SciNote

Research Data Management Guide: NIH 2023 and European Commission policies

Contents

SciNote electronic lab notebook is trusted by researchers at the FDA, NIH and European Commission.

Two major research financing bodies, the National Institutes of Health (NIH) in the USA and the European Commission (EC) in the European Union independently released policies that promote the management and sharing of scientific data generated from NIH-funded and EC-funded research.

Their main predicament is empowering the people, businesses, and organizations by ensuring they have access to the data as this is essential for innovation and growth of the economy and society. This goes in line with general trends for data-driven economies around the globe.

In this article, we'll have a look at how NIH and EC policies relate to data management practices for researchers and scientists and how electronic laboratory notebooks can help.

Investment in data management and data accessibility

By 2025, the total amount of genomics data alone is expected to equal or exceed the total amounts from the three major producers of large amounts of data: astronomy, YouTube, and Twitter.

Other scientific fields are not really lagging behind.

Data exists in many formats, which complicates the ability of researchers to find and use research data generated by others. In addition, data usually requires extensive data "cleanup". This contributes to data scientists having to spend most of their work time (about 80%) collecting existing data sets and organizing data. That leaves less than 20% of their time for innovation e.g. tasks like mining data for patterns that lead to new research discoveries. Without getting into the discussion about who data scientists are, it is safe to say that in smaller research groups, a data scientist and a researcher are often the same person.

"Data value chain" offers many opportunities for improving capture, access, sustaining, and reuse of high-value data in order to put it to optimal use. Open science initiatives are aiming to improve that and both NIH and EC boarded the open science train.

Artificial intelligence/machine learning (AI/ML) is starting to play an increasingly important role in science. For an AI/ML algorithm to deliver results and predictions, it has to be trained first. And training is performed on quality training data sets. The foundation of quality data sets is good data management practice.

In biomedicine, data is being generated with increased volume in many aspects: genome, epigenomics, gene expression (transcriptome), proteomics, metabolomics, functional and phenotypic measurements, and ecological and lifestyle properties. Each of these dimensions has its own data types that were obtained with specific analytical methods. In a way, we are examining one or a few data types at a time and are able to assemble only a part of the puzzle. The factors necessary to understand a complex biological phenomenon such as disease, cannot be captured by a single data type. Therefore much of the complexity in biology and medicine still remains unexplained and will remain unexplained if we rely on these single-data-type studies. It is thus critical to integrate diverse sources of information to obtain a comprehensive understanding of biology and medicine. Machine learning approaches that can integrate data from diverse sources are becoming available and are opening great opportunities. With such approaches, we will ultimately be able to predict how drugs affect the human body (drug-target interaction prediction, drug-drug interaction, and drug combination prediction), repurpose drugs, predict disease subtypes, and discover new biomarkers so that we could have more effective treatments for them.

For such approaches to materialize, data first needs to be accessible. Therefore any investment in data management and data accessibility is a valuable investment.

Learn more:

SciNote Data Protection White Paper

Focus towards open science

Sharing knowledge and research data have become an important foundation open science initiatives are based on. Major research financing bodies, i.e. the European Commission (EC) in the European Union and the National Institutes of Health (NIH) in the USA see it as an integral part of scientific research. Adherence to the FAIR principles (Findable, Accessible, Interoperable, and Reusable) seem to be at its core as well.

Let's have a look at both EC and NIH initiatives.

NIH policies - background

NIH is a major investor in USA biomedical research and has been promoting policies that make research available to the public from the early 2000s: the 2003 NIH Data Sharing Policy, the 2008 Genome-Wide Association Studies Policy, NIH's 2014 Genomic Data Sharing (GDS) Policy, the NIH Policy on the Dissemination of NIH-Funded Clinical Trial Information (Clinical Trials Policy). In 2015 NIH initiated the development of a more comprehensive data-sharing policy along with modernization of data-sharing infrastructure with its 2015 Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research.

NIH then started seeking feedback from the community to develop a new data sharing policy, which happened in several steps during the course of 2016 and 2018. Using the public feedback, NIH released a draft proposal for a future data management and sharing policy and requested comments on it in 2019 (84 FR 60398).

Finally, NIH considered all feedback and developed the final Data Management and Sharing Policy (DMS policy) and published it on 29th October 2020.

On January 25th 2023, the DMS policy will come into force, and replace the 2003 NIH Data Sharing Policy. Along with the DMS policy, NIH also released supplemental materials that could help researchers integrate effective data management and sharing practices into research.

NIH has also committed to the financing of data repositories for public data.

EC policies - background

The vision for an e-infrastructure that will ensure sharing and preserving of reliable data produced during the scientific process in the European Union dates back to 2010. The High Level Expert Group on Scientific Data emphasized the critical importance of such infrastructure and provided a vision and plan for it in the final report "Riding the wave: How Europe can gain from the rising tide of scientific data". The group recognized that science has a pivotal role in creating economic growth and a fairer, happier society. Therefore Europe must manage the digital assets its researchers generate. EC acted on this proposal and eventually a series of Open Access (OA) initiatives and infrastructural projects started.

One of them was the European Open Science Cloud (EOSC), envisaged in 2016 within the "European Cloud Initiative" and formally launched as a platform in 2018.

The main aim of EOSC is to connect the existing research data infrastructure in Europe and realize a web of FAIR (Findable, Accessible, Interoperable, and Reusable) data and supporting services for science, making research data interoperable and machine-actionable following the FAIR guiding principles. EOSC is becoming a fundamental enabler of Open Science and of the digital transformation of science in the EU. The EOSC infrastructure has been continuously improving and will open up gradually to the public sector and industry.

EC is taking Open Science and Open Access seriously. It has been enforcing open access to scientific peer-reviewed publications and research data that beneficiaries have to follow in projects funded or co-funded under Horizon 2020 (the EU funding program for research and innovation running between 2014 and 2020).

While the open access to data was still optional and part of an Open Research Data pilot program within Horizon 2020, that will no longer be the case in the next EU funding program, Horizon Europe (running between 2021 and 2027).

Pharma companies and open science initiatives

It is not only public research that is promoting open science, there are big changes happening in pharma research as well. Large pharma companies essentially depend on research that is filling up their pipelines and they invest substantial funding into it. It is no secret that knowledge lies in data, which is a source of untapped potential.

Pharma companies have recognized the importance of good data management practices such as FAIR principles and digitalization of their research activities in general. These companies are building centralized data management platforms across their global research sites, running digitalization initiatives within their research sites, and raising the awareness and importance of good data management in places where data is generated, including labs.

Although these data management initiatives in pharma companies are internal, they still resemble the open approach of NIH and EC, because many pharma companies do not have centralized research facilities. The challenges are very similar to research going on in NIH and EC-funded research projects. What is interesting is that large pharma companies such as Pfizer, Novartis, Merck, GSK, Jannsen are sharing their data management initiatives, at least at a high level, back to the community. In a way, there are many similarities with the open science initiatives that NIH and EC are promoting. And data is at the core of all of them.

What does NIH data management policy say?

When does the NIH data management policy come into force?

The official and final NIH Policy for Data Management and Sharing (DMS policy) was released on 29th October 2020 and will come into force on January 25th 2023.

Who will it apply to?

This means that for all the grant applications, proposals, contracts, etc. that are submitted to NIH on or after 25th January 2023 the DMS policy will apply. DMS policy applies to all research that results in the generation of scientific data and is funded or conducted in whole or in part by NIH.

What is the purpose of this policy?

The purpose of NIH's DMS policy is to make the results and outputs of NIH-funded research available to the public through effective and efficient data management and data sharing practices. DMS policy also provides some basic definitions of terms such as "scientific data", "data management", "metadata", "data sharing", and "data management and sharing plan".

What will researchers need to submit?

In its essence, DMS policy requires that researchers who are applying for NIH funding and are planning to generate scientific data, need to submit a Data Management and Sharing plan (DMS plan) as part of their application for funding. DMS plan should explain how scientific data generated by the research project will be managed and which of these scientific data and accompanying metadata will be shared.

By requiring DMS plans to be sent when applying for funding, NIH is trying to promote a culture in which data management and sharing are recognized by researchers as an integral part of biomedical research projects, rather than an administrative one.

What is defined as "scientific data"?

Since the policy is focusing on scientific data it is worth spending a few lines explaining what NIH considers as "scientific data" and what not. This is the official definition of "scientific data":

"The recorded factual material commonly accepted in the scientific community as of sufficient quality to validate and replicate research findings, regardless of whether the data are used to support scholarly publications. Scientific data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens".

This means that you do not need to provide a DMS plan for every experimental plan or note that you document in your notebook (hopefully that's an electronic lab notebook). However, for the data comprising the outcomes of your studies, whether being "positive" or "negative" results, whether being "publication-worthy" or not, you will need to disclose how you will manage them.

Is the use of repositories encouraged and what does it mean?

With DMS policy NIH strongly encourages the use of established repositories for preserving and sharing scientific data rather than keeping data by the researcher or institution and providing them on request. NIH does not specify in which repositories data should be preserved and shared but they did release the "Supplemental Information to the NIH Policy for Data Management and Sharing: Selecting a Repository for Data Resulting from NIH-Supported Research", which should help researchers to choose suitable repositories for the preservation and sharing of data. It also lists some repositories.

Does all scientific data need to be digitized?

Interestingly, one of the points that NIH decided not to include in the final policy after receiving public comments was the expectation that research organizations will make reasonable efforts to digitize all scientific data. NIH still encourages this, just outside the official policy as they do not want any technical difficulties related to digitization of scientific data to be counterproductive and result in limitation of data sharing.

When should data be made accessible?

DMS policy also states that the data should be made accessible as soon as possible but no later than the time of an associated publication, or the end of the performance period, whichever comes first. NIH ICO may also review the compliance with the DMS plan during regular reporting intervals during the funding period.

What if the needed actions will add additional costs for the researchers?

It seems that NIH does recognize that implementing DMS policy will come at some cost for the researchers and came up with the document "Supplemental Information to the NIH Policy for Data Management and Sharing: Allowable Costs for Data Management and Sharing" which explains eligible costs related to the execution of DMS policy.

In summary, NIH is trying to raise the importance of good data management practices in research and is trying to do so without creating uniform requirements for sharing all scientific data. On the contrary, they are expecting that it will be varied. NIH also sees this as a foundation for improving reproducibility and reliability of research findings, to which they are continuously committed (NIH Rigor and Reproducibility). NIH encourages data management and data sharing practices consistent with the FAIR data principles.

What do EC data management policies say?

What was required prior to 2021?

The European Commission started to enforce open access to scientific peer-reviewed publications and research data for all beneficiaries of projects funded or co-funded under Horizon 2020 program that was running between 2014 and 2020 in the following way:

1. Peer-reviewed publications: All peer-reviewed scientific publications arising from Horizon 2020 funding had to be made available in open access, which could be achieved either by publishing in an open access journal or by depositing the publication in a repository for scientific publications which ensured open access.

2. Research data: All Horizon 2020 projects were required to develop a Data Management Plan (DMP) and had to provide open access to research data supporting publications; other research data specified in the DMP, such as raw data, could be presented as well. The requirements for open access to research data were still a part of the so-called Open Research Data Pilot within Horizon 2020, into which all Horizon 2020 projects were enrolled automatically, but they could choose to opt-out without any effect on the proposal evaluation.

How does the 2021 - 2027 Horizon Europe funding program differ?

However, in the research and innovation funding program, Horizon Europe 2021 - 2027, EC is requiring Open Access for both research publications as well as research data by default. It still allows opting-out possibilities in duly justified cases under the principle “as open as possible, as closed as necessary”.

Which information should the data management plan include?

As part of making research data findable, accessible, interoperable and re-usable (FAIR), a Data Management Plan should include information on: the handling of research data during and after the end of the project, what data will be collected, processed and/or generated, which methodology and standards will be applied, whether data will be shared/ made open access and how data will be curated and preserved (including after the end of the project). DMP should be submitted within the first 6 months of the project and should be updated over the course of the project whenever significant changes arise.

Which platforms are helpful?

EC is strongly emphasizing open science and open access in Horizon Europe and is promoting OpenAIRE platform as an entry point to determine what repository to choose, EOSC as a virtual environment to store, share, process, and reuse research digital objects (like publications, data, and software) that are Findable, Accessible, Interoperable and Reusable (FAIR). EC also launched its own scholarly platform dedicated to publishing open access peer-reviewed publications for grants funded via Horizon 2020 and Horizon Europe in 2021 (Open research Europe platform). With these efforts, EC is trying to take away at least some burden of ensuring open access.

What if the needed actions will add additional costs for the researchers?

EC is recognizing costs associated with data management, including the creation of a Data Management Plan, as eligible costs in any Horizon 2020 or Horizon Europe grants.

What are your next steps?

Analyze how the work is being conducted in your lab. Map the process and the flow of data. Include the team and identify the potential bottlenecks. Are we spending time searching for data, notes, previous entries, files, or similar? How quickly can we find what we need? How quickly and accurately can we retrieve all related information to a specific result? Can we do that for things that date a few years back? How will we do it in the future? These are just starting points to define your lab's needs to save time and keep up with the pace and the direction the world is headed towards.

Electronic lab notebooks today cover the major part of the lab's operations related to project, team, inventory, and data management. Their task is to enable lab personnel to track, record, retrieve and build upon their research data.

Although NIH does not count the individual notebook entries as scientific data, it is important to understand that eventually, the conclusions of research will have to be made publicly available, e.g. publications with links to supplemental data, data uploaded to various data repositories, etc. Although not requiring the digitization of data, NIH is encouraging it.

Similarly, the EC is not so specific in requiring that research should or should not be conducted using digital tools. However, it is clear that all the open access to data that NIH and EC are promoting is, after all, access to digital data.

Having data in a digital form from the very start of your research work does not only increase the efficiency of the research process but is a step forward towards FAIR data practices and easier sharing of data.

As a researcher or a scientist, you are often very much involved in your work and devote a lot of effort and time to your research. You plan your work, document results, and when you're confident enough that what you've been working on really works you start presenting at conferences, and finally you publish the work. Somewhere along the line, it all starts to feel very personal and you start referring to the work as your work, to the results as your results, to the lab notebook as your notebook.

This immersive experience and devotion to so many details sometimes blur the bigger picture of your role as a scientist or researcher and the purpose of your contribution to science. Which is, in the end, a contribution to society, which in turn finances your work. This is the case of course for research funded in total or in part by national research agencies and initiatives.

For quite a while science and research have been very much centered around publishing scientific outcomes in the forms of scientific papers. Scientists' and researchers' efforts are so focused on publications that researchers think the job is done when the work is published. The focus quickly shifts towards the next publication, putting aside that someone else might be benefiting from the data. Not just published data, but supporting data such as detailed results, raw data, protocols, and metadata that can help others to either validate the work, build on top of it and perhaps complement it, find hidden patterns, new knowledge, new discoveries.

For these purposes, data management and data sharing policies the research financiers are enforcing make sense. It is vital that every researcher, every scientist understands that data management and sharing is an integral part of the research process, part of the research method.

It is the individual researcher's responsibility to address that in the best way they can as this is a very important part of science's contribution to society. You might go on and say, "but I'm not a data scientist, I'm good at discovering the signaling pathways in cancer stem cells, so how am I going to manage data?". Science and research are collaborative fields, now more than ever. You are relying on your colleagues or partner research institutions to perform a part of the experiments they are specialized in, e.g. high throughput sequencing or metabolome profiling with LC-HRMS, to help you put the pieces of the puzzle together for publication, in the same way, it is your obligation or your lab's obligation to take care of data management and sharing and recognize that this is also a vital piece of the puzzle that completes your research.

Looking back on your published work one might think, how did we put such an elaborate work together? But every work starts out by planning and doing small bits one at a time. In the end, it all adds up. The same is with data management. If you include it into your research work from the beginning and include it into your plans and then execution of your work, it becomes a part of regular work. And here is where we believe electronic laboratory notebooks can really help: as you can manage your data one bit at a time as you go along. And when you'll need to identify data you would like to publish into a repository, it will no longer be a daunting task.

By Matjaz Hren, VP of Product Management at SciNote LLC

SciNote