Data Harmonization through use of community standards in the Common Fund Data Ecosystem

Michelle Giglio[1], Suvarna Nadendla[1], Arthur Brady[1], Amanda Charbonneau[2], Karl Czajkowski[3], Jeremy Yang[4], Tom Gillespie[5], Philippe Rocca-Serra[6], Jeff Grethe[5], Raja Mazumder[7], Bernard de Bono[8], Jonathan Silverstein[9], Daniel Clark[10], Mark Musen[11], Owen White[1], and the CFDE Ontology Working Group

[1]Institute for Genome Sciences, University of Maryland School of Medicine, USA
[2]University of California, Davis, USA
[3]University of Southern California, Information Sciences Institute, USA
[4]School of Medicine, Department of Internal Medicine, Translational Informatics Division, University of New Mexico, USA
[5]University of California, San Diego, USA
[6]University of Oxford, United Kingdom
[7]George Washington University, USA
[8]University of Auckland, New Zealand
[9]University of Pittsburgh, USA
[10]Department of Pharmacological Sciences, Mount Sinai Center for Bioinformatics, Icahn School of Medicine at Mount Sinai, USA
[11]Stanford University, USA

Abstract

The NIH Common Fund has supported multiple programs that have resulted in the creation of numerous data coordination centers (DCCs) that house diverse data and resources. In order to facilitate the ability of researchers to find information across DCCs, the Common Fund Data Ecosystem (CFDE) was formed. The CFDE provides a centralized resource managed by the CFDE Coordinating Center where metadata about DCC data assets is stored. The CFDE Portal enables search of this metadata via web-based faceted queries. The Ontology Working Group within the CFDE has established a process for choosing standards to use for the capture of this metadata from DCCs. Multiple ontologies and controlled vocabularies were chosen and are now in active use by the DCCs to submit metadata to the CFDE centralized resource. As of this writing, there are ~4.5 million file records, 2,700 subject records, and more than 1.7 million biosample records linked to ontology or controlled vocabulary terms in the CFDE resource.

Keywords

ontology, Common Fund, metadata harmonization, metadata integration

ICBO 2022, September 25-28, 2022, Ann Arbor, MI, USA

EMAIL: mgiglio@som.umaryland.edu (A. 1); snadendla@som.umaryland.edu (A. 2); arthur.brady@gdit.com (A. 3); Achar@ucdavis.edu (A. 4); karlcz@isi.edu (A. 5); JJYang@salud.unm.edu (A. 6); tom.h.gillespie@gmail.com (A. 7); philippe.rocca-serra@oerc.ox.ac.uk (A. 8); jeffrey.grethe@gmail.com (A. 9); mazumder@gwu.edu (A. 10); b.debono@auckland.ac.nz (A. 11); J.c.s@pitt.edu (A. 12); daniel.clarke@mssm.edu (A. 13); musen@stanford.edu (A. 14); owhite@som.umaryland.edu (A. 15)
ORCID: 0000-0001-7628-5565 (A. 1); 0000-0003-3643-281X (A. 2); 0000-0002-8809-2477 (A. 3); 0000-0001-7376-7702 (A. 4); 0000-0002-9389-0633 (A. 5); 0000-0002-1476-6192 (A. 6); 0000-0002-7509-4801 (A. 7); 0000-0001-9853-5668 (A. 8); 0000-0001-5212-7052 (A. 9); 0000-0001-8823-9945 (A. 10); 0000-0003-0638-5274 (A. 11); 0000-0002-9252-6039 (A. 12); 0000-0003-3471-7416 (A. 13); 0000-0003-3325-793X (A. 14); 0000-0001-7406-7361 (A. 15)

1. Introduction

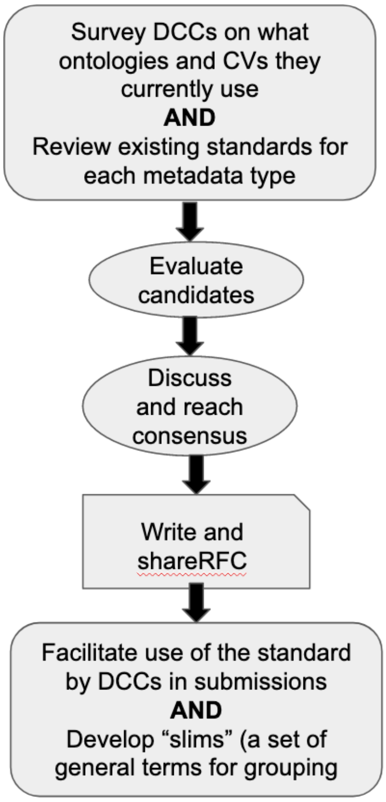
The NIH Common Fund was formed in 2006 to provide a mechanism to fund initiatives that do not fall under the purview of a single NIH institute or center. Common Fund programs must be transformative, catalytic, synergistic, cross-cutting, and unique (https://commonfund.nih.gov/). Past and present Common Fund programs have resulted in the creation of numerous data coordination centers (DCCs) that house the data and resources produced by a given Common Fund program. These DCCs generally provide tools for searching and viewing datasets produced by the program and often also provide analysis tools and other resources. In order to facilitate the ability of researchers to find information relevant to their research that might be housed in multiple DCCs, the Common Fund Data Ecosystem (CFDE) was formed. As of this writing, 11 Common Fund program DCCs participate in the CFDE. A current full list is maintained on the CFDE Portal (https://app.nih-cfde.org/) (1). The CFDE provides a centralized resource managed by the CFDE Coordinating Center where metadata about DCC data assets is stored. The CFDE Portal enables search of this metadata via web-based faceted queries that result in downloadable file manifests that can be imported into cloud-based analysis resources to facilitate the ability of researchers to find and use data from Common Fund programs (1). This serves to help make Common Fund data more FAIR, that is Findable, Accessible, Interoperable, and Reusable (2). In order for the metadata from diverse DCCs to be stored and effectively queried, metadata from DCCs should be harmonized before submission to the central repository. This was accomplished through the use of controlled vocabularies (CVs) and ontologies to capture many of the metadata elements and was managed by the CFDE Ontology Working Group (OWG). Here we describe the process used by the OWG to choose and implement the CVs and ontologies.

2. Goals and scope of harmonization effort

Our overarching goal throughout our efforts to develop a metadata capture system for the CFDE was to take a pragmatic approach to the collection of metadata from each DCC such that we could facilitate cross DCC queries. We did NOT want to attempt to unify all vocabularies or standards in use by any DCC as this is perhaps an impossible task and was certainly not in scope for our project. We also did NOT want to endeavor to capture every piece of metadata information stored at each DCC as this would have been duplicating the function of the individual DCCs, and that was not our mandate. What we wanted to do was find a way that all of the DCCs could contribute metadata that would be useful to a large swath of researchers at a level of granularity sufficient for users to identify datasets of interest. To accomplish this we had to accept that our capture of information would be imperfect and incomplete, keeping in mind that our goal is not to replicate the work of the DCCs, but rather to provide pointers and guideposts for researchers to find the resources provided by the DCCs.

3. The CFDE Ontology Working Group process

Figure 1: The OWG process for choosing ontologies and CVs for use in the CFDE central metadata repository. CV=controlled vocabulary; RFC=Request for Comment


Survey DCCs on what ontologies and CVs they currently use
AND
Review existing standards for each metadata type
Evaluate candidates
Discuss and reach consensus
Write and shareRFC
Facilitate use of the standard by DCCs in submissions
AND
Develop “slims” (a set of general terms for grouping

Our process for choosing which ontologies and CVs to use for metadata harmonization as well as for maximizing the utility of those ontologies and CVs for all DCCs included several steps which are described below and in Figure 1 above.

- **Choose which metadata types to focus on.** The types of metadata associated with datasets at each DCC are extensive. However, as mentioned above, our goal was not to capture everything available at each DCC but rather to focus on metadata elements that would provide the most utility as search criteria for the maximum number of researchers. Based on use cases established for various user profiles, we chose an initial 12 types of metadata to capture. These are listed in Table 1. Over time, we expect to expand the list of metadata types included in the CFDE.
- **Survey DCCs regarding current use of ontologies and CVs for those metadata types.** DCC representatives were asked to fill in an online survey that asked about their use of ontologies and CVs for the capture of metadata. Six out of nine DCCs (that were participating at that time, now the number of DCCs is 11) responded to the survey. The survey consisted of questions asking what data types, assay types, data formats, etc. were being captured by each center and what, if any, ontologies or controlled vocabularies were being used. DCCs provided answers as free text.
- **Identify candidate ontologies and CVs for those metadata types.** There are hundreds of ontologies being used in the biological research community, including many that cover the same conceptual areas. We employed the NCBO BioPortal tools (3,4) and the European Bioinformatics Institute (EBI) Ontology Lookup Service (OLS) (5) to assist in identifying ontologies and CVs that covered the metadata types of interest.
- **Evaluate the candidate ontologies and CVs based on our OWG criteria.** We established several criteria for assessing the suitability of an ontology or controlled vocabulary (CV) for use. These overlap with the ontology principles developed by the Open Biological and Biomedical Ontology (OBO) Foundry (6). Ideally, the ontology or CV should:
  - be stable, but not static,
  - be under active development,
  - have a mechanism for requesting new terms and ontology changes (e.g. a GitHub issue tracker),
  - be responsive to requests and questions,
  - have some level of community buy-in as measured by BioPortal “Acceptance Score” and GitHub issue activity (new issues being submitted recently),
  - conform to community conventions on ontology and vocabulary development,
  - provide mappings to other related ontologies/vocabularies, as relevant
- **Discuss and reach consensus.** Candidate ontologies were discussed in working group meetings to reach consensus.
- **Write and circulate a “Request For Comments” (RFC).** Once the working group reached a decision, an RFC was written that described the metadata type in question, the ontology/CV that was chosen, and any other information needed by DCCs for correct usage. The RFC was circulated first within the OWG and then throughout the entire CFDE for comment and revision before becoming a final policy. RFCs are versioned and, as needed, revisions to the RFCs will be made.
- **Facilitate use of the chosen ontologies and CVs.** Some DCCs had not used ontologies or CVs for storage of metadata or had been using different ontologies or CVs than those chosen through the process above. In addition, the submission of metadata to the CFDE central resources in the form of ontology or CV terms was a process new to the DCCs. Therefore, OWG members engaged in helpdesk activities to support DCCs as they converted (as needed) their metadata to the OWG standards and submitted them to the CFDE central repository. In addition, there were occasions when DCCs needed terms that did not yet exist in the chosen ontologies. In these cases, the OWG facilitated the

development of the needed terms by liaising with the ontology developers, shepherding new term requests through the development process, and tracking term status. In some cases, the chosen ontology or CV did not provide updated releases on a schedule rapid enough for the needs of DCCs. In that case, we employed the InterLex system to make provisional terms with in-house ids that can be used until new terms can be incorporated into official releases of the relevant ontologies or CVs (7). InterLex is a product of SPARC, one of the CFDE DCCs, and provides an online interface that allows one to create and edit new terms that can be linked into existing ontologies within the InterLex system. The system also provides means of tracking the status of terms with respect to their incorporation into the external ontologies. To date, 79 new terms, primarily in the Ontology for Biomedical Investigations (OBI) (8), have been developed for use in CFDE data submissions. Another 14 provisional terms for data types and file formats have been created within InterLex for internal CFDE use with the plan for their ultimate incorporation into the official external ontologies.

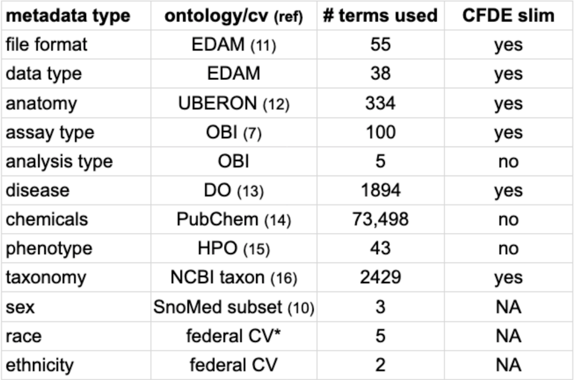
- **Build a "slim" for the ontology or CV.** DCCs are always encouraged to use the most granular terms that are applicable to their metadata as this provides the most accurate and specific information. However, visualization of hundreds or thousands of terms that have been used within a dataset can present challenges. Ontology "slims" can solve this difficulty by providing a way to see a more high-level view of a set of annotations (9). Generally, a slim is built using more general, less specific terms from an ontology representing broad classes within the ontology. Granular terms can then be mapped to the slim term under which they have parentage. Specific metadata term associations can then be binned into slim-term-based categories via those mappings. This is useful in comparing datasets to each other, creating visualizations of dataset annotations, and in searching. Therefore, we also developed CFDE specific "slims" for most of the ontologies and CVs used by the CFDE (Table 1). The assignment of slim terms is done automatically via mapping files after DCCs submit their metadata to the central repository. With regard to building slims, in some cases we modified existing slims provided by the ontology and in others we crafted the slims based on either a bottom-up or top-down approach using the existing CFDE metadata records as our guide.

4. Current Status of CFDE ontology and CV use

The current list of ontologies and CVs that have been adopted for metadata capture in the CFDE are listed in Table 1. According to the nature of the individual metadata types, some ontologies have dozens of terms in use, while others have thousands. Table 1 indicates how many unique terms are being used for each ontology as of July 2022. It also displays which ontologies we have built CFDE-specific slims for. For metadata type sex, a small selection consisting of four SnoMed-CT terms have been adopted (10). They are: 'Indeterminate', 'Female', 'Male', 'Intersex'. These were chosen to reflect values that DCCs currently have in their project metadata. Moving forward, we realize this field and associated vocabulary are inadequate for all potential needs (e.g. transgender people). Thus, we plan to revise this field and its allowed values in light of United States Core Data for Interoperability (USCDI) (https://www.healthit.gov/isa/united-states-core-data-interoperability-uscdi) and Health Level Seven (HL7) (https://www.hl7.org/) standards as they evolve to meet community needs. To capture race and ethnicity we will be using the Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity (https://www.govinfo.gov/content/pkg/FR-1997-10-30/pdf/97-28653.pdf) and will strive for continued alignment with the USCDI.

The CFDE central metadata repository captures metadata about subjects, biosamples, and files as core entities. Currently, there are ~4.5 million file records with links to OBI assay terms, EDAM format terms, and EDAM data terms. In addition, there are more than 2,700 subject records linked to Disease Ontology terms and more than 1.7 million biosample records linked to Uberon anatomy terms. This is just a sampling of the large quantities of metadata currently contained in the CFDE central repository and accessible through the CFDE Portal (https://app.nih-cfde.org/).

Table 1.
Ontologies and CVs in use by CFDE. Abbreviations: EDAM=Ontology of bioscientific data analysis and data management; UBERON= Uber-anatomy ontology; OBI=Ontology for Biomedical Investigations; DO=Human Disease Ontology; HPO=Human Phenotype Ontology; NA=Not Applicable. *Federal CV is described here: https://www.govinfo.gov/content/pkg/FR-1997-10-30/pdf/97-28653.pdf



metadata type	ontology/cv (ref)	# terms used	CFDE slim
file format	EDAM (11)	55	yes
data type	EDAM	38	yes
anatomy	UBERON (12)	334	yes
assay type	OBI (7)	100	yes
analysis type	OBI	5	no
disease	DO (13)	1894	yes
chemicals	PubChem (14)	73,498	no
phenotype	HPO (15)	43	no
taxonomy	NCBI taxon (16)	2429	yes
sex	SnoMed subset (10)	3	NA
race	federal CV*	5	NA
ethnicity	federal CV	2	NA

5. Ongoing work and future directions

Although the use of standards such as ontologies and CVs is crucial for data harmonization, it is not enough. Even when multiple parties are using the same ontology, there can still be inconsistencies in the use of terms for specific situations. For example, we found that even within a single DCC when looking at output files of the same kind, with the same format, and from the same software tool, different curators sometimes chose different data type terms. The problem becomes even more magnified when looking across DCCs. Therefore, in an effort to increase the consistent use of the standards chosen for the CFDE central metadata repository, we are working to identify inconsistent use of terms and then to build term-use guidelines that will address those inconsistencies across DCCs.

In the area of ontology/CV slim development, we will continue to revise the slims we have built for each ontology/CV as more metadata records are submitted to the central repository and as new DCCs join the CFDE so as to maintain the most useful set of general categories for each ontology/CV as possible. We will also build a slim for PubChem since, with more than 73,000 terms in use, a slim will be very advantageous for summary views and comparisons. We will build slims for additional ontologies used in CFDE as needed.

As work on the project continues, we hope to add additional metadata types to our central repository. We will use the above process to identify community standards to use for those datatypes as well.

6. Acknowledgements

We wish to acknowledge the NIH for funding much of this work through award OT3-OD02549. We also want to acknowledge all of the members of the Ontology Working Group: David Chen, Keyang Yu, Matt Roth - Baylor College of Medicine; Jared Nedzel - Broad Institute; Allison Heath, Deanne M Taylor, Eric Wenger, Taha M. Ahooyi - Children's Hospital of Philadelphia; Mano Ram Maurya - Department of Bioengineering, University of California, San Diego; Avi Ma'ayan, Eryk Kropiwnicki, John Erol Evangelista, Sherry L. Jenkins, Zhuorui Xie - Department of Pharmacological Sciences, Mount Sinai Center for Bioinformatics, Icahn School of Medicine at Mount Sinai; Daniel Lyman - George Washington University; Andrew Schroeder, Sarah Reiff - Harvard Medical School; Ellen Quardokus, Katy Borner - Indiana University; Robert Carter - Institute for Genome Sciences, University of Maryland School of Medicine; Asiyah Lin, Chris Kinsinger, Erika Kim, George Papanicolaou, Haluk Resat, Olga Vovk - National Institutes of Health; Brian Walsh - Oregon Health and Science University; Diane Eshelman, Phil Blood - Pittsburg Supercomputing Center; Christophe Lambert, Cristian Bologa, Jessica Binder, Vincent Metzger - School of Medicine, Department of Internal Medicine, Translational Informatics Division, University of New Mexico; Soha Hassoun - Tufts University, Medford; Bernard de Bono - University of Auckland, New Zealand; Amanda Charbonneau, Jeremy Walter, Saranya Canchi - University of California, Davis; Srinivasan Ramachandran - University of California, San Diego; Jie Liu, Yuanhao Huang - University of Michigan; Steve Mathias - University of New Mexico; Susanna-Assunta Sansone - University of Oxford.

7. References

[1] Charbonneau AL, Brady A, Czajkowski K, Aluvathingal J, Canchi S, Carter R, Chard K, Clarke DJB, Crabtree J, Creasy HH, D'Arcy M, Felix V, Giglio M, Gingrich A, Harris RM, Hodges TK, Ifeonu O, Jeon M, Kropiwnicki E, Lim MCW, Liming RL, Lumian J, Mahurkar AA, Mandal M, Munro JB, Nadendla S, Richter R, Romano C, Rocca-Serra P, Schor M, Schuler RE, Tangmunarunkit H, Waldrop A, Williams C, Word K, Sansone SA, Ma'ayan A, Wagner R, Foster I, Kesselman C, Brown CT, White O. Making Common Fund data more findable: catalyzing a data ecosystem. Gigascience. 2022 Nov 21;11:giac105. doi: 10.1093/gigascience/giac105. PMID: 36409836; PMCID: PMC9677336.

[2] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3:160018. doi: 10.1038/sdata.2016.18. Erratum in: Sci Data. 2019 Mar 19;6(1):6. PMID: 26978244; PMCID: PMC4792175.

[3] Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, Musen MA. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W541-5. doi: 10.1093/nar/gkr469. Epub 2011 Jun 14. PMID: 21672956; PMCID: PMC3125807.

[4] Martínez-Romero M, Jonquet C, O'Connor MJ, Graybeal J, Pazos A, Musen MA. NCBO Ontology Recommender 2.0: an enhanced approach for biomedical ontology recommendation. J Biomed Semantics. 2017 Jun 7;8(1):21. doi: 10.1186/s13326-017-0128-y. PMID: 28592275; PMCID: PMC5463318.

[5] JSimon Jupp, Tony Burdett, James Malone, Catherine Leroy, Matt Pearce, Julie McMurry, Helen Parkinson. A new Ontology Lookup Service at EMBL-EBI. In: Malone, J. et al. (eds.) Proceedings of SWAT4LS International Conference 2015\. 2015 http://ceur-ws.org/Vol-1546/paper_29.pdf

[6] Jackson R, Matentzoglu N, Overton JA, Vita R, Balhoff JP, Buttigieg PL, Carbon S, Courtot M, Diehl AD, Dooley DM, Duncan WD, Harris NL, Haendel MA, Lewis SE, Natale DA, Osumi-Sutherland D, Ruttenberg A, Schriml LM, Smith B, Stoeckert CJ Jr, Vasilevsky NA, Walls RL, Zheng J, Mungall CJ, Peters B. OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. Database (Oxford). 2021 Oct 26;2021:baab069. doi: 10.1093/database/baab069. PMID: 34697637; PMCID: PMC8546234.

[7] Monique C. Surles-Zeigler, Troy Sincomb, Thomas H. Gillespie, Bernard de Bono, Jacqueline Bresnahan, Gary M. Mawe, Jeffrey S. Grethe, Susan Tappan, Maci Heal, Maryann E. Martone. Extending and using anatomical vocabularies in the Stimulating Peripheral Activity to Relieve Conditions (SPARC) program. bioRxiv, December 19, 2021 doi: https://doi.org/10.1101/2021.11.15.467961

[8] Bandrowski A, Brinkman R, Brochhausen M, Brush MH, Bug B, Chibucos MC, Clancy K, Courtot M, Derom D, Dumontier M, Fan L, Fostel J, Fragoso G, Gibson F, Gonzalez-Beltran A, Haendel MA, He Y, Heiskanen M, Hernandez-Boussard T, Jensen M, Lin Y, Lister AL, Lord P, Malone J, Manduchi E, McGee M, Morrison N, Overton JA, Parkinson H, Peters B, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Schober D, Smith B, Soldatova LN, Stoeckert CJ Jr, Taylor CF, Torniai C, Turner JA, Vita R, Whetzel PL, Zheng J. The Ontology for Biomedical Investigations. PLoS One. 2016 Apr 29;11(4):e0154556. doi: 10.1371/journal.pone.0154556. PMID: 27128319; PMCID: PMC4851331.

[9] The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 2019 Jan 8;47(D1):D330-D338. doi:

10.1093/nar/gky1055. PMID: 30395331; PMCID: PMC6323945.

[10] Bodenreider O, Cornet R, Vreeman DJ. Recent Developments in Clinical Terminologies - SNOMED CT, LOINC, and RxNorm. Yearb Med Inform. 2018 Aug;27(1):129-139. doi: 10.1055/s-0038-1667077. Epub 2018 Aug 29. PMID: 30157516; PMCID: PMC6115234.

[11] Ison J, Kalas M, Jonassen I, Bolser D, Uludag M, McWilliam H, Malone J, Lopez R, Pettifer S, Rice P. EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. Bioinformatics. 2013 May 15;29(10):1325-32. doi: 10.1093/bioinformatics/btt113. Epub 2013 Mar 11. PMID: 23479348; PMCID: PMC3654706.

[12] Haendel MA, Balhoff JP, Bastian FB, Blackburn DC, Blake JA, Bradford Y, Comte A, Dahdul WM, Dececchi TA, Druzinsky RE, Hayamizu TF, Ibrahim N, Lewis SE, Mabee PM, Niknejad A, Robinson-Rechavi M, Sereno PC, Mungall CJ. Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. J Biomed Semantics. 2014 May 19;5:21. doi: 10.1186/2041-1480-5-21. PMID: 25009735; PMCID: PMC4089931.

[13] Schriml LM, Munro JB, Schor M, Olley D, McCracken C, Felix V, Baron JA, Jackson R, Bello SM, Bearer C, Lichenstein R, Bisordi K, Dialo NC, Giglio M, Greene C. The Human Disease Ontology 2022 update. Nucleic Acids Res. 2022 Jan 7;50(D1):D1255-D1261. doi: 10.1093/nar/gkab1063. PMID: 34755882; PMCID: PMC8728220.

[14] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res. 2021 Jan 8;49(D1):D1388-D1395. doi: 10.1093/nar/gkaa971. PMID: 33151290; PMCID: PMC7778930.

[15] Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, Danis D, Balagura G, Baynam G, Brower AM, Callahan TJ, Chute CG, Est JL, Galer PD, Ganesan S, Griese M, Haimel M, Pazmandi J, Hanauer M, Harris NL, Hartnett MJ, Hastreiter M, Hauck F, He Y, Jeske T, Kearney H, Kindle G, Klein C, Knoflach K, Krause R, Lagorce D, McMurry JA, Miller JA, Munoz-Torres MC, Peters RL, Rapp CK, Rath AM, Rind SA, Rosenberg AZ, Segal MM, Seidel MG, Smedley D, Talmy T, Thomas Y, Wiafe SA, Xian J, Yüksel Z, Helbig I, Mungall CJ, Haendel MA, Robinson PN. The Human Phenotype Ontology in 2021. Nucleic Acids Res. 2021 Jan 8;49(D1):D1207-D1217. doi: 10.1093/nar/gkaa1043. PMID: 33264411; PMCID: PMC7778952.

[16] Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, Mcveigh R, O'Neill K, Robbertse B, Sharma S, Soussov V, Sullivan JP, Sun L, Turner S, Karsch-Mizrachi I. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford). 2020 Jan 1;2020:baaa062. doi: 10.1093/database/baaa062. PMID: 32761142; PMCID: PMC7408187.