TraitBank: Practical semantics for organism attribute data

Abstrakt

Biologové po léta střádali a udržovali informace o morfologii organismů, jejich chování a jejich vývoji mnoha rozdílnými cestami. Tyto informace poskytují data, která popisují vlastnosti organismů (výskyt, barva květu, váha aj.). Většina z těchto informací přežila ve formě volného textu nebo datových tabulek v publikovaných dokumentech. V nedávné minulosti vědecké komunity začaly anotovat tyto dokumenty, extrahovat informace z textu a budovat pro tento účel speciální databáze. Data o vlastnostech organismů z těchto databází však nemohla být efektivně využita, protože chyběla jejich kvalitní standardizace.

Organizace Encyclopedia of Life (EOL)[1] proto vytvořila systém TraitBank pro správu, organizaci a standardizaci těchto dat. TraitBank[2] pracuje s daty z několika zdrojů, jako jsou např.: databáze organismů, přírodní sbírky, vědecké texty či projekty. Data TraitBank jsou pak uložena v relačních databázích EOL v kombinaci s databázovými strukturami pro RDF data.

RDF datový model TraitBank využívá pro popis vlastností organismů konceptů z mnoha známých slovníků a ontologií (např. Dublin Core[3], Darwin Core[4], ALA[5], SIO[6], UO[7], PATO[8]), které přiřazuje jednotlivým taxonům (taxon = taxonomická jednotka v taxonomické nomenklatuře organismů - druh, třída, rod apod.). Jeden záznam v TraitBank tedy obsahuje odkaz na taxon v EOL a k tomu navíc koncepty, definující danou vlastnost taxonu. Mezi koncepty se řadí atributy, hodnoty a metadata.

Pokud je do TraitBank přidán nový koncept, je zapotřebí jej přidat do TraitBank URI registru. Pokud se jedná o koncept typu atribut, pak je nutné ho zařadit do jedné z tematických kategorií (Ekologie, Chování, Evoluce a sémantika, Fyzický popis a další). Atributy je také potřeba ohodnotit z hlediska jejich významu pro uživatele EOL. Atributy s vyšší prominencí jsou poté zobrazovány v EOL přednostně.

Ne všechna data ze zdrojových systémů přicházejí do TraitBank se sémantickou anotací. V opačném případě taková data převezmou zaměstnanci EOL a analyzují význam jednotlivých atributů. Poté vyberou z patřičné ontologie atribut, který nejlépe odpovídá odhadnutému významu. Dobrou praktikou je zde využívat ontologie, které jsou aktivně vyvíjeny (OBO[9], GO[10], PO[11]). Ne však pro všechny nově zařazené koncepty v TraitBank existuje ekvivalentní koncept v nějaké existující ontologii či slovníku. Proto EOL přidává do ontologií jako jsou PATO, CHEBI[12] nebo RO[13] pravidelně

[1] http://eol.org/
[2] http://eol.org/traitbank
[3] http://dublincore.org
[4] http://rs.tdwg.org/dwc/
[5] http://www.ala.org.au
[6] http://semanticscience.org
[7] http://code.google.com/p/unit-ontology/
[8] http://purl.bioontology.org/ontology/PATO
[9] http://obofoundry.org
[10] http://geoontology.org
[11] http://www.plantontology.org
[12] http://www.ebi.ac.uk./chebi/

nové koncepty, aby tak navýšila a zpřesnila možnosti popsání vlastností organismů. Některé vlastnosti organismů jsou naopak problematické v tom, že jsou příliš komplexní a nelze si při jejich definici vystačit pouze s jednou onotologií. Pro tvorbu nových atributů k těmto vlastnostem, které se skládají z tříd z více ontologií, je k dispozici nástroj Term Genie[14], který dokáže kombinovat třídy a vytvářet tak nové atributy.

Atributy z TraitBank, jejich definice a URI jsou společně evidovány v TraitBank Data Glossary[15], který vychází z TraitBank URI registru. Pro importované atributy z jiných ontologií může sloužit Data Glossary jako záložní zdroj pro případ, že by došlo k migraci nebo k nedostupnosti původní ontologie.

Kvalita dat v TraitBank se velmi liší dle zdroje, ze kterého byly pořízeny. Vedle důvěryhodných dat (např. z ověřených vědeckých studií) zde mohou být i data poněkud „pochybného“ původu (nedůvěryhodné analýzy, statistiky bez ověřeného zdroje). Jelikož tato data mohou poskytovat informace, které nelze získat vědeckým výzkumem, nebo zde takováto data nejsou k dispozici nebo nemohou být použita, je všeobecně přijímán fakt, že alespoň pro některé aplikace platí, že je lepší mít nějaká data než žádná.

Dotazování a přístup k datům v TraitBank jsou možné přes rozhraní EOL nebo přes JSON-LD API. Rozhraní EOL představuje webovou aplikaci, přes kterou je možné zadávat dotazy na data z TraitBank, která jsou také přehlednou formou prezentována.

TraitBank k 27. lednu 2015 obsahoval 331 vlastností organismů.

[13] http://code.google.com/p/obo-relations/
[14] http://www.berkeleybop.org/software/termgenie
[15] http://eol.org/data_glossary

Reference

[1] Angeli, N. F., Otegui, J., Wood, M., and Gomez-Ruiz, E. P., A process to support species conservation planning and climate change readiness in protected areas, PeerJ PrePrints 2 (2014), e492v2. http://dx.doi.org/10.7287/peerj.preprin ts.492v2

[2] Baker, E., Rycroft, S., and Smith, V. S., Linking multiple biodiversity informatics platforms with Darwin Core Archives, Biodiversity Data Journal 2 (2014), e1039. http://dx.doi.org/10.3897/BDJ.2.e1039

[3] Balhoff, J. P., Dahdul, W. M., Kothari, C. R., Lapp, H., Lundberg, J. G., Mabee, P., Midford, P. E., Westerfield, M., and Vision, T. J., Phenex: Ontological annotation of phenotypic diversity, PLoS ONE 5(5) (2010), 10. http://dx.doi.org/10.1371/journal.pone.0010 500

[4] Barnagaud, J.-Y., Papaïx, J., Gimenez, O., and Svenning, J.-C., Dynamic spatial interactions between the native invader Brown-headed Cowbird and its hosts, Diversity and Distributions 21(5) (2014), 511– 522. http://dx.doi.org/10.1111/ddi.12275

[5] Bizer, C., Heath, T., and Berners-Lee, T., Linked data - the story so far, International Journal on Semantic Web and Information Systems 5(3) (2009), 1– 22. http://dx.doi.org/10.4018/jswis.2009081901

[6] Burleigh, J. G., Alphonse, K., Alverson, A. J., Bik, H. M., Blank, C., Cirranello, A. L., Cui, H., Daly, M., Dietterich, T. G., Gasparich, G., Irvine, J., Julius, M., Kaufman, S., Law, E., Liu, J., Moore, L., O'Leary, M. A., Passarotti, M., Ranade, S., Simmons, N. B., Stevenson, D. W., Thacker, R. W., C Theriot, E., Todorovic, S., Velazco, P. M., Walls, R. L., Wolfe, J. M., and Yu, M., Next-generation phenomics for the Tree of Life, PLoS Currents Tree of Life, 2013 Jun 26, Edition 1, 2013. http://dx.doi.org/10.1371/currents.tol.085 c713acafc8711b2ff7010a4b03733

[7] Buttigieg, P. L., Morrison, N., Smith, B., Mungall, C. J., Lewis, S. E., and the ENVO Consortium, The environment ontology: contextualising biological and biomedical entities, Journal of Biomedical Semantics 4 (2013), 43. http://dx.doi.org/10.1186/2041- 1480-4-43

[8] Caldwell, I. R. and Hart, E. M., Using Encyclopedia of Life's TraitBank to identify plant traits associated with vulnerability, PeerJ PrePrints 2 (2014), e491v1. http://dx.doi.org/10.7287/peerj.preprin ts.491v1

[9] Chapman, A. D., Numbers of living species in Australia and the World Report, 2nd edition, Commonwealth of Australia, Department of the Environment and Water Resources, 2009. http://www.environment.gov.au/node/13 876

[10] Chave, J., Coomes, D., Jansen, S., Lewis, S. L., Swenson, N. G., and Zanne, A. E., Towards a worldwide wood economics spectrum, Ecology Letters 12 (2009), 351– 366. http://dx.doi.org/10.1111/j.1461- 0248.2009.01285.x

[11] Courtot, M., Gibson, F., Lister, A. L., Malone, J., Schober, D., Brinkman, R. R., and Ruttenberg, A., MIREOT: The minimum information to reference an external ontology term, Applied Ontology 6 (2011), 23– 33. http://dx.doi.org/10.1038/npre.2009.3576.1

[12] Deans, A. R., Lewis, S. E., Huala, E., Anzaldo, S. S., Ashburner, M., Balhoff, J. P., Blackburn, D. C., Blake, J. A., Burleigh, J. G., Chanet, B., Cooper, L. D., Courtot, M., Csösz, S., Cui, H., Dahdul, W., Das, S., Dececchi, T. A., Dettai, A., Diogo, R., Druzinsky, R. E., Dumontier, M., Franz, N. M., Friedrich, F., Gkoutos, G. V., Haendel, M., Harmon, L. J., Hayamizu, T. F., He, Y., Hines, H. M., Ibrahim, N., Jackson, L. M., Jaiswal, P., James-Zorn, C., Köhler, S., Lecointre, G., Lapp, H., Lawrence, C. J., Le Novère, N., Lundberg, J. G., Macklin, J., Mast, A. R., Midford, P. E., Mikó, I., Mungall, C. J., Oellrich, A., Osumi-Sutherland, D., Parkinson, H., Ramírez, M. J., Richter, S., Robinson, P. N., Ruttenberg, A., Schulz, K. S., Segerdell, E., Seltmann, K. C., Sharkey, M. J., Smith, A. D., Smith, B., Specht, C. D., Squires, R. B., Thacker, R. W., Thessen, A., FernandezTriana, J., Vihinen, M., Vize, P. D., Vogt, L., Wall, C. E., Walls, R. L., Westerfeld, M., Wharton, R. A., Wirkner, C. S., Woolley, J. B., Yoder, M. J., Zorn, A. M., and Mabee, P., Finding our way through phenotypes, PLOS Biology 3(1) (2015), e1002033. http://dx.doi.org/10.1371/journal.pbi o.1002033

[13] Dumontier, M., Baker, C. J., Baran, J., Callahan, A., Chepelev, L., Cruz-Toledo, J., Del Rio, N. R., Duck, G., Furlong, L. I., Keath, N., Klassen, D., McCusker, J. P., QueraltRosinach, N., Samwald, M., VillanuevaRosales, N., Wilkinson, M. D., and Hoehndorf, R., The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery, Journal of Biomedical Semantics 5(1) (2014), 14. http://dx.doi.org/10.1186/2041-1480-5-14

[14] Faulwetter, S., Markantonatou, V., Pavloudi, C., Papageorgiou, N., Keklikoglou, K., Chatzinikolaou, E., Pafilis, E., Chatzigeorgiou, G., Vasileiadou, K., Dailianis, T., Fanini, L., Koulouri, P., and Arvanitidis, C., Polytraits: A database on biological traits of marine polychaetes, Biodiversity Data Journal 2 (2014), e1024. http://dx.doi.org/10.3897/BDJ.2.e1024

[15] Franz, N. M. and Thau, D., Biological taxonomy and ontology development: Scope and limitations, Biodiversity Informatics 7 (2010), 45–66.

[16] Gkoutos, G. V, Schofield, P. N., and Hoehndorf, R., The Units Ontology: a tool for integrating units of measurement in science, Database: The Journal of Biological Databases and Curation 2012 (2012), bas033. http://dx.doi.org/10.1093/database/bas 033

[17] Guisan, A., Biodiversity: Predictive traits to the rescue, Nature Climate Change 4(3) (2014), 175– 176. http://dx.doi.org/10.1038/nclimate2157

[18] Harfoot, M. and Roberts, D., Taxonomy: Call for ecosystem modelling data, Nature 505(7482) (2014), 160. http://dx.doi.org/10.1038/505160a

[19] Harmon, L. J., Baumes, J., Hughes, C., Soberon, J., Specht, C. D., Turner, W., Lisle, C., and Thacker, R. W., Arbor: Comparative analysis workflows for the Tree of Life, PLOS Currents Tree of Life, 2013 Jun 26, Edition 1, 2013. http://dx.doi.org/10.1371/currents.tol.099 161de5eabdee073fd3d21a44518dc

[20] Heidorn, P., Shedding light on the dark data in the long tail of science, in: Institutional Repositories: Current State and Future, Sarah Sheeves and Melissa Cragin, eds., Library Trends 57(2) (2008), 280– 299. http://hdl.handle.net/2142/9127

[21] Jaiswal, P., Ware, D., Ni, J., Chang, K., Zhao, W., Schmidt, S., Pan, X., Clark, K., Teytelman, L., Cartinhour, S., Stein, L., and McCouch, S., Gramene: development and integration of trait and gene ontologies for rice, Comparative and Functional Genomics 3(2) (2002), 132– 136. http://dx.doi.org/10.1002/cfg.156

[22] Jones, K. E., Bielby, J., Cardillo, M., Fritz, S. A., O'Dell, J., Orme, C. D. L., Safi, K., Sechrest, W., Boakes, E. H., Carbone, C., Connolly, C., Cutts, M. J., Foster, J. K., Grenyer, R., Habib, M., Plaster, C. A., Price, S. A., Rigby, E. A., Rist, J., Teacher, A., BinindaEmonds, O. R. P., Gittleman, J. L., Mace, G. M., and Purvis, A., PanTHERIA: a specieslevel database of life history, ecology, and geography of extant and recently extinct mammals, Ecology 90 (2009), 2648. http://dx.doi.org/10.1890/08-1494.1

[23] Kao, R. H., Gibson, C. M., Gallery, R. E., Meier, C. L., Barnett, D. T., Docherty, K. M., Blevins, K. K., Travers, P. D., Azuaje, E., Springer, Y. P., Thibault, K. M., McKenzie, V. J., Keller, M., Alves, L. F., Hinckley, E.-L. S., Parnell, J., and Schimel, D., NEON terrestrial field observations: designing continental-scale, standardized sampling, Ecosphere 3(12) (2012), 1–17. http://dx.doi.org/10.1890/ES12- 00196.1

[24] Kattge, J., Díaz, S., Lavorel, S., Prentice, I. C., Leadley, P., Bönisch, G., Garnier, E., Westoby, M., Reich, P. B., Wright, I. J., Cornelissen, J. H. C., Violle, C., Harrison, S. P., Van Bodegom, P. M., Reichstein, M., Enquist, B. J., Soudzilovskaia, N. A., Ackerly, D. D., Anand, M., Atkin, O., Bahn, M., Baker, T. R., Baldocchi, D., Bekker, R., Blanco, C. C., Blonder, B., Bond, W. J., Bradstock, R., Bunker, D. E., Casanoves, F., Cavender-Bares, J., Chambers, J. Q., Chapin Iii, F. S., Chave, J., Coomes, D., Cornwell, W. K., Craine, J. M., Dobrin, B. H., Duarte, L., Durka, W., Elser, J., Esser, G., Estiarte, M., Fagan, W. F., Fang, J., Fernández-Méndez, F., Fidelis, A., Finegan, B., Flores, O., Ford, H., Frank, D., Freschet, G. T., Fyllas, N. M., Gallagher, R. V., Green, W. A., Gutierrez, A. G., Hickler, T., Higgins, S. I., Hodgson, J. G., Jalili, A., Jansen, S., Joly, C. A., Kerkhoff, A. J., Kirkup, D., Kitajima, K., Kleyer, M., Klotz, S., Knops, J. M. H., Kramer, K., Kühn, I., Kurokawa, H., Laughlin, D., Lee, T. D., Leishman, M., Lens, F., Lenz, T., Lewis, S. L., Lloyd, J., Llusià, J., Louault, F., Ma, S., Mahecha, M. D., Manning, P., Massad, T., Medlyn, B. E., Messier, J., Moles, A. T., Müller, S. C., Nadrowski, K., Naeem, S., Niinemets, Ü., Nöllert, S., Nüske, A., Ogaya, R., Oleksyn, J., Onipchenko, V. G., Onoda, Y., Ordoñez, J., Overbeck, G., Ozinga, W. A., Patiño, S., Paula, S., Pausas, J. G., Peñuelas, J., Phillips, O. L., Pillar, V., Poorter, H., Poorter, L., Poschlod, P., Prinzing, A., Proulx, R., Rammig, A., Reinsch, S., Reu, B., Sack, L., Salgado-Negret, B., Sardans, J., Shiodera, S., Shipley, B., Siefert, A., Sosinski, E., Soussana, J.-F., Swaine, E., Swenson, N., Thompson, K., Thornton, P., Waldram, M., Weiher, E., White, M., White, S., Wright, S. J., Yguel, B., Zaehle, S., Zanne, A. E., and Wirth, C., TRY - a global database of plant traits, Global Change Biology 17(9) (2011), 2905– 2935. http://dx.doi.org/10.1111/j.1365- 2486.2011.02451.x

[25] Lanthaler, M. and Gütl, C., On using JSON-LD to create evolvable RESTful services, in: Proceedings of the 3rd International Workshop on RESTful Design WSREST 2012 at WWW2012, Rosa Alarcon, Cesare Pautasso, and Erik Wilde, eds., ACM Press, 2012, pp. 25–32. http://dx.doi.org/10.1145/2307819.2307827

[26] Mabee, P. M., Ashburner, M., Cronk, Q., Gkoutos, G. V, Haendel, M., Segerdell, E., Mungall, C., and Westerfield, M., Phenotype ontologies: the bridge between genomics and evolution, Trends in Ecology & Evolution 22(7) (2007), 345– 50. http://dx.doi.org/10.1016/j.tree.2007.03.013

[27] Mungall, C. J., Torniai, C., Gkoutos, G. V, Lewis, S. E., and Haendel, M. A., Uberon, an integrative multi-species anatomy ontology, Genome Biology 13 (2012), R5. http://dx.doi.org/10.1186/gb-2012-13-1-r5

[28] Pafilis, E., Frankild, S. P., Schnetzer, J., Fanini, L., Faulwetter, S., Pavloudi, C., Vasileiadou, A., Leary, P., Hammock, J., Schulz, K., Parr, C. S., Arvanitidis, C., and Jensen, L. J., ENVIRONMENTS and EOL: identification of Environment Ontology terms in text and the annotation of the Encyclopedia of Life, Bioinformatics 31(11) (2015), 1872– 1874. http://dx.doi.org/10.1093/bioinformatics/ btv045

[29] Park, C. A., Bello, S. M., Smith, C. L., Hu, Z.- L., Munzenmaier, D. H., Nigam, R., Smith, J. R., Shimoyama, M., Eppig, J. T., and Reecy, J. M., The Vertebrate Trait Ontology: a controlled vocabulary for the annotation of trait data across species, Journal of Biomedical Semantics 4(1) (2013), 13. http://dx.doi.org/10.1186/2041-1480-4-13

[30] Parr, C. S. and McClain, C. R., EOL-BHLNESCent Research Sprint Report, PeerJ PrePrints 2 (2014), e503v1. http://dx.doi.org/10.7287/peerj.preprin ts.503v1

[31] Parr, C. S., Sachs, J., Parafiynyk, A., Wang, T., Espinosa, R., and Finin, T., ETHAN: the Evolutionary Trees and Natural History Ontology, Tech Report, University of Maryland, Baltimore County, 2006. http://aisl.umbc.edu/get/softcopy/id/1025 /1025.pdf

[32] Parr, C. S., Wilson, N., Leary, P., Schulz, K. S., Lans, K., Walley, L., Hammock, J. A., Goddard, A., Rice J., Studer, M., Holmes, J. T. G., and Corrigan Jr., R. J., The Encyclopedia of Life v2: Providing global access to knowledge about life on Earth, Biodiversity Data Journal 2 (2014), e1079. http://dx.doi.org/10.3897/BDJ.2.e1079.

[33] Patterson, D. J., Faulwetter, S. and Shipunov, A., Principles for a names-based cyberinfrastructure to serve all of biology, Zootaxa 1950 (2008), 153–163.

[34] Payne, P. R. O., Chapter 1: Biomedical knowledge integration, PLoS Computational Biology 8(12) (2012), e1002826. http://dx.doi.org/10.1371/journal.pc bi.1002826

[35] Pereira, H. M., Ferrier, S., Walters, M., Geller, G. N., Jongman, R. H. G., Scholes, R. J., Bruford, M. W., Brummitt, N., Butchart, S. H. M., Cardoso, A. C., Coops, N. C., Dulloo, E., Faith, D. P., Freyhof, J., Gregory, R. D., Heip, C., Höft, R., Hurtt, G., Jetz, W., Karp, D. S., McGeoch, M. A., Obura, D., Onoda, Y., Pettorelli, N., Reyers, B., Sayre, R., Scharlemann, J. P. W., Stuart, S. N., Turak, E., Walpole, M., and Wegmann, M., Essential Biodiversity Variables, Science 339 (6117) (2013), 277– 278. http://dx.doi.org/10.1126/science.1229931

[36] Poelen, J. H., Simons, J. D. and Mungall, C. J., Global Biotic Interactions: An open infrastructure to share and analyze speciesinteraction datasets, Ecological Informatics 24 (2014), 148– 159. http://dx.doi.org/10.1016/j.ecoinf.2014.08. 005

[37] Quintero, E., Thessen, A. E., Arias-Caballero, P., and Ayala-Orozco, B., A statistical assessment of population trends for data deficient Mexican amphibians, PeerJ 2 (2014), e703. http://dx.doi.org/10.7717/peerj.703

[38] Roskov Y., Kunze T., Orrell T., Abucay L., Paglinawan L., Culham A., Bailly N., Kirk P., Bourgoin T., Baillargeon G., Decock W., De Wever A., and Didžiulis V., eds., Species 2000 & ITIS Catalogue of Life, 2013 Annual Checklist, Species 2000: Naturalis, Leiden, the Netherlands, 2013. Digital resource at http://www.catalogueoflife.org/annualchecklist/2013/

[39] Rotman, D., Procita, K., Hansen, D., Parr, C. S., and Preece, J., Supporting content curation communities: The case of the Encyclopedia of Life, Journal of the American Society for Information Science and Technology 63(6) (2012), 1– 29. http://dx.doi.org/10.1002/asi.22633

[40] Thessen, A. E. and Parr, C. S., Knowledge extraction and semantic annotation of text from the Encyclopedia of Life, PLoS ONE 9(3) (2014), e89550. http://dx.doi.org/10.1371/journal.pone. 0089550

[41] Vision, T., The Dryad Digital Repository: Published evolutionary data as a part of the greater data ecosystem, Nature Precedings (713) (2010), 1– 1. http://hdl.handle.net/10101/npre.2010.4595.1

[42] Urbani, J., Three laws learned from web-scale reasoning, in 2013 AAAI Fall Symposium Series, Semantics for Big Data, Frank van Harmelen, James A. Hendler, Pascal Hitzler, and Krzysztof Janowicz, Program Cochairs, eds., (FS-13-04), 2013. https://www.aaai.org/ocs/index.php/FSS/ FSS13/paper/view/7585

[43] Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertsib, T., and Vieglais, D., Darwin Core: An evolving community-developed biodiversity data standard, PLoS ONE 7(1) (2012), e29715. http://dx.doi.org/10.1371/journal.pone. 0029715

[44] Wright, C. and Seltmann, K., Usage patterns of blue flower color representation by Encyclopedia of Life content providers, Biodiversity Data Journal 2 (2014), e1143. http://dx.doi.org/10.3897/BDJ.2.e1143