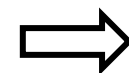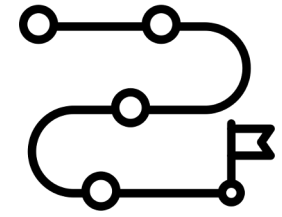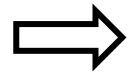# Increasing the 'Open' and 'Community' in an Open Community Data Standard, the OHDSI Standardized Vocabularies
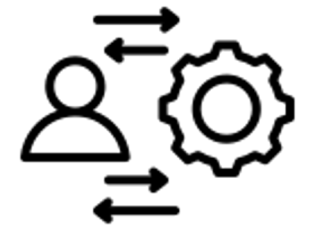
06-June-2023

Alexander Davydov
Anna Ostropolets
Christian Reich
& Vocabulary Team

Overview of the talk

Landscape assessment of community needs

Committee and improvement process

Roadmap and releases

Community involvement and contribution

Landscape assessment:
What does the community need?

Landscape assessment

Conducted in Jan - Feb | 2 parts: overall and data-source specific |Community commented on the draft in Mar - Apr 2023

What we asked about

- Which vocabularies you use in ETL, research or development

- Problems you encountered with Vocabularies completeness and correctness

- Problems you encountered with Vocabularies recency and updates (versioning, upload/download)

- What you like to see improved

More details in the document: Teams General -> Files -> OHDSI Standardized Vocabularies landscape assessment

Landscape assessment

Part I: 188 responses from 144 institutions | Part II: Vocabularies use in 60 data sources

FINDINGS

- 87% of the community feels confident about Vocabularies' integrity
  (to be used as a benchmark)

- Most used vocabularies: SNOMED, ICD 9/10(US and int. versions), MedDRA, ICDO3, ATC, RxNorm / RxE, ICD10PCS, ICD9Proc, CPT4, LOINC, CVX, HCPCS, UCUM, NDC, NAACCR, Cancer Modifier

- Most of the community updates their data annually or semi-annually

Vocabulary	Used in data	Used in research
SNOMED	57%	86%
RxNorm	33%	79%
ATC	45%	62%
ICD-10PCS	43%	51%
LOINC	68%	25%
ICD-9-Proc	48%	43%
CPT4	50%	40%
HCPCS	42%	33%
ICD-9(CM)	62%	<10%
ICDO3	32%	35%
ICD-10(CM)	57%	<10%
RxNorm Extension	<10%	53%
ICD-10 (int. versions)	48%	<10%
NDC	33%	<10%
MedDRA	10%	24%
ICD-9 (int. versions)	28%	<10%
NAACCR	17%	15%
CVX	18%	13%
UCUM	<10%	26%
Cancer Modifier	<10%	25%
HemOnc	10%	14%
OncoTree	10%	11%
Medicare Specialty	15%	<10%
Revenue Code (CMS)	15%	<10%
CMS Place of Service	13%	<10%
CDT	12%	<10%
DRG (CMS)	12%	<10%
Multum	12%	<10%
NCIt	12%	<10%
NUCC	12%	<10%
ABMS	10%	<10%
dm+d	10%	<10%
OncoKB	10%	<10%
ClinVar	<10%	10%
Nebraska Lexicon	<10%	10%
HGNC	<10%	10%

Landscape assessment

Part I: 188 responses from 144 institutions | Part II: Vocabularies use in 60 data sources

NEEDS

- Transparent release schedule, tools for vocabulary changes, versioning

- Transparent & robust quality assurance and control procedures

- Better coverage and comprehensive hierarchies

- More documentation, incl. educational materials

Vocabulary Committee

- Oversees the improvement activities
- Output and decisions are communicated to the community

Vocabularies process improvement activities

1. Landscape assessment of community needs

2. Roadmap and release schedule

3. Documentation & guidelines (developer, end-user, community contribution)

4. Quality framework and QA system

5. Vocabulary versioning

6. Common development infrastructure

7. Better code and automation (better machinery)

8. Vocabulary one-stop-shop and self-service

Roadmap:
Stable and transparent release schedule

Roadmap

Roadmap

1. Stable cadence

As most of the community members refresh the Vocabularies and the data annually or semi-annually, the cadence of releases is twice a year. Such a schedule has a higher productivity, transparency in the content of the releases, and better version alignment in the community. Two releases (August and February) correlate with the source release schedule. An intermediate release in May 2023 is planned for work already accomplished.

2. Combination of maintenance and process improvement

Vocabulary work balances:

- Routine maintenance,
- Automation, usually across concepts and vocabularies of one domain at a time (overhauls, machinery improvements, etc.),
- Process improvement (e.g., community contribution guidelines or version control).

3. Prioritization based on the needs of the community

The roadmap is based on a continuous need assessment of the community, both in terms of vocabulary maintenance as well as process improvement.

Improvement activities

Vocabulary-specific overhauls and improvements include:

1. SNOMED overhaul

- Stable domain and concept class id assignment.
- Alignment of the validity dates with the source.
- Fix of the problem with replacement relationships (such as "Concept replaced by") not having "Maps to" links that prevent the users from automatically following the "Maps to" relationships from non-standard to standard counterparts.

2. ICD family improvement

- Mapping re-use across ICD family to identify the discrepancies and similarities across different versions of ICD and improve the consistency of mappings.
- Incorporation of the mappings provided by SNOMED-CT and other sources.
- Fix of the source (CIAML) file processing to capture the ICD concepts currently missing.

Activities per release

The plan for 2023 Q1 - 2024 Q2 includes the refreshes of the following commonly used vocabularies: SNOMED-CT, ICD family, Read, RxNorm, CVX, LOINC, HCPCS, ICD10PCS,

MedDRA, MeSH, NAACCR, dm+d, as well as improvement activities tailored to the most commonly reported problems described above.

Table 1 outlines the vocabularies included in each release as per the roadmap above.

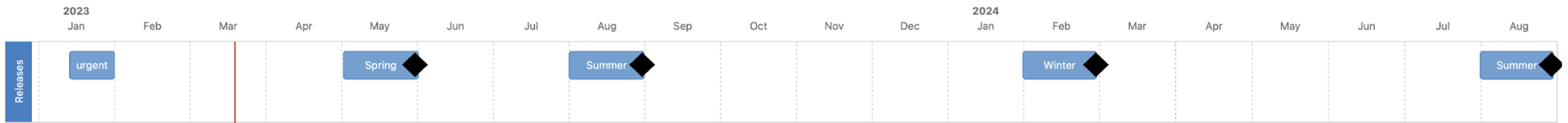**Table 1.** Vocabularies and activities included in each release.

Activity	Vocabulary version and modification	Name
Spring release, May 2023		
CVX	refresh (20230222 version) and refactored code	Maria, Timur
dm+d	refresh (20220927 version) and refactored code	Oleg, Timur
HCPCS	improvement + refresh (Apr 2023 version)	Masha, Timur
MeSH	refresh (2022 version) and refactored code	Timur
NAACCR	mapping addition	Vlad, Timur
NDC	refresh (20230319 version)	Oleg
RxNorm	refresh (20230306 version)	Oleg, Timur
RxNorm Extension	refresh (May 2023 version)	Oleg
Smoking hierarchy	mapping addition	Maria, Timur
SPL	refresh (20230319 version)	Oleg

11

Roadmap: stable cadence

Enables more transparency and allows users to better plan their activities

- Two major releases aligned with the source refresh schedule - **August and February** (planned on the roadmap now: August 2023, February 2024, August 2024)

Maintenance: Aug 2023 and February 2024 releases

August 2023
CPT4 refresh
LOINC refresh
NDC refresh
RxNorm refresh
RxNorm Extension update
SPL refresh
VANDF refresh

RxNorm update

February 2024	
CVX refresh	***SNOMED overhaul*** and refresh (UK, US, Int)
LOINC refresh	
NDC refresh	***ICD10 improvement*** + refresh (ICD10CM, ICD10, ICD10CN, ICD10GM, CIM10)
RxNorm refresh	
RxNorm Extension update	MedDRA ***improvement*** + refresh
SPL refresh	Read mapping refresh
HCPCS refresh	ICD10PCS refresh

August 2024: refreshes of vocabularies above + ***ATC overhaul***

More information: Vocabulary-v5.0 GitHub Wiki [bit.ly/43q8vc6](bit.ly/43q8vc6)

Roadmap: vocabularies

Vocabularies on maintenance: ATC, CPT4, CVX, HCPCS, ICDs (10, 10CM, 10GM, 10PCS, CIM10, KCD7), LOINC, MedDRA, OMOP Invest Drug, RxNorm, RxNorm Extension, SNOMED, VANDF + **administrative/supportive vocabularies (type, status)**.

Obsolete vocabularies: CCS, Cohort, Cohort Type, Condition Type, Cost Type, Death Type, Device Type, Drug Type, Episode Type, Meas Type, Multilex, NDFRT, Note Type, Obs Period Type, Observation Type, OXMIS, Procedure Type, SMQ, Specimen Type, Visit Type.

Vocabularies no longer maintained:

- **Oncology:** Cancer Modifier, CAP, HemOnc, ICDO3, NAACCR, Nebraska Lexicon, OncoTree, OMOP Genomic group (OncoKB, CGI, CIViC, ClinVar, JAX, HGNC, NCIt).
- **Source:** AMIS, AMT, BDPM, CCAM, CIEL, CTD, dm+d*, DPD, EDI, EU Product, GCN_SEQNO, Gemscript, GGR, GPI, JMDC, KDC, KNHIS, MeSH*, Multum, NCCD, OPS, PCORNet, SUS, UK Biobank, VA Class.
- **Standard/Classification:** APC, CDT, DRG, EphMRA ATC, ETC, Indication, MDC, NFC, OPCS4, OSM, US Census, SNOMED Veterinary.

* - refreshed one last time.

Community contribution

Community contribution

- Rolled out in two parts

- First part focuses on the use cases that can be incorporated through filling the templates

- We created documentation on GitHub, templates and a pipeline to follow

More info: bit.ly/42qOscr

Community contribution Part I: currently supported use cases

ADD NEW CONTENT

Does not impact the rest of the community
Can be incorporated quickly (no review)

- Adding *non-standard* concepts to existing vocabularies
- Adding synonyms to existing concepts
- Adding new mappings
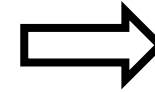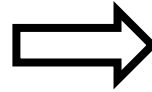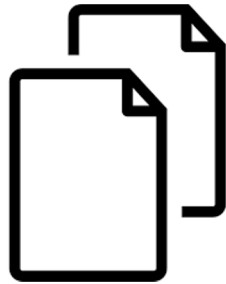- Adding *non-standard* vocabularies

MODIFY EXISTING CONTENT

Impacts community as already is used by others
Requires review of the Vocab Team & other WGs/broader community

- Modifying existing mappings
- Modifying concept attributes (e.g., domain)
- Promoting non-standard concepts to standard

More info: bit.ly/42qOscr

Community contribution: pipeline

Create issue on Vocabulary
GitHub
(cat: Add content; Modify content)
Describe your use case
and proposed contribution

Choose the template,
download and fill

- Content
- Meta-data
- Checklist (QA)

Submit the template
to Google Drive

Come to review session
if needed

Modify and re-upload if
needed

See the contribution in the
next release!

If submitted and approved
at least **2 months prior to
release (end of June for
the next one)**

More info: bit.ly/42qOscr

18

What to do…

If your use cases are supported?

Submit a GitHub issue!

If your use cases are YET not supported?

Talk to us!

What comes next?

- We are gathering requests for contribution for dry runs by the end of June!

- June 13 WG meeting: please come with your questions.

- Quality assurance and control improvements to follow.

All links:

Landscape assessment report	Teams General -> Files -> OHDSI Standardized Vocabularies landscape assessment
Roadmap	bit.ly/43q8vc6 (Vocabulary-v5.0 GitHub Wiki)
Community contribution guidelines	bit.ly/42qOscr (Vocabulary-v5.0 GitHub Wiki)