UniProt, a FAIRness assessment

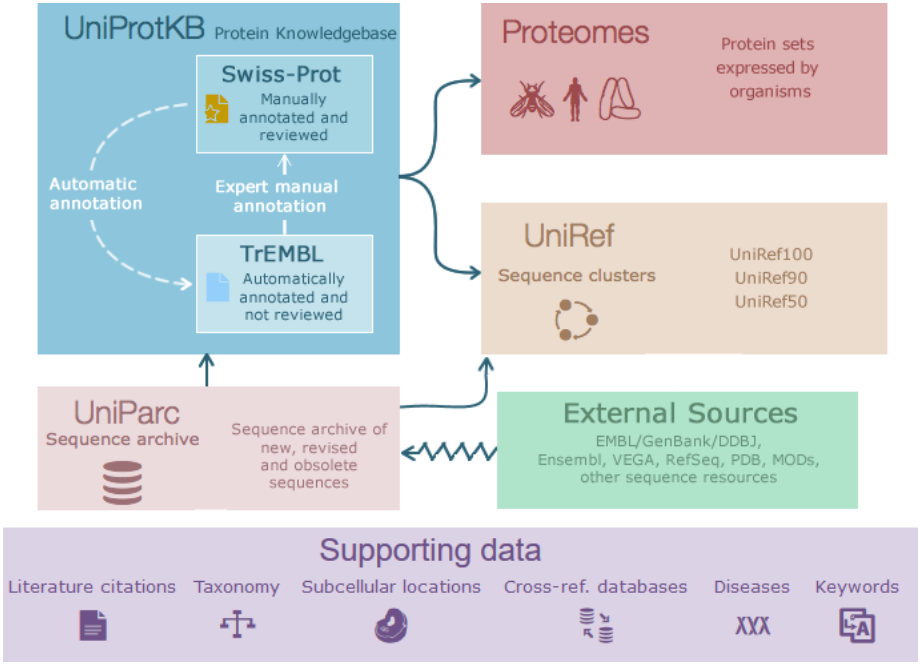
UniProt datasets

Leyla Garcia

Protein Function Development Team

EMBL-EBI

UniProt at a glance



Universal Protein Resource

- Comprehensive resource for protein sequence and annotation data.
- Datasets: UniProt Knowledgebase (UniProtKB), UniProt Reference Clusters (UniRef), UniProt Archive (UniParc) and Proteomes
- Supporting data: Diseases, taxonomy, keywords, subcellular location, citations, cross-references

Our FAIR story starts from the website

schema.org + Bioschemas

UniProt datasets, nearly 100% FAIR

Metric	Assessment
FM-F1A Identifier uniqueness	Stable and unique identifiers → accessions
FM-F1B Identifier persistence	PURL identifiers
FM-F2 Machine readability of metadata	RDF format + VOID file FASTA files + headers XSD for XML
FM-F3 Resource identifier in metadata	PURL included in RDF Accession always provided
FM-F4 Indexed in a searchable resource	Indexed by Google, FAIRsharing, identifiers.org

Metric	Assessment
FM-A1.1 Access protocol	HTTPS for webpages & FTP for downloads
FM-A1.2 Access authorization	Public datasets, not authorization required
FM-A2 Metadata longevity	Entry history always available Archive specialized dataset

UniProt datasets, nearly 100% FAIR

Interoperable

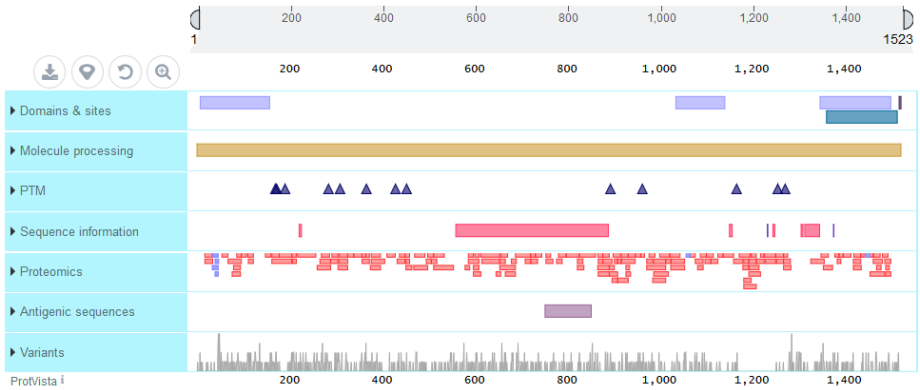
Metric	Assessment
FM-I1 Use a knowledge representation language	UniProt ontology for RDF
FM-I2 Use FAIR vocabularies	Uses well-known ontologies whenever possible
FM-I3 Use qualified references	“seeAlso” or “cross-reference” is commonly used

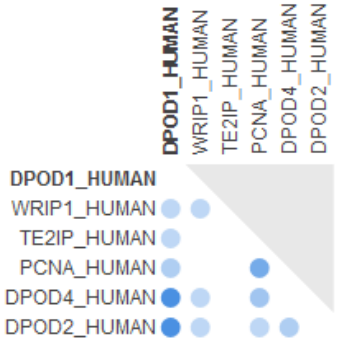
Reusable

Metric	Assessment
FM-R1.1 Accessible usage license	Creative commons attribution (CC BY 4.0)
FM-R1.2 Detailed provenance	Provided on help pages Machine-readable via VOID for RDF On entries via Evidence Codes
FM-R1.3 Meets community standard	What certification body should be used?

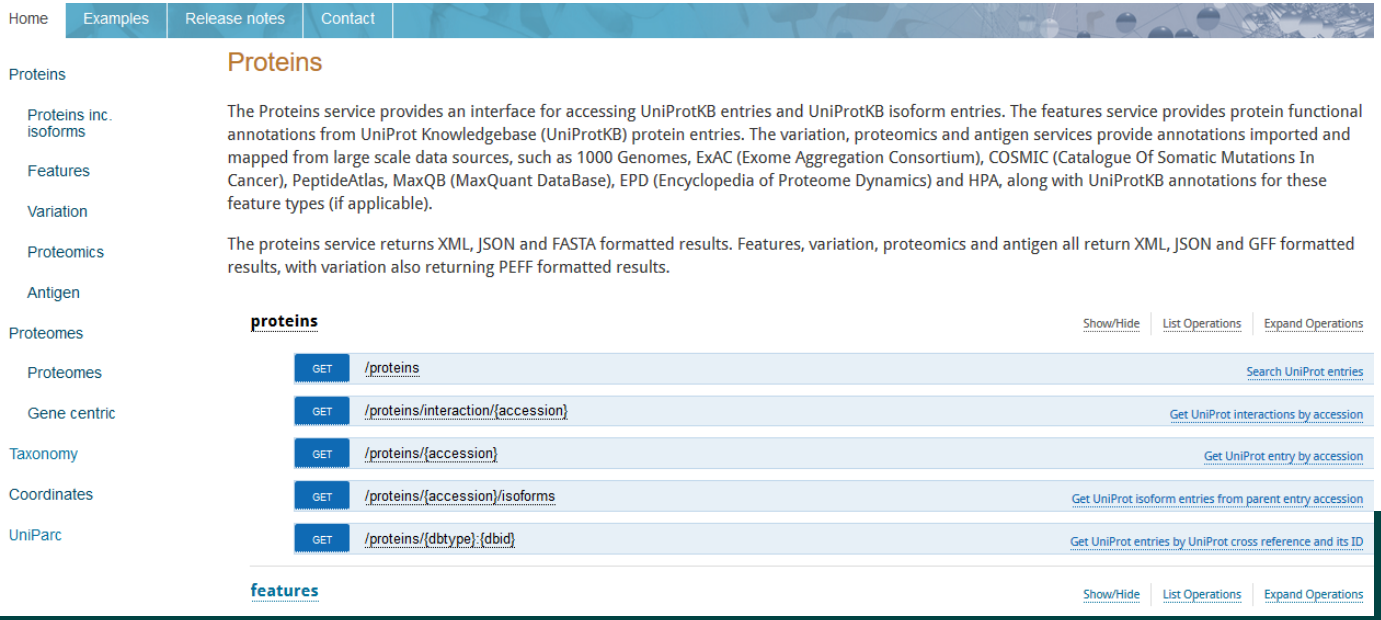
Tools, looking for FAIRness

ebi-webcomponents & ebi-uniprot





Protein API → https://www.ebi.ac.uk/proteins/api/doc/



EMBL-EBI

Conclusions

- FAIRness assessment is not easy
  - Principles are broad
  - Metrics help but are not always clear
  - Certification/validation mechanisms could help but where are they?

UniProt Team

PIs: Alex Bateman, Alan Bridge, Cathy Wu

Key staff: Cecilia Arighi (Curation), Lionel Breuza (Curation), Elisabeth Coudert (Curation), Hongzhan Huang (Development), Damien Lieberherr (Curation), Michele Magrane (Curation), Maria Martin (Development), Peter McGarvey (Content), Darren Natale (Content), Sandra Orchard (Content), Ivo Pedruzzi (Curation), Sylvain Poux (Curation), Manuela Pruess (Coordination), Shriya Raj (Coordination), Nicole Redaschi (Development)

Content / Curation: Lucila Aimo, Ghislaine Argoud-Puy, Andrea Auchincloss, Kristian Axelsen, Emmanuel Boutet, Ramona Britto, Hema Bye-A-Jee, Cristina Casals-Casas, Anne Estreicher, Livia Famiglietti, Marc Feuermann, John S. Garavelli, Penelope Garmiri, George Georghiou, Arnaud Gos, Nadine Gruaz, Emma Hatton-Ellis, Ursula Hinz, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Guillaume Keller, Kati Laiho, Philippe Lemercier, Yvonne Lussi, Alistair MacDougall, Patrick Masson, Anne Morgat, Klemens Pichler, Sandrine Pilbout, Catherine Rivoire, Karen Ross, Christian Sigrist, Elena Speretta, Andre Stutz, Shyamala Sundaram, Michael Tognolli, Nidhi Tyagi, C. R. Vinayaka, Qinghua Wang, Kate Warner, Lai-Su Yeh, Rossana Zaru

Development: Emanuele Alpi, Leslie Arminski, Parit Bansal, Delphine Baratin, Teresa Batista Neto, Benoit Bely, Mark Bingley, Jerven Bolleman, Borisas Bursteinas, Chuming Chen, Yongxing Chen, Beatrice Cuche, Alan Da Silva, Edouard De Castro, Tunca Dogan, Leyla Garcia Castro, Elisabeth Gasteiger, Sebastien Gehant, Leonardo Gonzales, Alexandr Ignatchenko, Rizwan Ishtiaq, Vishal Joshi, Dushyanth Jyothi, Arnaud Kerhornou, Vicente Lara, Thierry Lombardot, Jie Luo, Mahdi Mahmoudy, Xavier Martin, Andrew Nightingale, Joseph Onwubiko, Monica Pozzato, Sangya Pundir, Guoying Qi, Rabie Saidi, Tony Sawford, Edward Turner, Preethi Vasudev, Vladimir Volynkin, Yuqi Wang, Tony Wardell, Xavier Watkins, Hermann Zellner, Jian Zhang

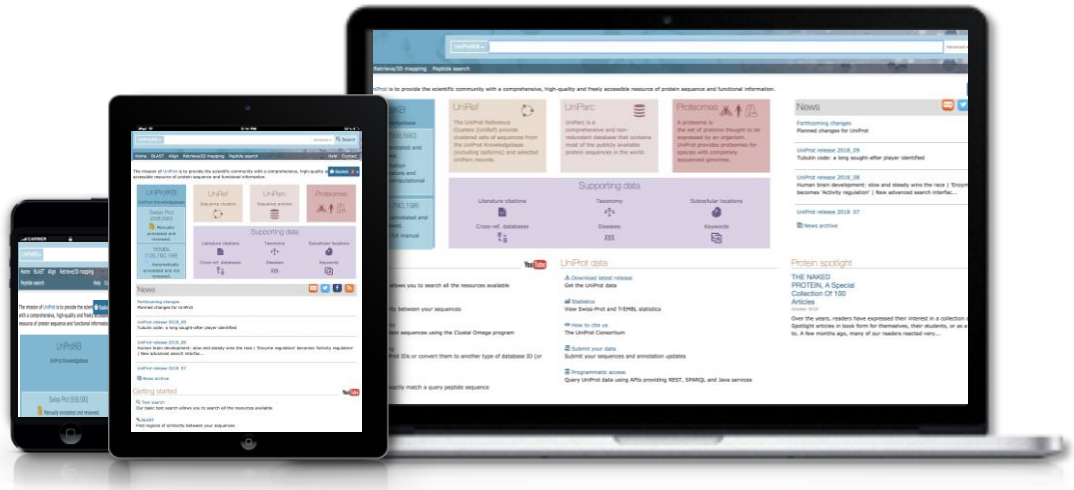
European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, UK

Protein Information Resource (PIR), Washington DC and Delaware, USA

SIB Swiss Institute of Bioinformatics (SIB), Geneva, Switzerland

Your Feedback Matters!

Help us improve the UniProt Website.



Contact us:

Register : http://bit.ly/helpuniprot

Email us on: help@uniprot.org

https://twitter.com/uniprot

https://www.facebook.com/uniprot.org

https://www.youtube.com/user/uniprotvideos