- To check concepts' semantic match

Integration testing:

- Code - description uniqueness among the whole project

OHDSI

ODYSSEUS
DATA SERVICES INC

Custom mapping process and Q&A steps

Source testing:

- source_code uniqueness
- CDM constraints: length of the fields

Length of the fields	
concept_code	varchar(50)
vocabulary_id	varchar(20)
concept_name	varchar(255)

Unit testing + semantic testing:

- target_concept_id is standard and valid
- *Maps to* is mapped to right domain/concept_class
- 2 Observations/Measurements for one value and vice versa
- *Maps to value* without *Maps to*
- License issue check
- Concepts that can't be used without *Maps to value*
- Key term lost

Integration testing:

- Code - description uniqueness among the whole project

MANUAL CROSS-REVIEW!

OHDSI

ODYSSEUS
DATA SERVICES INC

Source checks

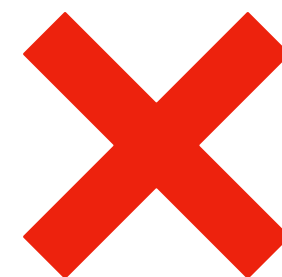| Test	Pitfall	Solution
Check source_code uniqueness among vocabulary	Data extraction error Source_code definition error	Use another field or even concatenation of some fields as a source_code
Check the length of fields	CDM constraints violation	Use another field or cut to the needed length. Extend CDM field length

Example:

type_of_reaction	substance
Allergy	Penicillin
Intolerance	Penicillin

source_code	source_code_description
Penicillin	Penicillin
Penicillin	Penicillin

source_code	source_code_description
Allergy|Penicillin	Allergy|Penicillin
Intolerance|Penicillin	Intolerance|Penicillin

Source

OHDSI

ODYSSEUS
DATA SERVICES INC

Unit tests

Test	Pitfall	Solution
Check if a target concept has the same values in the concept table and are valid and standard	Each ETL conversion runs on a specific OMOP CDM Vocabulary version. Re-run on updated Vocabulary version may result in some concepts are not Valid/Standard anymore. Manual mapping mistakes can lead to corrupted target_concept_id and semantic evaluation of a different concept.	Use the same Vocabulary version in mapping and ETL. Amend corrupted target_concept_id.

Example:

source_code_description	relationship_id	target_concept_id	target_concept_name	target_vocabulary_id	target_domain_id	valid_end_date
Kleefstra syndrome	Maps to	44805996	Kleefstra syndrome	Snomed	Condition	1/30/2018
Kleefstra syndrome	Maps to	37110119	Kleefstra syndrome	Snomed	Condition	12/31/2099

OHDSI

ODYSSEUS
DATA SERVICES INC

Unit tests

Test	Pitfall	Solution
Maps to mapping to abnormal Domain/Concept class	Each CDM table and field has its purpose so the list of possible concept Domains/Classes is predefined. E.g., mapping to Unit, Meas Value, Specimen Domains must not be used if ETL rules are adjusted to event_concept_id fields.	Amend mapping. ⸤SEP⸥⸤SEP⸥ Change ETL rules.

Example:

source_code_description	relationship_id	target_concept_id	target_concept_name	target_vocabulary_id	target_domain_id	target_concept_class_id
Asthma	Maps to	45877009	Asthma	LOINC	Meas Value	Answer
Asthma	Maps to	317009	Asthma	Snomed	Condition	Clinical Finding

Unit tests

Test	Pitfall	Solution
Value ambiguous mapping (2 Observation/Measurement concepts for 1 value or vice versa)	This leads to duplication of records in CDM.	Amend the multiple mapping.⌞SEP⌟⌞SEP⌟Skip if duplication is required, e.g. in Allergy data.

Example:

source_code_description	relationship_id	target_concept_id	target_concept_name	target_vocabulary_id	target_domain_id	target_concept _class_id
Allergy to house dust	Maps to	4304626	House dust RAST	Snomed	Measurement	Procedure
Allergy to house dust	Maps to value	9191	Positive	Snomed	Meas Value	Qualifier Value
Allergy to house dust	Maps to	4048168	Allergy to house dust	Snomed	Observation	Clinical Finding
Allergy to house dust	Maps to	4048168	Allergy to house dust	Snomed	Observation	Clinical Finding

Unit tests

Test	Pitfall	Solution
Maps to value without 'Maps to'	Value_as_concept_id field cannot be populated if event_concept_id is not defined.	Add 'Maps to' relationship mapping.

Example:

source_code_description	relationship_id	target_concept_id	target_concept_name	target_vocabulary_id	target_domain_id	target_concept _class_id
Allergy to phytosterols	Maps to value	19044812	Phytosterols	RxNorm	Drug	Ingredient
Allergy to phytosterols	Maps to	4169307	Allergy to substance	Snomed	Observation	Clinical Finding
Allergy to phytosterols	Maps to value	19044812	Phytosterols	RxNorm	Drug	Ingredient

OHDSI

ODYSSEUS
DATA SERVICES INC

Unit tests

Test	Pitfall	Solution
Used vocabularies	Most vocabularies are used only in certain circumstances⸬SEP⸬SEP⸬Some vocabularies are license-required	Amend the mapping⸬SEP⸬SEP⸬Confirm the license

Example:

source_code_description	relationship_id	target_concept_id	target_concept_name	target_vocabulary_id	target_domain_id	target_concept_class_id
implant /abutment supported fixed denture for partially edentulous arch - mandibular	Maps to	944898	implant /abutment supported fixed denture for partially edentulous arch - mandibular	CDT	Observation	CDT
implant /abutment supported fixed denture for partially edentulous arch -			IMPLANT/ABUTMENT SUPPORTED FIXED DENTURE FOR PARTIALLY			

Unit tests

Test	Pitfall	Solution
Concepts that can't be used without 'Maps to value' link	Includes 'History of', 'Disease suspected' and other concepts that don't make sense without 'Maps to value'.	Add 'Maps to value' mapping. Remove unnecessary mapping

Example:

| source_code_description	relationship_id	target_concept_id	target_concept_name	target_vocabulary_id	target_domain_id	target_concept _class_id
Family history of viral pneumonia	Maps to	4083519	Family history of disorder	Snomed	Observation	Context-dependent
Family history of viral pneumonia	Maps to	4083519	Family history of disorder	Snomed	Observation	Context-dependent
Family history of viral pneumonia	Maps to value	261326	Viral pneumonia	Snomed	Condition	Clinical Finding

Unit tests

Test	Pitfall	Solution
Key terms loss/misuse	Acute, recurrent, suspected, chronic, left/right and other attributes might be lost or misused	Add/amend mapping

Example:

source_code_description	relationship_id	target_concept_id	target_concept_name	target_vocabulary_id	target_domain_id	target_concept _class_id
Acute arthritis of left knee joint	Maps to	4159739	Arthritis of knee	Snomed	Condition	Clinical Finding
Acute arthritis of left knee joint	Maps to	759891	Arthritis of left knee	Snomed	Condition	Clinical Finding
Acute arthritis of left knee joint	Maps to	4000634	Acute arthritis	Snomed	Condition	Clinical Finding

Mapping algorithm

1. Source data analysis and source code definition, excluding junk and meaningless source terms. Source testing

2. Automated term match in Standardized vocabularies - subsequent matching by following pathways:
(a) concept_name of Standard concepts;
(b) concept_synonym_name of Standard concepts;
(c) concept_name of non-Standard concepts + corresponding 'Maps to' link;
(d) concept_synonym_name of non-Standard concepts + corresponding 'Maps to' link;
(e) recursive PL/pgSQL function for removing word endings;
(f) concept_code for UCUM vocabulary.
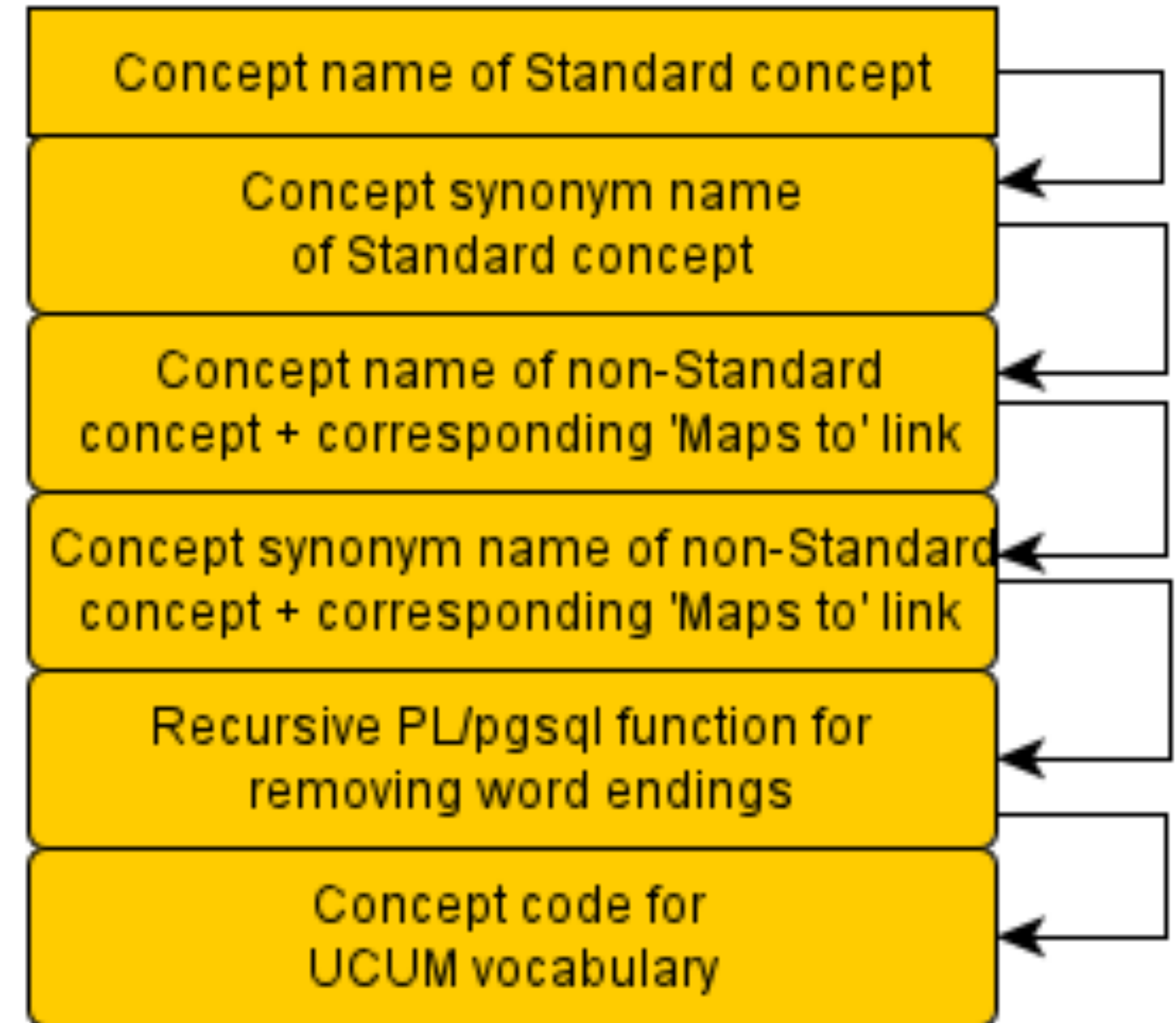
Match with a references

USAGI (https://github.com/OHDSI/Usagi) automated mapping and manual

Conclusion

- The QA/QC tests and algorithm that were developed may improve mapping accuracy and effectiveness of the process.

- We recommend implementing both automated tests and those that require further expert review.

- The current mapping rate achieved with the help of the provided mapping algorithm is around 50%.

- We believe that further improvements are possible with the implementation of Natural language processing (NLP) and an extensive increase in the number of references.