

BFFO: An Open, Machine-Actionable Knowledge Base for Bioinformatics File Formats

Nishad Thalhath¹  , Mitsuharu Nagamori²  , Deepa Kasaragod³  , and
Takeya Kasukawa¹  

¹RIKEN Center for Integrative Medical Sciences, Yokohama, Japan, ²Faculty of Library, Information and Media Studies, University of Tsukuba, Tsukuba, Japan, ³Hiroshima University, Higashi-Hiroshima, Japan

Abstract

BFFO (Bio File Formats and Ontology, <https://bffo.org>) is an open, curated knowledge base for bioinformatics file formats, currently in active development. It puts the format itself at the centre, recording what each format is, how it relates to other formats, which tools work with it, and where to find its specification. That information is published for both humans and software agents. This poster reports work in progress: the gap the project aims to fill, the design intentions, the state of the data and ontology drafts, and the road ahead.

1. The Gap

A typical bioinformatics pipeline reads and writes a dozen file formats in a single run, and the field as a whole spans sequence work, imaging, single-cell, structural biology, and proteomics. The formats are not independent (BAM is the BGZF-compressed binary representation of SAM, BAI and CSI both index BAM, CRAM is an alternative to BAM), but this network is rarely recorded in any one place a person or a program can consult.

The existing resources each cover part of the picture. EDAM [1] has good ontological structure and documents over 600 formats, but it does not link directly to tools, and its format relationships stop at the prose: its BAM entry, for instance, calls BAM a compressed version of SAM yet records no link between them. The bio.tools registry [2], [3] is canonical, but the tool is the first-class entity, not the format. Wikidata [4] has scattered entries of uneven depth. Outside the life sciences, PRONOM and the Library of Congress format descriptions model typed relationships, but between them reach barely a dozen bioinformatics formats. Galaxy records conversions, but only those its own runtime performs. Specifications themselves live as PDFs, HTML pages, and GitHub READMEs. The same gap is felt by AI coding assistants now generating pipelines: an agent can guess that BAM and CRAM are related, but cannot reliably check which index goes with which container, or what the current specification URL is.

2. The Project

BFFO is a curated knowledge base whose first-class entity is the file format. Each entry brings together identity (name, abbreviation, identifier), classification (category, research domain, biological data type, pipeline role), technical character (encoding, compression, coordinate system, indexability), relationships to other formats (binary encoding, indexing, containment, alternation, conversion), associated tools, version history, and a short encyclopaedic overview written for humans. Each entry is also a

Poster

Available Aug 01, 2026

Correspondence to
Nishad Thalhath
nishad.thalhath@riken.jp



Copyright © 2026 Thalhath *et al.*
This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license, which enables reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator.

node in a network: format-to-format relationships are typed, and each type declares its inverse or is marked symmetric, so a reader (or a program) can move from FASTQ to BAM to CRAM and understand what kind of step each link represents.

The accompanying ontology is a small application profile that reuses Dublin Core, schema.org, SKOS, DCAT, ADMS, and FOAF wherever an existing term fits, and only mints new terms in the `bffo:` namespace where the existing vocabularies do not cover the concept: the classification axes, the technical properties, and the format-to-format relationship vocabulary. Controlled vocabularies are SKOS concept schemes; validation constraints are SHACL.

The work is published openly: the ontology is CC0 and the curated data CC-BY, and the identifiers follow stable URI patterns (`bffo.org/format/BAM`, `bffo.org/ontology/`). Entries are served as web pages today, with the machine-facing channels still to come.

3. Aim and Direction

BFFO is not meant to replace EDAM or bio.tools but to fill the format-shaped hole between them, in a way linked-data tooling can use directly. Three aims shape the work:

- *A reference for people*, organised the way a developer or curator actually thinks about formats, starting from the format rather than from a tool or an ontology branch.
- *Structured knowledge for AI agents* writing or auditing bioinformatics pipelines: stable identifiers, typed relationships, explicit tool roles, verified specification links.
- *A FAIR, citable, open resource*, with mappings to EDAM, Wikidata, and FAIRsharing where verified mappings are available, and deliberately empty fields where they are not, since fabricated identifiers in a reference do more harm than missing ones.

4. Status

A first curation round has produced 373 format entries and 865 typed relationships between them; 188 entries carry a verified EDAM mapping. Coverage is densest in genomics, proteomics, sequencing, and imaging, thinner in structural biology and single-cell work, with a long tail of niche and emerging formats. The ontology is a 0.1 editor's draft. The data website (<https://bffo.org>) and the ontology documentation (<https://ontology.bffo.org>) are publicly available as previews. What remains is broader coverage, a 1.0 ontology release, the RDF, JSON, and agent-facing interfaces, a contribution model, and registration with linked-data registries.

References

- [1] J. Ison *et al.*, "EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats," *Bioinformatics*, vol. 29, no. 10, pp. 1325–1332, 2013, doi: [10.1093/bioinformatics/btt113](https://doi.org/10.1093/bioinformatics/btt113).
- [2] J. Ison, "The bio.tools registry of software tools and data resources for the life sciences," *Genome Biology*, vol. 20, no. 1, p. 164, 2019, doi: [10.1186/s13059-019-1772-6](https://doi.org/10.1186/s13059-019-1772-6).

- [3] A. I. S. Mendes, “bio.tools: an expanded web service for research software in the life sciences,” *Nucleic Acids Research*, vol. 54, no. W1, pp. W4–W12, 2026, doi: [10.1093/nar/gkag420](https://doi.org/10.1093/nar/gkag420).
- [4] D. Vrandečić and M. Krötzsch, “Wikidata: a free collaborative knowledgebase,” *Communications of the ACM*, vol. 57, no. 10, pp. 78–85, 2014, doi: [10.1145/2629489](https://doi.org/10.1145/2629489).

UNCORRECTED
PROOF