Challenges on data and knowledge integration, analysis, life cycle and reproducibility in life sciences

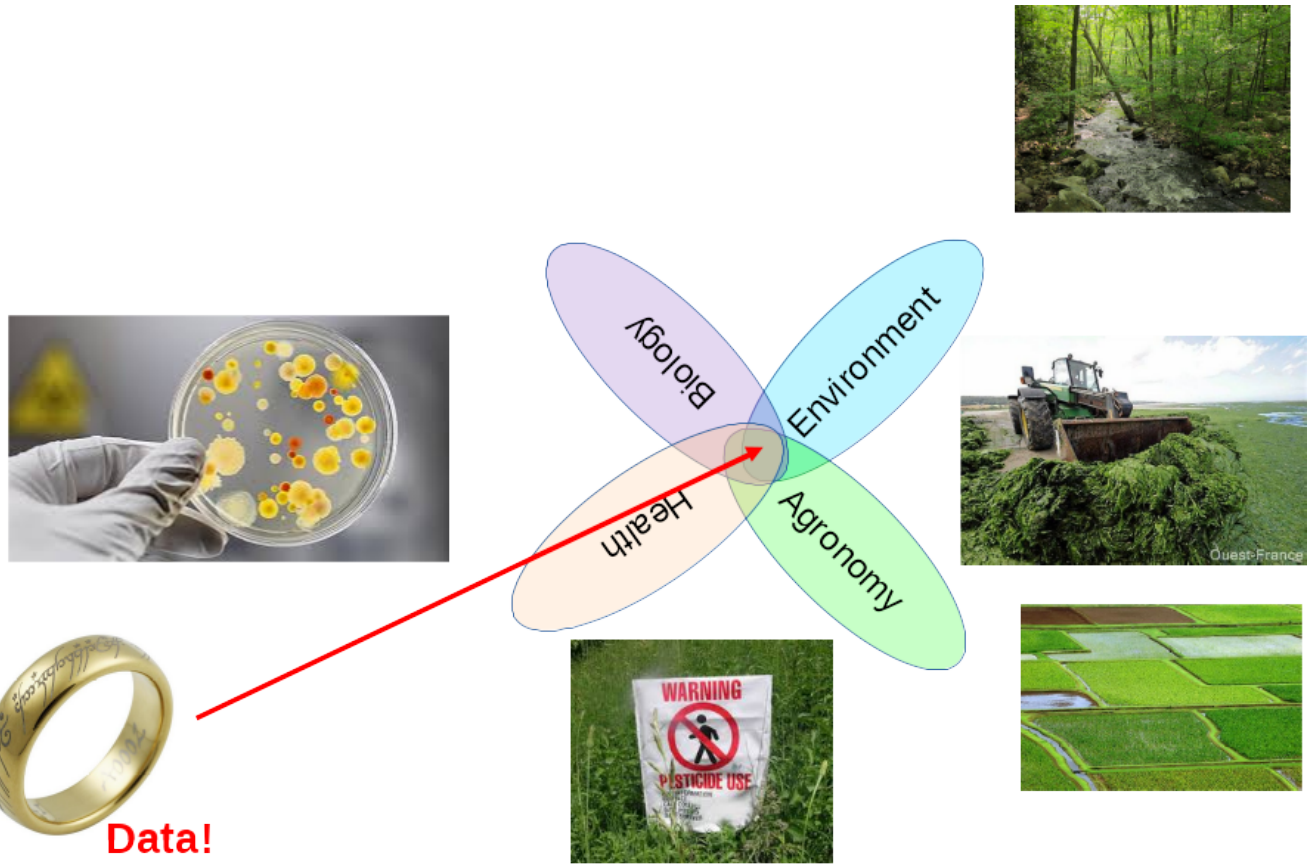
Olivier Dameron
https://orcid.org/0000-0001-8959-7189

Université de Rennes

2025-10-15

Life science data

Life science : a large, complex, inter-dependent domain...



... that stands out among other experimental sciences

"Biology has become an information science" [T. Lenoir, 1998 Stanford]

What to expect for 2025 ?

Our estimation is that genomics is a "four-headed beast" – it is either on par with or the most demanding domain [...] in terms of

- data acquisition
- data storage
- data distribution
- data analysis

Big Data: Astronomical or Genomical?

Zachary D. Stephens[1], Skylar Y. Lee[1], Faraz Faghri[2], Roy H. Campbell[2], Chengxiang Zhai[3], Miles J. Efron[4], Ravishankar Iyer[1], Michael C. Schatz[5]*, Saurabh Sinha[3]*, Gene E. Robinson[6]*

PLOS Biology | DOI:10.1371/journal.pbio.1002195 July 7, 2015

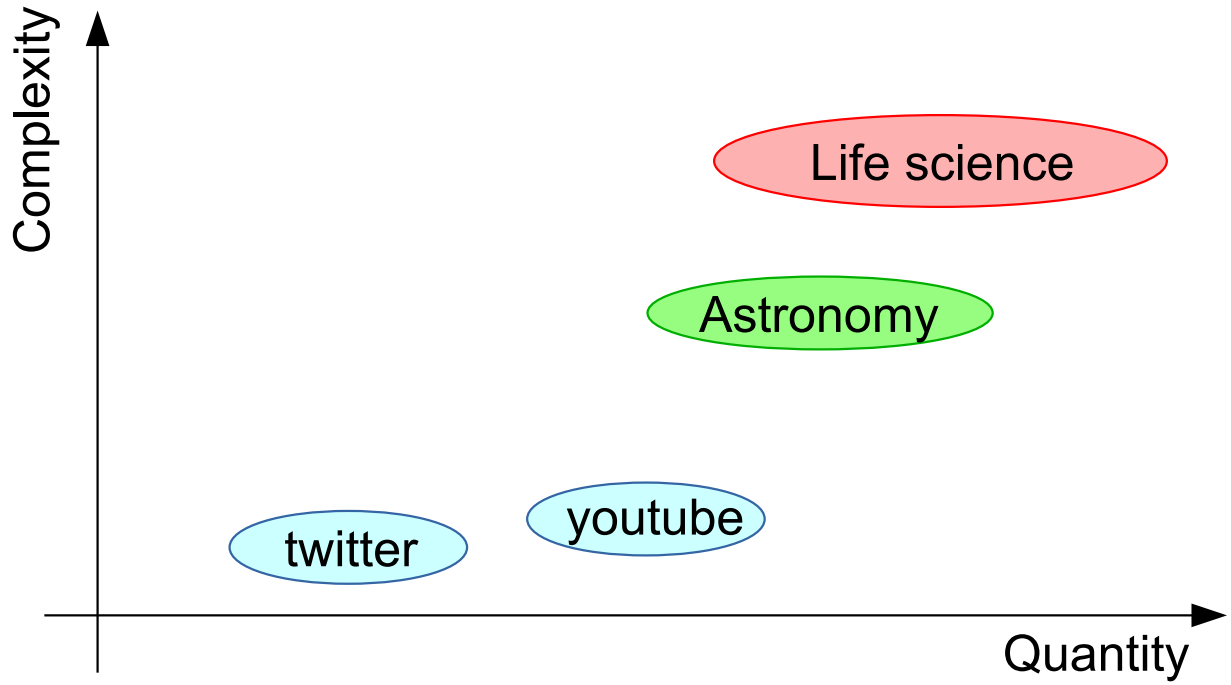
Table 1. Four domains of Big Data in 2025. In each of the four domains, the projected annual storage and computing needs are presented across the data lifecycle.

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001

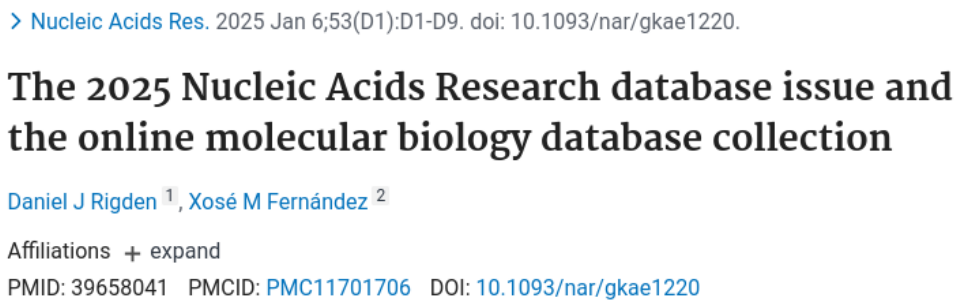
Degrees of data complexity

- multiple scales (heterogeneity)
- (highly) interdependent at each scale
- interdependent between scales
- variability
- incompleteness
- distributed (and lack of interoperability)
- evolving



Life science : beyond "data science" ...

- we have accumulated a trove of data
- we store and share these data
  (in 2.000+ reference databases [Rigden2025])



> Nucleic Acids Res. 2025 Jan 6;53(D1):D1-D9. doi: 10.1093/nar/gkae1220.

The 2025 Nucleic Acids Research database issue and the online molecular biology database collection

Daniel J Rigden [1], Xosé M Fernández [2]

Affiliations + expand

PMID: 39658041 PMCID: PMC11701706 DOI: 10.1093/nar/gkae1220

Life science : a domain that stands out

- by its complexity
- by the scarcity of its unifying laws
- by its long history of knowledge description and formalization

... and toward "knowledge(-based) science" ?

We needs a framework to support this transition

Attempt at defining some underlying notions

Database (e.g. Gene Expression Omnibus)

structured description of sets of homogeneous (as in "of the same class") instances and the relations between them.

Ontology (e.g. GeneOntology, ChEBI, HPO)

formal description of the general concepts (as in "the classes of things") of a domain and the relations between them.
(support general inferences).

Knowledge base (e.g. UniProt, Reactome, Rhea)

combines elements of databases and ontologies
(supports domain-based inferences about instances).

Knowledge graph (e.g. Reactome)

knowledge base that emphasizes graph-based capabilities.

Requirements for coping with life science data complexity

- **Requirement 1 :** **identify** resources with interoperable identifiers
- **Requirement 2 :** **describe** resources
  - their characteristics (e.g. start and end position of a gene,...)
  - their relations to other entities (e.g. the transcripts associated to a gene, the transcription factors that regulate it,...)
  - the categories they belong to
- **Requirement 3 :** **combine** descriptions from different origins, different points of view, different granularity levels
- **Requirement 4 :** **query** these descriptions
- **Requirement 5 :** support **semantically-rich** querying and reasoning (because of the inner complexity) using domain knowledge (this is required for capturing *expertise*)
- **Requirement 6 :** cover the whole **data life cycle**
- **Requirement 7 :** enforce **reproducibility**

If only the solutions to all these requirements were compatible !

The Semantic Web : a framework for integrating seamlessly (data,) metadata and knowledge, and querying and reasoning over it

Semantic Web and Linked (Open) Data

Semantic Web offers a unified framework to Linked Data

- **URIs** for identiying entities
- **RDF** for representing and aggregating entities descriptions
- **RDFS+OWL** for representing domain knowledge (and combine it with data descriptions)
- **SPARQL** for querying everything (possibly from multiple repositories)

SPARQL endpoints offer unified query access to RDF repositories
ex : Fuseki, Virtuoso, QLever...

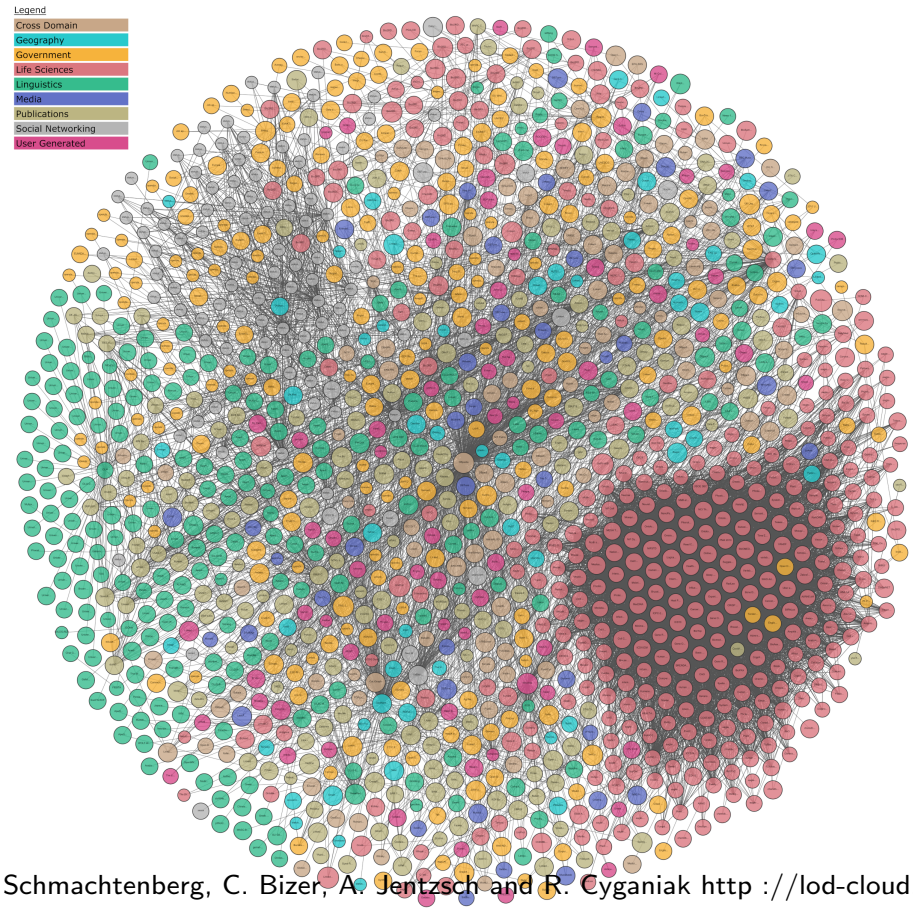
Linked Open Data : a federation of RDF repositories

LODStats (http://lodstats.aksw.org/) [Ermilov2016]

- 9960 datasets ; 149.10^9 triples
- general scope ; Life sciences = major field (size+density)

Linked open data (in 2025-09-02)

- RDF repositories can be queried in SPARQL via endpoints
- data from one endpoint can make references to data from another endpoint



Linked open data cloud, by M. Schmachtenberg, C. Bizer, A. Jentzsch and R. Cyganiak http ://lod-cloud.net/

So... problem solved ?

Some successes

- Semantic Web technologies have been around for 20+ years
- Most reference knowledge bases are part of the LOD cloud

Review › NPJ Digit Med. 2019 Sep 10;2:90. doi: 10.1038/s41746-019-0162-5. eCollection 2019.

Enabling Web-scale data integration in biomedicine through Linked Open Data

Maulik R Kamdar [1], Javier D Fernández [2 3], Axel Polleres [2 3], Tania Tudorache [1], Mark A Musen [1]

Affiliations + expand

PMID: 31531395 PMCID: PMC6736878 DOI: 10.1038/s41746-019-0162-5

› Sci Data. 2021 Jan 21;8(1):24. doi: 10.1038/s41597-021-00797-y.

An empirical meta-analysis of the life sciences linked open data on the web

Maulik R Kamdar [1 2], Mark A Musen [3]

Affiliations + expand

PMID: 33479214 PMCID: PMC7819992 DOI: 10.1038/s41597-021-00797-y

So... problem solved ?

Some successes

- Semantic Web technologies have been around for 20+ years
- Most reference knowledge bases are part of the LOD cloud

Review › NPJ Digit Med. 2019 Sep 10;2:90. doi: 10.1038/s41746-019-0162-5. eCollection 2019.

Enabling Web-scale data integration in biomedicine through Linked Open Data

Maulik R Kamdar [1], Javier D Fernández [2 3], Axel Polleres [2 3], Tania Tudorache [1], Mark A Musen [1]

Affiliations + expand

PMID: 31531395 PMCID: PMC6736878 DOI: 10.1038/s41746-019-0162-5

› Sci Data. 2021 Jan 21;8(1):24. doi: 10.1038/s41597-021-00797-y.

An empirical meta-analysis of the life sciences linked open data on the web

Maulik R Kamdar [1 2], Mark A Musen [3]

Affiliations + expand

PMID: 33479214 PMCID: PMC7819992 DOI: 10.1038/s41597-021-00797-y

But world domination is not there yet

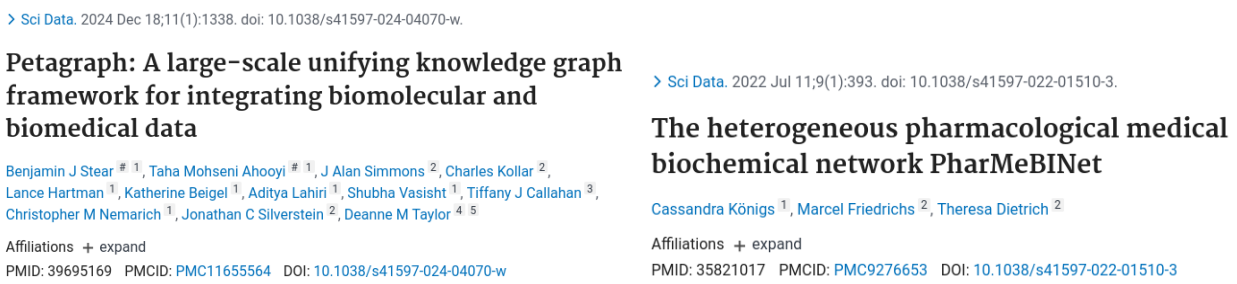
- Most reference knowledge bases... ... but not all of them
  - no Gene Ontology SPARQL endpoint
  - The EBI SPARQL endpoints have closed
  - SIB (Switzerland) and NBDC+DBCLS (Japan) are the new hotspots
- Adoption by both users and developers is still far away
- Some are even abandoning SW for more fancy solutions that address immediate needs but miss the big picture
  - integration : OmniPath
  - querying : BioCypher

Challenges

Challenge : data and knowledge integration

- 2.000+ reference knowledge bases
- countless project-specific databases and knowledge base

- each with its own, complex data schema
- they are here and yet, we seldom (re-)use them
- the reflex still too often is to integrate snaphots, expose it and let it become soon out of date
- the decentralized web is still a (far away) vision



› Sci Data. 2024 Dec 18;11(1):1338. doi: 10.1038/s41597-024-04070-w.

Petagraph: A large-scale unifying knowledge graph framework for integrating biomolecular and biomedical data

Benjamin J Stear [# 1], Taha Mohseni Ahooyi [# 1], J Alan Simmons [2], Charles Kollar [2], Lance Hartman [1], Katherine Beigel [1], Aditya Lahiri [1], Shubha Vasisht [1], Tiffany J Callahan [3], Christopher M Nemarich [1], Jonathan C Silverstein [2], Deanne M Taylor [4 5]

Affiliations + expand

PMID: 39695169 PMCID: PMC11655564 DOI: 10.1038/s41597-024-04070-w

› Sci Data. 2022 Jul 11;9(1):393. doi: 10.1038/s41597-022-01510-3.

The heterogeneous pharmacological medical biochemical network PharMeBINet

Cassandra Königs [1], Marcel Friedrichs [2], Theresa Dietrich [2]

Affiliations + expand

PMID: 35821017 PMCID: PMC9276653 DOI: 10.1038/s41597-022-01510-3

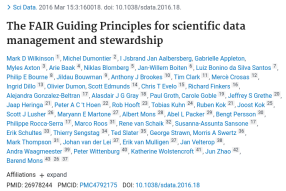
Challenge : data and knowledge analysis

Beyond the reductionist approach

- not quite sure what questions to ask
- not even in the mindset of (most) end-users
- how to do it ? before, after, before and after integration ?
- where to run the analysis ? Infrastructure like clusters, Galaxy are available but their user base is far from spanning the life science community
- what to do with the results ? there are so many of them, we still need methods for making sense out of them

Challenge : data and knowledge life cycle

- the FAIR principles have been around for 10 years [Wilkinson2016]



> Sci Data. 2016 Mar 15:3:160018. doi: 10.1038/sdata.2016.18.

The FAIR Guiding Principles for scientific data management and stewardship

> Front Plant Sci. 2016 May 12:7:641. doi: 10.3389/fpls.2016.00641. eCollection 2016.

Publishing FAIR Data: An Exemplar Methodology Utilizing PHI-Base

Alejandro Rodríguez-Iglesias [1], Alejandro Rodríguez-González [2], Alistair G Irvine [3], Ane Sesma [1], Martin Urban [4], Kim E Hammond-Kosack [4], Mark D Wilkinson [1]

Affiliations + expand

PMID: 27433158 PMCID: PMC4922217 DOI: 10.3389/fpls.2016.00641

> F1000Res. 2017 Aug 31:6:1618. doi: 10.12688/f1000research.12344.2. eCollection 2017.

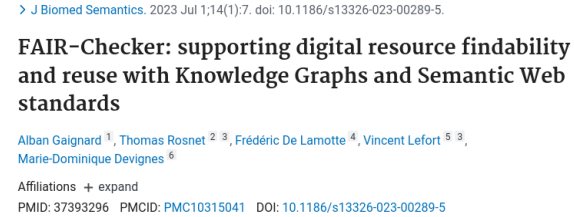
Best practice data life cycle approaches for the life sciences

Affiliations + expand

PMID: 30109017 PMCID: PMC6069748 DOI: 10.12688/f1000research.12344.2

- Few datasets actually follow them (but support is developping)

  https://fair-checker.france-bioinformatique.fr



> J Biomed Semantics. 2023 Jul 1;14(1):7. doi: 10.1186/s13326-023-00289-5.

FAIR-Checker: supporting digital resource findability and reuse with Knowledge Graphs and Semantic Web standards

Alban Gaignard [1], Thomas Rosnet [2 3], Frédéric De Lamotte [4], Vincent Lefort [5 3], Marie-Dominique Devignes [6]

Affiliations + expand

PMID: 37393296 PMCID: PMC10315041 DOI: 10.1186/s13326-023-00289-5

> Sci Data. 2025 Feb 24;12(1):328. doi: 10.1038/s41597-025-04451-9.

Applying the FAIR Principles to computational workflows

Affiliations + expand

PMID: 39994238 PMCID: PMC11850811 DOI: 10.1038/s41597-025-04451-9

FAIRify

- data
- software
- workflows + workflow runs

Will generate even more data

Challenge : reproducibility

How to ensure the reproducibility ?

- of the data we produce
- of the software we develop
- of the analyses we perform
- of the conclusion we draw

Will generate even more data

		Data	
		Same	Different
Code	Same	Reproducible	Replicable
	Different	Robust	Generalisable

[Whitaker2016]

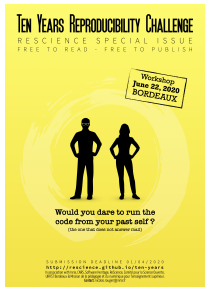
Review > Patterns (N Y). 2021 Sep 10;2(9):100322. doi: 10.1016/j.patter.2021.100322.

The role of metadata in reproducible computational research

Jeremy Leipzig [1], Daniel Nüst [2], Charles Tapley Hoyt [3], Karthik Ram [4], Jane Greenberg [1]

Affiliations + expand

PMID: 34553169 PMCID: PMC8441584 DOI: 10.1016/j.patter.2021.100322



> PLoS Comput Biol. 2022 Mar 9;18(3):e1009935. doi: 10.1371/journal.pcbi.1009935. eCollection 2022 Mar.

Urgent need for consistent standards in functional enrichment analysis

Kaumadi Wijesooriya [1], Sameer A Jadaan [2], Kaushalya L Perera [1], Tanuveer Kaur [1], Mark Ziemann [1]

Affiliations + expand

PMID: 35263338 PMCID: PMC8936487 DOI: 10.1371/journal.pcbi.1009935

A road map to improve software sharing practices

Roberto Di Cosmo, Sabrina Granger, Konrad Hinsen, Nicolas Jullien, Daniel Le Berre, Violaine Louvet, Camille Maumet, Clémentine Maurice, Raphaël Monat & Nicolas P. Rougier

The road ahead

Open questions

How to promote usage by users ?

- How to make data + knowledge + softwares + workflows available ?
- How to shield them from the technical difficulties ?
- How to handle the complexity of data schema that is a consequence of Life Science intrinsic complexity ?
- How to accomodate the various kinds of reasoning that users (legitimately) want to perform but are beyond the scope of Semantic Web (boolean networks, path finding, temporal reasoning,...) ?

Open questions

How to promote usage by users ?

- How to make data + knowledge + softwares + workflows available ?
- How to shield them from the technical difficulties ?
- How to handle the complexity of data schema that is a consequence of Life Science intrinsic complexity ?
- How to accomodate the various kinds of reasoning that users (legitimately) want to perform but are beyond the scope of Semantic Web (boolean networks, path finding, temporal reasoning,...) ?

How to promote contribution by users ?

- How to facilitate integration of project-specific datasets with SW ?
- How to facilitate exposing their datasets ?
  (hopefully in a FAIR perspective and workflow-compatible way)

Open questions

How to promote usage by users ?

- How to make data + knowledge + softwares + workflows available ?
- How to shield them from the technical difficulties ?
- How to handle the complexity of data schema that is a consequence of Life Science intrinsic complexity ?
- How to accomodate the various kinds of reasoning that users (legitimately) want to perform but are beyond the scope of Semantic Web (boolean networks, path finding, temporal reasoning,...) ?

How to promote contribution by users ?

- How to facilitate integration of project-specific datasets with SW ?
- How to facilitate exposing their datasets ?
  (hopefully in a FAIR perspective and workflow-compatible way)

Beyond *Homo sapiens* : how to connect with ML ?