The ultimate guide to biocuration

Biocuration is a cornerstone of data management. Its aim is to enable scientists, both in academia and R&D organisations, to extract valuable knowledge from biological data.

1

Know the subject area and assemble a team of experts

Domain knowledge is a pre-requisite. For example, one should be familiar with typical RNA-seq workflows in the laboratories in order to curate the data generated. Try to have a mixture of expertise from different subject domains to help resolve ambiguous curatorial cases.

2

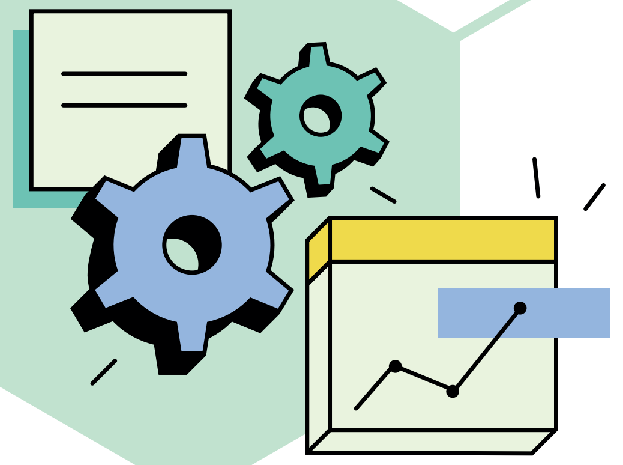
Clearly define the intended use of the curated data

Identify the 3 Ws:

- what knowledge the curation aims to capture,
- when in the data stream the curation should be carried out,
- who will consume the curated data.

This helps to create the data model and curation guidelines.

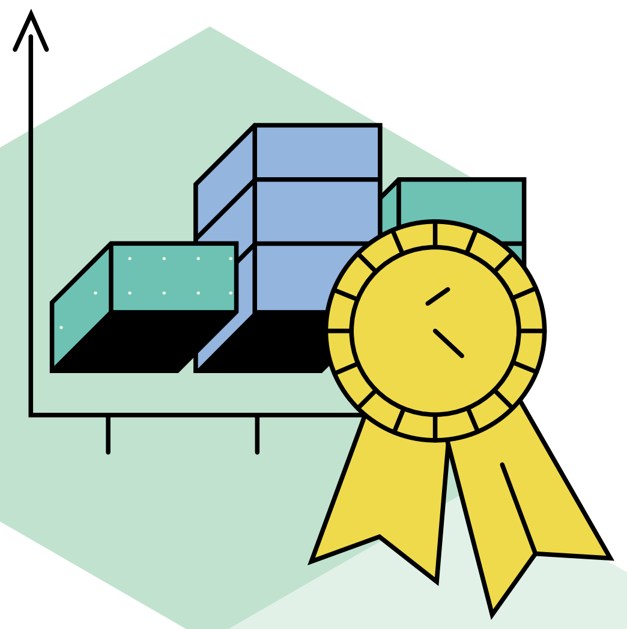
3



Automate as much curation as possible

Automate monotonous or mundane curation tasks to free up time for inspecting edge cases, reduce manual (human) error, and promote metadata standardisation.

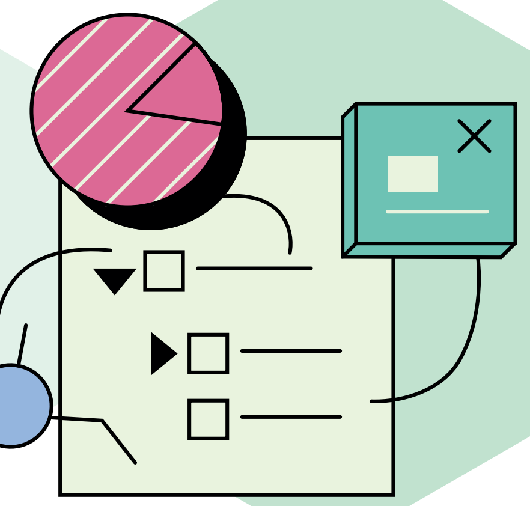
4



Share your data in a standard structure

To ensure data are easily reusable without any format conversion or transformation, curated (meta)data should be shared in widely accepted formats, accompanied by a metadata schema that documents the mandatory metadata fields.

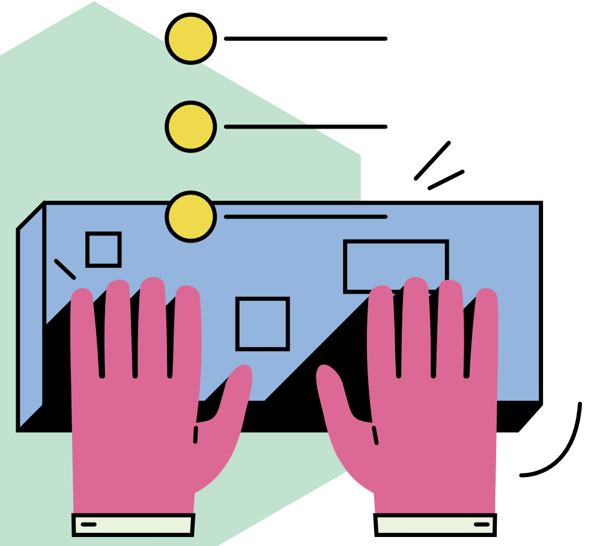
5



Use ontologies and persistent identifiers to annotate your data

Since human language is very flexible, often there are different ways to say the same thing (e.g. "pavement" and "sidewalk"). Use ontologies (controlled vocabulary with relationships defined amongst the terms) and persistent identifiers to disambiguate language and uniquely identify any entity (e.g. a gene, a journal article).

6



Develop robust curation guidelines

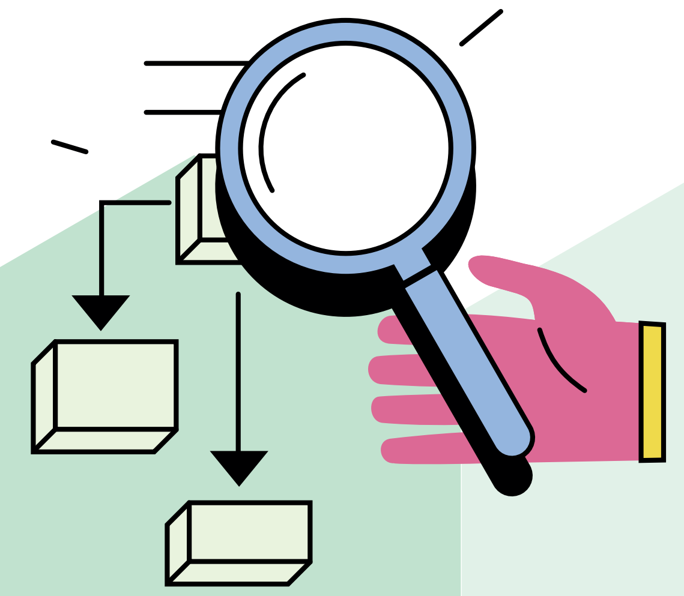
Curation guidelines do not only promote consistency (e.g. have the "organism" field as the first column in the metadata table), but can also ensure data provenance is captured (e.g. recording the data source and any transformation performed on the data).

7

Curate early, stay cozy with the data

Like with many things in life, it is easier if you start early. Curation is the same. The closer to the point of data generation, the easier it is to seek clarification about the data and curate, when the research work is still fresh in the mind of the data provider.

8



Commit to maintaining data

Curation should be regarded as an ongoing and iterative process, not an end-point. Because of new biological knowledge or updated standards, most curated data will require updating or re-curation in the future, and this should not come as a surprise.

9

Learn basic programming for ad hoc data wrangling

Once in a while, as a curator, you will encounter tasks which are best performed with a computer script for efficiency and accuracy. For example, extracting particular columns from a file and concatenating them into one column. Being able to write basic scripts will be a huge advantage in completing the task quickly without having to seek third-party assistance.

10

Persits the data and provide it in multiple formats

It would be a huge pity if after spending so much effort in curating data, they are lost due to the lack of maintenance or accidents. Therefore, it is important to make sure the curated data persist by storing them in some sort of databases. An added benefit of using a database is to be able to export data in different formats which are readily consumable by a wider audience.

Genestack

At Genestack we develop software to help life science researchers and R&D scientists manage multi-omics data. To learn more about us and Omics Data Manager, our enterprise software solution to make data FAIR, visit **www.genestack.com**

References: Tang YA et al. (2019) Ten quick tips for biocuration. PLoS Comput Biol 15(5): e1006906