How to Choose
the Best System
to Manage Your Data

Genestack

Proper management of your research data is half the battle for success. A good data management platform can significantly speed up scientific research and save a lot of work. Although choosing one is no easy task, there are a few things you definitely can not overlook.

66

Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process."

The FAIR Guiding Principles for scientific data management and steward-ship. Sci Data 3, 160018 (2016)

Genestack

In this E-book, we shall walk you through some of the most important considerations when choosing a data management system.

We will walk through why it is important to:

- Make sure your platform is easy to understand for non-technical users,

- Ensure it is possible to connect with other services and streamline your work-flows.

- Make your metadata, molecular data, and phenotypic data searchable.

- Ensure that your new platform makes curating, sharing, and versioning large amounts of data effortless and less error-prone.

- Work with a vendor who provides a professional service and support team as standard

Genestack

Table of Contents

Genestack

Easy To Use

Easy for researchers

Remember that your system will not only be used by advanced technical users. Researchers should also find it easy to use. After all, they are the main people responsible for creating new data. So make sure the system is user-friendly and does not require too much learning.

A well-designed user interface means less time is spent and fewer mistakes are made. If the platform looks like a flight control panel, your researchers will likely continue to use their old solutions and keep all their data in spreadsheets and ELNs.

" We spent 6 months just looking for data that we knew we had!"

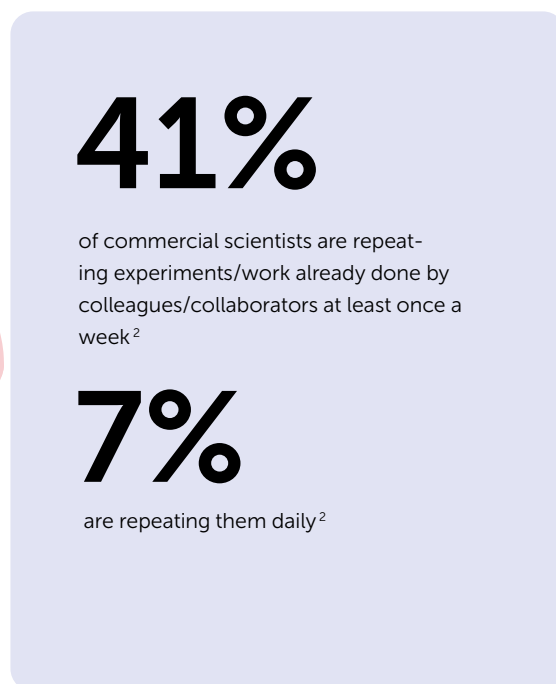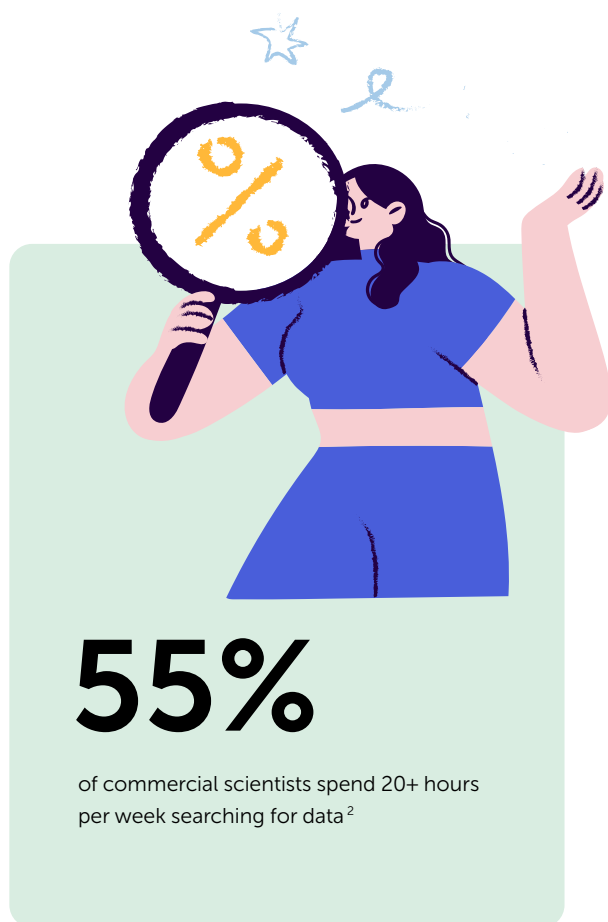Director of Data Integration & Ontologies, Novo Nordisk, Biodata 2021

Simple user interface

Another reason you should pay attention to your platform's user interface: The simpler it is, the less time you'll have to spend on mundane tasks. Our research shows that 55% of commercial scientists spend 20 hours or more every week just searching for data.[2] Imagine spending all that time generating valuable ideas and how you can accelerate your research by simply spending less time searching and organising your data.

Searchable data

The impact of making your data searchable with ease by all cannot be understated though. Being able to quickly understand what data is available can help to reduce your climate impact through reducing experimental redundancy. Many scientists are repeating experiments that have already been done by colleagues simply because they are unaware or cannot find the data. This not only means increased use of single-use plastics and harmful chemicals but also slows down your pipeline and research progress significantly.

41%

of commercial scientists are repeating experiments/work already done by colleagues/collaborators at least once a week [2]

7%

are repeating them daily [2]

55%

of commercial scientists spend 20+ hours per week searching for data [2]

Helps To Clean
Your Data

Importance of well-curated data

Why do data scientists end up cleaning up so much data though? Simply put, many of the researchers who produce the data do not always appreciate the impact that having accurate and complete descriptive data around each experiment can have. There is therefore a degree of education and cultural change that is often required to help them see the potential benefits to gain from well-curated data.

Quick curation

Regardless of who performs the curation, entering the descriptive data around thousands of samples is never a fun task. Everyone is more focused on getting insights out of the experiment immediately rather than its future potential. Your platform should therefore do the curation as quickly as possible and suggest corrections to you to make the process as painless as it can be.

The ability to map to standard scientific ontologies is also extremely helpful, especially for disease recording and taxonomy.

This will save annoyance at people being corrected for using "AML" instead of "Acute Myeloid Leukaemia" (and avoid all the variations that come from mispellings).

Imagine you need to record gender and you have "female" in one study and "F" in another.

How would you search for such data?

- The best platform will automatically suggest "female" for "F" when you load the study

- and highlight all possible problems with your metadata,

- or even better automatically correct it as it is loaded.

Curating your data at scale

Your platform should not require you to change each entry manually either. Bulk or automated tools that curate your data at scale should be a standard part of the platform.

$3.1
trillion per year

US Businesses lose due to poor data quality.

$9.7
million per year

The average financial impact of poor data quality on organisations

Genestack

Saves Older Versions
Of Your Data

Cleaning data is not only tedious but sometimes dangerous. You may delete or change some important attributes and not notice it immediately. The implications of such accidents can have a significant impact on downstream analysis and, in the long run, the success of your research.

Imagine you accidentally switch the labels between treatment and control samples. The impact of your analysis could potentially be significant. How would you know that this has even occurred? Imagine the consequences if you didn't notice the alteration!

The impact goes beyond just scientific research, however. In such cases of accidental change, loss or deletion of data you have to start again with the raw version (or even worse redo the experimental work itself) - costing both time and money.

Built-in versioning feature

So make sure you choose a platform that has a built-in versioning feature so you can roll back in time when you need to, spot and correct such errors or recover lost data in just a moment.

$586.000
a year

average business costs due to data loss [7]

1%

Average human error rate In manual data entry [13]

Has A Powerful API

Connect all your existing systems

API allows you to connect and leverage existing systems in your data landscape. Keep in mind that your platform should become a SPoT (single point of truth), from which all other systems can be accessed. This is impossible to do without the API to connect all your existing systems to a new platform.

Automate the mundane

Your data scientists do not have to spend hours/days processing data when they only need to make a few API calls to retrieve all the information they need. Many boring and mundane tasks can be automated and saved as a ready-to-use script, so you do not have to do the same repetitive task over and over again. and done a lot faster.

Changing "F" to "Female" in all your studies will no longer take half of the working day. With a good API, it can be done in a few minutes. Therefore, a good API is a must.

> **APIs enhance the productiveness of development teams via maximising reusability and enforcing consistency in new purposes"**
>
> SDK.Finance (2018) How to Spot a Good API

With fewer errors

Automating your tasks helps do things not only faster but also with fewer errors. Any small mistake in research data can affect the result and lead to a wrong conclusion.

674% ROI

can be achieved by a business when APIs are executed and utilised properly [9]

Makes Molecular and Phenotypic Data Searchable

Research shows that the volume of data is getting bigger each year and data management systems will be a key component of processing biological knowledge in the coming decades; being able to quickly search all your data is key.[10]
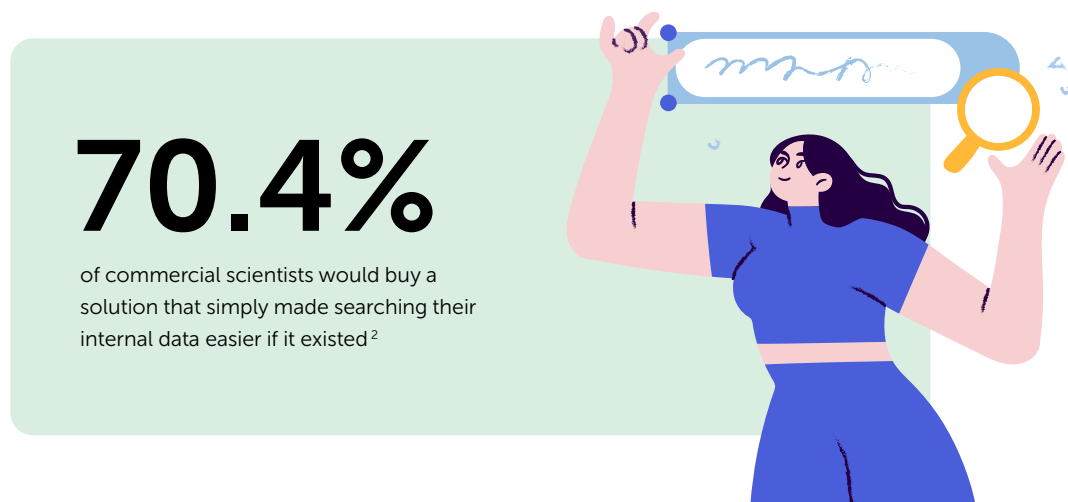
Complex search parameters...

Imagine that for your research, you need PD1 gene expression and genomic variants for all female patients with BRCA cancer who have reached menopause. To do that you need to analyse several individual studies with thousands of samples.

With "the majority of the studies exploring gene expression data result in one or more gene signatures"[11] and their growing use for the classification of patients and samples it is becoming increasingly important to be able to identify samples with complex search parameters.

...in seconds

But what if you could dissect the data with a single query and export the result within seconds?

Most existing platforms do not support searching by a combination of metadata, high-throughput omics data (expression, variant) and phenotypic data (age, sex, disease, etc.). Make sure your platform of choice allows for both UI searching with facets and synonyms, as well as automated queries through the API. This will save your data scientists hours/days and make your work much more efficient.

70.4%

of commercial scientists would buy a solution that simply made searching their internal data easier if it existed [2]

Genestack

Controls Access
To Data

Remember that good data must be FAIR - findable, accessible, interoperable and reusable.[1]

Easily share your data

Sharing a study with a particular department in your organisation should only require a few clicks, not a plethora of emails and attached files. Although this kind of sharing seems simple at first glance, it only leads to confusion and clutter. Recipients have to load the data onto their platform before they can use it. And a lot of precious data can easily get lost. Worse still people can accidentally allow access to data outside of their permissions - leading to potential regulatory breaches.

Control access

Make sure your data management platform allows you to control access to each of your studies separately and share them with a specific group of people. These groups can consist of members of a department in your company or a laboratory. This will not only ensure that people don't get overwhelmed with data but also that confidential and highly sensitive data can be only shared with the appropriate teams. With the right controls, there shouldn't be a need for further complex steps to share the data with a group of people.

Genestack

Assign privileges

Another feature to look out for is the ability to create custom user groups with different privileges. For example, let it be a curator and user groups to start with. This way you can allow a limited number of people to edit the data and thus exercise more control over it. Until it is ready for wider consumption. Control is key both from a regulatory compliance perspective and from a research efficiency perspective.

"

With greater demand for effective communication between information systems, via data integration, the need for data security in information systems increased significantly."

Moghaddasi H, Sajjadi S, Kamkarhaghighi M. Reasons in Support of Data Security and Data Security Management as Two Independent Concepts: A New Model. [14]

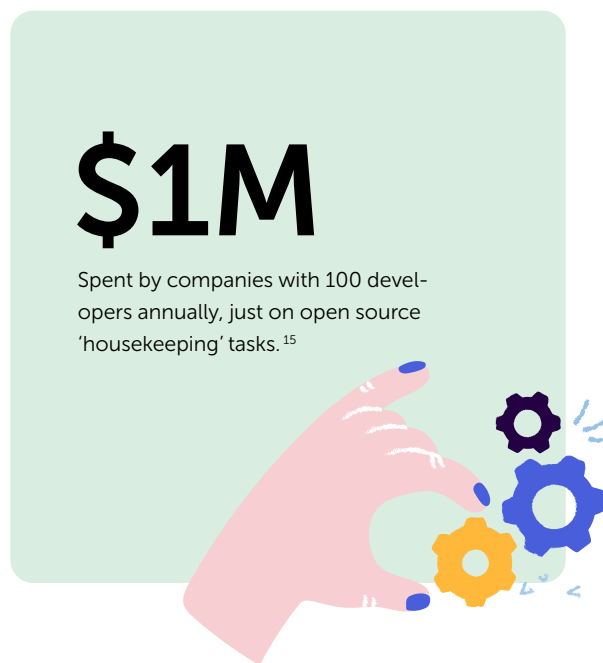# Comes With Dedicated Vendor Support

Using an open source solution looks like a great way to save money, but let's think about the reality.

Development pacing

Life sciences companies and their workflows are complex, have different landscapes, and evolve at different rates. New features may need to be added to the solution that a few months ago didn't seem to be required. The pace of development is therefore often too slow with open source tools when compared to a commercial offering. Whilst the initial outlay might seem more without good support from your vendor, you will be forced to spend much of your time customising the platform rather than using it - ultimately a false economy.

Multi-user support

You need to customise the system so that it works for each of your user-type specific use cases. Ready-made solutions rarely work for everyone, so they need to be customised. Most open source tools are designed with a single user persona in mind whilst commercial tools recognise the need for multi-user support - again beware the false economy of "free" tools.

$1M

Spent by companies with 100 developers annually, just on open source 'housekeeping' tasks.[15]

Supported learning

You need good training for your staff, which your vendor should provide as part of your subscription, so that you can use the system with minimal effort. This should also help you load and maintain the initial data so you can see the benefits and ROI as quickly as possible.

Genestack

At Genestack, once we realised that none of the existing platforms adequately addressed the above requirements, we decided to build our own.
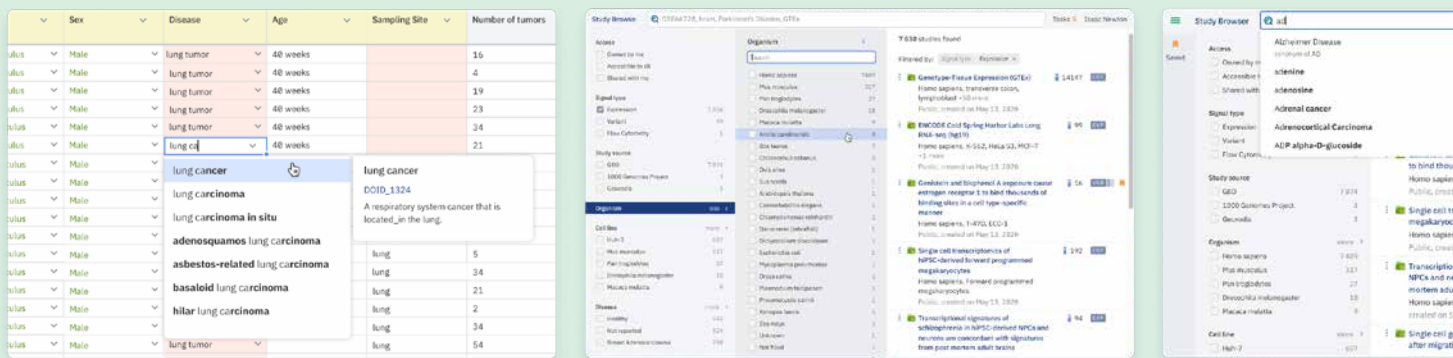
- The Genestack Platform caters to the needs of both non-technical and technical users.

- Non-technical users can use user-friendly GUI to search, explore, and curate data.

- Data scientists/engineers also have access to a powerful API to flexibly browse and slice the data, create custom scripts/apps, and connect with other systems.

- Data cleansing is done in a few clicks and can be easily automated.

- We also have a versioning system that allows you to track any changes to your metadata. Once you have the study ready, you can share it with just a few clicks.

- But most importantly, we have a professional services team that can help you at every stage of data management should you need it.

So if you're looking to upgrade your data management experience to one that makes your data management platform a help rather than a hindrance, get in touch or simply click the "Request demo" button on our website and let us show you how we can make your data work for you.

Genestack.com

References

1. Wilkinson, M., et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

2. 2021 Genestack Industry Market Research Study.

3. Merck: Smartlab Exchange Europe 2021

4. Gartner (2018) How to Create a Business Case for Data Quality Improvement, https://www.gartner.com/smarterwithgartner/how-to-create-a-business-case-for-data-quality-improvement

5. IBM (2016) Four vs Big data Infographic, http://www.ibmbigdatahub.com/infographic/four-vs-big-data

6. Jennifer Bryan (2018) Excuse Me, Do You Have a Moment to Talk About Version Control?, The American Statistician, 72:1, 20-27, DOI: https://doi.org/10.1080/00031305.2017.1399928

7. Acronis (2014) The True Cost of Lost — or Nearly Lost — Data, https://www.acronis.com/en-gb/blog/posts/true-cost-lost-or-nearly-lost-data/

8. SDK.Finance (2018) How to Spot a Good API, https://sdk.finance/how-to-spot-a-good-api/

9. Forrester (2017) Forrester TEI Study Results Show 674% ROI, https://community.ibm.com/community/user/integration/blogs/alan-glickenhouse1/2017/03/13/forrester-tei-study-results-show-674-roi

10. Michener WK (2015) Ten Simple Rules for Creating a Good Data Management Plan. PLoS Comput Biol 11(10): e1004525. https://doi.org/10.1371/journal.pcbi.1004525

11. Cantini, L., Calzone, L., Martignetti, L. et al. Classification of gene signatures for their information value and functional redundancy. npj Syst Biol Appl 4, 2 (2018). https://doi.org/10.1038/s41540-017-0038-8

12. Jim Gray, David T. Liu, Maria Nieto-Santisteban, Alex Szalay, David J. DeWitt, and Gerd Heber. 2005. Scientific data management in the coming decade. SIGMOD Rec. 34, 4 (December 2005), 34–41. DOI:https://doi.org/10.1145/1107499.1107503

13. Heikki Laurila (2020) Manual Data Errors, https://blog.beamex.com/manual-data-entry-errors#What-about-the-1-percent-error

14. Moghaddasi H, Sajjadi S, Kamkarhaghighi M. Reasons in Support of Data Security and Data Security Management as Two Independent Concepts: A New Model. Open Med Inform J. 2016;10:4-10. Published 2016 Oct 28. doi:10.2174/1874431101610010004

15. Helad11 (2020), Open Source is free - but expensive, https://dev.to/helad11/open-source-is-free-but-expensive-3h8a

Genestack