Benchmarking Report

European insights into research data centres

October 2020

tech4Germany_

unter der Schirmherrschaft des Chefs des Bundeskanzleramt, Prof. Dr. Helge Braun

Bundeskanzleramt

Intro

Tech4Germany project in the German research centre for health data

Deep Dives

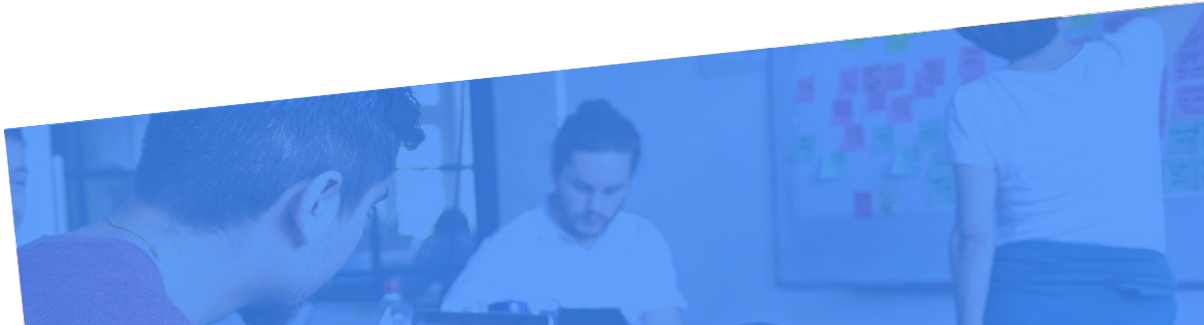
Addressing the interests of citizens, offering technical data access and permitting

Best Practices

European best practices from research centres

Vision

Outlook & recommendations



tech4Germany_

Intro (1/2)

What is Tech4Germany?

Tech4Germany is the technology task force for the federal government under the patronage of the head of the Federal Chancellery. Over a period of 12 weeks, the program brings together digital talents and those responsible in the public administration. Tech4Germany fellows work in an interdisciplinary team with representatives of the ministries to realize a digitization project.

The project "Forschungsdatenzentrum"

Four of the Tech4Germany fellows worked with representatives of the BfArM Institute and the Federal Ministry of Health to advance the access of researchers to health data. The German research centre for health data "Forschungsdatenzentrum" was incorporated into the BfArM Institute in 2020 and is currently being redesigned.

Project results

On the one hand, a prototype was developed for the new Forschungsdatenzentrum. A new portal should support the researchers to make specific requests for access to the health data. Interviews were held with many researchers and research institutes to better understand their current problems. In order to solve these problems, a few functionalities of the new portal were then designed in a user-centric approach. User feedback was incorporated into the design of functionalities.

Furthermore, further interviews were held with European research centres. The goal was to support the redesign of the "Forschungsdatenzentrum" by presenting European data centre best practices.

Intro (2/2)

Relevance of the benchmarking

It is crucial to learn from the practices of other research centres, especially from European neighbors with similar challenges and the legal framework of the GDPR. We can learn from it and see concretely, how different interpretations of legal frameworks are put into practice. It should be taken into consideration that every research centre is designed and set up in a different way and owns a different kind of data set. Every example is bound to national regulations and expressions of cultural norms.

For the benchmarking, a few examples were chosen through existing contacts and desk research. This report does not cover a conclusive market analysis. It is not giving a complete picture of all options, but rather reflects our punctual deep insights gained from exchanges with European research data centres.

Structure of the report

First, the following ***three deep dives*** will show how a challenge is approached differently by diverse research centres.

// Addressing the interests of citizens

// Technical access to the data

// Forms of access authorization

Then, detailed views of further ***best practices*** will be given at the example of four research data centres.

// Health Data Hub France

// CBS Netherlands

// Findata Finland

// Statistics Denmark

Finally, a ***vision*** will be given in the form of a brief outlook and a few recommendations.

Deep Dive: Addressing citizens (1/7)

Addressing the interests of citizens

The public interest regarding research in health is essential for its success. Within the EU, different levels of critical analysis and diverse experiences regarding electronic identities and the analysis of health data can be observed on a national level. According to these national contexts, the interests of citizens regarding the research with their health data are addressed differently. While Sweden or Denmark have had a central citizen ID for many years, which is used as a basis for health data research, other countries have developed a distinct communication strategy for addressing possible concerns of citizens.

Example UK

The National Health Service in the UK has a practice of researching the public's opinion towards health data research. As can be seen in their **public explanations**, they carried out public dialogue exercises for example. These allowed citizens to voice their opinion on how data should be handled. A key findings was that the public would be "happy to share personal data (...) if they are given a clear explanation of how their data would be used". Based on surveys results from citizens, they also published a **guide for proportionate consent** back in 2017. The importance of giving consent remains high, as citizens "were open to the idea of research nurses having access to patient notes, with the proviso that patients are informed and have the ability to opt-out". Even under the provision of GDPR research exemptions, the NHS approaches for respecting the patient's consent are valuable.

Deep Dive: Addressing citizens (2/7)

Example France

The Health Data Hub contains a variety of health related databases and will make them available to researchers over the next years. They consider it important to address their citizens by giving ***in-depth explanations*** on why and how the data is used and being made available on their ***website*** and in the ***FAQ*** section. They published the Health Data Hub's **commitments to civil society**, which elaborates their commitment to fostering research which benefits society. It shows that they transparently address concerns which citizens might have. They also entertain a ***social media presence*** on LinkedIn and Twitter.

Survey Tech4Germany

Adding to many existing and more scientific surveys, we also conducted a short survey in Germany. We wanted to understand how citizens felt about their health data being used for research purposes as well as which reasons were most convincing. Overall, 70% supported the research with their health data, while only 10% strongly disagreed, and 79% trust researchers to handle their data in a trustworthy manner. In order to build trust, the top three topics to communicate to the public are details about the data use (75%), measures for data protection (59%) and the legal framework (47%). Especially those who were more critical were most convinced by the argument that research could improve their own health, the medical care in Germany or if better drugs could be developed.

Deep Dive: Technical data access (3/7)

Limiting research datasets

The European research data centers we interviewed (Findata, CBS Netherlands, Statistics Denmark etc.) follow the data minimization principle. This means that researchers only get access to the limited dataset they need for their specific research requests. In order to do that the researcher has to state his research question in a research proposal and then discuss it with an expert, e.g. in a 1-hour video call. They agree on which datasets and variables are actually needed in order to properly investigate the research question. This data is then provisioned for the researcher's project from the main database (aka a limited copy of the data is drawn). The process until giving the digital access to the dataset takes a few weeks (CBS) up to 2 months. Once the needed data is available the researcher usually has access until the research question is answered, which can take a couple of months.

Overall, this approach helps in:

- ***Data leak safety:*** Avoiding a massive data leak as only the necessary data is provided remotely
- ***Traceability:*** Controlling which data is given to whom and have traceability in case of a partial data leak
- ***Faster computing:*** Minimizing computing resources as queries are only run on the limited data set

Deep Dive: Technical data access (4/7)

Remote data access

After the dataset for the research question is provisioned, the research centers enable remote access to this dataset. This means that the researchers receive remote desktop access over a secure VPN. Within this secure environment, the researcher has all the analysis tools needed for the research. These tools can include R, SPSS, STATA, Python, SAS. During remote access, everything is logged so the research data center maintains control over what is happening. Additionally, the researcher cannot export the dataset due to the protected environment, hence, the data is kept on the premise of the research data center at all times. With this approach, the risks remain that a researcher takes a screenshot or manually writes down individual data records. Therefore the system is based on trust. Still, there have not been any known problems or cases of misuse, even though the system has been in place for several years e.g. in the Netherlands.

The advantages of this approach for the researcher are:

- ***Planability:*** Researchers can do the analysis in their own pace and timeline
- ***Independence:*** Researchers can do the research iterations without involvement of the research data center in every step
- ***Comfort:*** Researchers can use the analysis tools they are used to
- ***Autonomy:*** Researchers can do it remotely just requiring internet access

Deep Dive: Technical data access (5/7)

Exporting aggregated result sets

After a researcher finished his research on the data he got access to, he can export his aggregated result sets for his publication. He does this by copying the needed information in a shared folder accessible in the remote access environment. This folder is then checked by the research data center (either manually or by an algorithm) and then shared with the researcher by email or a secure hoster to download the data. The main purpose of the check is that no individual records are exported but only aggregated data which is used for the publication and therefore minimizing the risk of reidentification of individuals. The aggregation limitation varies from country to country between a K value of 3-10 which means that at least 3 -10 data points must be aggregated per exported record. Values below the limit need an explanation that is manually assessed by the datacenter. The process of exporting a resultset is quite fast and often takes just a couple of days.

The advantages of such an approach are:

- ***Fast checks:*** The result set is not very big, can therefore be checked quickly and be used for publications in the end
- ***Minimal data export:*** As only the final result set is exported, minimal data leaves the research centers
- ***Minimal resources:*** By checking only the final results after all research iterations are done, the controlling overhead is minimal

Deep Dive: Access authorization (6/7)

Deciding who gets what kind of access

A central question to answer for each research center is who can request access to what kind of data under which verification criteria. Countries have different ways of increasing the accountability of researchers to treat sensitive health data in an adequate manner. Many countries already give access to private organizations as well, as long as there is no commercial purpose in the research project. The more extensive the given dataset in a research center is and the more standardized its approval and access process is, the more research requests can be expected. As a result, the scalability of the access authorization process plays a crucial role in increasing the number of completed research projects for the benefit of society.

Example Denmark

Only such organizations can apply for research access at Statistics Denmark which has the infrastructure and expert knowledge to handle the data, such as statistical or data security expertise. Researching institutions have to appoint a responsible person vouching for adequate conditions. In addition, each research project has to be approved according to set standards. Private organizations can apply for research access as well, as long as they prove the value for society of their project.

Deep Dive: Access authorization (7/7)

Example Finland

At Findata, private as well as public organizations can apply for access. Research proposals are authorized separately. According to the law, valid purposes are are scientific research, statistics, teaching, steering and supervision by authorities, authorities' planning and reporting duties promoting national health or social security. Findata receives research requests from all over Europe. To increase the transparency of data use, all scientific research results gained from the data have to be published.

Example France

At the Health Data Hub in France, data access is only allowed for public interest research, with a strictly defined project duration and a limited scope. Any private actor requesting access to the data will have to prove that the project is of public interest, for the benefit of citizens, in the same way as public actors. Furthermore, every proposal goes goes through an independent ethics and scientific committee and needs an approval by the french data protection authority.

Example Netherlands

At CBS Statistics Netherlands, institutions have to be authorized, and can then suggest researchers and projects for approval. The institution itself is accountable for what their researchers do. Hence, in case of misconduct in handling the health data, the whole institution will receive consequences. In preparation, the institution has to co-sign a confidentiality agreement for every research proposal which is suggested to CBS.

Best practices - CBS Netherlands (1/4)

Institution CBS Netherlands (Centraal Bureau voor de Statistiek) is the governmental institution for statistics in the netherlands. Their mission is to publish reliable and consistent statistical information, that responds to society's demands in this respect.

Country 🇳🇱

Topic Building a scalable self sustaining research data center

Expert Video Call

In a research proposal you need to outline which data sets you need. After submission you have a 1 hour video call with a expert of these data sets before you get a remote access.

Mandatory Publication

After finishing your research you have to publicize your results which doesn't give companies a competitive edge and promotes research use

Superior sign-off

The superior of a researcher needs to sign-of as he is also responsible for appropriate use of the data and is liable for misconduct

Researchers absorb cost

As this service is not funded by the government the researcher pays for all cost for the service which included consulting fees and infrastructure cost

Dedicated Helpdesk

CBS employs people in the team which are dedicated to just answer questions of researchers and therefore enable other people to just do the data provisioning for example

Best practices - Findata (2/4)

Institution Findata is the Health and Social Data Permit Authority. They collect health and social data from different institutions and promote secondary use of health and social data, facilitate data permit processing and improve data protection for individuals.

Country 🇫🇮

Topic Providing a central interface for secondary data for researchers

Collecting data sources

Findata collects and links data from different institutions like municipalities & hospitals and provisions pseudonymized datasets for researcher

Scalable computing power

For your remote access you can pay for different infrastructure setups where you can choose how many cores and RAM you need for your analysis

European collaboration

Findata is relatively new but cooperates with different european health data centers and enables access to the data to researchers out of Finland

IT by external provider

The IT-Services and Infrastructure are not maintained by themselves but are taken care of by a company owned by the government

Transparent statistics

On the findata website you can see up to date statistics which reflect the number of request which have been approved and denied as well as pending applications

Best practices - Health Data Hub (3/4)

Institution The Health Data Hub is aimed at boosting and facilitating the use of available health data for research projects, by both private and public entities. Being a unique gateway for researchers it is both an infrastructure as well as a health database catalogue.

Country 🇫🇷

Topic Medical data analysis thought big

Public Funded

The french data hub receives 76 million Euro in public funding for the first 4 years. For-profit actors could be charged for access in the future.

Ethics and data privacy board

Data access is only allowed for public interest research, with a strictly defined project duration and a limited scope after approval by the Scientific and Ethics Committee and the national Data protection agency

Health data ecosystem

The french health data hub creates a ecosystem around the data by organizing talks, challenges and events to connect researchers with each other

Transparency

The HDH aims to guarantee transparency towards civil society. The citizens-related information is made available, in order to empower citizens by making them aware of their individual rights.

Platform for other data

External providers can put their data sources on the platform as well and hence foster the research insights from different data sources

Best practices - Statistics Denmark (4/4)

Institution

Statistics Denmark gives researchers access to health data via data collected in registries. These are based on health data connected to each citizens central ID.

Country 🇩🇰

Topic

Health data registries, self-service downloads for researchers, private sector access to data

Health data with central citizen ID

Health data are made accessible for research and can be combined with other social registers via the central ID of Danish citizens

Self-service research

Once given permission, the researchers can access their pseudonymized micro data set in a save remote environment as well as download analysis results (without reidentifiable data) autonomously

Punishment for misuse

In the case that deidentified microdata is downloaded, the whole institution of the researcher loses access for at least 1-2 months; as more data protection measure are put into place, leakages happen very rarely

Private sector access

Private health organisations can also request access for specific research projects; they also pay more and hence co-finance public research requests

Responsibility & trust

For each institution a responsible person is appointed for the authorization ensuring all researchers know the rules for accessing data under the research arrangement

Vision

Outlook

The European health data market is increasingly merging. This process is accelerated by the current COVID-19 crisis. There is a high priority on making health data accessible for more researchers in a faster and more convenient manner. Connecting databases and correlating them with findings from clinical studies is something that Northern European countries already do. The French Health Data Hub is planning to become the most comprehensive health data centre in Europe, connecting over 50 institutes and database sources. Other countries like Sweden will rely on the health data registries with data from over a decade. European neighbors are required to collaborate in order to combat a crisis, while their health systems are increasingly entering competition as the EU market merges. Making health data accessible for research is a key step in this journey.

Recommendations

For the German research centre for health data we recommend to consider the following aspects.

- The long-term success of the research centre will be correlated to its communication strategy with citizens. Especially the high value given to data protection in Germany should play a crucial role.
- As formerly enacted, data leakages should be prevented. At the same time a strategy should be in place for when data leakages happen including adequate consequences to be drawn while keeping the FDZ up and running to deliver its value for the society.
- Remote access to relevant and limited data sets should be given for each specific research question, as this minimizes exposure of reidentifiable data as well as maximizes the researchers convenience.

With special thanks to

Data center	Contact	E-Mail
CBS Netherlands	Ruurd Schoonhoven	r.schoonhoven@cbs.nl
Findata	Antti Piirainen	antti.piirainen@thl.fi
French Health Data Hub	Louisa Stuwe	louisa.stuwe@health-data-hub.fr
Statistics Denmark	Jørn K. Petersen	JKP@dst.dk

For further questions please contact

Forschungsdatenzentrum	Dr. Steffen Heß	steffen.hess@bfarm.de

tech
4Germany_

unter der
Schirmherrschaft
des Chefs des
Bundeskanzleramt,
Prof. Dr. Helge
Braun

Bundeskanzleramt

Project Forschungsdatenzentrum

Michael Rotarius
Jonas Pohlmann
Vanessa Robayo Ladino
Carolin Bednarz