Seminal OHDSI Vocabulary Paper

OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization

PMC10873827

Alerts | Author Instructions | Submit | About
JAMIA
Journal of the American Medical Informatics Association

J Am Med Inform Assoc. 2024 Mar; 31(3): 583–590.
Published online 2024 Jan 4. doi: 10.1093/jamia/ocad247

PMCID: PMC10873827
PMID: 38175665

Journal Article

OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization

Christian Reich, MD,✉ Anna Ostropolets, PhD, Patrick Ryan, PhD, Peter Rijnbeek, PhD, Martijn Schuemie, PhD, Alexander Davydov, MD, Dmitry Dymshyts, MD, and George Hripcsak, MD

▸ Author information ▸ Article notes ▸ Copyright and License information PMC Disclaimer

Intro

- Observational research needs large scale for sample size and diverse populations
- That needs standardization, all major data networks (Sentinel, PCORNet, OHDSI) adopted a standard
- Standardization of format and representation (vocabularies)
- Standardization of medical vocabularies can be done ad-hoc or through a central reference
- OHDSI chose the central system
- Supports
  - Cohort definition
  - Covariate construction
  - Large-scale analytics
  - Result reporting
- UMLS is such a repository of vocabularies, but not focused on our use cases

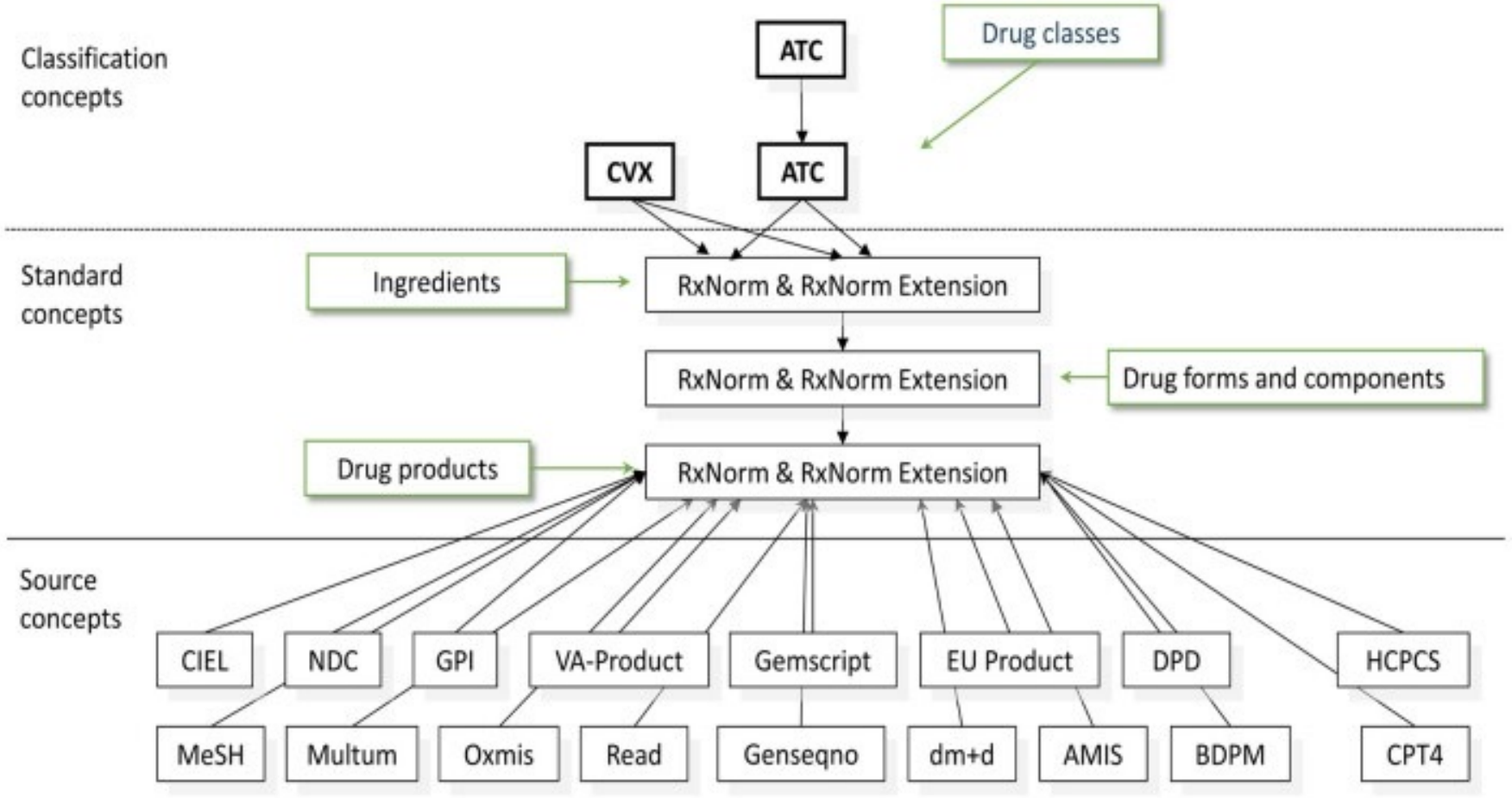
Requirements

Requirement	Definition
Standard concepts	Unique concepts of fully pre-coordinated medical entities, to be stated as fact, no negations of facts, no reference to the past, and no flavors of null (unknown, not reported, etc.)
Concept domains	Assignment of concepts to domain categories (condition, drug, visit, etc.)
Comprehensive coverage	In each domain, standard concepts must cover all possible entities and mappings from terms and codes used in databases around the world
Polyhierarchies	Precalculated hierarchies organizing concepts
Efficiency	Computationally efficient data model
Use case focus	Storing and analyzing patient-level data for evidence generation

Methods

- Vocabularies and concepts
- Domains
- Standardization of concepts
- Mapping, hierarchical, and other relationships between concepts
- Life cycle and distribution
- Quality assurance

Structure


Classification concepts
ATC
Drug classes
CVX
ATC
Standard concepts
Ingredients
RxNorm & RxNorm Extension
RxNorm & RxNorm Extension
Drug forms and components
Drug products
RxNorm & RxNorm Extension
Source concepts
CIEL
NDC
GPI
VA-Product
Gemscript
EU Product
DPD
HCPCS
MeSH
Multum
Oxmis
Read
Genseqno
dm+d
AMIS
BDPM
CPT4

Overall content

- 136 vocabularies
  - 101 from external sources
- 10,574,359 concepts
  - 8,761,976 valid ones
    - 40.5% standard ones
    - 50.1% non-standard ones
    - 9.4% classification (mostly Drug and Measurement)
- 28 million valid relationships
  - 38.3% Is a
  - 14.1% Maps to (covering 66.8% of non-standard concepts)

Equivalence relationships per type and domain

Type of "Maps to" relationship, % (*n*)				
Domain	**One-to-one**	**Many-to-one**	**One-to-many**	**Many-to-many**
Condition	1.9% (54 671)	10.3% (292 507)	<0.1% (38)	3.2% (90 034)
Device	6.1% (172 216)	3.6% (101 774)	<0.1% (4)	<0.1% (10)
Drug	18.1% (515 360)	42.6% (1 208 579)	<0.1% (181)	1.6% (44 780)
Measurement	0.5% (14 502)	0.4% (11 580)	1.2% (33 655)	1.2% (33 489)
Observation	3% (86 014)	2.4% (69458)	<0.1% (3)	0.2% (4579)
Procedure	1.1% (29 973)	2.5% (70 974)	<0.1% (16)	0.2% (5581)

Discussion

- Work in progress, will never end
  - Errors need constantly be addressed
  - Workgroups are taking care of the system
  - Standard concepts particularly challenging – UMLS helps but not enough
  - Quality system being built
- Not a Knowledge Base
  - Lateral relationships only adopted “lazily”