Improving FAIR Access to Open DNA Collections

Felipe Xavier Buson, Yan Kay Ho, Jenny Molloy*
Department of Chemical Engineering and Biotechnology, University of Cambridge
*Corresponding author: **jcm80@cam.ac.uk**

OPEN BIOECONOMY LAB

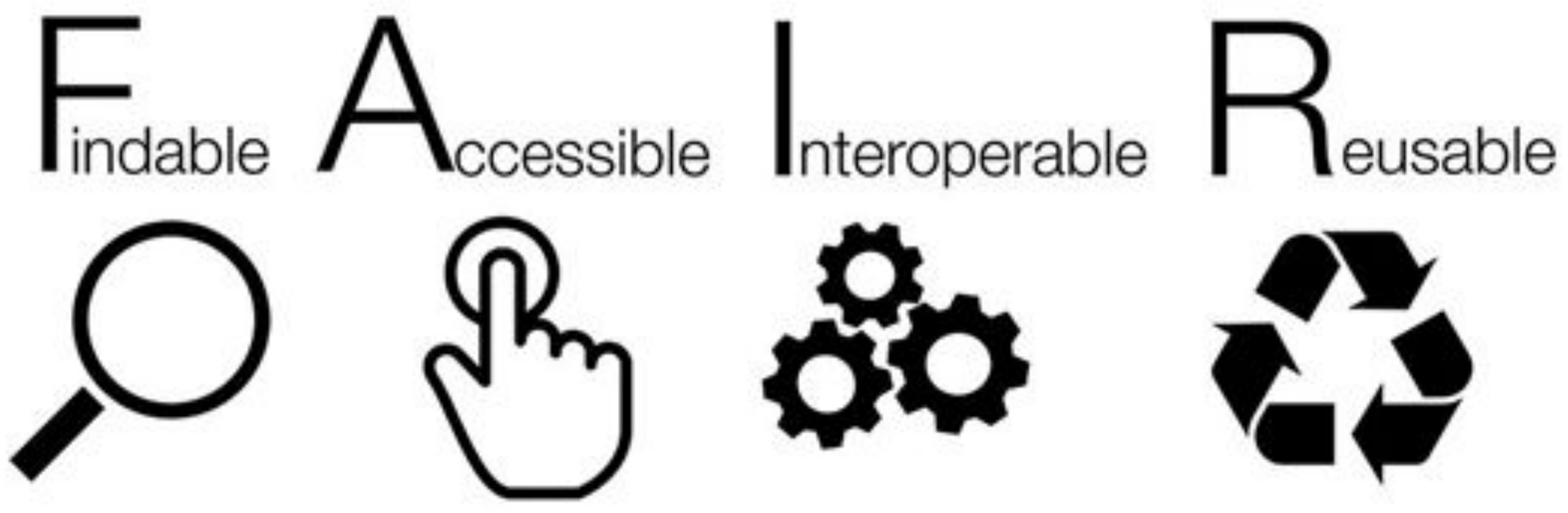
UNIVERSITY OF CAMBRIDGE

Abstract

DNA constructs are used ubiquitously in synthetic biology to engineer a host of solutions for addressing global challenges, but the **current platforms to find and share DNA sequences vary both in quality and openness**. This makes it difficult to determine whether parts are reliable and fit-for-purpose. Building on the joint OpenPlant initiative[1], coupled with the Open DNA Collections created by the Open Bioeconomy Lab[2] and other efforts such as the FreeGenes Project[3], as currently stewarded by Reclone.org[4], **we are developing a platform to host DNA constructs that better adhere to FAIR principles** (Findable, Accessible, Interoperable, and Reusable)[5].

To make our collections **F**indable and **I**nteroperable with existing software, we are making a curated pilot collection using the suite of open SBOL[6] tools developed by the community, such as SynBioHub[7], PySBOL[8], Excel2SBOL[9], and SeqImprove[10]. To make our platform (**Re**)Usable and **A**ccessible to researchers, we attach the high-quality data generated by this workflow to expertise in web design and user experience[11] to ensure that our collections are delivered in a user-friendly package, bespoke to the needs of synthetic biology users.

With our pilot collection of well-curated parts as a model, and a user-friendly, standardised workflow for researchers to use, improve, and add to the collections, we hope to welcome and encourage further adoption of this system to produce a better place for the free sharing and collaboration of this FAIR DNA Data Repository.

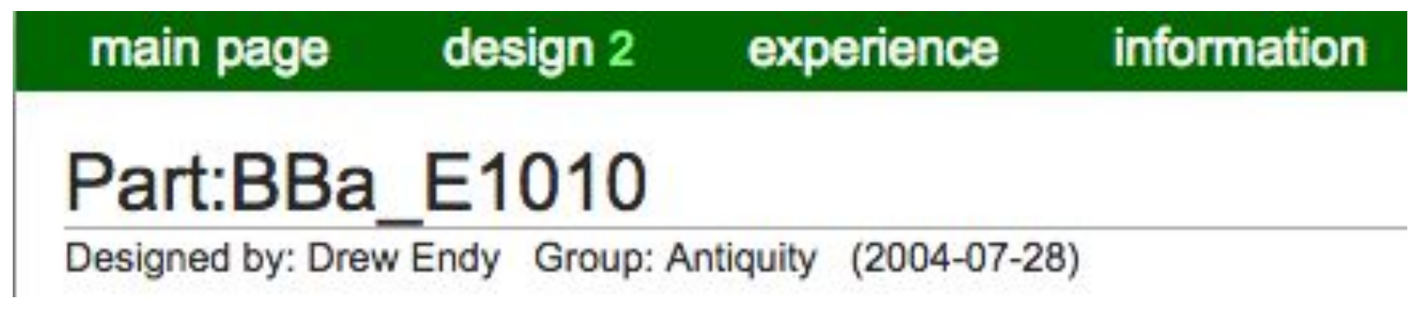


What's wrong?

Cataloguing of parts for synthetic biology has been carried out mostly through online databases and repositories such as the iGEM Registry of Standard Biological Parts[12], SynBioHub[7], JBEI's Inventory of Composable Elements (ICE)[13], and Addgene[14]. These have emerged from initiatives with different goals for searchability, quality standards, data types, and scope of parts.

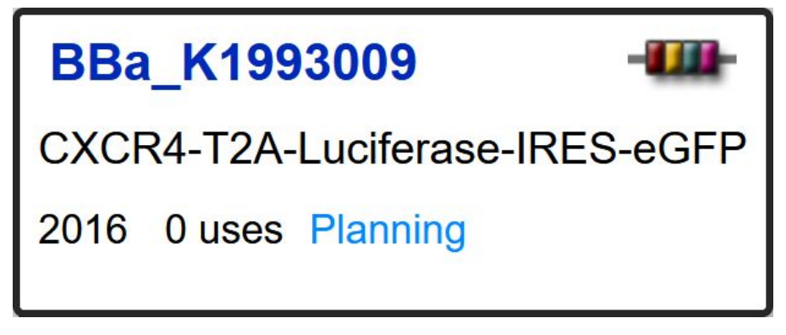
The most relevant repositories exhibit limitations in searchability, due to an overwhelming number of components and **no adoption of standards for part metadata**. The truly useful information is diluted among the rest, and it takes more time and specific knowledge/experience to identify the best components. Variability on the quality of sequence data itself is a hurdle for using these repositories, since broken or poorly annotated files may lead to problems in their use in other software, and potential misinterpretation by users.

Using IDs as part names

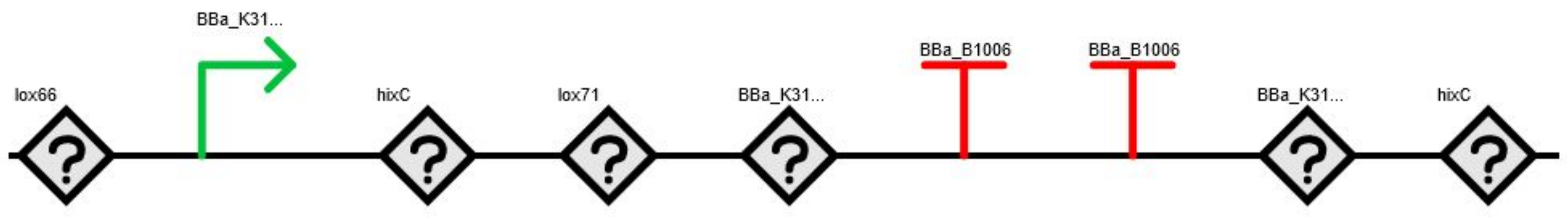

main page | design 2 | experience | information

Part:BBa_E1010
Designed by: Drew Endy Group: Antiquity (2004-07-28)

Name as list of components


BBa_K1993009
CXCR4-T2A-Luciferase-IRES-eGFP
2016 0 uses Planning

Broken/poorly-annotated files

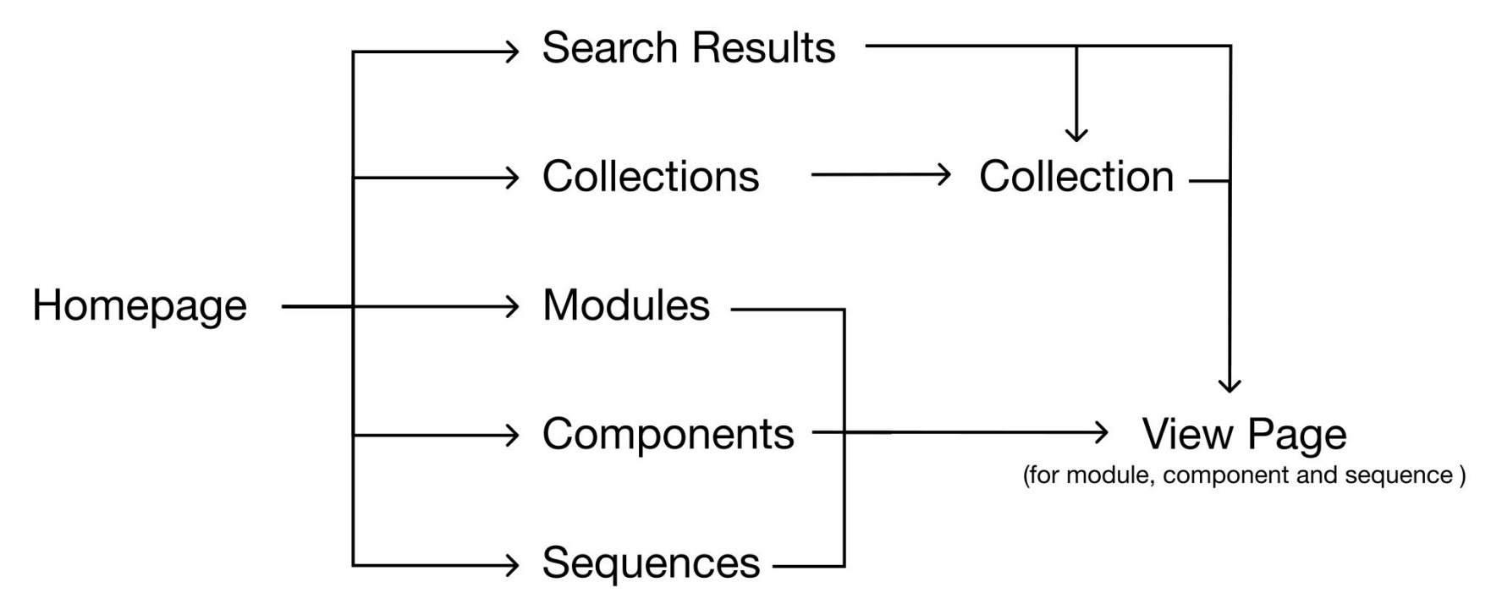


Additionally, the platforms don't necessarily **prioritize user experience** for browsing through their components. Important information about parts is often omitted or poorly displayed, and it's hard to filter out relevant parts. In some cases, the repository's data structure is presented instead of a human-friendly approach, making it so only experienced users are able to navigate through the entries efficiently.

Enforcing data quality

Our efforts to improve searchability and interoperability in DNA part repositories start from **using the SBOL[6] standard and adopting a semi-automated workflow** to generate the repository objects. Simply transferring our sequences from working Genbank files to SBOL and hosting them on SynBioHub[7] is already an improvement, but human intervention is necessary to ensure the correct data structure and annotation. We are currently using Excel2SBOL[9] and custom Jupyter Notebooks to facilitate this transition by identifying common annotations in a collection, and prompting users to pick which annotations should be translated to the final SBOL objects.

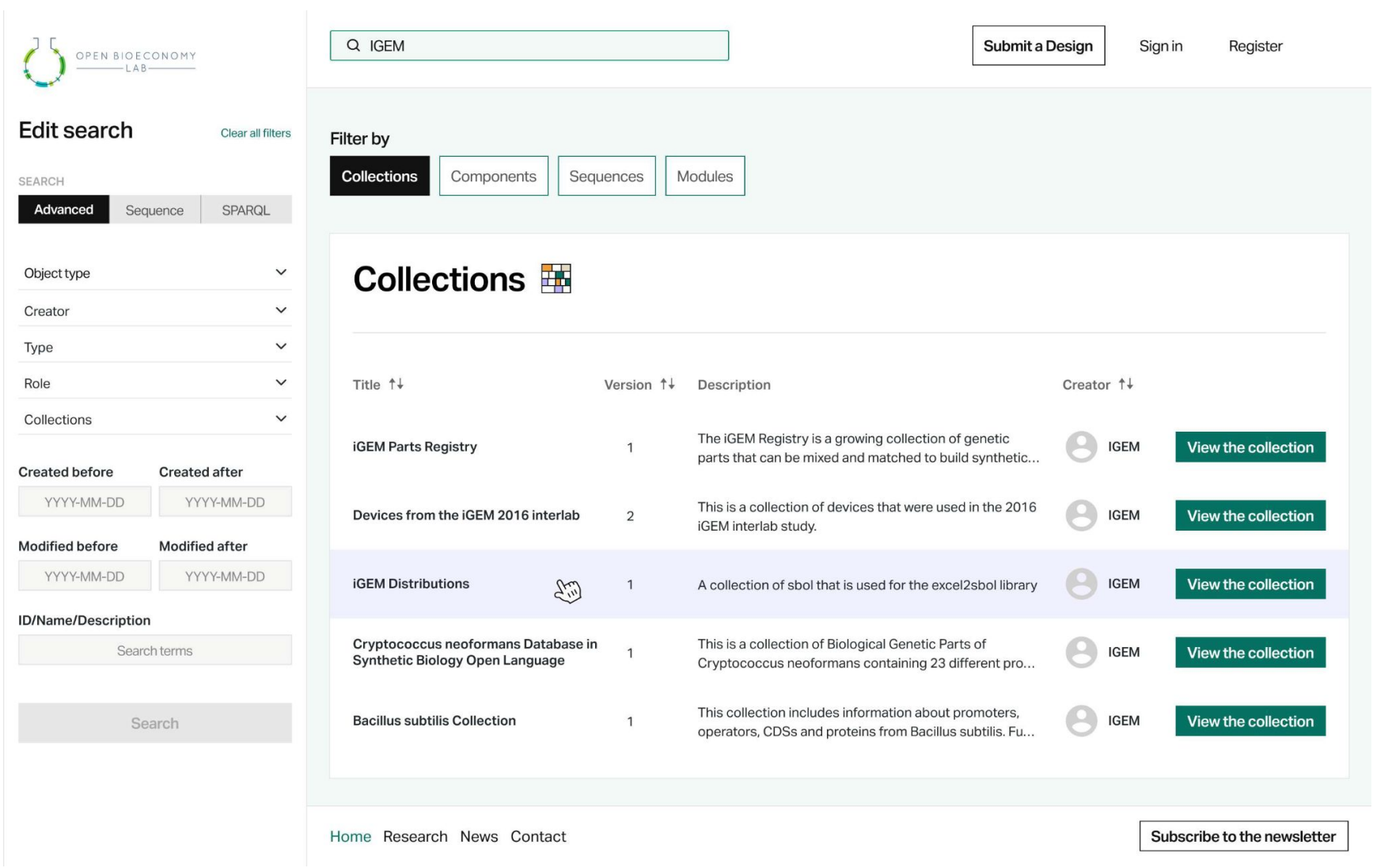
Another key change is searching and browsing for parts with a **collection-first approach**. Collections are useful packages for parts where users can better identify their purpose and how well the whole collection is characterized.



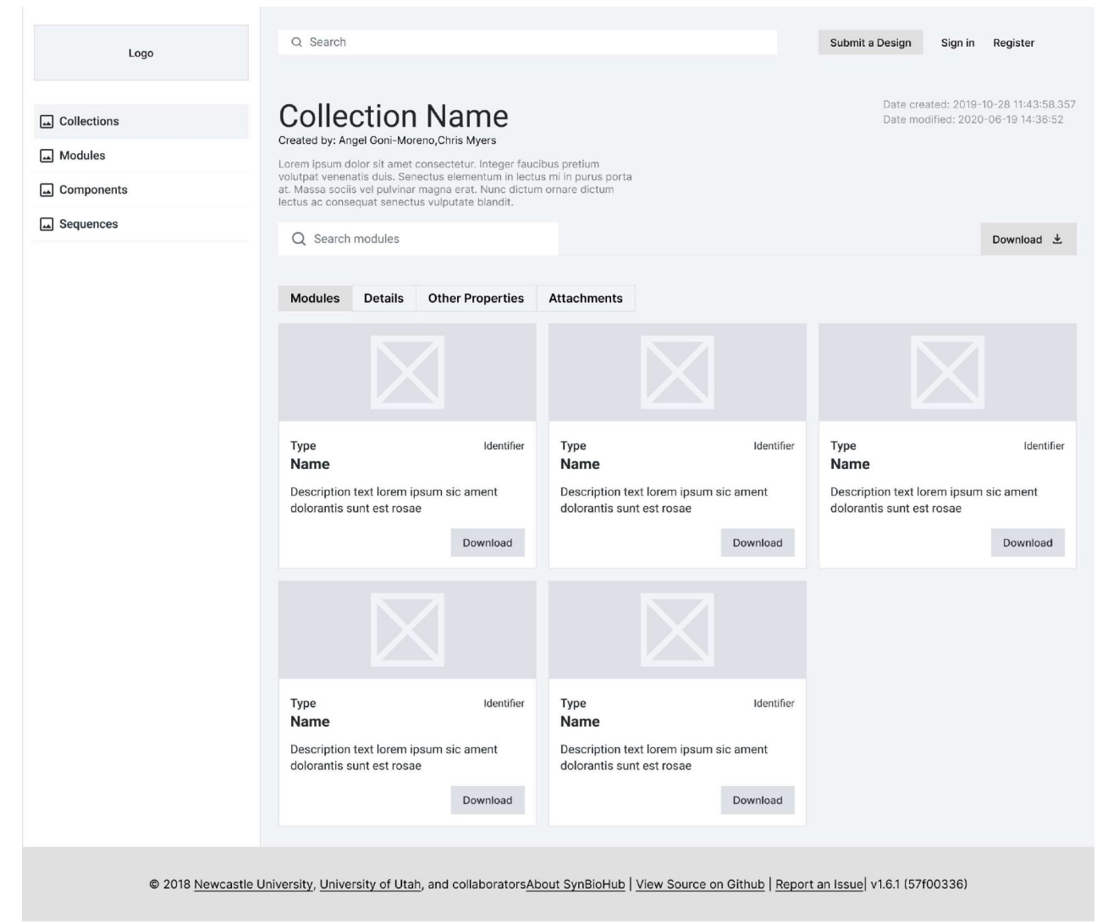
Enhancing visualization / UX

To improve the user experience, we're collaborating with the data visualization experts at **Accutat.it**

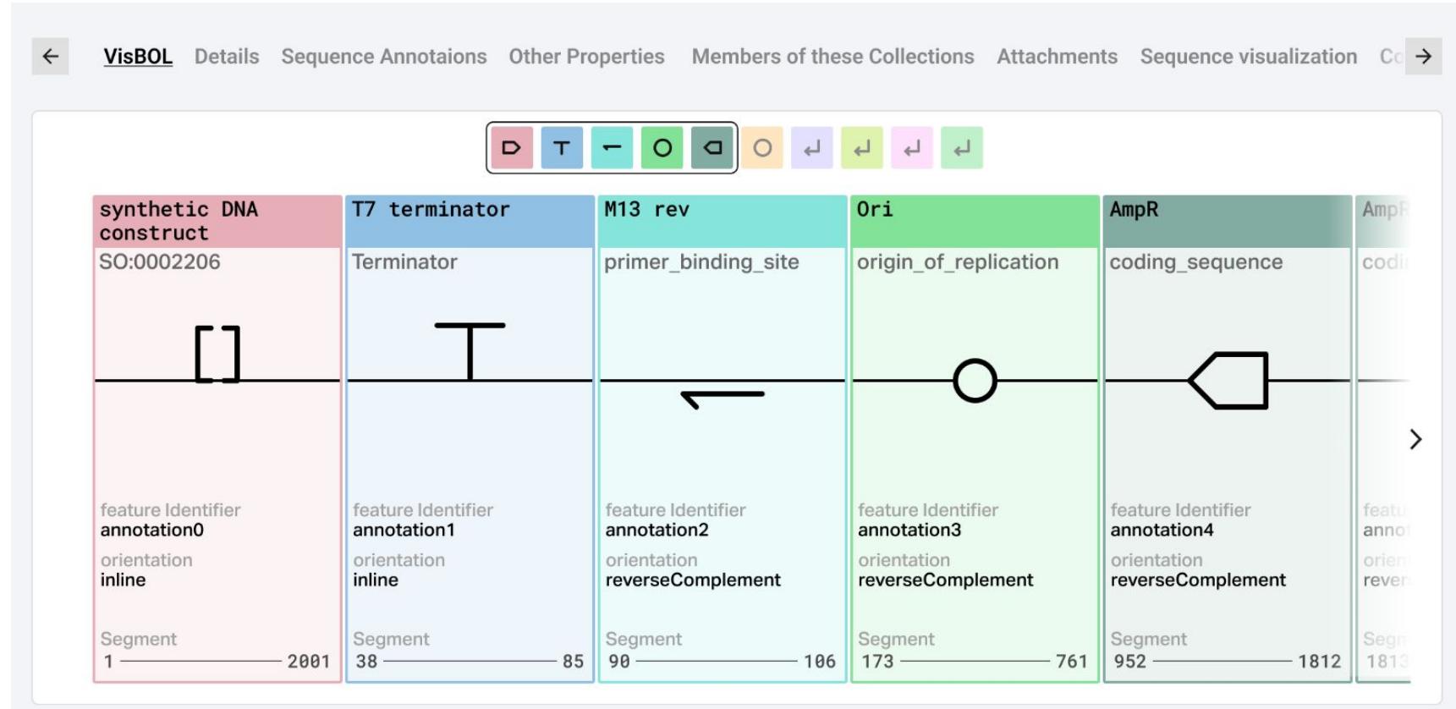
New search results interface



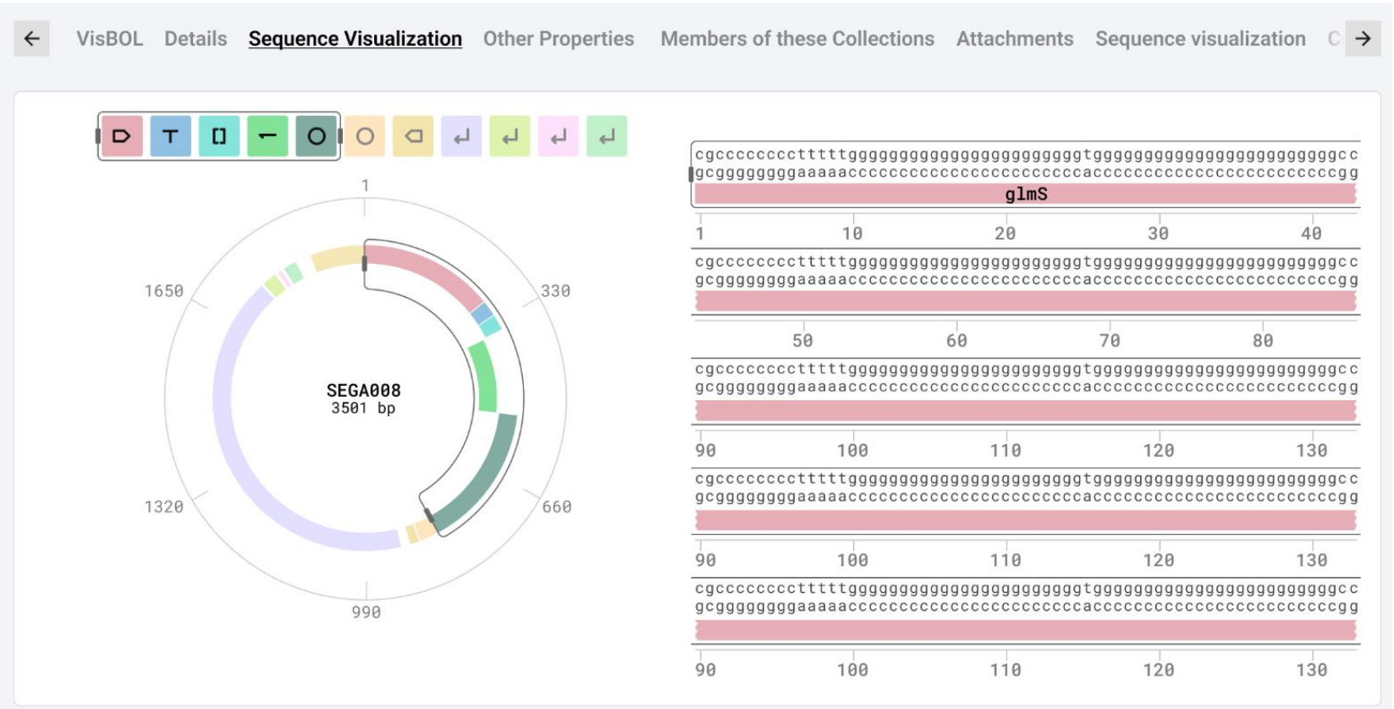
More info on collection pages



VisBOL-like plugin for SBOL objects



Sequence visualization plugin



References

1. https://www.openplant.org/open-tools-and-technologies
2. https://openbioeconomy.org/projects/open-enzyme-collections/
3. https://stanford.freegenes.org/
4. https://reclone.org/reagents/
5. Wilkinson, Mark D., et al. **The FAIR Guiding Principles for scientific data management and stewardship.** *Scientific data* 3.1 (2016): 1-9.
6. McLaughlin, James Alastair, et al. **The Synthetic Biology Open Language (SBOL) version 3: simplified data exchange for bioengineering.** *Frontiers in Bioengineering and Biotechnology* 8 (2020): 1009.
7. McLaughlin, J. A. et al. **SynBioHub: A Standards-Enabled Design Repository for Synthetic Biology**. ACS Synth. Biol. 7, 682–688 (2018).
8. Mante, Jeanet, et al. **Excel–SBOL Converter: Creating SBOL from Excel Templates and Vice Versa.** *ACS Synthetic Biology* 12.1 (2023): 340-346.
9. Bartley, Bryan A., et al. **pySBOL: a python package for genetic design automation and standardization.** *ACS synthetic biology* 8.7 (2018): 1515-1518.
10. Mante, Jeanet, Zach Sents, and Chris J. Myers. **SeqImprove: Machine Learning Assisted Creation of Machine Readable Sequence Information.** *bioRxiv* (2023): 2023-04.
11. https://accurat.it/
12. https://parts.igem.org
13. Ham, Timothy S., et al. **Design, implementation and practice of JBEI-ICE: an open source biological part registry platform and tools.** *Nucleic acids research* 40.18 (2012): e141-e141.
14. https://www.addgene.org/collections/

Give us your feedback!

We're continually developing this project, but would love to hear from you about whether this is FAIR and fit-for-purpose. Would you use it?

Give us your feedback by emailing Jenny (**jcm80@cam.ac.uk**), Felipe (**fxb23@cam.ac.uk**), and Reclone Team (**coordination@reclone.org**).

In collaboration with

OpenPlant sharing tools for a sustainable future
Earlham Institute
John Innes Centre

Funded by

UKRI Biotechnology and Biological Sciences Research Council
UKRI Engineering and Physical Sciences Research Council