Produce Candidate Definitions for DOIDs without them

J. Allen Baron

2021-11-02

Contents

1 PURPOSE

To obtain candidate definitions from MESH and NCI Thesaurus for DOIDs without definitions.

2 Data Setup (Manual Execution)

2.1 Download Data

1. Copy doid.owl (release: v2021-10-11) from HumanDiseaseOntology git repo to data.
2. Download NCI Thesaurus OWL file (asserted version; release 21.09d) to data from https://evs.nci.nih.gov/evs-download/thesaurus-downloads.
    - NOTE: .zip can't be read by `virtuoso` bulk database uploader, so I unzipped and recompressed it as .gz file.
3. Download MESH N-triples file to data from https://hhs.github.io/meshrdf/ ftp site (release: 2021-10-28, updated nightly).

2.2 Setup a Virtuoso Database

I installed Virtuoso using `scripts/setup_virtuoso.R` and default values. That script and SPARQL queries herein rely on the `virtuoso` R package.

I chose to load the 3 rdf files as individual graphs so that I could explore them separately and avoid unexpected results from queries finding hits in multiple data sources.

3 DOIDs without definitions

I extracted the DOIDs without definitions, excluding obsolete terms, along with all annotated xrefs for further analysis. I did not use the `src/sparql/extra/DO_no_defs.rq` SPARQL query (it has the same terms but results in multiple rows per term because it represents parents). Instead, I directly excluded obsolete terms and did not extract parent information.

3.1 General Statistics

There are **2,535 DOIDs without definitions**. Of these, 40 have NO xrefs.

Here's a quick count of the xrefs generally:

xref_ns	n
UMLS_CUI	2593
NCI	1803
SNOMEDCT_US_2021_03_01	1779
ICD9CM	886
MESH	875
ICD10CM	846
GARD	221
OMIM	162
NA	40
ORDO	16
EFO	14
ICDO	14
SNOMEDCT_US_2020_03_01	1

And lastly, the number of DOIDs without definitions **with or without NCI/MESH xrefs** is:

has_nci_mesh_xref	n
FALSE	518
TRUE	2017

That's about 1/5 of the DOIDs without definitions that NCI & MESH can't help with. Among those it appears most have xrefs to one of the ICD's or SNOMED (via UMLS). I'm not sure any of those will be accessible for definitions.

xref_ns	n
UMLS_CUI	478
ICD9CM	446

xref_ns	n
SNOMEDCT_US_2021_03_01	431
ICD10CM	285
NA	40
GARD	7
EFO	2
OMIM	2

3.2 [Potentially] Actionable DOID w/o Def Lists

3.2.1 Need xrefs

Here are the 40 DOIDs without definitions that have NO xrefs.:

doid	do_name	xref
DOID:66	muscle tissue disease	NA
DOID:3378	conventional central osteosarcoma	NA
DOID:3666	cutaneous solitary mastocytoma	NA
DOID:3814	extraskeletal chondroma	NA
DOID:4286	sebaceous basal cell carcinoma	NA
DOID:4294	adenoid basal cell carcinoma	NA
DOID:4295	follicular basal cell carcinoma	NA
DOID:4299	infiltrative basal cell carcinoma	NA
DOID:4302	cystic basal cell carcinoma	NA
DOID:4490	malignant peritoneal solitary fibrous tumor	NA
DOID:490	hemangioma of lung	NA
DOID:505	hobnail hemangioma	NA
DOID:5161	Monckeberg arteriosclerosis	NA
DOID:5507	clear cell ependymoma	NA
DOID:5569	malignant syringoma	NA
DOID:5859	periosteal chondrosarcoma	NA
DOID:5889	anaplastic ependymoma	NA
DOID:6571	non-invasive bladder urothelial carcinoma	NA
DOID:7222	gallbladder pleomorphic giant cell adenocarcinoma	NA
DOID:749	active peptic ulcer disease	NA
DOID:7571	malignant cystic nephroma	NA
DOID:7718	osteoclast-like giant cell neoplasm of the pancreas	NA
DOID:7891	testicular spermatocytic seminoma	NA
DOID:7902	adult extraosseous chondrosarcoma	NA
DOID:8256	olfactory neural tumor	NA
DOID:8303	congenital granular cell tumor	NA
DOID:8642	Hodgkin's paragranuloma	NA
DOID:8651	Hodgkin's granuloma	NA
DOID:10869	fourth cranial nerve palsy	NA
DOID:11132	prostatic hypertrophy	NA
DOID:13730	malignant renovascular hypertension	NA
DOID:14202	adult dermatomyositis	NA
DOID:1460	atheroembolism of kidney	NA
DOID:2061	nodular hidradenoma	NA
DOID:3033	colon signet ring adenocarcinoma	NA
DOID:3281	combined thymoma	NA

doid	do_name	xref
DOID:60004	malignant cystadenoma	NA
DOID:60006	benign vascular tumor	NA
DOID:9080	macroglobulinemia	NA
DOID:9912	hydrocele	NA

3.2.2 [Potentially] Missing UMLS xrefs

I assumed that most NCI and MESH terms are also in UMLS. Of the 2,017 DOIDs with NCI or MESH xrefs there are 10 DOIDs *WITHOUT* UMLS xrefs. Those DOIDs are:

doid	do_name	xref
DOID:3021	acute kidney failure	MESH:D058186
DOID:3318	epithelioid type angiomyolipoma	NCI:C38151|SNOMEDCT_US_2020_03_01:733836008
DOID:5284	retroperitoneal leiomyosarcoma	NCI:C27904
DOID:5469	biliary tract intraductal papillary mucinous neoplasm	NCI:C37215
DOID:5509	childhood ependymoma	NCI:C8578
DOID:5974	renal pelvis transitional cell carcinoma	NCI:C7355
DOID:6607	nervous system hibernoma	NCI:C6997
DOID:7533	subareolar duct papillomatosis	NCI:C9008
DOID:7922	benign mediastinal neurilemmoma	NCI:C6625
DOID:8170	fibroepithelial polyp of the anus	NCI:C5604

3.2.3 [Potentially] Get Defs from EFO/GARD/OMIM

The small number of DOIDs without definitions with xrefs to EFO/GARD/OMIM are:

doid	do_name	xref
DOID:10211	cholelithiasis	EFO:0004799
DOID:10719	toxic diffuse goiter	GARD:6549
DOID:1089	tethered spinal cord syndrome	GARD:4018
DOID:11817	urachus cancer	GARD:7836
DOID:11843	coronary artery anomaly	GARD:1534
DOID:12070	Dieulafoy lesion	GARD:10930
DOID:1875	impotence	EFO:0004234
DOID:4455	hereditary renal cell carcinoma	GARD:9571
DOID:5806	stork bite	OMIM:163100
DOID:9378	glaucomatocyclitic crisis	GARD:10737
DOID:9822	partial central choroid dystrophy	OMIM:613105

4 Extraction of Definitions from NCI & MESH

4.1 SPARQL Queries to RDF Files

I attempted to extract definitions from the MESH & NCI rdf files. Though the SPARQL query works, only NCI definitions are returned. It appears that there are *NO* `meshv:scopeNote` (definition) triples for the MESH terms of interest in the file.

I attempted to identify a definition (`meshv:scopeNote`) for a specific MESH record that I know has one (MESH:D010211, "Papilledema", https://meshb.nlm.nih.gov/record/ui?ui=D010211), by submitting queries to MeSH's SPARQL API endpoint. As in my attempts to get definitions from the mesh.nt.gz file download, there is no `meshv:scopeNote` triple available via the SPARQL API endpoint, with or without inference on.

SPARQL query submitted to the API:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX meshv: <http://id.nlm.nih.gov/mesh/vocab#>
PREFIX mesh: <http://id.nlm.nih.gov/mesh/>
PREFIX mesh2021: <http://id.nlm.nih.gov/mesh/2021/>
PREFIX mesh2020: <http://id.nlm.nih.gov/mesh/2020/>
PREFIX mesh2019: <http://id.nlm.nih.gov/mesh/2019/>

SELECT *
FROM <http://id.nlm.nih.gov/mesh>
WHERE {
  ?class ?p ?o .
  VALUES ?class { mesh:D010211 }
}
```

I have not been able to identify a downloadable or queryable source, besides the MESH rdf resources, from which to get definitions. The last option I can think of is to try the Entrez API.

4.2 MESH Entrez API Queries

I extracted the MESH codes and checked to find out what access to MESH the Entrez API has.

MESH is an available database and is searchable by code (i.e. NLM MeSH Browser Unique ID).

```
##  [1] "pubmed"          "protein"         "nuccore"         "ipg"
##  [5] "nucleotide"      "structure"       "genome"          "annotinfo"
##  [9] "assembly"        "bioproject"      "biosample"       "blastdbinfo"
## [13] "books"           "cdd"             "clinvar"         "gap"
## [17] "gapplus"         "grasp"           "dbvar"           "gene"
## [21] "gds"             "geoprofiles"     "homologene"      "medgen"
## [25] "mesh"            "ncbisearch"      "nlmcatalog"      "omim"
## [29] "orgtrack"        "pmc"             "popset"          "proteinclusters"
## [33] "pcassay"         "protfam"         "biosystems"      "pccompound"
## [37] "pcsubstance"     "seqannot"        "snp"             "sra"
## [41] "taxonomy"        "biocollections"  "gtr"
```

```
##  DbName: mesh
##  MenuName: MeSH
##  Description: MeSH Database
##  DbBuild: Build211101-0310.1
##  Count: 348617
##  LastUpdate: 2021/11/01 03:36
```

```
## Searchable fields for database 'mesh'
##   ALL     All terms from all searchable fields
##   UID     Unique number assigned to publication
##   FILT    Limits the records
##   TN      Tree Number
##   MESH    MeSH Terms
##   SUBS    Substance Name
##   WORD    Free text
##   ALSO    See Also
##   PREV    Previous Indexing
##   NOTE    Scope Note
##   REG     Registry Number
##   MULT    Multi
##   TYPE    Record type - main heading, subheading, pharmacological action, substance name, publicatio
##   MHUI    NLM MeSH Browser Unique ID
```

Hopefully, the definitions available in the online MeSH search are also accessible via the Entrez API.

I thought the definitions would most likely be included in the summary info so I attempted to access the summaries but nothing was returned. After some investigation I discovered the Entrez API is using a different internal unique identifier for their records (UID instead of MHUI; not visible on search result pages).

Online searches didn't work, so I used an Entrez API search with MHUIs to get the UIDs, and then was able to obtain the MeSH summary information, which included Scope Notes (ie. definitions!!).

Some of the NLM MeSH Browser Unique IDs (MHUI) I passed to the API did not return Unique IDs (UIDs; 836 records for 875 MHUIs were returned). I checked to see which MHUIs did not have matches and found none.

```
## character(0)
```

After a little more exploration, I discovered that I'd included duplicates in the list to the API and that there are *not* actually any missing. So the data looks complete! ... in the sense that there is a record for every xref I have.

I merged this MESH data in with the DOIDs and NCI definitions and cleaned it up a bit to make it more human-readable.

4.3 Xref Definition Statistics

For those DOIDs with NCI and/or MeSH xrefs, the overall count of DOIDs with candidate definitions from both NCI & MeSH, just one of them, or none is:

def_source	doid_n
NCI & MESH	580
NCI	1082
MESH	287
none	68

The count of NCI/MeSH xrefs without their own definitions is:

xref_ns	n
MESH	4
NCI	86

And those xrefs without any definitions are:

```
## `summarise()` has grouped output by 'doid'. You can override using the `.groups` argument.
```

doid	do_name	xref
DOID:10289	prostate malignant phyllodes tumor	MESH:C549759
DOID:1171	hyperlipoproteinemia type V	NCI:C35645
DOID:11832	visual epilepsy	NCI:C3980
DOID:12016	frontal lobe neoplasm	NCI:C5572
DOID:13313	pancreatic mucinous ductal ectasia	NCI:C5717
DOID:13407	hypercalcemic sarcoidosis	NCI:C35807
DOID:14125	abducens nerve neoplasm	NCI:C5826
DOID:1726	partial of retinal vein occlusion	NCI:C35341
DOID:1760	facial nerve neoplasm	NCI:C5827
DOID:2135	temporal lobe neoplasm	NCI:C5567
DOID:2410	skin granular cell tumor	NCI:C5617
DOID:2550	tactile epilepsy	NCI:C4687
DOID:2566	corneal dystrophy	NCI:C34513
DOID:2668	mesenchymoma	NCI:C3233
DOID:2682	intracystic papillary adenoma	NCI:C4191
DOID:3177	verrucous papilloma	NCI:C4101
DOID:3184	spinal cord oligodendroglioma	NCI:C4535
DOID:3186	adult oligodendroglioma	NCI:C9376
DOID:3198	hypoglossal nerve neoplasm	NCI:C5830
DOID:337	spinal accessory nerve neoplasm	NCI:C5829
DOID:3417	glossopharyngeal nerve neoplasm	NCI:C5828
DOID:3428	granulomatous myositis	NCI:C27575
DOID:3639	spinal cord intramedullary teratoma	NCI:C5428
DOID:3641	conus medullaris neoplasm	NCI:C5443
DOID:3688	plexopathy	NCI:C27744
DOID:3828	chromophobe adenoma	NCI:C2857
DOID:3843	diencephalic neoplasm	NCI:C5126 | NCI:C5125
DOID:3850	hemangiopericytic tumor	NCI:C7076
DOID:3968	papillary follicular thyroid adenocarcinoma	NCI:C7380
DOID:4030	eosinophilic gastritis	NCI:C27052
DOID:4035	lymphocytic gastritis	NCI:C27051
DOID:4201	peroneal neuropathy	NCI:C27596
DOID:4542	basophil adenoma	NCI:C2856
DOID:4548	extraskeletal mesenchymal chondrosarcoma	NCI:C27481
DOID:4586	familial meningioma	MESH:C537443
DOID:4690	childhood mediastinal neurogenic tumor	NCI:C5429
DOID:4693	nerve plexus neoplasm	NCI:C5822
DOID:4707	cervicomedullary junction neoplasm	NCI:C5423
DOID:479	angiokeratoma	NCI:C4488
DOID:4846	cauda equina intradural extramedullary astrocytoma	NCI:C5408
DOID:4847	cauda equina neoplasm	NCI:C5479
DOID:4855	diencephalic astrocytoma	NCI:C5128

doid	do_name	xref
DOID:4858	pineal gland astrocytoma	NCI:C8274
DOID:5155	multiple mucosal neuroma	NCI:C6559
DOID:5222	acute necrotizing encephalitis	NCI:C35383
DOID:5341	pineal region yolk sac tumor	NCI:C6752
DOID:5392	acidophil adenoma	NCI:C6780
DOID:5469	biliary tract intraductal papillary mucinous neoplasm	NCI:C37215
DOID:5484	fibrous synovial sarcoma	NCI:C6533
DOID:5510	pineal dysgerminoma	NCI:C7169
DOID:5553	pineal region choriocarcinoma	NCI:C6759
DOID:5758	malignant mesenchymoma	NCI:C4268
DOID:5838	extragonadal seminoma	NCI:C7327
DOID:5861	myxoid chondrosarcoma	NCI:C4303
DOID:5874	retroperitoneal germ cell neoplasm	NCI:C6447
DOID:5893	childhood malignant mesenchymoma	NCI:C8097
DOID:5894	adult malignant mesenchymoma	NCI:C7947
DOID:5907	penis non-invasive verrucous carcinoma	NCI:C27791
DOID:5913	brachial plexus neoplasm	NCI:C5823
DOID:5948	angiokeratoma of mibelli	NCI:C3927
DOID:5949	angiokeratoma circumscriptum	NCI:C7751
DOID:5996	blunt duct adenosis of breast	NCI:C6941
DOID:6016	adult central nervous system mature teratoma	NCI:C27400
DOID:6018	adult central nervous system immature teratoma	NCI:C27401
DOID:6098	thalamic neoplasm	NCI:C6221 | NCI:C4576
DOID:6335	bilateral meningioma of optic nerve	MESH:C000608854
DOID:6858	pineal region immature teratoma	NCI:C6755
DOID:7179	mixed eosinophil-basophil adenoma	NCI:C4148
DOID:7237	pancreatic non-invasive mucinous cystadenocarcinoma	NCI:C41245
DOID:7315	Jewett-Marshall bladder cancer	NCI:C9368
DOID:7380	squamous cell papilloma of skin	NCI:C4462
DOID:7381	lymphohistiocytoid mesothelioma	NCI:C27779
DOID:7441	chronic metabolic polyneuropathy	NCI:C35602
DOID:7533	subareolar duct papillomatosis	NCI:C9008
DOID:7558	glossopharyngeal motor neuropathy	NCI:C27212
DOID:7559	asymmetric motor neuropathy	NCI:C27953
DOID:7574	pancreatic intraductal papillary-colloid carcinoma	NCI:C5725
DOID:7735	pancreatic colloid cystadenoma	NCI:C5718
DOID:7825	chronic toxic polyneuropathy	NCI:C35603
DOID:7867	adult central nervous system germinoma	NCI:C5792
DOID:7922	benign mediastinal neurilemmoma	NCI:C6625
DOID:8170	fibroepithelial polyp of the anus	NCI:C5604
DOID:8340	endocervical type cervical mucinous adenocarcinoma	NCI:C40202
DOID:8389	lumbar plexus neoplasm	NCI:C5824
DOID:8538	reticulosarcoma	NCI:C27824
DOID:8593	chronic monocytic leukemia	NCI:C34774
DOID:910	occipital lobe neoplasm	NCI:C5574
DOID:9574	choanal atresia	MESH:C562435

5 Save Candidate Definitions

Lastly I saved the definitions data to a Google Sheets file in `Disease_Ontology/DO_definitions/Sets of terms to Define/2021-Missing_Definitions`.

```
## v Writing to "DO-MESH_NCI_defs".
```

```
## v Writing to sheet 'definitions'.
```