How open databases turn out to be

crucial in the fight against Covid-19

CHRISTINE STANSBERG
node coordinator ELIXIR Norway, Department of Informatics, University of Bergen

NILS P. WILLASSEN
head of ELIXIR Tromsø
Department of Chemistry
University of Tromsø

GARD O. THOMMASSEN
deputy director of the University of Oslo Center for Information Technology, USIT

EIVIND HOVIG
head of ELIXIR Oslo
Department of Informatics, University of Oslo

INGE JONASSEN
head of ELIXIR Norway, Department of Informatics, University of Bergen

During the last nine months, we have become used to following the fight against the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and the resulting coronavirus disease (Covid-19) daily through news and social media. We hear about new mutations and what these might mean for the disease. We also hear about possible breakthroughs in vaccine developments, and in finding new drugs that may help in the fight against the virus. It is easy for us to picture clinicians and experimental biologists at work in hospitals and at lab benches trying to understand how the virus works, and how the disease hits. It is maybe harder for us to picture computer scientists at work as equally important players in this quest, but in reality, most of the findings and breakthroughs that we rely on in fighting Covid-19 (and other diseases) depend on the availability of numerous computational tools and collections of biological data that are built and maintained by computational biologists and bioinformaticians. This text will mainly deal with the databases, although the computational tools are just as important in this fight.

We first learned that the mysterious new pneumonia observed in China in December 2019 was caused by a new virus that resembled the virus from the 2002 SARS outbreak. This identification could be done quickly because researchers all over the world have been gathering all available information about viruses (and all living beings) into databanks (or databases) for several decades. Some databases gather all the available genetic information of the beings, while others gather information about their proteins. Yet other databases combine information from many different sources, for example genetic information from a Covid-19 patient combined with other information about the patient that may explain why they are badly affected, such as age, gender, and clinical data such as blood pressure, other diseases etc, making it possible to see the whole picture. And the data should be represented using standardized vocabularies in a machine-readable form, enabling integration and further processing in other contexts.

Databases containing all of the building blocks of SARS-CoV-2

In order to identify the new virus as a SARS-like virus, the genetic information of the virus, i.e. represented by the nucleotide sequence, had to be compared with the sequences of all previously known viruses that had been stored in the nucleotide databases, for example the European Nucleotide Archive (ENA)[1]. In its simplest explanation, this happens by sending the nucleotide sequence of the new virus to a tool that searches the database and returns the nucleotide sequences that are most similar to it. The comparison tools that are used will also give a similarity score that tells us how similar the virus is to the various other viruses that have already been stored in the database, and in this case it returned the previously described 2002 SARS virus, or the SARS-CoV to be correct. Using other bioinformatics tools, we can map out which parts of the new virus that are very similar to the previous SARS-CoV, and which parts that are different. We can track the changes in the nucleotide sequence, or the mutations, as the new virus spreads, using variations of the same methods, and we can use yet other bioinformatics tools to predict what these changes might mean for the different virus strains, whether they for example are likely to make them more infectious, or severe, or maybe both.

When it comes to the bioinformatics comparison- and prediction tools, we are quite lucky that these have been openly available and perfected for as long as we have been able to sequence nucleotide sequences, and therefore we did not have to start all over and make brand new tools for this particular situation. The same goes for the databases, we had a head start against the new virus because we immediately were able to see that it was quite similar to its relative, the SARS-CoV, which we now after two decades of research know quite a bit.

As it turns out, 'open' is a key word in the fight against SARS-CoV-2 and Covid-19, and most, if not all, of the efforts taken with regard to understanding the new virus rely on having previous data and computational methods publicly available. This may seem obvious but is in fact not. Maintaining a database and making sure that it is up and running and that everything inside is correct and up-to-date and at the same time structured in such a way that it is easy to use, alone or in combination with other databases and tools, requires enormous efforts and depends on having highly qualified personnel available that knows exactly how to do this. These people of course want to be paid for the important job that they do, and in addition, there are costs associated with storing all of the data on servers.

The easiest way to solve this would be to ask all users of the database to pay a fee to cover the costs associated with sustaining it. But researchers tend to not want to pay for services they do not use. We could quickly get a situation with isolated domain-specific databases that did not 'talk' to each other, where databases that did not currently cover a hot topic, lost users and died out due to lack of funding. On a global scale, relatively few researchers probably would have paid subscription fees for a corona virus-specific database in the long term, and in a world where this was the general funding model of biological databases, it could very well have suffered this fate. Biology is very unpredictable, and although epidemiologists have long warned us about the probable emergence of global pandemics, we were completely unprepared for the causative agent being a coronavirus.

BOX 1
About ELIXIR and a few selected ELIXIR Core Data Resources

elixir	The European research infrastructure for Life Science data
ENA European Nucleotide Archive	Nucleotide sequences, raw sequencing data, sequence assemblies, functional annotation
UniProt	Protein sequences, functional information
e!Ensembl	Vertebrate genome browser
EUROPEAN GENOME-PHENOME ARCHIVE	Personally identifiable genetic and phenotypic data from biomedical research projects
PDBe Protein Data Bank in Europe	Biological macromolecular structures

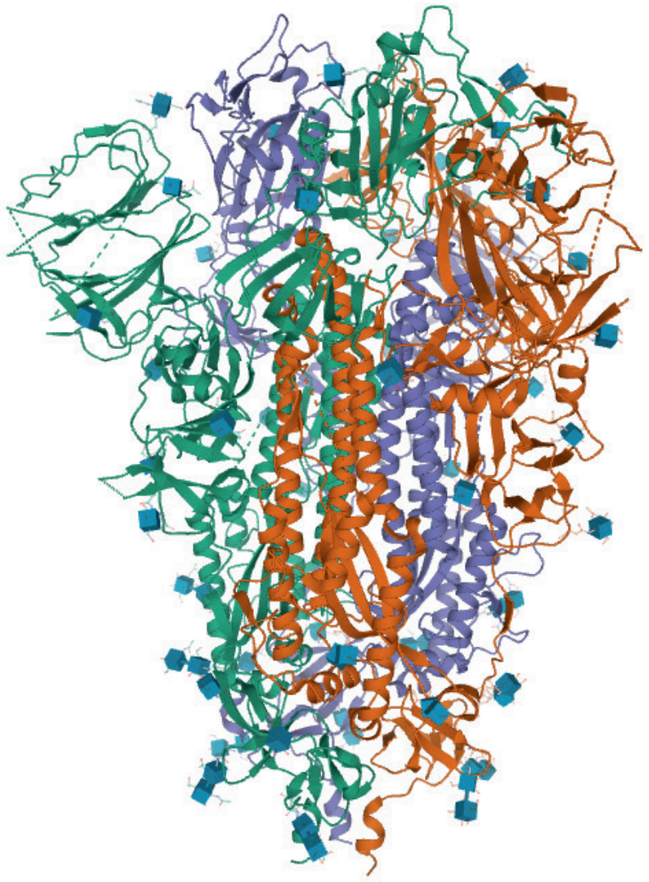

FIGURE 1.
The structure of the SARS-CoV2 Spike protein as modelled in submission 6X6P, openly available through a quick browse in the Protein Data Bank Europe9 (Herrera NG., Biorxiv 2020)

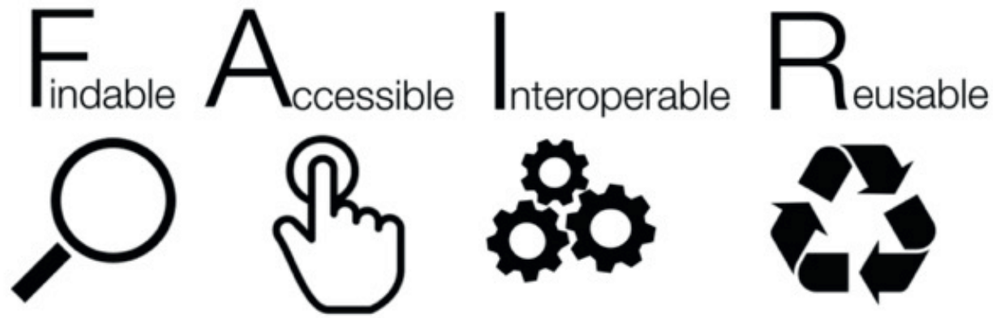

FIGURE 2.
The FAIR principles[12]

In reality, most biological databases initially emerged as a result of research projects funded by a national or international research agency. Some databases have proved to be so important for so much ongoing research that they have continued to attract public and private funding over many years, whereas others struggle and rely to a large degree on the volunteer work of the few researchers behind the resource.

ELIXIR is the home for the most crucial European databases

On this backdrop, ELIXIR, the European research infrastructure for life science data was established in 2013, to bring together and safe-guard life science resources developed and maintained all over Europe[2]. The ELIXIR collection of resources is not limited to databases, but also includes computational tools, training material, cloud storage and access to supercomputers. The goal of ELIXIR was, and still is, to coordinate all of these resources so that it is easier for researchers to find and share data, exchange expertise and agree on best practices. The bioinformatics communities in Norway have collaborated on providing similar services to Norwegian researchers for almost two decades, originally funded by the FUGE programme of the Research council, and this structure was the basis for the formation of the Norwegian Node of ELIXIR in 2013[3].

Having a database included in the list of ELIXIR resources generally increases the visibility and usage of the resource and gives national funders an extra incentive to keep on funding it. In addition to several tools, ELIXIR Norway currently provides LiceBase[4], a database including all available genomic information about the Salmon louse, and the Marine Metagenomic Portal[5], that collects all available genetic information about marine microbes, including a collection of marine viruses. We have recently opened a call to include other Norwegian services[6] and hope to see this list grow in the future.

On top of this collection of services, the ELIXIR community has started a process to identify databases that are absolutely crucial for the everyday research of the majority of European, and also international, life scientists. As a result of this, a number of ELIXIR Core Data Resources have been selected through a careful evaluation process[7] (Box 1) and will receive extra attention from ELIXIR to make sure that these always are updated, are up and running, and remain open for life scientists and industry also in the future. In part inspired by the effort by ELIXIR, a global coalition of funders is being formed these days. One of the main objectives is to identify core data resources on the global level and work to ensure their sustainability. The Research Council of Norway is one of the founding members of this Global Biodata Coalition[8].

As outlined above, the fight against Covid-19 and SARS-CoV2 relies on many of these resources, such as comparing virus sequences with all available sequences present in ENA, and using the Protein Data Bank (PDB)[9] to make models of the 3D structure of the proteins (Figure 1) in order to understand the working of the virus, and to help develop vaccines and drugs. In fact, the European Bioinformatics Institute (EBI) has made an effort to measure how much the society in general benefits from having these databases openly available to all and found that the value for the society equals about 20 times the investment it takes to operate and maintain these databases[10]. Some of the issues taken into account in these calculations is the fact researchers save time from not having to generate all the data and tools themselves each time.

Why it is only FAIR to have open data

The current Covid-19 pandemic has spread extremely rapidly throughout the world, and has effectively shown how important it is that researchers and clinicians are able to immediately share the results and knowledge that they have gathered with international colleagues, and similarly that they are able to access what others have produced. When new mutations are observed in local outbreaks, like the recent one observed in Trondheim in October[11], it is urgent to compare the sequence of this particular strain to all other available sequences, both to understand what the mutations may mean biologically, and also to understand how this strain may have spread through the population.

The databases (and all services) provided by ELIXIR follow a set of principles to ensure that data can be found, interpreted and used together with data from other databases, and with the user's own experimental data. These principles are called FAIR[12] and call for all researchers to make their data Findable, Accessible, Interoperable and Reusable for others (Figure 2). That means that we should not only allow our colleagues to Find our data in a publication, but also to make sure they may Access them and Reuse them in combination with other data types in other settings, meaning that they are Interoperable. In order for this to happen, standards need to be developed and utilized across scientific fields and national borders.

In the case of SARS-CoV-2, we observe that a lot of the sequences are stored in the GISAID[13] database that was initially established to ensure that we were able to prepare for the annual Influenza outbreaks on a global level. For historical and commercial reasons, the data submitted to GISAID are not FAIR. The circumstances around Covid-19 are very different from those of the annual influenza outbreak and because of the high level of urgency, we want to encourage all Norwegian Covid-19 researchers to publish their sequences in an open database, for example ENA, and we are eager to help researchers to do this.

Human data need to be treated in a special way

In order to understand how the virus attacks us, and why the attacks become more

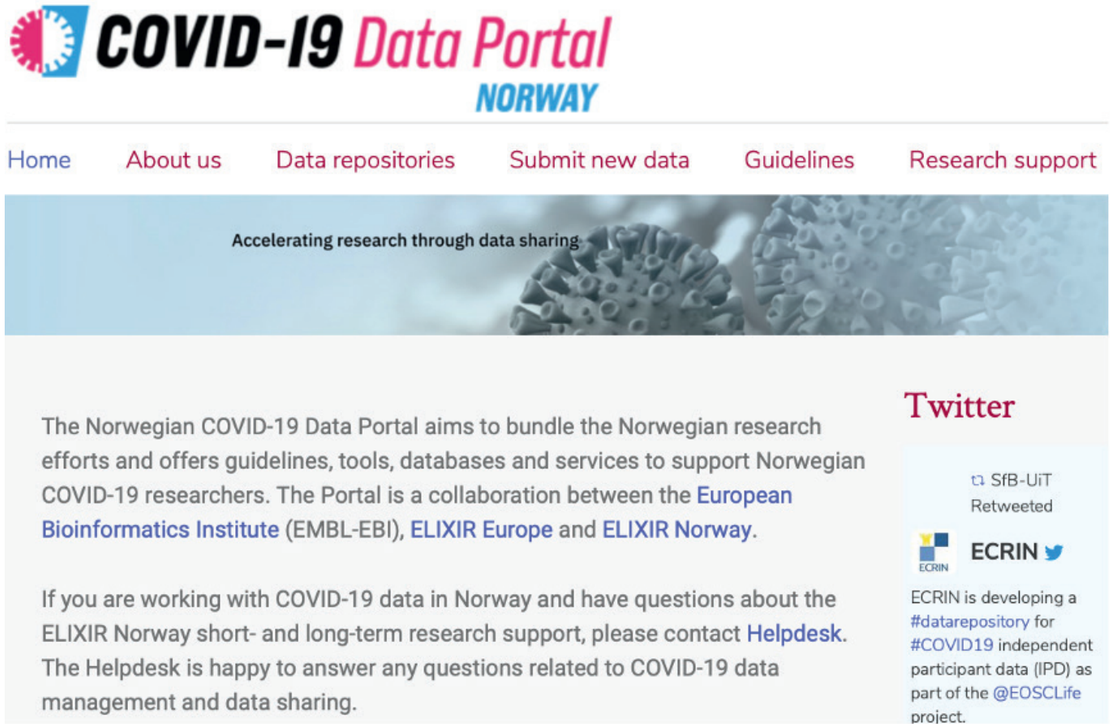

FIGURE 3.
A snapshot of the Norwegian Covid-19 Data Portal17, with quick links to data submission, repositories and support.

severe for some more than others, we also need to study patients who have caught the disease. We need to compare the genetic sequence, or genome, of the patients who have become critically ill, with the ones who have only been mildly hit or not affected at all. Since each study is limited in size, it is useful to use data from several studies to get a better picture. Again, the FAIR principles come into play. However, working with human data and human genomes gives a whole new set of challenges that we have to face. Privacy legislations ensure that personal information cannot be openly shared and re-used for other applications than for those purposes for which the consent originally was given. Handling these personal data securely is further complicated by the enormous magnitude of genomic data.

One of the ELIXIR Core Data Resources addresses exactly this challenge. The European Genome Phenome Archive (EGA)[14] is a secure databank for human genomes and other data associated with these (phenotype/clinical data and metadata). Whilst this information is securely stored in a FAIR manner in one of two 'vaults' in Barcelona or in Cambridge, European researchers may get access to search in the bank for information that may be relevant for them, and then to apply for access if they find something of interest. However, Norwegian privacy legislation imposes restrictions on sensitive Norwegian data leaving the country. In ELIXIR Norway, we are therefore collaborating with the rest of ELIXIR to make national 'branches' of this human data bank that will store national genomes. As a result, EGA becomes federated, and each country operates a node of this. In this way, we can store data from Norwegian individuals safely in the Norwegian 'branch' of the bank and make available high-level information about each study in our 'branch' through the central EGA databank. Researchers outside Norway can apply for access to a Norwegian data set and if this is granted by a data access committee assigned by the data owners, they can review and analyse the data within the Norwegian 'branch' without the data ever having to leave the branch or the country, or the EGA node.

In ELIXIR Norway, we hope to be able to launch the Norwegian EGA node very soon, as one of the very first in Europe. In order to do this, we collaborate closely with the University of Oslo Centre for Information Technology, USIT. USIT has, in collaboration with Uninett Sigma2 AS, developed a national service for sensitive data that maintains the strict privacy legislations for handling sensitive human data, and this service, TSD[15], will host the Norwegian node of EGA.

A common portal to access everything Covid-19

Even if lots of results from Covid-19 research are made openly available, it may still not necessarily be easy to find or share data. In order to make these steps a little bit easier, the European Commission, the European Molecular Biology Laboratory (EMBL) and ELIXIR have collaborated on building a common portal that collects all openly available information about SARS-CoV-2, and how we react to the virus. The resulting Covid-19 Data Portal[16] guides visitors towards a large number of viral sequences that have been read all over the world, and also to a considerable number of sequences from humans and other hosts. There are also pointers to information on how the proteins of the virus might function, and on how human host cells try to fight the infection.

Similar portals are now also being set up on a national level to guide national researchers towards the joint European efforts, and also to guide them on how they may get practical support locally ELIXIR Norway recently launched the Norwegian Covid-19 Data Portal[17] as one of the first national nodes (Figure 3). This national site quickly directs visitors to the research support provided by the Helpdesk of ELIXIR Norway and also guides them on how to share and find Covid-19 data. In fact, at the onset of the pandemic, researchers at the Tromsø node of ELIXIR Norway quickly set up a portal that collects all publicly available SARS-CoV-2 sequences, by reusing the framework of the portal that they previously established for marine viruses in the Marine Metagenomic Portal described earlier in this text. The SARS-CoV-2 portal18 also adds all available information about the context in which this particular virus was found (metadata) in a process called curation. Examples of such metadata can be when and where the virus was found, the age of the host, whether they were female or male and the organ from where the sample was made. Through the SARS-CoV-2 portal, we may also visualise the countries that have provided these sequences openly.

A safe space for Covid-19 clinical trial results set up in Norway

All over the world, potential vaccines and therapeutics for Covid-19 are now being tested through clinical trials. Many life-science oriented research infrastructures in Europe already collaborate to build an environment where life science researchers may analyse, share and store data in the European Open Science Cloud (EOSC-Life)[19]. In response to Covid-19, this part of the cloud is now being rigged so that it may also support sharing of data on clinical trials of vaccines and therapeutics against Covid-19. To enable the use of data from several clinical trials together in an optimal way, data on individual patients need to be shared. In order to do so, a secure infrastructure solution is needed to protect the data vault that contains the human data. We mentioned above that the solutions developed by TSD and USIT and ELIXIR Norway for the Norwegian branch of the EGA database are quite advanced on the European level. On this background, TSD and ELIXIR Norway were recently challenged by ECRIN[20], the European research infrastructure for clinical trials, to develop a secure repository that will allow storage and sharing of results from Covid-19 clinical trials – and at a later stage also data from other clinical trials. This work is progressing nicely, and a beta version is expected to launch February 1st, 2021.

To conclude, the Covid-19 pandemic has proven to be an effective use case to demonstrate the importance of sharing research results openly in a way that these are easy to find and reuse for other researchers. In many ways, it is also an incentive to intensify work to enable effective solutions for enrichment and reuse of data resources. We are confident that the efforts that have been put down in the bioinformatics world to fight Covid-19 will allow us to respond even more strongly to the next emerging pandemic.

REFERENCES:

1. https://www.ebi.ac.uk/ena/browser/home
2. https://elixir-europe.org
3. https://elixir.no
4. https://licebase.org
5. https://mmp.sfb.uit.no
6. https://elixir.no/news/41/63/Interested-in-contributing-a-bioinformatics-service-to-ELIXIR https://f1000research.com/articles/5-2422/v2
7. https://globalbiodata.org/
8. https://www.ebi.ac.uk/pdbe/
9. https://beagrie.com/static/resource/EBI-impact-report.pdf
10. 10 - https://www.dagbladet.no/nyheter/ny-variant-pavist-i-trondheim/72966602
11. https://www.nature.com/articles/sdata201618
12. https://www.gisaid.org
13. https://ega-archive.org
14. https://www.uio.no/tjenester/it/forskning/sensitiv/
15. https://www.covid19dataportal.org/
16. https://covid19dataportal.no
17. https://covid19.sfb.uit.no
18. https://www.eosc-life.eu
19. https://ecrin.org