HHS Public Access

Author manuscript

J Proteome Res. Author manuscript; available in PMC 2022 May 07.

Published in final edited form as:

J Proteome Res. 2021 May 07; 20(5): 2182–2186. doi:10.1021/acs.jproteome.1c00177.

A Second Look at FAIR in Proteomic Investigations

J. Harry Caufield,
Department of Physiology and NHLBI Integrated Cardiovascular Data Science Training Program (iDISCOVER), University of California, Los Angeles, California 90095, United States

John Fu,
NHLBI Integrated Cardiovascular Data Science Training Program (iDISCOVER), University of California, Los Angeles, California 90095, United States

Ding Wang,
Department of Physiology, University of California, Los Angeles, California 90095, United States

Vladimir Guevara-Gonzalez,
NHLBI Integrated Cardiovascular Data Science Training Program (iDISCOVER), University of California, Los Angeles, California 90095, United States

Wei Wang,
Department of Computer Science, Department of Computational Medicine, Scalable Analytics Institute (ScAi), and Bioinformatics Interdepartmental Graduate Program, University of California, Los Angeles, California 90095, United States

Peipei Ping
Department of Physiology, NHLBI Integrated Cardiovascular Data Science Training Program (iDISCOVER), Scalable Analytics Institute (ScAi), Bioinformatics Interdepartmental Graduate Program, Department of Biomedical Informatics, and Department of Medicine/Cardiology, University of California, Los Angeles, California 90095, United States

Abstract

Proteomics is, by definition, comprehensive and large-scale, seeking to unravel ome-level protein features with phenotypic information on an entire system, an organ, cells, or organisms. This scope consistently involves and extends beyond single experiments. Multitudinous resources now exist to assist in making the results of proteomics experiments more findable, accessible, interoperable, and reusable (FAIR), yet many tools are awaiting to be adopted by our community. Here we highlight strategies for expanding the impact of proteomics data beyond single studies. We show how linking specific terminologies, identifiers, and text (words) can unify individual data points across a wide spectrum of studies and, more importantly, how this approach may potentially

Corresponding Author Peipei Ping – Department of Physiology, NHLBI Integrated Cardiovascular Data Science Training Program (iDISCOVER), Scalable Analytics Institute (ScAi), Bioinformatics Interdepartmental Graduate Program, Department of Biomedical Informatics, and Department of Medicine/Cardiology, University of California, Los Angeles, California 90095, United States; ppingucla@gmail.com.

Author Contributions

J.H.C., J.F., D.W., V.G.-G., and P.P. wrote the paper. J.H.C., D.W., and V.G.-G. designed the figures. W.W. and P.P. reviewed the paper.

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.1c00177

The authors declare no competing financial interest.

reveal novel relationships. In this effort, we explain how data sets and methods can be rendered more linkable and how this maximizes their value. We also include a discussion on how data linking strategies benefit stakeholders across the proteomics community and beyond.

Keywords

data sharing; FAIR principles; ontologies; knowledgebases; standardization

1. INTRODUCTION

Proteomics is unquestionably a data-rich and data-driven field. From the immediate, raw spectra output of the mass spectrometers to the organized, identified, analyzed, and deposited peptides, every step of this analytical platform concerns voluminous and multifaceted observations. This poses a challenge not only in managing and interpreting each singular data set but in connecting them across multiple studies. Building upon the solid foundations established by proteomics standards[1] and shared resources[2–4] for data set accessibility and interoperability, considerable energy is necessary to find contextual as well as ontology-defined connections among observations and results across multiple proteomic studies. The opportunity to democratize proteomics results and render them truly "linkable" is within our grasp, and it may be obtained through a collection of practical as well as parallel community-wide actions.

Major steps toward making data more findable, accessible, interoperable, and reusable (i.e., FAIR[5]), are gaining acceptance in proteomics but are not yet commonplace. Every upload to ProteomeXchange constitutes another opportunity to find and access data. The observations made about individual proteins are not limited to entries in a data table: Findings of interest are included in the manuscript text. This text data offers further chances to link observations, whether directly or indirectly through curated knowledgebases. Here we briefly touch upon practical approaches to rendering concepts described in proteomics publications more linkable, including their accompanying data and methods. We also provide justification for these efforts. Encouraging FAIR data and linkable manuscripts is not simply a community-building effort but rather a set of behaviors that can massively elevate the value of any completed proteomics project while requiring comparatively few resources.

How may we identify the areas in which the FAIRness of research products may be improved? One straightforward strategy is to consult the FAIRshake[6] rubrics. These metrics are not intended as an assessment of a manuscript or the resource quality but rather permit a summarization of a product's potential FAIRness limitations. When the FAIRshake data set rubric is applied to our own work on post-translational modifications in the cardiac hypertrophy proteome,[7] we find that several crucial elements are in place (e.g., the data set is freely available on ProteomeXchange, with clear contact details), but two are missing: Our data set is not described with metadata using a formal, broadly applicable vocabulary and no licensing details are provided. (The full assessment is available at https://fairshake.cloud/digital_object/811994/assessments/.) Explicitly using keywords matching a clearly defined ontology (e.g., MeSH, or one of the more domain-specific ontologies described in the next

section) would improve our data set's findability *and* ensure that it would be found along with conceptually similar data. Ideally, these linkable concepts should be clearly mentioned in the accompanying manuscript as well. Comparing these results with the assessment of one of our nonproteome data sets,[8] we find a similar limitation (with the full assessment at https://fairshake.cloud/digital_object/811995/assessments/): Linkability is limited because we have provided just a few informal keywords. At a minimum, we could have aligned with more MeSH terms by selecting keywords such as "Data Curation" or "Medical Records" to ensure that our data could be linked to any similar resources.

Here we define making research linkable as any effort resulting in direct connections between research products or between products and consistent, unique identifiers. Much in the same way as how citations identify the source of claims, ideas, and methods, researchers may identify the exact concepts and properties they refer to, reducing ambiguity and contributing to networks of conceptual relationships. This has the added benefit of maintaining the accuracy of published work over time: Concepts linked to knowledgebases can be immediately looked up in those resources. Should a human gene's canonical name change, for example, it need only be changed in the linked knowledgebase. More broadly, we refer to the FAIR principles, or efforts to ensure that data (including their accompanying descriptions, whether as formal manuscripts or otherwise) are findable, accessible, interoperable, and reusable.[5] The result is tangible benefits to the value of individual projects *and* the broader proteomics data ecosystem (Figure 1).

2. LINKING TERMS AND IDENTIFIERS

In general, linking specific observations across manuscripts in proteomics is a matter of one primary factor: Concepts must be linked to consistent, unique identifiers. This is a common but not universal practice when referring to proteins. Because most well-studied proteins correspond to UniProtKB entries, including a UniProt accession along with the first mention of a protein not only clearly defines the protein being mentioned and avoids confusion but also creates an opportunity for readers to compare the manuscript with all others mentioning the same protein. The corresponding knowledgebase entry may also assist with verifying the protein nomenclature and spelling. Beyond proteins, virtually any type of molecular, disease, pathway, or model organism or even more general concepts such as experimental procedures may be linked to knowledgebases or ontologies (Table 1). Organism and species names are especially valuable candidates for linking because many research questions involve questions of evolution,[9] for example, "How broadly conserved is the protein expression pattern we observe?" Identifiers may be provided in-line by using a compact identifier format containing both a short name denoting a source and an accession code.[10] Canonical source names can be retrieved through the Identifiers.org project.[11]

Several resources exist to assist with the term linking process. The PubReCheck tool addresses these issues,[12] essentially working as a spell-check and a system for finding undefined acronyms or identifiers. Malone et al. provide a convenient guide to selecting an appropriate ontology.[13] The BioPortal ontology collection, while somewhat imposing as a collection of more than 1000 categorization systems, also provides a recommender function (https://bioportal.bioontology.org/recommender) for suggesting suitable identifiers

given a segment of manuscript text.[14] Publishers may assist by presenting relevant terms and identifiers in an organized manner: ACS journals such as the *Journal of Proteome Research* organize articles through an internal list of topics in addition to keywords, although it is up to authors to provide identifiers for specific concepts.

3. LINKING DATA

Data may be used, reused, reprocessed, and repurposed, potentially as part of a single project or as part of numerous studies. Each data set therefore presents another opportunity for impactful linking. The construction of dedicated repositories for storing and indexing proteomics data was one of the primary accomplishments in efforts to establish a data-sharing infrastructure for the field.[16] Data indexed in public resources such as ProteomeXchange[2] is accessible and is accompanied by standardized metadata, although stored file formats and storage conventions vary. Open formats developed by the Proteomics Standards Initiative assist with data linkability by ensuring interoperability, a key element in comparing any two or more data collections.[17] Because studies increasingly cover multiple data types (as well as multiple types of omics data, such as both proteomics and transcriptomics), the sundry data sets may be appropriate for generalist repositories.[18] A 2020 NIH workshop produced a table comparing these generalist data repositories.[19]

How we store, format, and identify data affects both the interoperability and findability of proteomic data, which, in turn, determines the data linkability. To improve the interoperability of a proteomic data set, it is advisable to store the data in an open file format as well as to arrange any tabular data as “tidy” data. Open file formats are types of files that are readable on a variety of computers and operating systems.[20] Tidy data is tabular data where all variables are encoded in their own column and each observation is exactly one row. No additional data properties are stored in visual properties, such as color or font format; nesting columns or denoting group membership through spacing is also avoided. Improving the findability of a proteomic data set can be done by assigning a data set a unique identifier such as Digital Object Identifier (DOI), Archival Resource Key (ARK), or a persistent URL (PURL). For any data set that is meant to be updated, assigning a version number after every update allows for a quick way to differentiate between data sets.

With the number of proteomic data sets available as well as the size of the data sets increasing, manually validating proteomics data sets can be a difficult if not impractical task. To ensure the correctness of the data being linked, we suggest including a simple hash of the data, such as an md5 hash, along with all metadata. Hashing turns data of an arbitrary size into a fixed bit size value, such that the same data always returns the exact same fixed bit size value if unchanged. This provides a quick, automated way of ensuring that the data linked is not corrupted.[20]

4. LINKING METHODS

Just as the establishment of proteomics data repositories has proved crucial for the repurposing of data,[16] the onset of method-linking platforms has introduced the ability to securely develop and share reproducible methods.[21] Platforms such as protocols.io[21]

and Sage Synapse[22] provide the tools necessary to create projects, organize findings and protocols, and ultimately share novel research with other scientists.[21,22] Both protocols.io and Sage Synapse are available to researchers at no cost, but protocols.io provides additional services, such as method development collaboration in private workspaces and protocol execution records tracking, for a monthly subscription fee.[21] The research benefits provided by these platforms are perhaps none more significant than the benefits reaped from the tagging of research methods. Sage Synapse gives researchers the capacity to mint a DOI on assets such as methods, data, code, and analyses, providing an accessible avenue to referencing these objects in publications. Likewise, protocols.io grants investigators a DOI for published methods and archives each of these methods to ensure the long-term preservation of knowledge.[21]

The benefits of linkable methods stem from providing clear descriptions of research protocols. This includes, but is not limited to, the names and versions of the software used, the location and accession codes for data resources, and adaptations made to ensure that data processing completed appropriately (e.g., did file formats require conversion?). The provenance of all data should be clearly stated. DOIs are convenient ways to provide access to both data sets and their corresponding methods. Consistency and clarity in describing data-intensive methods catalyzes research that is not only actionable but also impactful.

5. BENEFITS OF IMPROVED FAIRness AND LINKABILITY

Linking research products has distinct benefits for multiple stakeholders, with perhaps the most value granted to researchers themselves. Because many proteomics investigators now pursue both independent projects and concurrent collaborative efforts spanning institutions and borders, it is paramount that their research products are connected and interoperable. They can gain added value from linked data and literature by using it as aggregate data sources, granting access to more comprehensive, accurate data and, most importantly, data on a scale far beyond what any single lab could produce. Expanded *in silico* cohorts and higher statistical power become realizable in this scenario. Fully linked and unambiguously written manuscripts render them eminently compatible with biomedical natural language processing approaches, supporting much more intuitive search operations and comprehensive knowledge organization[23] while massively enhancing each work's visibility.[24] Indeed, all journal readers stand to benefit from the enhanced accessibility afforded by improved linkage. Unfamiliar or ambiguous gene names are connected with standardized resources with limited additional effort. Articles can reach broader audiences. The benefits to authors are substantial: Other researchers can more easily test, validate, and integrate results. Publishers will also gain from the increased adoption of linkage. Articles will be more impactful and citable, enhancing much-needed quality metrics. All will benefit from better, more integrated, and more interoperable science.

6. CONCLUDING REMARKS

Proteomics data is a valuable resource to the broad biomedical community and, when empowered with FAIR standards, nurtures creative new discoveries. FAIRness transcends the practical efforts that are necessary to reveal the relevance with results from others and

to add value to individual data sets. When text data sets and methods jointly enrich and contribute to the proteomics data ecosystem, our entire community stands to benefit.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health (NIH) grants R35 HL135772, T32 HL139450, and R01 HL146739 and the UCLA Laubisch Endowment to P.P.

REFERENCES

(1). Taylor CF; Paton NW; Lilley KS; Binz P-A; Julian RK; Jones AR; Zhu W; Apweiler R; Aebersold R; Deutsch EW; Dunn MJ; Heck AJR; Leitner A; Macht M; Mann M; Martens L; Neubert TA; Patterson SD; Ping P; Seymour SL; Souda P; Tsugita A; Vandekerckhove J; Vondriska TM; Whitelegge JP; Wilkins MR; Xenarios I; Yates JR; Hermjakob H The Minimum Information about a Proteomics Experiment (MIAPE). Nat. Biotechnol 2007, 25 (8), 887–893. [PubMed: 17687369]

(2). Deutsch EW; Bandeira N; Sharma V; Perez-Riverol Y; Carver JJ; Kundu DJ; García-Seisdedos D; Jarnuczak AF; Hewapathirana S; Pullman BS; Wertz J; Sun Z; Kawano S; Okuda S; Watanabe Y; Hermjakob H; MacLean B; MacCoss MJ; Zhu Y; Ishihama Y; Vizcaíno JA The ProteomeXchange Consortium in 2020: Enabling 'Big Data' Approaches in Proteomics. Nucleic Acids Res 2019, gkz984.

(3). Perez-Riverol Y; Bai M; da Veiga Leprevost F; Squizzato S; Park YM; Haug K; Carroll AJ; Spalding D; Paschall J; Wang M; del-Toro N; Ternent T; Zhang P; Buso N; Bandeira N; Deutsch EW; Campbell DS; Beavis RC; Salek RM; Sarkans U; Petryszak R; Keays M; Fahy E; Sud M; Subramaniam S; Barbera A; Jiménez RC; Nesvizhskii AI; Sansone S-A; Steinbeck C; Lopez R; Vizcaíno JA; Ping P; Hermjakob H Discovering and Linking Public Omics Data Sets Using the Omics Discovery Index. Nat. Biotechnol 2017, 35 (5), 406–409. [PubMed: 28486464]

(4). Poux S; Arighi CN; Magrane M; Bateman A; Wei C-H; Lu Z; Boutet E; Bye-A-Jee H; Famiglietti ML; Roechert B; UniProt Consortium T. On Expert Curation and Scalability: UniProtKB/Swiss-Prot as a Case Study. Bioinformatics 2017, 33 (21), 3454–3460. [PubMed: 29036270]

(5). Wilkinson MD; Dumontier M; Aalbersberg Ij. J.; Appleton G; Axton M; Baak A; Blomberg N; Boiten J-W; 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 AJG; Groth P; Goble C; Grethe JS; Heringa J; 't Hoen PAC; 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 S-A; 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, 3, 160018. [PubMed: 26978244]

(6). Clarke DJB; Wang L; Jones A; Wojciechowicz ML; Torre D; Jagodnik KM; Jenkins SL; McQuilton P; Flamholz Z; Silverstein MC; Schilder BM; Robasky K; Castillo C; Idaszak R; Ahalt SC; Williams J; Schurer S; Cooper DJ; de Miranda Azevedo R; Klenk JA; Haendel MA; Nedzel J; Avillach P; Shimoyama ME; Harris RM; Gamble M; Poten R; Charbonneau AL; Larkin J; Brown CT; Bonazzi VR; Dumontier MJ; Sansone S-A; Ma'ayan A FAIRshake: Toolkit to Evaluate the FAIRness of Research Digital Resources. Cell Systems 2019, 9 (5), 417–421. [PubMed: 31677972]

(7). Wang J; Choi H; Chung NC; Cao Q; Ng DCM; Mirza B; Scruggs SB; Wang D; Garlid AO; Ping P Integrated Dissection of Cysteine Oxidative Post-Translational Modification Proteome During Cardiac Hypertrophy. J. Proteome Res 2018, 17 (12), 4243–4257. [PubMed: 30141336]

(8). Caufield JH; Zhou Y; Garlid AO; Setty SP; Liem DA; Cao Q; Lee JM; Murali S; Spendlove S; Wang W; Zhang L; Sun Y; Bui A; Hermjakob H; Watson KE; Ping P A Reference Set of Curated Biomedical Data and Metadata from Clinical Case Reports. Sci. Data 2018, 5, 180258. [PubMed: 30457569]

(9). Sarkar IN Biodiversity Informatics: Organizing and Linking Information across the Spectrum of Life. Briefings Bioinf 2007, 8 (5), 347–357.

(10). Smith B; Ashburner M; Rosse C; Bard J; Bug W; Ceusters W; Goldberg LJ; Eilbeck K; Ireland A; Mungall CJ; Leontis N; Rocca-Serra P; Ruttenberg A; Sansone S-A; Scheuermann RH; Shah N; Whetzel PL; Lewis S The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration. Nat. Biotechnol 2007, 25 (11), 1251–1255. [PubMed: 17989687]

(11). Wimalaratne SM; Juty N; Kunze J; Janée G; McMurry JA; Beard N; Jimenez R; Grethe JS; Hermjakob H; Martone ME; Clark T Uniform Resolution of Compact Identifiers for Biomedical Data. Sci. Data 2018, 5 (1), 180029. [PubMed: 29737976]

(12). Leaman R; Wei C-H; Allot A; Lu Z Ten Tips for a Text-Mining-Ready Article: How to Improve Automated Discoverability and Interpretability. PLoS Biol 2020, 18 (6), No. e3000716. [PubMed: 32479517]

(13). Malone J; Stevens R; Jupp S; Hancocks T; Parkinson H; Brooksbank C Ten Simple Rules for Selecting a Bio-Ontology. PLoS Comput. Biol 2016, 12 (2), No. e1004743. [PubMed: 26867217]

(14). 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. Journal of Biomedical Semantics 2017, 8 (1), 21. [PubMed: 28592275]

(15). The UniProt Consortium. UniProt: The Universal Protein Knowledgebase. Nucleic Acids Res 2017, 45 (D1), D158–D169. [PubMed: 27899622]

(16). Vaudel M; Verheggen K; Csordas A; Raeder H; Berven FS; Martens L; Vizcaíno JA; Barsnes H Exploring the Potential of Public Proteomics Data. Proteomics 2016, 16 (2), 214–225. [PubMed: 26449181]

(17). Deutsch EW; Orchard S; Binz P-A; Bittremieux W; Eisenacher M; Hermjakob H; Kawano S; Lam H; Mayer G; Menschaert G; Perez-Riverol Y; Salek RM; Tabb DL; Tenzer S; Vizcaíno JA; Walzer M; Jones AR Proteomics Standards Initiative: Fifteen Years of Progress and Future Work. J. Proteome Res 2017, 16 (12), 4288–4298. [PubMed: 28849660]

(18). Martone M; Stall S NIH Workshop on the Role of Generalist Repositories to Enhance Data Discoverability and Reuse: Workshop Summary https://datascience.nih.gov/data-ecosystem/NIH-data-repository-workshop-summary (accessed 2021-02-26).

(19). Stall S; Martone ME; Chandramouliswaran I; Crosas M; Federer L; Gautier J; Hahnel M; Larkin J; Lowenberg D; Pfeiffer N; Sim I; Smith T; Van Gulick AE; Walker E; Wood J; Zaringhalam M; Zigoni A Generalist Repository Comparison Chart. Zenodo 2020,(https://zenodo.org/record/3946720)

(20). Hart EM; Barmby P; LeBauer D; Michonneau F; Mount S; Mulrooney P; Poisot T; Woo KH; Zimmerman NB; Hollister JW Ten Simple Rules for Digital Data Storage. PLoS Comput. Biol 2016, 12 (10), No. e1005097. [PubMed: 27764088]

(21). protocols.io. https://www.protocols.io (accessed 2020-08-05).

(22). Sage Bionetworks. Synapse. https://www.synapse.org/ (accessed 2020-04-02).

(23). Krallinger M; Valencia A; Hirschman L Linking Genes to Literature: Text Mining, Information Extraction, and Retrieval Applications for Biology. Genome Biol 2008, 9 (S2), S8.

(24). Rinaldi A. For I Dipped into the Future. EMBO Rep 2010, 11 (5), 345–349. [PubMed: 20428107]

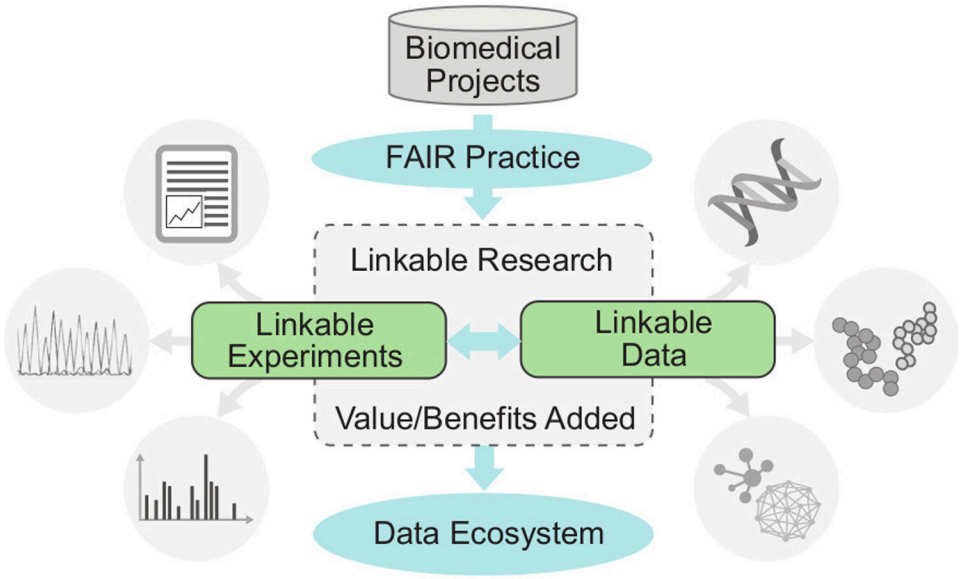

Figure 1.

Overview and benefits of improved linkability in proteomics investigations. Improved linkability of individual proteomics projects may be achieved through practices encouraging findability, accessibility, interoperability, and reusability of project data and text as well as metadata accompanying all project elements. Improved linkability has tangible benefits for individual projects (e.g., entities such as protein names may be more clearly and accurately identified). As more studies become more linkable, they improve the overall linkability of the proteomics data ecosystem. Taken together, these efforts will render proteome data more accessible, informative, and comprehensive.

Table 1.

Entity Types, Selected Knowledgebases, and Examples of Clear Entity Linking with Compact Identifiers[a]

type of entity	knowledgebase	example
protein	UniProtKB[15]	human troponin I, cardiac (UniProt: P19429)
gene	NCBI Gene	human myoglobin (NCBIGene: 4151)
disease	Disease Ontology	Tetralogy of Fallot (DOID: 6419)
chemical	CHEBI	ATP (CHEBI: 15422)
pathway	Reactome	mitochondrial biogenesis (Reactome: R-HSA-1592230)
drug	DrugBank	isoprenaline (DrugBank: DB01064)
model organism	Alliance of Genome Resources (MGI)	BALB/cJ (MGI: 2159737)
experimental methods and devices	Ontology for Biomedical Investigations	liquid chromatography mass spectrometry platform (OBI: 0000051)
proteomics standards and data formats	Mass Spectrometry Ontology	Skyline mzQuantML converter (MS: 1002546)
general concepts	MeSH	MALDI (MESH: D019032)

[a] In some cases, such as Reactome, identifiers are species-specific.