Sense makers: How standards are enabling data reuse in the life sciences

March 20 2024, by Oana Stroe and Dorota Badowska

Data standards transform unstructured information into well-organised databases, like turning pages into books and systematically cataloguing them in libraries so they're easily searchable by keywords. Credit: Karen Arnott/EMBL-EBI

A bit like sorting a messy pile of clothes into a neatly arranged closet, minimal information standards transform unstructured data from journal articles into structured databases. This enables researchers to "mine" across multiple datasets, reuse data, and gain new insights.

Minimum information standards are guidelines and formats for reporting scientific data generated by high-throughput methods, such as genome sequencing. They ensure all datasets are structured in the same way, making them easy to find, verify, and analyze by researchers worldwide. Standards also provide context for datasets—for example, when, where, and how the data were generated, or what species they describe.

Public molecular databases, such as the ones managed by EMBL, ensure that data generated once can be reused again and again to ask new research questions, rather than information being 'hidden away' on the servers of individual laboratories.

This is an efficient approach to capturing data generated by publicly funded science, making them easy to access. In a way, it's similar to turning paper piles into books, and systematically cataloging them at the public library where anyone can access them. Just like libraries play a role in knowledge sharing, public data resources and minimal information standards enable researchers to access and use data generated outside their own labs.

What makes a good minimal information standard?

"You have to strike a balance between what is possible and what is practical," explained Alvis Brazma, Senior Team Leader at EMBL-EBI, and co-author of some of the first minimal information standards published.

"The people generating the data will probably say the standard requires too much information, and the people analyzing the data will say it's not enough. So they have to meet somewhere in the middle.

"But importantly, you need to try and understand what is needed for reanalysis now and try to predict what might be needed in the future. It's not an easy task! In my experience, it's best to start with a minimum, and keep adding to it once the community is on board," says Brazma.

Minimum information standards typically have two parts. First, there is a set of reporting requirements—typically presented as a table or a checklist. Second, there is an agreed data format. Information about an experiment needs to be converted into the appropriate data format for it to be submitted to the relevant database.

Driving the development of new methods

Standardized data are key to developing new methods. Every bioinformatic research method, whether it be to predict new disordered proteins, to interpret the effect of protein modifications, or to analyze bioimaging data, critically hinges on the availability and unambiguity of the data used to train the methods.

"Minimum information standards provide context that stitches together scientific outputs into the unknowable fabric of 'big data,'" said Cy Jeffries, Staff Scientist at EMBL Hamburg and the curator of the Small Angle Scattering Biological Data Bank (SASBSB). "It means that results from different scientific disciplines can be linked together, reused, and openly shared to find new patterns that we have not thought of yet, but future AI might."

"In the age of AI, minimal information standards and standardized databases are more important than ever because they open up the data to machine learning and AI algorithms," explained Jo McEntyre, Deputy Director of EMBL-EBI. "Take AlphaFold, for example—Google DeepMind's AI system that can accurately predict protein structures. The development of AlphaFold simply wouldn't have been possible without the decades-worth of organized, annotated public protein structure and function data in the Protein Data Bank in Europe, and UniProt. As with many research methods, what you get out is only as good as the data you put in."

Many flavors of standards

EMBL scientists and colleagues have contributed to the development of many minimal information standards for different data types. The standards usually follow developments in technology and improved accessibility, which result in an increase in the volume of data produced.

Below are a few examples of minimal information standards that are now widely being used in the scientific community:

- MIAME—Minimum information about a microarray experiment : Dating back to 2001, MIAME is one of the first data standards. Microarray technology has been used for a variety of purposes in research and clinical studies, including measuring gene expression and detecting specific DNA sequences.
- MIAPE—Minimum information about a proteomics experiment: To encourage the standardized collection and dissemination of proteomics data, the Human Proteome Organization's Proteomics Standards Initiative developed guidance modules for reporting the use of techniques such as gel electrophoresis and mass spectrometry.
- REMBI—Recommended Metadata for Biological Images: Developed in 2021 to enable the reuse of microscopy data in biology, which is particularly important as technological developments and improved accessibility to bioimaging are resulting in increased microscopy data.
- MIADE—Minimum Information About Disorder Experiments: Published in 2023 to support research on proteins that

continuously change their shape. About one-third of all known proteins are considered to be disordered.

"Community consultations and buy-in are key for the success of data standards," explained Sandra Orchard, Protein Function Content Team Leader at EMBL-EBI. "The standard has to be functional, so it's adopted worldwide and ideally supported by publishers and reviewers. And of course, the generation and public sharing of research data needs to be recognized as a valuable contribution to science, along with other outputs such as publications, the development of software tools, and knowledge sharing."

Data standards are helping to capitalize on the vast amount of data being generated in the life sciences. Although submitting research results to public data resources and abiding by minimal information standards can be time-consuming and onerous, it's an important step in the research process and can help data remain useful long after a paper has been published.

After all, you might not enjoy tidying up your closet, but it feels good once you've done it.

Provided by European Molecular Biology Laboratory

Citation: Sense makers: How standards are enabling data reuse in the life sciences (2024, March 20) retrieved 27 April 2024 from https://phys.org/news/2024-03-makers-standards-enabling-reuse-life.html