



One Thousand and One Software for Proteomics: Tales of the Toolmakers of Science

Tsiamis, Vasileios; Ioan Ienasescu, Hans; Gabrielaitis, Dovydas; Palmblad, Magnus; Schwämmle, Veit; Ison, Jon

Published in:
Journal of Proteome Research

Link to article, DOI:
[10.1021/acs.jproteome.9b00219](https://doi.org/10.1021/acs.jproteome.9b00219)

Publication date:
2019

Document Version
Peer reviewed version

[Link back to DTU Orbit](#)

Citation (APA):
Tsiamis, V., Ioan Ienasescu, H., Gabrielaitis, D., Palmblad, M., Schwämmle, V., & Ison, J. (2019). One Thousand and One Software for Proteomics: Tales of the Toolmakers of Science. *Journal of Proteome Research*, 18(10), 3580-3585. <https://doi.org/10.1021/acs.jproteome.9b00219>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

A Thousand and One Software for Proteomics: Tales of the Toolmakers of Science

Vasileios Tsiamis, Hans-Ioan Ienasescu, Dovydas Gabrielaitis,
Magnus Palmblad, Veit Schwämmle, and Jon Ison

J. Proteome Res., **Just Accepted Manuscript** • DOI: 10.1021/acs.jproteome.9b00219 • Publication Date (Web): 20 Aug 2019

Downloaded from pubs.acs.org on August 23, 2019

Just Accepted

“Just Accepted” manuscripts have been peer-reviewed and accepted for publication. They are posted online prior to technical editing, formatting for publication and author proofing. The American Chemical Society provides “Just Accepted” as a service to the research community to expedite the dissemination of scientific material as soon as possible after acceptance. “Just Accepted” manuscripts appear in full in PDF format accompanied by an HTML abstract. “Just Accepted” manuscripts have been fully peer reviewed, but should not be considered the official version of record. They are citable by the Digital Object Identifier (DOI®). “Just Accepted” is an optional service offered to authors. Therefore, the “Just Accepted” Web site may not include all articles that will be published in the journal. After a manuscript is technically edited and formatted, it will be removed from the “Just Accepted” Web site and published as an ASAP article. Note that technical editing may introduce minor changes to the manuscript text and/or graphics which could affect content, and all legal disclaimers and ethical guidelines that apply to the journal pertain. ACS cannot be held responsible for errors or consequences arising from the use of information contained in these “Just Accepted” manuscripts.

A Thousand and One Software for Proteomics: Tales of the Toolmakers of Science

Vasileios Tsiamis¹, Hans-Ioan Ienasescu², Dovydas Gabrielaitis³, Magnus Palmblad⁴, Veit Schwämmle^{1*} and Jon Ison²

*veits@bmb.sdu.dk

¹ Department of Biochemistry and Molecular Biology and VILLUM Center for Bioanalytical Sciences, University of Southern Denmark, Campusvej 55, 5230 Odense, Denmark

² National Life Science Supercomputing Center, Technical University of Denmark, Building 208, 2800 Kongens Lyngby, Denmark

³ IT Service, Technical University of Denmark, Kemitovet, Building 208, 2800 Kongens Lyngby, Denmark

⁴ Center for Proteomics and Metabolomics, Leiden University Medical Center, Postzone S3-P, Postbus 9600, 2300 RC Leiden, The Netherlands

Abstract

Proteomics is a very active field driven by frequent introduction of new technological approaches, leading to high demand for new software tools and the concurrent development of many methods for data analysis, processing and storage. The rapidly changing landscape of proteomics software makes finding a tool fit for a particular purpose a significant challenge. The comparison of software and the selection of tools capable to perform a certain operation on a given type of data relies on their detailed annotation using well-defined descriptors. However, finding accurate information including tool input/output capabilities can be challenging and often heavily depends on manual curation efforts. This is further hampered by a rather low half-life of most of the tools, thus demanding the maintenance of a resource with updated information about the tools. We present here our approach to curate a collection of 189 software tools with detailed information about their functional capabilities. We furthermore describe our efforts to reach out to the proteomics community for their engagement, which further increased the catalogue to >750 tools being about 70% of the estimated number of 1,097 tools existing for proteomics data analysis.

Descriptions of all annotated tools are available through <https://proteomics.bio.tools>

1
2
3
4
5 *Keywords:* Catalogue, community, software tools, curation, data analysis, collection, annotations, developers,
6
7 registry, descriptions
8
9
10

11 12 13 Introduction

14
15 Proteomics is a dynamic field. New experimental procedures continually emerge, driving the development of
16
17 new analytical methods and software tools, and the adaptation of existing methods and tools for new types of
18
19 data. Over the last two decades, hundreds of academic groups and individual researchers have produced, we
20
21 estimate (see *Sources of information* below), some 75,000 publications and at least 1000 software tools,
22
23 servicing a far smaller number of common operations in proteomics data analysis. While this great wealth is
24
25 welcome, it presents the scholar with significant challenges to find, compare and select appropriate tools for
26
27 the problem at hand, giving rise to various attempts to describe, organise and present the available offerings.

28
29 In the early 1990s, a manually curated Gopher or Web link collection was the place to find and share information
30
31 about online services in a particular domain. For bioinformatics there was “Pedro's BioMolecular Research
32
33 Tools”, a list of software and databases maintained by Dr. Pedro M. Coutinho, until he graduated from Iowa
34
35 State University in December 1996. “Pedro’s List” went permanently off-line sometime in late 2006 or early
36
37 2007, but inspired many other efforts in its wake.^{1,2} For example, the ms-utils.org wiki (<https://ms-utils.org>) was
38
39 created by M. Palmblad in 2006 to collect and curate free (gratis) software for analysis of mass spectrometry
40
41 data in proteomics. It includes a variety of useful information, including supported data formats, programming
42
43 language, and links to source code. Tools are sorted into categories such as peak picking, protein quantitation
44
45 and mass spectrometry imaging, and subcategories where needed. ms-utils.org is manually curated by a
46
47 community effort, including many of the software authors, and currently (Mar 2019) contains 244 annotated
48
49 tools and 141 literature references,
50

51
52 But what about Pedro’s biomolecular research tools? How many are still available at their original location? Not
53
54 many, as it turns out. Of the 184 tools (excluding mirrors) listed in the “Part 1: Molecular Biology Search and
55
56 Analysis” of Pedro’s list (last updated June 15, 1996) only 6 (or 3%) could be found at their original URL 23 years
57
58 later, suggesting a tool *URL* half-life of approximately 4.7 years; a link rot of 14% per year. If one discounts the
59
60 11 listed Gopher or 5 FTP resources (none still available in March 2019), and 6 tools which could be located by

1
2
3 URL redirection, the half-life is 6.0 years, assuming exponential decay. The services that are still findable include
4 several on ExPASy³ maintained by the Swiss Institute of Bioinformatics. PredictProtein⁴, one of the first Internet
5 servers in molecular biology, is still available, albeit through an automatic redirection since it was relocated in
6 2013 (after 21 years of service). However, these conscientiously maintained tools are rare exceptions. Things
7 are more heartening in the case of mass spectrometry-based proteomics. We extracted 55 publications
8 describing at least one new software tool (excluding databases and tool reviews) by searching PubMed for "mass
9 spectrometry" and "software" in the title, abstract or full text from 2000, 2005, 2010 and 2015. Two out of five
10 tools from 2000 were still available in 2019 (one more replaced by a tool with a different name), as were 3/7
11 from 2005, 6/10 from 2010 and 29/33 of the tools published in 2015. This gives a tool half-life of around 12
12 years, twice that for Pedro's list. This may be explained in part by mass spectrometry still being used in similar
13 ways as in the late 1990s, whereas genome sequencing methods have changed dramatically, making older tools
14 obsolete.

15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Clearly, there is an enduring requirement for ways to describe and catalogue tools, to ensure the legacy of
development is not lost, and in turn, promote the repurposing of tools to avoid reinvention. It is precisely this
need that motivates the *bio.tools* initiative⁵ to provide, under a convenient portal (<https://bio.tools>), a registry
of standardised information for over 12,000 bioinformatics tools, providing end-users a convenient means to
find, understand, compare and select appropriate tools for their research. New software tools, web applications
and databases for proteomics constantly appear, and the recently published^{5,6} first Special Issue of the *Journal
of Proteome Research*^{5,6} provides an accessible resource of tools selected for their applicability and ease of
adoption. Here we describe an effort (Figure 1) where the *ms-utils.org* and *bio.tools* developers have joined
forces, to strive for high quality descriptions and comprehensive coverage of the prevalent proteomics research
software. We describe our curation method and summarise the 754 annotated proteomics tools currently
available in *bio.tools*. The work is assisted by the fledgling Proteomics Community⁷ under the auspices of ELIXIR,
the European Infrastructure for Biological Information. The latter organization will provide ongoing support as
part of its infrastructure to ensure *bio.tools* content will continuously be updated and the service further
improved. The process is community-driven and we invite and welcome the broader proteomics community to
participate.

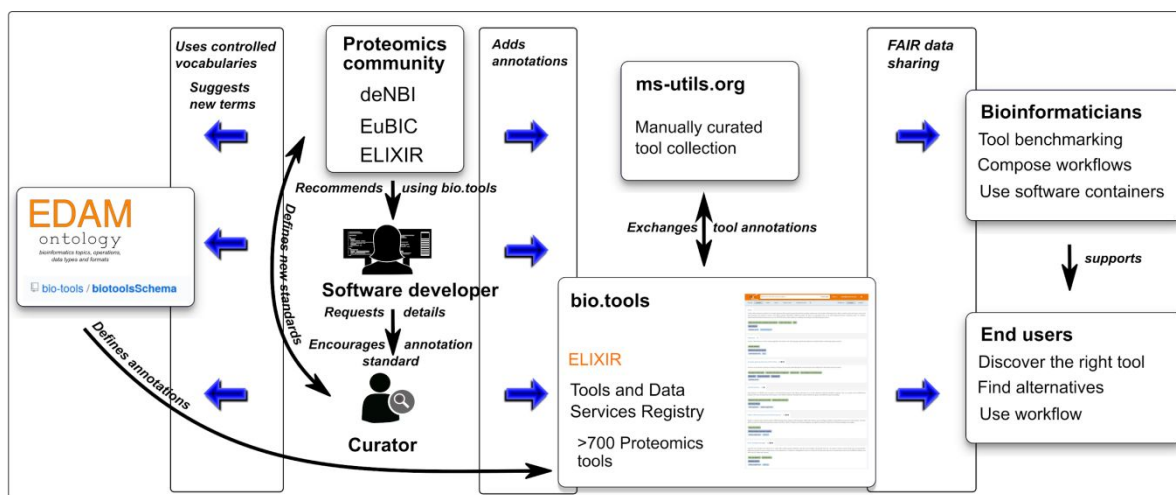
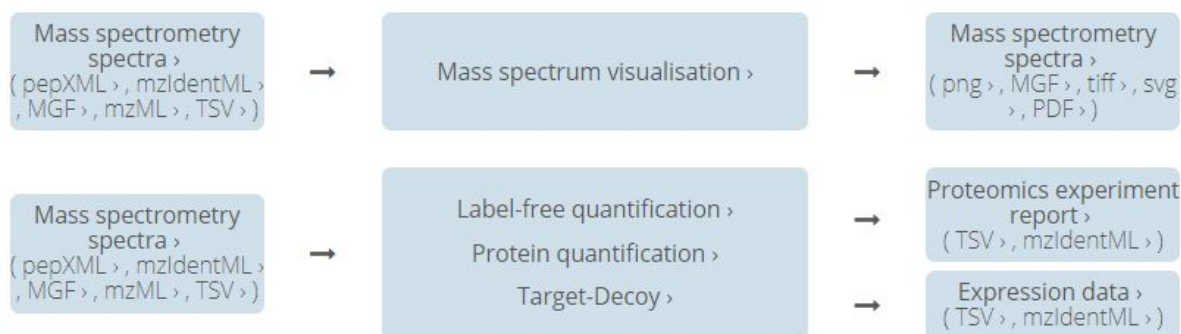


Figure 1. The relationships between the EDAM ontology, *bio.tools*, ms-utils.org, the proteomics community and tool curators.

Methods

Software descriptors

biotoolsSchema (<https://github.com/bio-tools/biotoolschema>) defines a controlled vocabulary and precise syntax for some 50 general software attributes such as name, description, software license and programming language. *bio.tools* uses a simplified model of software, defined by biotoolsSchema, where a tool (Figure 2) can have one or more basic functions or modalities, each expressed in terms of input and output data types, specific operations and supported data formats. The schema was recently extended to allow the annotation of tool relationships, for example single-operation tools being part of a multi-functional software suite.



1
2
3 **Figure 2.** Extract from *bio.tools* Tool Card for PeptideShaker (<https://bio.tools/peptide-shaker>) summarising the
4 tool's basic functionality. The tool has two major modes. The figure includes types of data and supported data
5 formats for input (left of figure) and output (right), and specific operations performed (centre).
6
7
8
9

10
11 When describing a tool, it's necessary to identify the distinct functions and the individual operations, inputs and
12 outputs (including types of data and supported formats) associated with each one. These attributes are covered
13 by the **EDAM ontology** ^{7,8} (<http://edamontology.org>) which provides a precise nomenclature for scientific
14 characterisation of tools, including *topic* (a category within the life sciences, *e.g.* "Proteomics"), *operation* (*e.g.*
15 "Peptide identification"), *data* (*e.g.* "Mass spectrometry spectra") and *format* (*e.g.* "Thermo RAW"). Finally, an
16 emerging information standard (<https://bio-tools.github.io/Tool-Information-Standard/>) used by *bio.tools*
17 defines 5 tiers (from "Sparse" to "Comprehensive") of progressively richer annotation that may be provided for
18 a tool. For example, the "Basic details" tier mandates annotations for tool name, description, homepage, unique
19 ID, tool type, scientific topic, publication and support (contact) information.
20
21
22
23
24
25
26
27
28
29
30

31 Sources of information

32
33 A search in PubMed for publications throughout the last two decades returns 76,689 hits for the search term
34 "proteomics", and 1,097 hits for the query ("new software" OR "new algorithm" OR "new web service" OR "novel
35 software" OR "novel algorithm" OR "novel web service") AND "proteomics" - giving a reasonable estimate of the
36 number of proteomics publications and software tools. Thus, there is a wealth of information that can be mined
37 when producing tool descriptions. When registering tools in *bio.tools*, we used various sources of information,
38 listed below in approximate order of priority when annotating a tool:
39
40
41
42
43
44

- 45 1. existing annotations in ms-utils.org
- 46 2. the primary publication about the tool
- 47 3. the official website of the tool and/or online repository in cases of open source tools
- 48 4. official documentation
- 49 5. other text materials, *e.g.* publications demonstrating tool usage, supplementary files, tutorials *etc.*
- 50 6. source code and testing of software functionality after installation
- 51
- 52
- 53
- 54
- 55
- 56
- 57
- 58
- 59
- 60

1
2
3 Particular attention was made of passages and sentences that are highly descriptive of the tool's functionality
4
5 as these were used later, when identifying relevant EDAM terms.
6
7

8 9 Registration of software information

10 Anyone with a *bio.tools* account can add new entries, or edit existing entries subject to edit rights which are
11 shareable between users. *bio.tools* supports manual registration via the *bio.tools* registration user interface,
12 and registration via API by submission of a file in JSON or XML format, which must conform to biotoolsSchema.
13
14 The ms-utils.org wiki has been built up over more than a decade by several contributors adding new tools as
15 they appear or were discovered in the literature. Defunct but historically important tools are kept in the list,
16 though dead links are removed, distinguishing these tools from those still available. The annotations are all done
17 manually by editing the wiki page directly, requiring basic familiarity with the wiki format.
18
19
20
21
22
23
24
25

26 27 Curation process

28 We sought to optimize the curation of proteomics tools by producing high-quality annotations on a subset of
29 189 tools from the full proteomics collection. The curation of this corpus of proteomics tools involved a multi-
30 step process:
31
32

- 33
34 1. *Compile list of proteomics resources.* We took as a starting point what was available in ms-utils.org and
35 *bio.tools*, augmenting this with lists of tools harvested from Web searches and by searching the
36 scientific literature. The tool name, homepage URL and publication DOI (where available) were
37 recorded: for convenient co-editing a shared spreadsheet was used, which included, for later use,
38 controlled vocabularies (from EDAM and biotoolsSchema) for attributes of interest.
- 39
40 2. *Establish curation priorities.* Our main priority was to provide information corresponding to the
41 "Detailed" tier from the Tool Information Standard for every tool: tool name, description, homepage,
42 tool type, scientific topic, publication (DOI), support information (a link to a helpdesk, issue tracker or
43 mailing list, or an email or URL for a contact person), scientific operation, documentation (URL),
44 operating system, programming language and license. A secondary priority was to annotate the data
45 formats supported by tools.
- 46
47 3. *First curation pass.* We recorded whatever attributes were immediately obvious from superficial
48 inspection of the tool publication and homepage, typically not spending more than 10 minutes per tool.
49
50
51
52
53
54
55
56
57
58
59
60

Where EDAM annotations could not immediately be assigned, passages or sentences describing the tool functionality were recorded, to be used later as a source of key words for assignment or creation of new EDAM topic, operation, data or format terms.

4. *Further curation passes.* We recorded attributes from a deeper inspection of the tool materials, progressively enriching the metadata according to tiers from the Tool Information Standard. Various browsers and utilities (Table 1) were used to make EDAM annotations, noting any new required terms. To provide reliable, consensus annotations, tools were independently annotated and annotations cross-validated, recording any disagreements that warranted further investigation.
5. *Refinement of EDAM annotations and ontology.* We inspected EDAM annotations across the corpus of tools, making adjustments where necessary to ensure consistency and finalise a definite list of new terms. EDAM was extended with the new concepts (terms), synonyms of existing concepts and various other changes to improve the concept hierarchy and usability.
6. *Update of ms-utils.org and bio.tools.* The spreadsheet was transformed to biotoolsSchema-compatible XML, the data validated against the biotoolsSchema and uploaded to *bio.tools* via API. In some cases, where pre-existing entries were not owned by us, the new data was carefully merged into the existing records. Some entries were subsequently polished using the *bio.tools* registration interface. *ms-utils.org* was updated with new information and entries by manual editing. Links were added, from *bio.tools* to *ms-utils.org* for every tool listed in both resources, and from *ms-utils.org* subject categories to EDAM, where possible.

Utility	Description	URL
OLS	Ontology Lookup Service from EMBL-EBI	www.ebi.ac.uk/ols/ontologies/edam
BioPortal	Ontology browser from NCBO	bioportal.bioontology.org/ontologies/EDAM
EDAM Browser	EDAM ontology browser from IFB	ifb-elixirfr.github.io/edam-browser
EDAM Annotator	Emerging utility for annotating tools using EDAM	people.binf.ku.dk/vzn529/eta
EDAM Map	Utility for mapping text (terms, phrases and free text) to EDAM	biit.cs.ut.ee/edammap

Table 1. Browsers and utilities that are useful when annotating software tools

Community engagement

The work built upon lists of tools, preliminary EDAM terms and annotations provided by contributors to *ms-utils.org*, independent editors of *bio.tools* for example from the de.NBI infrastructure (<https://www.denbi.de/>), and by participants at a hackathon of the EuBIC Winter School 2017.⁹ This “long tail” of contribution from the proteomics research and tool user community presented an invaluable starting point for the work. The bulk of subsequent curation work was achieved through a *bio.tools* studentship (<http://biotools.readthedocs.io/en/latest/studentships.html>); a scheme run by the Danish ELIXIR node to support early career stage scholars to contribute to *bio.tools* and gain experience with the ELIXIR infrastructure. Following broad ranging discussions, we distilled two primary goals. First, the entire tools corpus should be collated and described in at least a basic level of detail. Second, for a subset of high priority tools, the annotations should enable the construction of data analysis workflows, which implies a greater level of detail including supported data formats. The work is in context, and but a small part of, a broader sweep of action in scope of the emerging ELIXIR Proteomics Community (<https://www.elixir-europe.org/communities/proteomics>) including data integration with other omics data, automatized data processing and standardization including the description of software in *bio.tools*.

Results and discussion

bio.tools includes a total of 754 tools with the EDAM Topic annotation of “Proteomics” (EDAM:topic_0121), “Proteomics experiment” (EDAM:topic_3520), or synonyms of these terms. These tools are, for convenience, associated with a *bio.tools* subdomain available for browsing at <https://proteomics.bio.tools>. The tool descriptions include a total of 30,187 annotations, of which 6,350 are EDAM annotations. The metadata richness of these tools (Table 2) according to the Tool Information Standard shows that 81% of the collection have “Detailed” (or richer) annotation. Of this corpus, this work contributed 189 new tool registrations in *bio.tools*, of which 167 are annotated to at least “Detailed” level. Of these 189 tools, 93% (176) have an input or output defined, with a total of 562 EDAM Data annotations and 963 EDAM Format annotations. The corpus of 754 tools includes command-line tools (276, 31%), web applications (224, 25%), desktop applications (170, 19%), libraries (97, 11%) and a long tail of other tool types defined in *biotoolsSchema* (note that a single tool can be annotated

as of more than one basic type). The ms-utils.org wiki has been structured according to the EDAM ontology, linking most tool categories to EDAM operations. As a single-page wiki, a user can conveniently search the list for keywords appearing in the subheaders or tool descriptions, or look through the tools in a particular category.

Tier of Tool Information Standard	# tools
Sparse	34
Basic details	105
Detailed	589
Highly detailed	5
Comprehensive	21

Table 2. Metadata richness of curated proteomics tools. The number of *bio.tools* entries compliant to different tiers in the Tool Information Standard is shown.

At the outset, EDAM only contained a few terms for proteomics data analysis; many additions and changes were needed, and these were made progressively over multiple EDAM releases. Changes included adding new concepts where these were missing, ensuring the preferred label reflected the vernacular, and adding common synonyms of this term. The conceptual hierarchy (concept subsumption relationships) was also extensively revised, to make navigation of EDAM and term picking easier in ontology browsers. As an illustration, Figure 3 shows spectral analysis operations (EDAM:operation_3214). A particular effort was focussed on the curation of proteomics data formats as these have high practical value in applications such as workflow composition.¹⁰ 30 new mass spectrometry data formats were added, including proprietary formats created by companies to support specific machines and specific commercial software, such as Thermo RAW format (EDAM:format_3712) supported by Xcalibur, and open source data formats such as mzXML (EDAM:format_3654). EDAM overlaps with a few of the concepts within the PSI-MS ontology¹¹ which is designed to describe a mass spectrometry experiment. Collaborative efforts between the maintainers of both ontologies will be intensified to ensure the interoperability of these ontologies, for example by cross-referencing equivalent concepts.

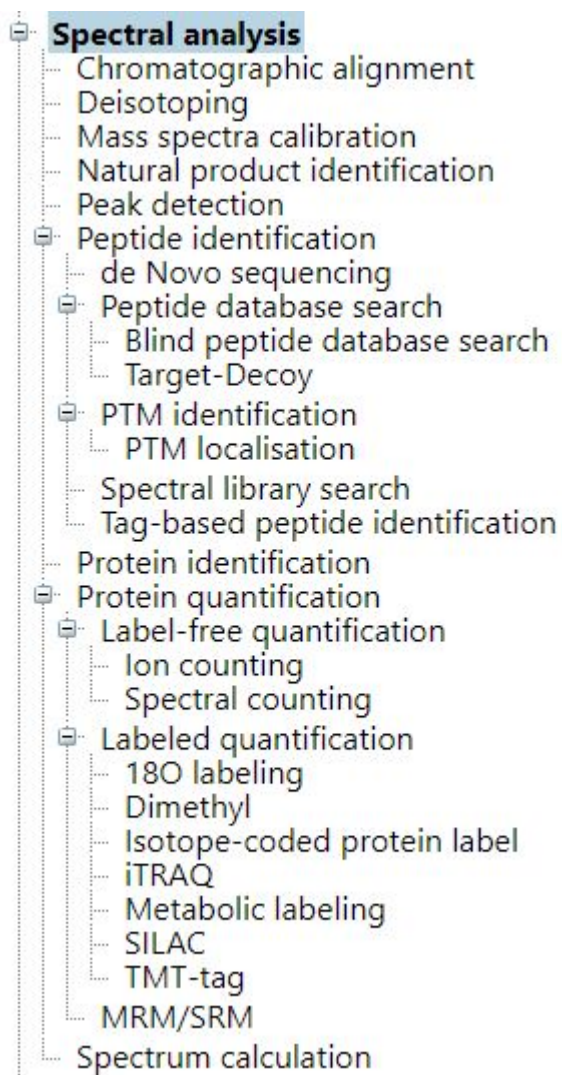


Figure 3. Extract from the EDAM ontology showing operations for spectral analysis (EDAM:operation_3214).

Early community engagement helped to prioritise the curation effort and focus on annotations of high practical value to tool interoperability, such as input and output data formats. The emerging Tool Information Standard was helpful to structure the curation effort on a technical level. It was surprising that in many cases, crucial usage information such as programming language, software license and terms of use, were not easy to find. Basic input and output data types and formats were also often not stated in an explicit and clear manner. As a last resort, supported data formats were identified through inspection of the source code or by testing the tools at the command-line. Such challenges in finding information present a barrier to end-users' efficient discovery and use of tools, and underlines the need for resources such as *bio.tools* and *ms-utils.org*. To provide consistent search and discovery, these resources benefit greatly from the use of controlled vocabularies defined within *biotoolsSchema* and the EDAM ontology. The numerous EDAM concepts about proteomics are well described

1
2
3 and documented, and able to represent almost all the functionality found in any proteomics tool, which in
4 combination with ontology browsers such as OLS, BioPortal and EDAM Browser, greatly facilitates the manual
5 annotation of tools.
6
7

8
9 The proteomics community benefits from consistent and detailed tool annotations in various ways. *bio.tools*
10 allows researchers to query and find appropriate tools with algorithms that fulfill a certain task, with correctly
11 described input and output file compatibilities, and with valuable references to documentation, tutorials and
12 training. Annotations can furthermore be used to compose workflows comprising multiple operations as
13 previously shown¹⁰. Collections of single tools and workflows executing identical operation(s) for the analysis of
14 proteomics data can be created and benchmarked by comparing the results using ground-truth data sets. We
15 envision that future efforts will result in a feedback loop where valuable information about tool performance
16 will continuously be updated by monitoring their usage and therefore enhance the annotation, presentation
17 and discovery of optimally performing tools.
18
19
20
21
22
23
24
25
26
27
28

29 Conclusion

30 Organized catalogues of expert-annotated software, like Pedro's list, that describe software using a standardized
31 vocabulary on one platform, facilitate what can be a daunting search for tools that suit a particular scientific or
32 technical purpose. Providing better and more permanent tool findability should automatically lead to longer
33 half-lives, assuming the software to be functional and maintained. We have summarised a successful curation
34 effort that has enriched *bio.tools*, ms-utils.org and the EDAM ontology, and rendered a significant proportion of
35 all proteomics analysis software more findable, accessible, interoperable and reusable, *i.e.* more FAIR.^{10,12} While
36 detailed annotation of fine-grained details such as data formats are costly, the effort is warranted where it
37 supports valuable scientific applications such as tool interoperability and workflow composition. Detailed tool
38 annotations including input/output data and format will open the door for identifying novel workflows of
39 compatible tools and for implementing alternative workflow components to benchmark their performance.
40 Such efforts can leverage software containers, for example those being registered in Biocontainers¹³ in
41 collaboration with *bio.tools*, with the hope to greatly simplify deployment of full data analysis pipelines on local
42 high-performance machines or on the cloud. BioContainers is providing an infrastructure to create, deploy and
43 maintain software containers using Conda and Docker technologies. *bio.tools* and BioContainers are coordinated
44 under the ELIXIR Tools Platform (<https://elixir-europe.org/platforms/tools>). As an example, systematic efforts
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 are ongoing to ensure that all tools which have been containerised are registered in bio.tools, and conversely,
4
5 bio.tools is being used to provide metadata for exposure in the BioContainers registry.
6

7 The difficulties we encountered in finding sometimes even basic information about tools point to a pressing
8
9 requirement for the promotion of better standards of information for life science software generally. There is a
10
11 need for upstream provision of richer and more consistent software metadata that can be conveniently reused
12
13 by efforts such as *bio.tools*. On the cataloguing side, there is no free lunch; high quality content requires an
14
15 investment of time and manual effort. The curation effort would benefit greatly from more automated ways to
16
17 harvest trivial annotations such as software license, for example by text mining the literature, and by a closer
18
19 integration of the ontology construction and tool registration processes, for example harvesting missing terms
20
21 and synonyms at tool registration time. The quality of annotations would benefit from more powerful and
22
23 convenient means for term selection, which itself is a significant challenge given the sizeable vocabularies that
24
25 the typical end-user, pressed for time, may be unfamiliar with.
26

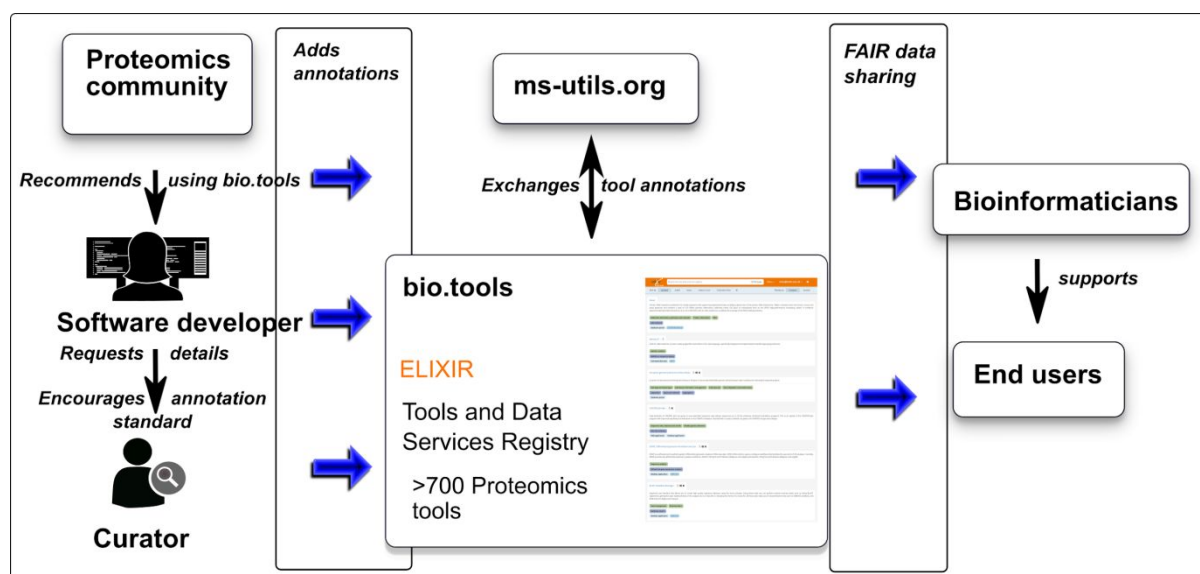
27 We approximate that of the potential volume of 1,097 tools existing for proteomics (see *Sources of information*),
28
29 *bio.tools* captures 68% (754) with 56% (615) annotated to “Detailed” level or better. Thus, there is more curation
30
31 work to do, and we can expect many new resources to appear in the future, which will also reflect new analytical
32
33 methods, types of data and data formats. We recently contacted the developers of tools in the proteomics
34
35 corpus for which contact details were available in *bio.tools*, and hope this will lead to community adoption and
36
37 maintenance of the corpus in the long-term. We hope the efforts described here for proteomics will stimulate
38
39 similar efforts in other domains, which are also witness to a large rise in the volume of new tools and data
40
41 resources.¹⁴ The anchoring of *bio.tools* within the ELIXIR infrastructure will ensure *bio.tools* is maintained in the
42
43 long term and we encourage the whole proteomics community to collaborate with us on further improving the
44
45 corpus of tool descriptions.
46
47
48
49
50

51 Acknowledgements

52 We acknowledge with gratitude the support of our funders: The Danish Ministry of Higher Education and Science;
53
54 ELIXIR-EXCELERATE under the European Union's Horizon 2020 research and innovation programme (grant
55
56 agreement number 676559).
57
58
59
60

References

- (1) Matthiesen, R. Useful Mass Spectrometry Programs Freely Available on the Internet. *Methods Mol. Biol.* **2007**, *367*, 303–305.
- (2) Deutsch, E. W.; Lam, H.; Aebersold, R. Data Analysis and Bioinformatics Tools for Tandem Mass Spectrometry in Proteomics. *Physiol. Genomics* **2008**, *33* (1), 18–25.
- (3) Artimo, P.; Jonnalagedda, M.; Arnold, K.; Baratin, D.; Csardi, G.; de Castro, E.; Duvaud, S.; Flegel, V.; Fortier, A.; Gasteiger, E.; et al. ExpASY: SIB Bioinformatics Resource Portal. *Nucleic Acids Res.* **2012**, *40* (Web Server issue), W597–W603.
- (4) Yachdav, G.; Kloppmann, E.; Kajan, L.; Hecht, M.; Goldberg, T.; Hamp, T.; Hönigschmid, P.; Schafferhans, A.; Roos, M.; Bernhofer, M.; et al. PredictProtein--an Open Resource for Online Prediction of Protein Structural and Functional Features. *Nucleic Acids Res.* **2014**, *42* (Web Server issue), W337–W343.
- (5) Ison, J.; Rapacki, K.; Ménager, H.; Kalaš, M.; Rydza, E.; Chmura, P.; Anthon, C.; Beard, N.; Berka, K.; Bolser, D.; et al. Tools and Data Services Registry: A Community Effort to Document Bioinformatics Resources. *Nucleic Acids Res.* **2016**, *44* (D1), D38–D47.
- (6) Weintraub, S. T.; Hoopmann, M. R.; Palmblad, M. Special Issue on Software Tools and Resources: Acknowledging the Toolmakers of Science. *J. Proteome Res.* **2019**, *18* (2), 575.
- (7) Vizcaíno, J. A.; Walzer, M.; Jiménez, R. C.; Bittremieux, W.; Bouyssié, D.; Carapito, C.; Corrales, F.; Ferro, M.; Heck, A. J. R.; Horvatovich, P.; et al. A Community Proposal to Integrate Proteomics Activities in ELIXIR. *F1000Res.* **2017**, *6*.
<https://doi.org/10.12688/f1000research.11751.1>.
- (8) 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**, *29* (10), 1325–1332.
- (9) Willems, S.; Bouyssié, D.; David, M.; Locard-Paulet, M.; Mechtler, K.; Schwämmle, V.; Uszkoreit, J.; Vaudel, M.; Dorfer, V. Proceedings of the EuBIC Winter School 2017. *J. Proteomics* **2017**, *161*, 78–80.
- (10) Palmblad, M.; Lamprecht, A.-L.; Ison, J.; Schwämmle, V. Automated Workflow Composition in Mass Spectrometry-Based Proteomics. *Bioinformatics* **2019**, *35* (4), 656–664.
- (11) Montecchi-Palazzi, L.; Beavis, R.; Binz, P.-A.; Chalkley, R. J.; Cottrell, J.; Creasy, D.; Shofstahl, J.; Seymour, S. L.; Garavelli, J. S. The PSI-MOD Community Standard for Representation of Protein Modification Data. *Nat. Biotechnol.* **2008**, *26* (8), 864–866.
- (12) Wise, J.; de Barron, A. G.; Splendiani, A.; Balali-Mood, B.; Vasant, D.; Little, E.; Mellino, G.; Harrow, I.; Smith, I.; Taubert, J.; et al. Implementation and Relevance of FAIR Data Principles in Biopharmaceutical R&D. *Drug Discov. Today* **2019**.
<https://doi.org/10.1016/j.drudis.2019.01.008>.
- (13) da Veiga Leprevost, F.; Grüning, B. A.; Alves Aflitos, S.; Röst, H. L.; Uszkoreit, J.; Barsnes, H.; Vaudel, M.; Moreno, P.; Gatto, L.; Weber, J.; et al. BioContainers: An Open-Source and Community-Driven Framework for Software Standardization. *Bioinformatics* **2017**, *33* (16), 2580–2582.
- (14) Editorial: The 16th Annual Nucleic Acids Research Web Server Issue 2018. *Nucleic Acids Res.* **2018**, *46* (W1), W1–W4.



For TOC only