Update on OHDSI Standardized Vocabularies

29-Aug-2023

Alexander Davydov
Anna Ostropolets
Oleg Zhuk
Christian Reich
& Vocabulary Team
Overview

1. Recap of the plan for Vocabularies improvement

2. Accomplishments so far
   ○ New process & Vocabulary Committee
   ○ Landscape assessment
   ○ Roadmap and releases
   ○ Community contribution
   ○ August release
   ○ Experience with community contribution
   ○ Vocabulary WG
   ○ Quality management system

3. What is missing, issues faced and ways to improvement
Recap of Vocabularies improvement: What was planned?
Vocabularies process improvement

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
Vocabulary Committee

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

- 87% of the community feels confident about Vocabularies’ integrity

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

- Most update data annually or semi-annually

NEEDS

- Transparent release schedule

- Vocabulary changes, versioning

- Transparent QA

- Better coverage and hierarchies

- More documentation and educational materials
This page provides you with the planned maintenance and improvement activities around the OHDSI Standardized Vocabularies. This is to be treated as a forecast. Below you can find the content of each release and an overview of the planned improvement activities (detailed content to be posted separately).

Roadmap 2023 Q1 - 2024 Q2:

More information: Vocabulary-v5.0 GitHub Wiki bit.ly/43q8vc6
Vocabularies on maintenance: ATC, CPT4, CVX, HCPCS, ICDs (10, 10CM, 10GM, 10PCS, CIM10, KCD7), LOINC, MedDRA, NDC, 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Vocabularies no longer maintained: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, HemOne, 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.
## Maintenance: Aug 2023 and February 2024 releases

### August 2023

- CPT4 refresh
- LOINC refresh
- NDC refresh
- RxNorm refresh
- RxNorm Extension update
- SPL refresh
- VANDF refresh

### February 2024

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

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

More information: Vocabulary-v5.0 GitHub Wiki [bit.ly/43q8vc6](https://bit.ly/43q8vc6)
What have you got so far?
**Community contribution Part I:**

currently supported use cases

---

**ADD NEW CONTENT**

Does not impact the rest of the community
Can be incorporated quickly (little review)

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

---

**MODIFY EXISTING CONTENT**

Requires review of the Vocab Team & other WGs/broader community

- Modifying existing **mappings**
- Modifying **concept fields** (e.g., description, domain, but not code)
- **Promoting non-standard** concepts to standard

---

Community contribution: pipeline

Issue on GitHub ➞ Template
  • Content
  • Meta-data
  • Checklist (QA) ➞ Submission
➤ Review ➞ Release

More info: bit.ly/42qOscr
August release
# August 2023 release: vocabularies on maintenance

<table>
<thead>
<tr>
<th>Vocabulary</th>
<th>Version planned</th>
<th>Version released</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPT4</td>
<td>Spring 2023</td>
<td></td>
<td>Masha Khitrun</td>
</tr>
<tr>
<td>LOINC</td>
<td>2.74</td>
<td>2.75 version</td>
<td>Maria Rogozhkina</td>
</tr>
<tr>
<td>NDC</td>
<td>August 2023</td>
<td></td>
<td>Oleg Zhuk, Timur Vakhitov</td>
</tr>
<tr>
<td>RxNorm</td>
<td>August 2023</td>
<td>July 2023</td>
<td>Oleg Zhuk, Timur Vakhitov</td>
</tr>
<tr>
<td>RxNorm Extension</td>
<td>August 2023</td>
<td></td>
<td>Anna Ostropolets, Timur Vakhitov</td>
</tr>
<tr>
<td>SPL</td>
<td>August 2023</td>
<td></td>
<td>Oleg Zhuk, Timur Vakhitov</td>
</tr>
<tr>
<td>VANDF/ VA class</td>
<td>20230306</td>
<td>20230807</td>
<td>Varvara Savitskaya</td>
</tr>
</tbody>
</table>

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

August release

• One of the core vocabularies
• Released 2023-08-15 by Regenstrief Institute - only two weeks before the end of August
• LOINC-SNOMED hierarchy improved
• One-to-many mappings fixed
• SARS-2 related lab tests incorporated into the existing hierarchy
LOINC

• New concepts

<table>
<thead>
<tr>
<th>vocabulary_id</th>
<th>domain_id</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOINC</td>
<td>Measurement</td>
<td>2453</td>
</tr>
<tr>
<td>LOINC</td>
<td>Observation</td>
<td>1217</td>
</tr>
<tr>
<td>LOINC</td>
<td>Procedure</td>
<td>402</td>
</tr>
<tr>
<td>LOINC</td>
<td>Meas Value</td>
<td>215</td>
</tr>
<tr>
<td>LOINC</td>
<td>Note</td>
<td>114</td>
</tr>
</tbody>
</table>

• New relationships

<table>
<thead>
<tr>
<th>vocabulary_id</th>
<th>vocabulary_id</th>
<th>relationship_id</th>
<th>count</th>
<th>percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOINC</td>
<td>Snomed</td>
<td>Is a</td>
<td>76</td>
<td>0,432%</td>
</tr>
<tr>
<td>LOINC</td>
<td>Snomed</td>
<td>Maps to</td>
<td>-1</td>
<td>-0,19%</td>
</tr>
<tr>
<td>LOINC</td>
<td>Omop Extension</td>
<td>Is a</td>
<td>7</td>
<td>5,512%</td>
</tr>
</tbody>
</table>
VANDF / VA Class

- Veterans Health Administration (VHA) produces the National Drug File (NDF)

- Updated overdue: v 2021-10-04 currently in Athena

- Huge changes to the algorithm of domain assignments. Previously, all the concepts from VANDF vocabulary were Drugs. Now we are separating Devices using internal hierarchy

Before

<table>
<thead>
<tr>
<th>Item</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>BANDAGE TUBULAR ELASTIC SZ C</td>
<td>Drug</td>
</tr>
<tr>
<td>TRACHEOSTOMY TUBE HOLDER WELCON #2740</td>
<td>Drug</td>
</tr>
<tr>
<td>DRESSING,SORBACELL FOAM 4IN X 4IN #84344</td>
<td>Drug</td>
</tr>
</tbody>
</table>

After

<table>
<thead>
<tr>
<th>Item</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>BANDAGE TUBULAR ELASTIC SZ C</td>
<td>Device</td>
</tr>
<tr>
<td>TRACHEOSTOMY TUBE HOLDER WELCON #2740</td>
<td>Device</td>
</tr>
<tr>
<td>DRESSING,SORBACELL FOAM 4IN X 4IN #84344</td>
<td>Device</td>
</tr>
</tbody>
</table>
VANDF / VA Class

Note: mapping provided by the source can be inaccurate in OMOP sense

<table>
<thead>
<tr>
<th>VANDF code</th>
<th>VANDF name</th>
<th>relationship_id</th>
<th>RxNorm code</th>
<th>RxNorm name</th>
</tr>
</thead>
<tbody>
<tr>
<td>4003882</td>
<td>CEPHAPIRIN NA 20GM/VIL INJ</td>
<td>Maps to</td>
<td>313899</td>
<td>cepapirin 417 MG/ML Injectable Solution</td>
</tr>
<tr>
<td>4004136</td>
<td>GENTAMICIN SO4 10MG/ML (PF) INJ</td>
<td>Maps to</td>
<td>1870676</td>
<td>2 ML gentamicin 10 MG/ML Injection</td>
</tr>
<tr>
<td>4004199</td>
<td>MEPIVACAINE HCL 3% INJ</td>
<td>Maps to</td>
<td>1305263</td>
<td>1.7 ML mepivacaine hydrochloride 30 MG/ML Cartridge</td>
</tr>
<tr>
<td>4004573</td>
<td>MEZLOCILLIN NA 1GM/VIL INJ</td>
<td>Maps to</td>
<td>311689</td>
<td>mezlocillin 250 MG/ML Injectable Solution</td>
</tr>
<tr>
<td>4005823</td>
<td>DOPAMINE HCL 80MG/ML INJ</td>
<td>Maps to</td>
<td>1743953</td>
<td>5 ML dopamine hydrochloride 80 MG/ML Injection</td>
</tr>
<tr>
<td>4006008</td>
<td>MANNITOL 100MG/ML INJ</td>
<td>Maps to</td>
<td>1791395</td>
<td>500 ML mannitol 100 MG/ML Injection</td>
</tr>
<tr>
<td>4007046</td>
<td>RANITIDINE HCL 150MG GRANULES,EFFERVSC</td>
<td>Maps to</td>
<td>827179</td>
<td>ranitidine 1 MG/ML Oral Solution</td>
</tr>
</tbody>
</table>
CPT4

- Updated to UMLS 2023AA version
- Visit concepts have been back rolled back to their original domains (after debates with the CDM WG)
- Deduplication of semantic counterparts between CPT4 categories and SNOMED
- Relationships between CPT and SNOMED were enriched by mappings and hierarchical ‘Is a’
- SNOMED-CPT crosswalks are now extracted directly from the source
CPT4

- **New concepts**

- **Domain change**

- **New relationships**

<table>
<thead>
<tr>
<th>vocabulary_id</th>
<th>domain_id</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPT4</td>
<td>Procedure</td>
<td>142</td>
</tr>
<tr>
<td>CPT4</td>
<td>Measurement</td>
<td>83</td>
</tr>
<tr>
<td>CPT4</td>
<td>Observation</td>
<td>23</td>
</tr>
<tr>
<td>CPT4</td>
<td>Drug</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>vocabulary_id</th>
<th>old_domain_id</th>
<th>new_domain_id</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPT4</td>
<td>Visit</td>
<td>Observation</td>
<td>175</td>
</tr>
<tr>
<td>CPT4</td>
<td>Visit</td>
<td>Procedure</td>
<td>8</td>
</tr>
<tr>
<td>CPT4</td>
<td>Visit</td>
<td>Measurement</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>vocabulary_id_1</th>
<th>vocabulary_id_2</th>
<th>relationship_id</th>
<th>count</th>
<th>percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPT4</td>
<td>SNOMED</td>
<td>Is a</td>
<td>1344</td>
<td>212,995%</td>
</tr>
<tr>
<td>CPT4</td>
<td>SNOMED</td>
<td>Maps to</td>
<td>1562</td>
<td>159,877%</td>
</tr>
<tr>
<td>CPT4</td>
<td>SNOMED</td>
<td>CPT4 - SNOMED cat</td>
<td>86</td>
<td>0,516%</td>
</tr>
<tr>
<td>CPT4</td>
<td>SNOMED</td>
<td>CPT4 - SNOMED eq</td>
<td>-105</td>
<td>-0,644%</td>
</tr>
</tbody>
</table>
Other changes

• RxNorm is updated to July 2023 version.
• NDC/SPL are updated with August 2023 version
• Refined Smoking-related concept mappings: HCPCS, Read, UK Biobank, ICD10CM
• Family of previously proprietary CO-CONNECT vocabularies is now distributed under an open-source license - thank you for your contribution!
• Minor bug fixes
Overview

1. Recap of the plan for Vocabularies improvement

2. Accomplishments so far
   - New process & Vocabulary Committee
   - Landscape assessment
   - Roadmap and releases
   - Community contribution
   - August release
   - Experience with community contribution
   - Vocabulary WG
   - Quality management system

3. What is missing, issues faced and ways to improvement
Community contribution in more details
Community contribution part I

• **11** contributions submitted so far:
  
  - Addition of content: 2 new non-standard vocabulaires
  - Modification of content: 1 with domain change, 3 with mapping changes
  - Complex: 2 with deduplication, 1 with hierarchy, 2 with new standard terms

• **4** contributions submitted and approved *before* the end of June and are *incorporated* into the August release

• **4** contributions submitted *after* the end of July and will be incorporated into the February release

• **3** contributions are not fully baked yet
Community contributions in the August release

- 2 new non-standard vocabularies added with mappings (NHS Ethnic Category and NHS Place of Service)
- 4 mapping bugs fixed in existing vocabularies
- RxNorm Extension drug concepts added to support the local source drug system (Z-index)
Community contributions going into February release

Issues 850, 856: mapping changes

Issue 872: concept attribute change

Issue 795: most of the concepts are new SNOMEDs absent in vocabularies => postponed till February
Not fully baked community contribution

Issue 846: ongoing discussion in Oncology WG

Issue 860: complex case (new hierarchy) with more discussion needed

Issue 851: ongoing discussion

Issue 871: to be reviewed and discussed
How first run of community contribution went

1. We incorporated all contributions submitted and approved by the end of June (many thanks to our contributors!)

1. Contributions can be made through templates but require involvement on both sides (and we advise to create an issue first!)

1. Quality assurance on the side of contributors and resolvement of inconsistencies on the Vocabularies development side are the areas for improvement
Examples of issues to watch out for in the checklist

Wrong dates

Different vocabulary_ids in different files, referring to the same concepts

Violating constraints: vocabulary_id longer than varchar(20)

Wrong dates

...and it gets harder for more complex cases (new drugs, new hierarchies)
Community contribution: summary

1. Community is getting increasingly involved in collaborative Vocabularies development

1. Current template approach supports advertised use cases well but requires time

1. Further steps include more education (e.g., demos, Vocabulary WG) and building automated pipelines
Current QA-System

- Look in Github Vocabulary-v5.0 repo under Wiki
- QA consists of:
  1. Quality issue tracking and resolution in Forum or as Github Issue
  2. Scripts
     - Quality control scripts, mostly conformance and logic checks, a few semantic checks
     - Quality enforcing scripts, overwriting deviating content
  3. Manual content checks
Future QA-System: Quality Management System

• QMS is a formalised approach with documented objectives, processes, procedures, responsibilities
• Consists of quality (i) planning, (ii) assurance, (iii) control, (iv) improvement
• We need:
  • Organization
  • Design
  • Documentation
  • Complaint handling and defect tracking
  • Corrective and preventive actions (CAPA)
  • Training
Vocabulary WG

• Changing focus to familiarize community with ongoing efforts and create a collaborative environment to share and discuss needs and ideas and to support other WGs

• Increased attendance (up to 40 people)

• Topics covered so far: community contribution, tracking vocabulary changes, SNOMED-LOINC hierarchy (check out our Teams subchannel!)

• Please come to raise questions and discuss your needs
Summary and next steps

1. We have implemented landscape assessment, established Vocabulary Committee, created roadmap, improved on documentation (incl. QA) and ran community contribution part I

1. More work is needed to establish working pipelines for community contribution, including common development environment

1. More engagement from community is needed (message Vocab Team, come to Vocab WG calls!)

1. Quality Management System is a crucial element and is currently in progress