A modern approach to IARC's data

Request for support from the Governing Council Special Fund

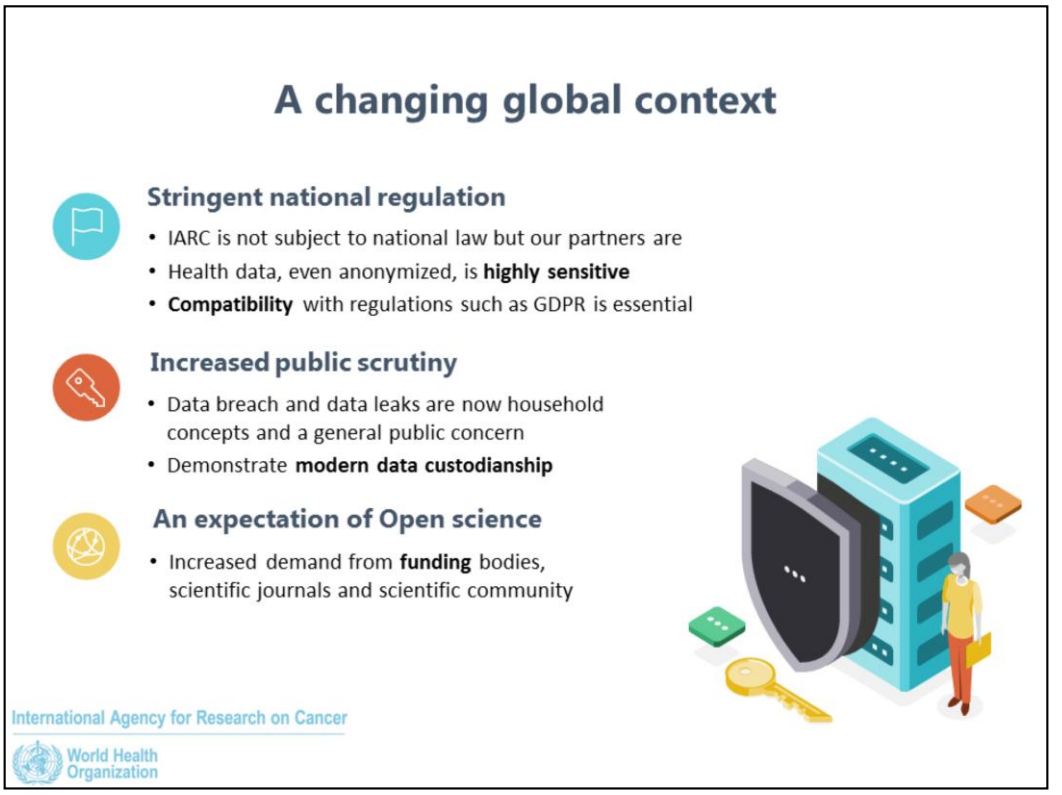
International Agency for Research on Cancer
Lyon, France

International Agency for Research on Cancer
World Health Organization

In today's digital world **data** has become an incredibly powerful resource. Commercial companies and private organizations strive to accumulate data about individuals. Personal data, has become a currency for many.

However, the world in which IARC operates sees **data** in a different manner, data at IARC is the resource we use to try and help reduce the world's cancer burden.

Since 2016 we have **increased** our efforts and with this proposal we aim to continue that evolution by introducing a modern approach to IARC's scientific data



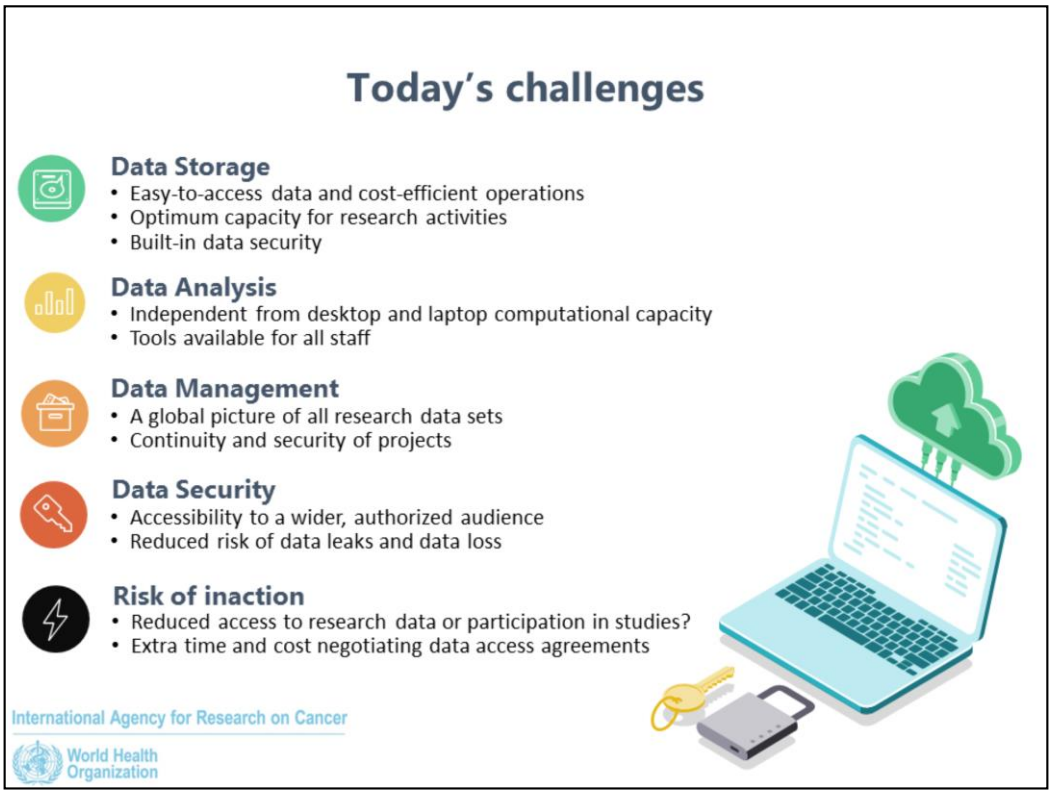
Since the enforcement of GDPR in May 2018 we have seen increasing national data protection regulation efforts, most notable but not exclusively in Europe.

IARC, as an inter-governmental organisation is not bound by these national laws, however and **critically** most of our partners are. How can we provide our partners with the assurances they need to share **their** data with us when we can not be held accountable in the same manner as them? It is evident that health data is one of the most sensitive categories of data that exists and quite reasonably IARC must be able to demonstrate it's **compatibility** with these national regulations and best practices.

Terms such as data protection, data breach and data leak are no longer reserved for those working within Information Security or Data Protection, these are now **household concerns**. This increased public awareness makes it vital that IARC shows it's modern data **custodianship** practices to the wider world.

Discussion within our Committee for Information Security Oversight (CISO) and our Computational Biology, Bioinformatics and Biostatistics Committee (C3B) have show us the importance to continually evaluate the way we work and how best to adapt to the changes around us.

We recognize that data protection and information security is a continuous process of evaluation and evolution and that we must always be trying to improve. With the help of our Scientific, IT and Administrative colleagues we have been able to identify some of the challenges we currently face.



Data storage needs to be **easily** accessible by all staff members who have be granted access, it should be easily maintainable by IARC's IT team and the capacity should not be a **limiting** factor for our research activities. The architecture should enable built-in data security.

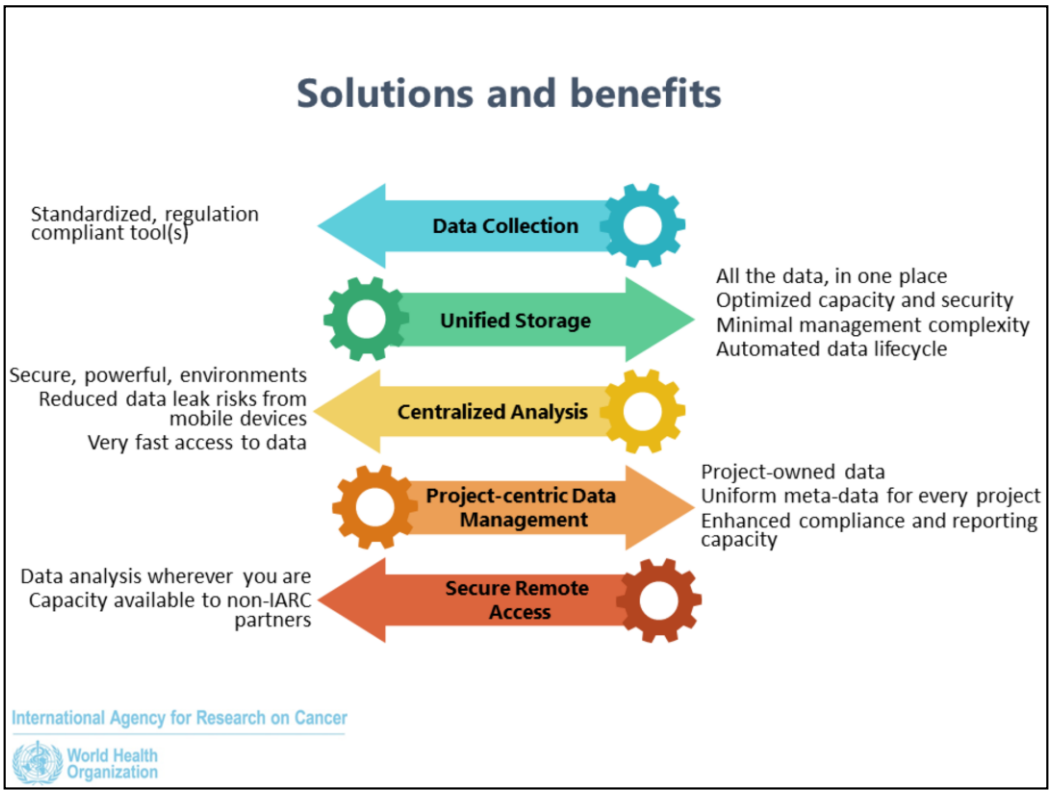
The **analysis** of our data should be carried out in the most appropriate place for the type of analysis being performed. Our staff should not be limited by the computational resources of desktop or laptop computers, neither the availability of the actual tools themselves.

The **management** of our research data should help enable a **global** overview of all of the data held at IARC; the type of data set, the provenance of the data and the area of study are a just few examples of the meta-data we should attach to it.

Data security should enable IARC to provide **access** to a wider authorized audience of scientific colleagues and it should compliment our existing best practices in order to help us **continue** to eliminate any risks of data leaks or data loss.

So, how do we propose to tackle these challenges?

We envisage the creation of chain of **custodianship** based on modern technologies and best practices which will provide security and protection of data at IARC at every step of it's lifecycle.



We aim to provide IARC scientists with a set of standardized, **regulation complaint** tools and to build capacity through education and training.

Data will stored in a single **centralized** repository. This will require us to invest in and increase our total storage capacity, increase the usage efficiency by implementing automated data lifecycle management and provide a long term archive for our historical data sets.

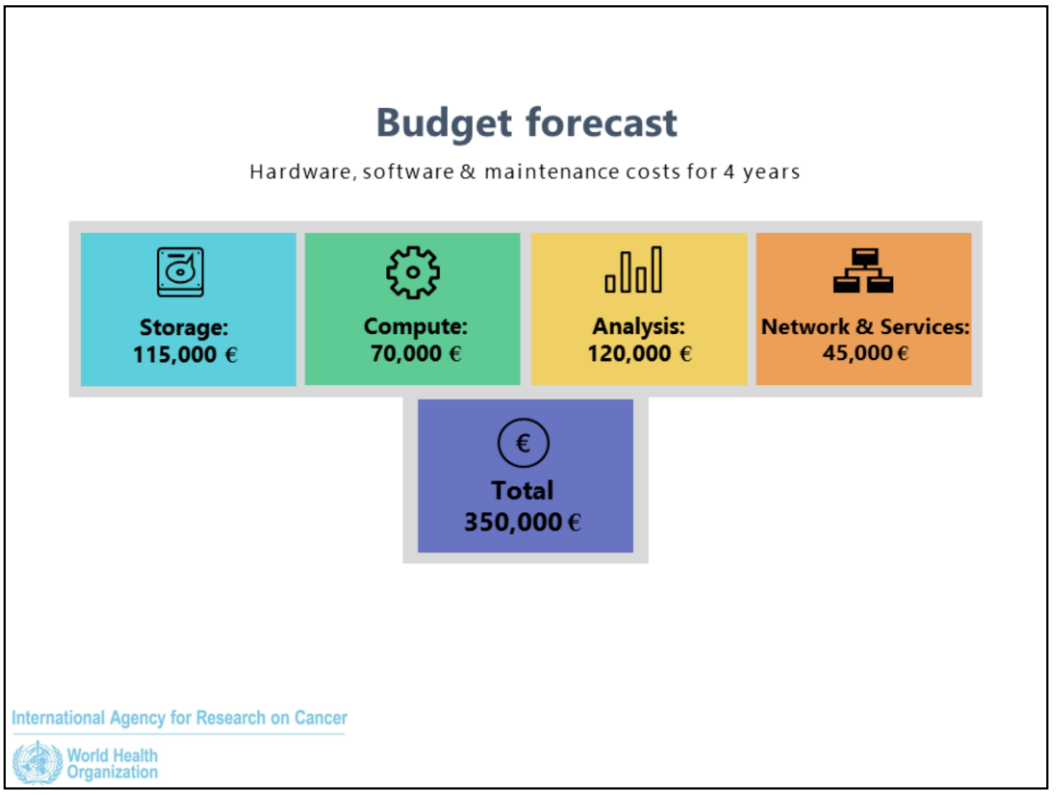
We must provide the ability to analyse our data without the need to move it to a different environment, helping ensuring we retain control over the **security** and **protection** of the data. By providing dedicated servers for this analysis we will remove any limitations of current desktop and laptop computers where some analysis may currently be taking place as well benefiting from the proximity of the data to the computational environments.

Building on the success of our High Performance Computing data management practices we will ensure that each data set is stored in a project construct with associated **metadata** such as PI, project description , data manager, data users, DTA reference, Project portal reference, and so forth.

By following this p**roject-centric data** approach we can more easily ensure that access and the use of the data is in compliance with any scientific, ethical or contractual obligations, as well as providing senior IARC staff with a **global overview** of all the data we have at the Agency.

Combining these elements will provide us with the technical tools to allow IARC scientists the ability to access and perform analysis on the data wherever in the world they are, in a secure and efficient manner without the need to move the data from it's central location.

Coupled with the legal expertise of IARC colleagues we can envisage that under certain, strict, conditions this capacity could be opened to our collaborators around the world. Allowing them too to benefit from our **secure analysis environments** whilst ensuring rigid data access controls.

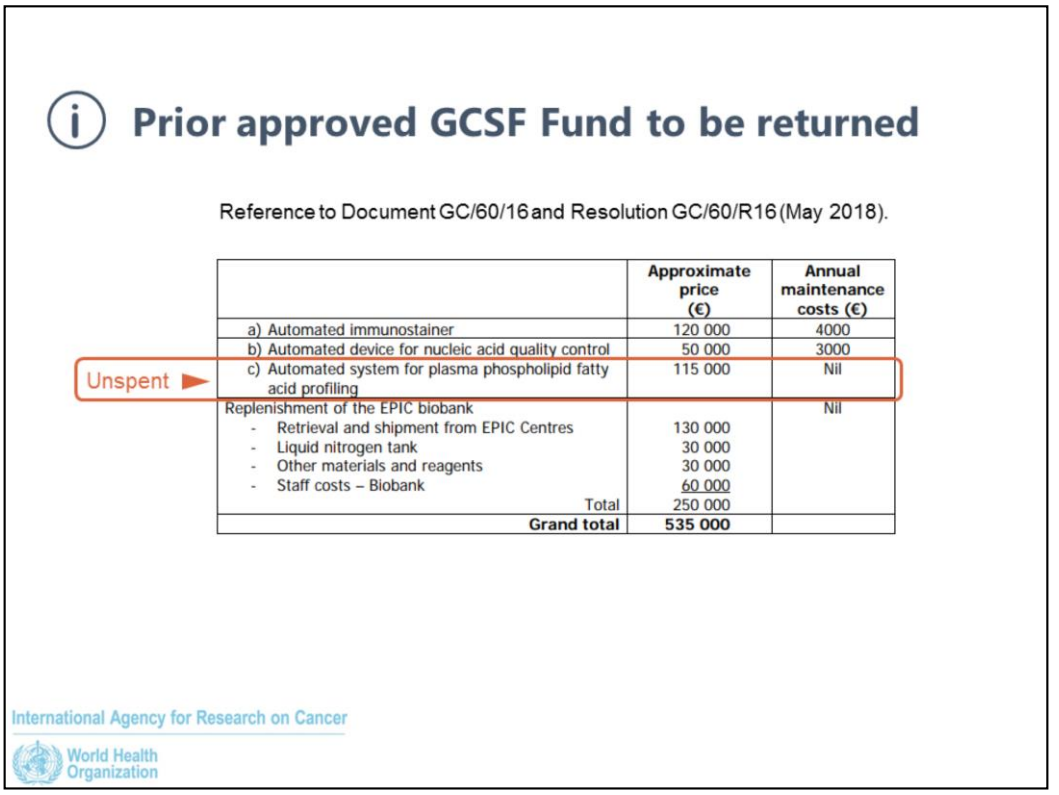


How does this proposal translate into physical equipment and software? In order to provide the centralised **repository** we will need to invest in storage systems and capacity, ensuring the modularity of the chosen system in order to grow capacity as required.

For the centralised analysis environments we will need to invest in both dedicated servers to undertake the analysis, as well as the software tools, such as SAS Server for example, and their associated licenses providing all staff with **equitable** access to these tools.

Investment in network technologies to interconnect these environments **plus** professional services where required is the last brick in our proposed budget.

In conclusion, we believe that this proposal will enable IARC to build on it's current data protection practices and allow us to implement a secure end-to-end solution that will provide our partners with the **reassurances** they require to continue sharing data with us, show our willingness to be **compatible** with national regulations and provide IARC scientists, and potentially our collaborators, with a powerful, secure and fit for purpose **Scientific IT platform**.



	Approximate price (€)	Annual maintenance costs (€)
a) Automated immunostainer	120 000	4000
b) Automated device for nucleic acid quality control	50 000	3000
c) Automated system for plasma phospholipid fatty acid profiling	115 000	Nil
Replenishment of the EPIC biobank		Nil
- Retrieval and shipment from EPIC Centres	130 000	
- Liquid nitrogen tank	30 000	
- Other materials and reagents	30 000	
- Staff costs – Biobank	60 000	
Total	250 000	
Grand total	**535 000**	

In 2018, Governing Council supported the request from Secretariat to purchase an automated system for plasma phospholipid fatty acid profiling and approved the budget of 115,000 euros for this purchase.

Due to the change of personnel and reprioritization of research, the concerned Group (NEP) felt that the purchase of such equipment would not be necessary hence returned the fund to be used for other purposes.