PHC Catalyst

Sharing is Caring?

Towards a Data Sharing Platform for Personalized Healthcare

Feasibility study

March 16, 2020

This report is developed by the members of the PHC Catalyst Alliance and has been made possible by a financial contribution of Roche Nederland B.V.

Personalised Healthcare Catalyst

Introduction by chairman PHC Catalyst Foundation & Alliance

Since 2018, the Personalized Healthcare Catalyst (PHC Catalyst) has developed into an alliance of more than 40 organizations to accelerate the transition towards personalized healthcare in the Netherlands. To maximise the impact of breakthroughs in biomedical science and to be able to achieve the 4P's of personalized healthcare: Predictive, Precise, Participatory and Personalised, we need data. A lot of data! Very different data. The average patient doesn't exist and a patient is much more than a body with a defect. It is a complex condition in a complex body in a complex society. This is a call for systems thinking. To characterize and understand the situation and to identify the best way to maintain, improve or control someone's health, we therefore need the power of data science. But the feedstock for developing and application of algorithms is data from various sources. And the value is not a single data source, but in the combination of these diverse datasets.

The PHC Catalyst started with an idea to perform a hackathon. The focus of this ImmunoPRO Hackathon is on transforming the fight against cancer using Big Data & AI. And the main questions are: Who will benefit from immunotherapy? Can we develop a pre-treatment predictive model? In other words: why is one person cured by immunotherapy and the other isn't? Interestingly, it turned out that finding, accessing and combining the data, turned out to be challenges by themselves, which proves the point that reshaping the system with its shared, but also opposite ambitions and interests, is key. For access and combination of data, we need the right environment.

This means that we need a culture of sharing and collaboration, we need to have the right processes and this has to be supported by the right technical infrastructure. Processes are necessary to identify relevant information, interoperability between datasets, to deal with legal and compliance issues and to facilitate agreements between sharing parties. Such a platform, a Data Sharing Platform, which has many commonalities with exchanges (e.g. for stocks or commodities), is crucial for the practical implementation of recommendations that result from the various investigations by and for the PHC Catalyst, as well as many other research and innovation projects elsewhere in the area of Personalised Healthcare.

This has inspired us to ask PNA, as 'knowledge engineers', to explore the various options for the concept of a 'Data Sharing Platform' and discuss with field parties the various scenarios. This report gives the highlights of the research and an overview of the options that were considered, as well as a clear recommendation, including some next steps. We are pleased that it was possible to develop a point of view for the collaborative infrastructure that will be necessary to reap the benefits from biomedical ánd data science. Now it is up to everyone, to jointly find out what elements of this platform can be developed and what kind of culture, processes, rules and regulations need to be in place to avoid the famous formula NT+OS=EOS (New Technology in an Old Society results in an Expensive Old Society).

As a first next step, the PHC Catalyst is supporting the development of a prototype of a Data Sharing Platform, combining a set of medical data with social-economic and demographic data from Statistics Netherlands (CBS). This is a good example of our primary process, 'Combinatoric Innovation': United we stand, apart we fall!

Paul Iske, Chairman PHC Catalyst

Management Summary

Data sharing, being the ultimate prerequisite for personalized healthcare, is hampered by various issues. During the past months, we investigated the feasibility of a data sharing platform (DSP) for personalized healthcare to alleviate these issues. The study comprised of a literature review, an analysis of existing platforms, interviews with subject matter experts and a survey among members of the alliance.

First and foremost, we found that control over shared data should always stay with the data provider, that access to the DSP should be on an invitation-only basis, and that access to metadata should be available to all DSP users.

Our study shows, that the DSP should have the following characteristics:

1. The DSP has **a centralized data store** containing both unstructured and structured data vaults. In a later stage, the vaults can also **distributed**, i.e. located with the data providers. The vaults are, as indicated above, fully controlled by the providers of the content of the data vault.

2. Standardization of data is facilitated through a **glossary** with clear and specific descriptions for the variables, containing strictly PHC-related data. **Governance** of the standard is handled by a platform member or a third-party.

3. **Access to the platform** is granted through invitation, whereas **access to data** is granted by data providers on a per-project basis. Access to the platform takes place via a server with limited connections. Ownership of intellectual property should be determined by the involved members.

4. Both a **marketplace** and a **collaborative platform** are feasible platform styles.

DSP Design: Summary

Prerequisites

Ownership of data	Ownership of the source data always stays with the original owner, unless the ownership of the source data is transferred to a different owner.
Permission to the platform	An invitation-only platform.
Access to metadata	Open sharing of metadata among platform users.

Platform parameters

Platform styles

Data storage	A centralized data store containing both unstructured and structured private data vaults	OP	MP	CP
Standardization	A glossary with clear and specific descriptions of the variables. Strictly PHC-related data.	OP	MP	CP
Data governance	Access to the platform granted through invitation, access to data granted by data providers. Governance of the standard by either a platform member or third-party.	OP	MP	CP
Data permission	Access to, and use of, data granted on a per-project basis.	OP	MP	CP
Data distribution	Access to data via a server with limited connections.	OP	MP	CP
Resulting IP	IP ownership should be determined by the involved members and included in the data sharing request (and data sharing agreement).	OP	MP	CP

Sharing is Caring?

Introduction

Background

Design of the investigation

Reading guide

Background

Personalized healthcare is a way of treating a patient based on the genetic characteristics of the disease and of the person themselves. Personalized healthcare has sparked the emergence of a multidisciplinary field combining genomics, healthcare and big data analytics. Healthcare professionals try to achieve this by combining data from various different sources and disciplines. The Personalized Healthcare Catalyst (PHC Catalyst) is an alliance of more than 40 members to accelerate the achievement of personalized healthcare in the Netherlands. The PHC Catalyst alliance tries to remove barriers and hurdles, by combining the power of all members. Combining different disciplines and data from different sources is essential for realizing personalized healthcare. However:

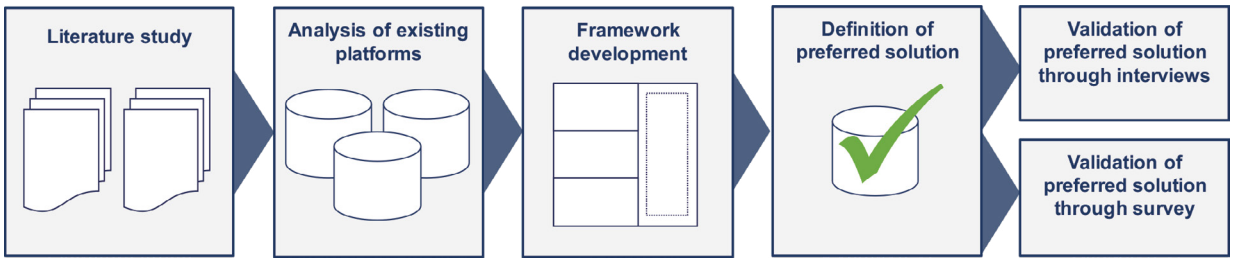
- **The existence of data silos does not enable the consolidation or combination of different data sources, and the "right" data is either not stored or not stored in a homogeneous format.** Data from different sources are often stored separately and in a wide range of formats. Hospitals, pharmaceutical companies, general practitioners, researchers and regulators all have their own data silo and even within one group, data are usually not stored in a homogeneous format. Discrepancies exist between the variables that are recorded. Hospitals store different information according to their own needs. There is not one standard to store the different types of data.

- **There is reluctance to share data outside the organization.** Sharing data between organizations (and often also within organizations) is a sensitive topic. Firstly, organizations fear the use of their data by others and are afraid to lose control. Legal frameworks do not fully cope with this discomfort. And secondly, data are a relatively new asset for organizations. Most organizations tend to not yet fully understand how their data can be used. This makes it difficult to value data as an asset. If one cannot value their own data, it is even harder to value the data of others. As a consequence, it is hard to predict what the outcome will be when organizations share their data. Therefore, trust is a major factor in data sharing.

Data sharing platforms (DSPs) might play a role in alleviating these hurdles. In this document we report on our investigation of a data sharing platform for the PHC Catalyst.

Design of the investigation

We condicted this feasibility study during Q4 2019 and Q1 2020. The study has the following steps, partly carried out in parallel:



The outcomes of the literature study and analyis of existing platforms were input for the development of a solution framework and the choice and definition of a preferred solution. This preferred solution has been discussed with experts in the field, and has been validated in a survey among the members of the PHC Catalyst Alliance.

Reading guide

First we discuss the **key issues that affect** (the realization of) a DSP for personalized healthcare.

Secondly we define a **scope** for the development of a framework for a data sharing platform for the PHC Catalyst Alliance.

Within the defined scope we then describe three **platform styles** that can act as the foundation for the DSP.
These three styles are not to be seen in isolation but aspects of each can be combined to build the overall foundation of the platform.

On top of the foundations we identified six parameters that will act as **building blocks** for platform development.
Each parameter can be tailored towards the chosen platform style.

Within the above framework we finalize the investigation with an advice on how the framework for personalized healthcare could be implemented.

Sharing is Caring?

Key issues

Key issue types that affect a personalized healthcare data sharing platform.

Legal

Cultural

Technical

Valuation

Key issues

The implementation of a DSP is complex and pushes boundaries in many directions. Therefore there are many hurdles to overcome during the realization of a fully functional DSP. In this section we will discuss the impact different issues will have on the realization of a DSP for personalized healthcare. Based on a literature study, and validated by interviews, the following types of key issues that will affect a personalized healthcare DSP have been identified:

- Legal issues
- Cultural issues
- Technical issues
- Valuation issues

Legal issues

Key questions

What are the legal issues to take into account (or to address or to resolve) when sharing medical data? Can you trade medical data and if so, how?

Prerequisites

Data sharing within a community requires a sound legal basis.

Issues

It is a given that a DSP has to act within the current AVG (Dutch implementation of the GDPR) legislation. The AVG allows to realize a proper DSP. There is however more legislation that needs to be taken into account. Copyright laws and database rights are other forms of legislation that have to be respected.
The current legislation is vague and incomplete when it comes to data exchange in the medical sector. Legislation is lagging behind; there is no sound legislative foundation. As a consequence it is unclear what is possible and what is not.
Current developments in the Netherlands on the patient secret ("patiëntgeheim", https://www.patientenfederatie.nl/themas/patientgeheim) are striving for greater restrictions on the disclosure of patient data. This could have fundamental negative impact on the possibilities of a DSP for personalized healthcare.

Potential solutions

Involving parties that are in charge of writing legislation could help setting a clearer legislative foundation for a DSP. The Ministry of Health, Welfare and Sport is one of those parties as is the "Patiëntenfederatie". A closer collaboration with the Dutch Data Protection Authority ("Autoriteit Persoonsgegevens") and getting them involved could also help with acceptance of a DSP. Timing and talking to the right people is crucial.

Cultural issues

Key questions

How to create a culture in which sharing of data is the right thing to do?

Prerequisites

Data sharing within a community requires a trusting mindset.

Issues

Most companies have not figured out all the uses of their data and do not have a full understanding of what their data could mean to the company as a whole. In addition to this, it is unclear to them what other parties could do with their data.
This lack of clarity leads to lack of trust.
Also, the public opinion on sharing an individual's data being shared is not very supportive, also due to data leakage incidents and general distrust of large enterprises using data for solely commercial purposes.

Potential solutions

Things that might help alleviate cultural issues include:
Clear contracts that describe what can be done with the shared data and what happens to resulting intellectual property
Emphasizing why the data is being shared is a good indication of legitimacy
Working towards a common goal or having a mutual interest
Better insight into the full range of use cases of a given dataset

Technical issues

Key questions

How to implement a system that is able to combine many different types of data? How will that system allow for a safe data exchange?

Prerequisites

Combining data from different sources is essential for personalized healthcare.

Issues

There is a lack of readily available data.
The data from the different sources are often stored separately and in a wide range of diverse formats.
Discrepancies often exist between the variables that are recorded.

Due to biases inherent to datasets, results can often not be replicated with a different dataset containing the same parameters.

Security is also a major issue. Medical data is sensitive data.

Potential solutions

In an ideal world for a data sharing platform you want all the data in the platform to be FAIR:

- Findable
- Accessible
- Interoperable
- Reusable

Valuation issues

Key questions

How to determine the value of data?

Prerequisites

One of the essential elements of a data marketplace is the valuation of data.

Issues

Data is different from other traditional assets due to its reproducibility (non-consumable asset).
There is no common standard yet for the valuation of data.
Most of the existing frameworks to estimate the value of data adhere to their own standard.

Potential solutions

There are measures that can be used in a data valuation framework to compare the relative value of different datasets within the same domain: foundational measures and financial measures.
Foundational measures define the value of the data itself, they are derived from the factors illustrated on the right and include:

- Intrinsic value: how correct, complete and scarce is the data?
- Business value: how relevant is the data?
- Performance Value: how does the data affect business drivers?

Financial measures define and include:

- Market value: what will the market bear for selling this asset?
- Income value: what income stream will this asset generate?
- Cost value: what would it cost to replace this asset if lost?

Objective
Accuracy
Integrity
Consistency
Completeness
Accessibility
Precision
Timeliness

Subjective
Relevance
Usability
Believability
Clarity
Objectivity
Scarcity

Sharing is Caring?

DSP Framework

To enable comparison and development of DSPs, we have developed a framework that defines the scope, the foundation and the configuration for a DSP. It consists of three components:

Prerequisites

Platform Styles

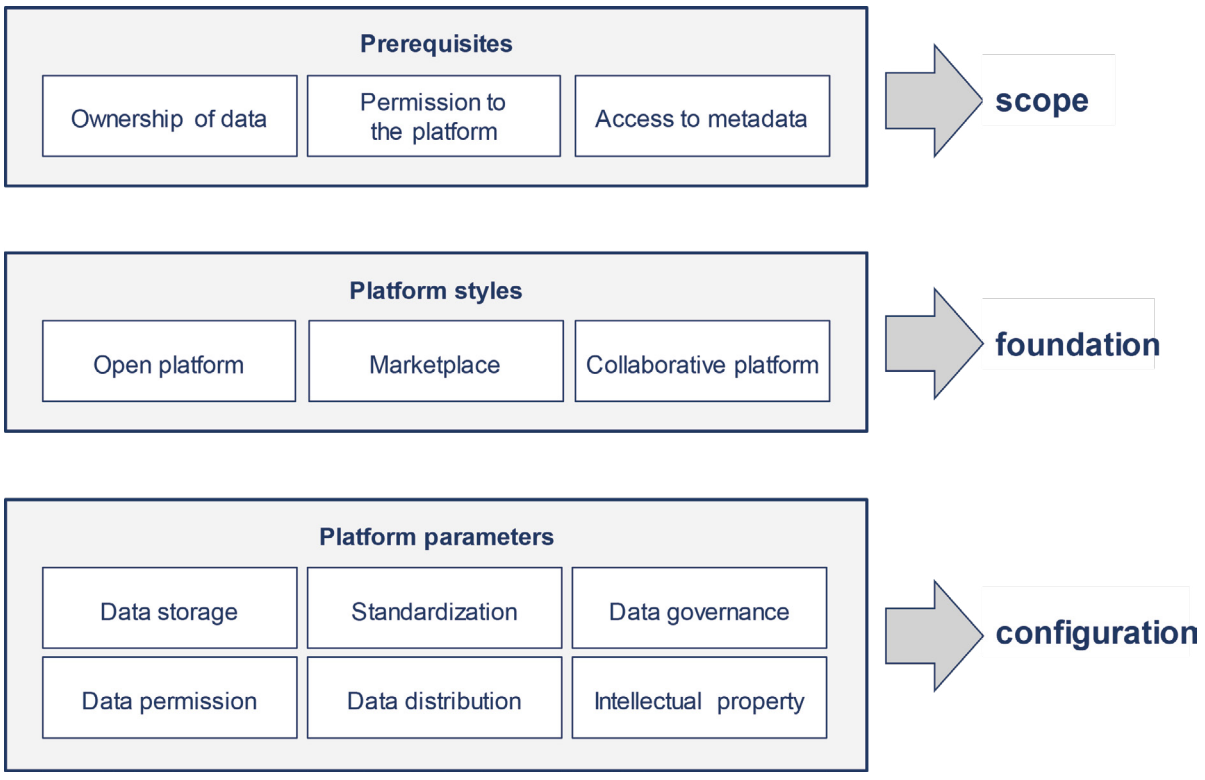
Platform Parameters

DSP framework

In order to develop a framework for a data sharing platform within the PHC alliance a scope will be set by three prerequisites on the implementation of the platform.

Within this scope we will define three platform styles that act as the foundation for the sharing platform. The three styles are not to be seen in isolation but aspects of each can be combined to lay the foundation of the platform.

On top of the foundations we have identified six parameters. These parameters will act as building blocks for configuration of the platform. Each parameter can be tailored towards the chosen platform style.



DSP Framework

●○○ Prerequisites

In order to develop a framework for a data sharing platform within the PHC alliance a scope is defined by three prerequisites for the implementation of the platform.

Ownership of data

Permission to the platform

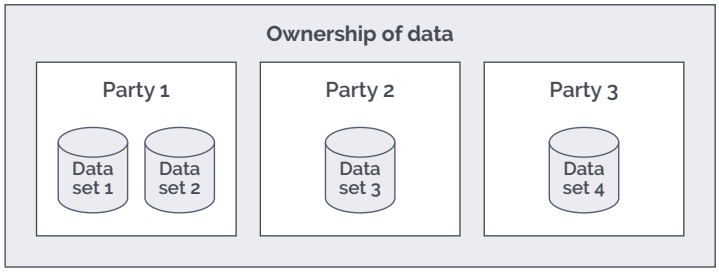
Access to metadata

Ownership of data

Data ownership is defined as control over who gets access to data

Companies are hesitant to give out their data freely or without care, let alone pass on ownership of the data. Even though data is one of the most valuable assets of a business, companies tend to not know the real value their data brings to the business. It is difficult to measure how important the data is to their organization or what competitors could do with the data. In addition, data is not protected by intellectual property laws.

It would be hard to convince members of the alliance to share or pass on ownership of these valuables, even when it might mean they get more value in return. That is why the first prerequisite is that ownership of the source data shared within the platform always stays with the original provider, unless the ownership of the original is transferred to a different owner. An alternative would be that all the data shared on the platform would be owned by a third party that would also handle governance, speeding up the data sharing process.



Platform permission

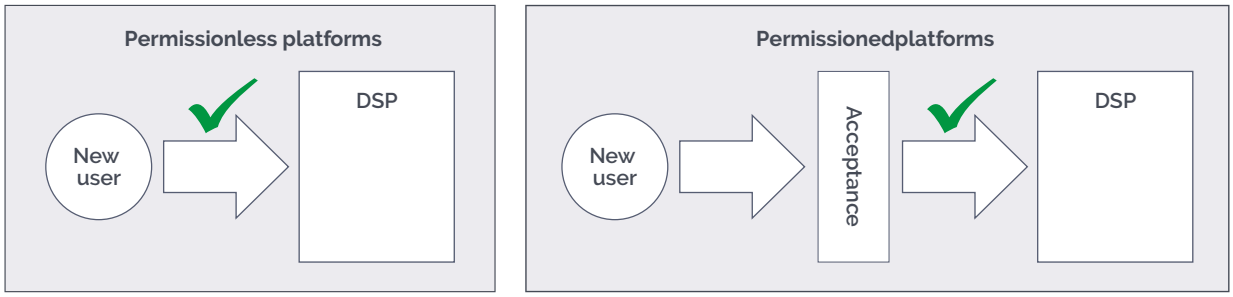
Permissionless platforms are platforms that users can freely join without a need for prior screening or for an invite

Permissioned platforms first require a new user to go through an acceptance procedure before they are accepted to participate on the platform

Permissionless platforms have the benefit of the masses. If anyone can access the platform, anyone can contribute to the shared data; the more popular the platform, the more input the platform will ultimately receive.

However, the data shared on the platform is sensitive and, as mentioned before, it is difficult to measure the real value of the data. Due to this, opening up to the masses seems ill-advised. Closing down the platform will prevent competitors from walking away with valuable knowledge or prevent parties with malicious intent from getting their hands on delicate data.

That is why we consider an invitation-only platform to be the best solution at this point in time.



Metadata access

Metadata is a set of data that describes and gives information about other data

Even though we consider an invitation-only platform the best approach to protect the interests of the data providers, we do however recommend open sharing of metadata among platform users.

Allowing all participants of the platform to see metadata of the information shared on the platform enables users to come up with ideas on how to combine or process different datasets based on their characteristics without actually getting access to all of the data. By allowing free access only to the metadata, the underlying data, and thus the value, stays within the hands of the data provider. It does however allow companies to see what data is already available on the platform. Consequently, they can potentially save the investment needed to collect the data themselves and request access to the already available data.

DSP Framework

○●○ Platform styles

Within the defined scope we describe three platform styles that act as the foundation for the DSP. These styles can be combined to build the overall foundation of the platform.

Open platform

Marketplace

Collaborative platform

Open platform

An open platform means that anyone who can access the platform, can contribute data and use data freely. Open access leads to high accessibility of the available data. There are no hurdles preventing users from sharing or accessing data. In order for an open platform to succeed however, the existing trust issues need to be resolved. If parties do not trust each other, no data will be shared.

Characteristics

- Anyone may add data
- Anyone may access data
- High accessibility
- High trust barrier

Considerations

On an open platform, all participants may share their data and access any data that other participants have shared already. In other words: there are no restrictions on who may use what data. Open platforms are widely used in collaborative environments where every member needs to have access to their shared files, documents and data. Good comparisons would be collaboration applications like Microsoft Teams and GitHub but also streaming services like Netflix and Spotify.

An open setup would lead to very high accessibility since there are no hurdles to overcome in order to access a specific dataset; all data on the platform is always available to the users at any given time.

Trust is an important precondition however. Participants need to be sure that the data they share will not affect them in a negative way. Otherwise there is no reason for them to share that data. Without mutual trust, one will most likely end up with an empty platform or a platform with only trivial and useless data.

Marketplace

A marketplace means that users can trade their own data and gain access to others'. Marketplaces have the benefit that even though your data is shared with another party, the data provider gets some immediate value from sharing. This will alleviate a lot of the trust issues that arise. Accessibility is a greater problem in a marketplace. There is a paywall in order to gain access. For some parties this might be a deal breaker due to the lack of financial flexibility.

Characteristics

- Trade data to share
- Trade data to access
- Low accessibility
- Low trust barrier

Considerations

On a marketplace participants may make their data available for trade. Any data available for trade can be acquired by other members of the platform. When a piece of data is traded the acquiring party will have access to this data.

There are different ways of defining how datasets are traded. The most obvious way would be using real world currencies. One could also think of implementing a platform currency, which could be acquired through sharing data or through simply purchasing it. A third way is to use data itself as a currency.

Obviously, the valuation of data is a big issue on a data marketplace. The easiest way of tackling this is for the participants themselves to define how much they think the data is worth. A more complex solution to the problem is to develop a system that values a dataset based on a predefined set of heuristics or a data valuation framework.

Trust issues will be less prevalent on a marketplace. Gaining immediate value from the data you share alleviates quite a number of the negative consequences you might have when sharing your data with competitors. This means that trust is no longer a necessity. Mutual trust between participants is, of course, still very beneficial.

By having a paywall in place an extra barrier is created for participants wanting to access data. This leads to having way lower accessibility.

Collaborative platform

On a collaborative platform data is shared through 'data sharing agreements'. In these agreements clear arrangements are made about:

- Which data is being used
- How this data is being used
- What is done with the resulting IP

By having a clear written agreement between two (or more) parties, a collaborative platform alleviates trust issues. Having to set up and sign an agreement might lead to a hurdle to overcome in order to access data. Standardized agreements might help to alleviate this.

Characteristics

- Sharing of data and access to data is arranged and documented in 'data sharing agreements'
- Medium accessibility (able to circumvent this through standardized agreements)
- Medium trust barrier

Considerations

A collaborative platform requires participants to sign so called 'data sharing agreements' when sharing data between two or more members. In these agreements it is clearly defined what the purpose of the data sharing is, what can be done with the shared data and what happens with potential intellectual property gained through the data sharing. In these agreements parties could also agree on (monetary) compensations given to the party sharing the data. This would keep some of the aspects of a marketplace without having the necessity of a paywall.

Having a written agreement that clearly states what is to happen with the data and anything resulting from it, tackles a substantial number of the trust issues that arise when talking about data sharing. Of course, when signing contracts or agreements, some form of mutual trust is still required.

Even though setting up a sharing agreement is a relatively minor hurdle that has be crossed when accessing a certain dataset, it is much less of a barrier than having a paywall in place. This hurdle can even be further reduced by having standardized sharing agreements available for the users of the platform.

DSP Framework

○○● Parameters

Building blocks within the set foundation.

Data storage

Standardization

Data governance

Data permission

Data distribution

Resulting IP

Data storage

Where and how data is stored are two very relevant questions one could ask themselves when designing a data sharing platform.

Where is the data stored?

We consider two options for the location of the data storage:

- **Centralized storage**
  This means that all data is stored at a single location. Each of the data providers would get their own digital safe in which the data they provided would be stored. Most of the current implementations of similar platforms use this approach. By centralizing data storage, data integrity could be optimized. All the provided data would be part of the same database. It would also be easier to access data from multiple different data providers at once. Centralizing the data would require a, third-party broker. Said broker should be trusted by all parties as they would have access to all the datasets (granted, these datasets could be encrypted).

- **Decentralized storage**
  This means that each data provider has their own place where they store the data they provide to the platform. Decentralizing the data would have the benefit of trust. The data providers would not need to trust a broker to securely store their valuable data, but rather they would do this themselves. Decentralized storage has the potential disadvantage, however, to be more complicated, involving extra steps for the automatic retrieval of datasets due to different database structures.

How is the data stored?

Regarding how the data is stored we found three possibilities:

- **An unstructured data store**
  This would be the quickest way to store all the data at once, without any need for standardization. However it would severely hinder the retrieval process. Unclear, unstandardized definitions and structures would make automation of the retrieval and analysis of data near impossible. Due to this data cleansing efforts would be needed every time data is exchanged, reducing its value.

- **A structured database with a standardized model**
  This would take a higher initial effort due to the need to develop such standard and convert all of the existing data into a format that satisfies said model. However in the long run it would greatly increase the efficiency and value of the data sharing platform.

- **A hybrid system with an unstructured store and structured dataset**
  This would mean that data can be freely added to the unstructured store without the need for standardization which would speed up the storage process while also allowing for a standard to speed up the retrieval and sharing of said data. If required an effort could be made to transform data from the unstructured store to the structured database manually or automatically.

Advice

A centralized hybrid system dataset.

We consider that the optimal storage strategy is centralized on a common data store. With this, access to data can be sped up and standardized on a platform level, facilitating exchanges and allowing clear and strict control over the access to the given datasets. In a centralized system you can make sure incoming data is standardized upon delivery, meaning you do not force data providers to store their data in a certain format or structure. Implementation of a centralized storage system will also be faster as it doesn't require a certain infrastructure from the data providers. As time progresses an additional decentralized storage system could be considered. But since decentralized storage asks for a lot of initial effort to ensure scalability, it is not advised as a first implementation.

We found that in most implemented solutions there was a centralization of datasets for the same reasons.

This data store would be a hybrid system composed of an unstructured file storage system and a standardized structured database system. This will have a two-fold effect simplifying and speeding up the submission of data while including a structured option that allows companies to submit data that follows specified standards. The standardized data could then be more easily shared among members reducing the time needed to clean and adapt discordant datasets.

On the researched implemented solutions different levels of standardization were required and so we considered a hybrid system ideal due to its ability to cope with those differences.

Standardization

A big challenge when sharing and combining data is the lack of standardization. The data can be stored in different types of databases, different formats, different variables or with mismatching data types. This often leads to silos of data that aren't interoperable without requiring additional resources to convert all the data to an agreed upon format. There is also the question of which data is relevant for the PHC and which data is not. The main trade-off that we find here is between breadth and depth. We want a platform with a clearly defined scope but that still has enough breadth to develop innovative techniques that deliver brand new insights.

What data is accepted into the platform?

Regarding what data should be accepted four options arise:

- **Allow only a specific set of defined variables**
  This would facilitate the standardization and set a defined scope. However by defining such a closed scope stakeholders could miss out on other types of potentially valuable data.

- **Personalized healthcare related data**
  Any data that directly links a specific group of people with the effect of a disease/treatment on them. This could range from omics data to sociological data.

- **Any medical data**
  That is, no matter if linked with specific groups or just general medical data.

- **Determined by the PHC alliance**
  In this case the PHC alliance would get to decide per dataset if it should be accepted or not.

What does the standard determine?

Regarding what the data standard determines we can go from the least to the most defined standard:

- Standard only defines the database structure where data should be stored.This would solve some of the technical issues related to retrieval of data.

- A joint glossary with clear and specific descriptions for each of the variables of the different datasets.Adding this to the standard would speed up the process of both data retrieval and analysis, also drastically increasing the value of a data sharing platform and its contents. However, such a joint glossary would require time and effort to be developed.

- A golden standard that combines the two options above. This entails a standard for the database structure and standardized definitions for all concepts used. A golden standard would also define the expected quality of provided data.

- A golden standard makes sure that the quality of the data in the data sharing platform is the best it can be for analysis. The included glossary helps mitigate confusion and miscommunication. And the structured database allows for quick and efficient data retrieval.

Advice

A glossary with clear and specific descriptions for the variables. Only strictly PHC-related data.

One of the biggest hurdles in personalized medicine is the lack of reproducibility of the results. This happens due to biases inherent to datasets that lead to results that cannot be replicated in new datasets.

One initiative that is currently being implemented is FAIR data. FAIR describes a series of guidelines for scientific data, focusing on four needed properties for data: findability, accessibility, interoperability and reusability. These trends highlight the importance of a clear data standard within the platform.

This standard should contain a glossary with clear and specific descriptions for each of the variables belonging to the different datasets. It should also specify the structure and content metadata should have to facilitate the search of different datasets.
Regarding what type of data should be accepted into the sharing platform we conclude that it is important that strictly PHC-related data can be included into the platform, especially in the starting stages.

This approach is in line with the found platforms, which had a narrow and well-defined scope and made efforts to standardize the data.

Data governance

This parameter determines who decides on who is allowed into the sharing platform, who gets access to what dataset and who defines the standard and enforces it.

Who decides who gets access to the platform?

First, we need to determine who determines who can join the sharing platform. To join the following mechanisms could be implemented:

- **No specific access control, since anyone who requests access may access the platform**
  This option would attract the most potential users to the platform but would not be preferred by data providers, since these users would get instant access to their metadata.

- **New members can be invited to the platform by active members of the PHC alliance**
  This would speed up the process of acceptance with members being able to introduce new potential partners.

- **A third-party outside of the PHC Catalyst alliance is chosen by current alliance members**
  This third-party decides who gets accepted based on credentials and/or potential contributions to the platform.The biggest issues with this are that a third, trustworthy party should be chosen and determining the potential contributions would be a subjective decision that could lead to conflict.

- **Establishing a governance body composed of members of the DSP that decides who gets accepted into the platform**
  This option would be the best at representing all of the stakeholders' interests but could lead to a slow bureaucratic admission process that could hamper the growth of the platform.

Who decides who gets access to the datasets?

Once users are in the data sharing platform we also need to determine who decides if a user can get access to a specific dataset. Two options arise:

- **Allowing the data providers themselves to allow or reject access to datasets**
  This option would probably be the preferred one by data providers since it keeps their power of decision over what happens with their data. However, this could also lead to some data providers never accepting any data exchange.

- **Establishing a governance body composed of member so the DSP that gets to give or deny access to the datasets based on the proposed projects**
  This would help foster a sharing environment but data providers would be highly unlikely to give up their control over their datasets.

Who makes sure the standard is adhered to?

- A third party makes sure that the data is relevant to the platform and complies with the established standards, or

- A governance body: ensures relevancy and compliance with the established standards

Advice

Access to the platform granted through invitation, access to data granted by data providers. Governance of the standard by either a platform member or third-party.

The PHC alliance contains many and diverse members and as the project evolves the expectation is to add on to this diversity. Therefore it's critical to have a structure that allows all parties' interests and concerns to be covered.

Therefore for governance we propose a model where access to the sharing platform is granted through invitation by a member of the PHC institute. This allows for a pre-filter to make sure that added members are trustworthy while keeping the process quick and simple which facilitates the inclusion of members without complex bureaucratic processes. Thanks to that researchers will have an easier access to the information they need to develop their project proposals.

This still keeps the data safe since platform users will only have access to the metadata.

Access to the data will be granted by each data provider, not asking members to give control over their data would increase their likelihood to join the sharing platform as contributors. Even though this also means companies could reject any request sent to them we feel the sharing environment would already increase their likelihood to give access to their data.

This approach is shared by most of the found solutions.

Regarding governance of the data standard we consider that a dual system should fit the storage dual system. Therefore sharing platform users should be able to freely submit their data to the unstructured storage system. However, to submit to the structured database a check should be performed by either a different member or a third-party to ensure the data complies with the specified standard.

Data permission

How do users get access to the data?

Data permission could be granted in the following ways:

- **On a per-project basis: users present a detailed project proposal that can be accepted or rejected**
  This method would allow to ensure no malicious projects are accepted and provide a legal basis for the protection of data.

- **On a subscription basis: users pay a fixed amount of money to be part of the platform and with that fixed contribution they get access to a specific amount (or value) of data**

- **Open access to the data for all users of the platform**
  This would facilitate research the most but as mentioned in previous open parameters, it would put at risk the value of the data contributed.

Advice

Access granted on a per-project basis.

Taking into consideration the governance structure and the objectives of the PHC alliance together with the need for the data providers to maintain data ownership and control we considered data access through a per-project basis to be the optimal solution.

This procedure enables data providers to allow or deny access to their datasets depending on the project at hand, the people involved, the time period in which they will have access and many other conditions. With this freedom parties will be more likely to find agreements that foster advancements in the field.

These data sharing agreements could also serve as a legal protection for the members involved.

Data distribution

How is the data distributed to permissioned users?

As important as who gets access to data is how this data is distributed to users. Different solutions imply different levels of data security. The possibilities are:

- **Direct download of the data after access has been granted**
  This option is the most direct but also the least secure. A legally-binding contract could be put into place to avoid users sharing data or misusing it.

- **Access to data through an APIThis option would be similar to that of direct download but giving a common API would facilitate downloads**

- **Access to data through a secured server with limited connections**
  This option increases the security while also increasing the complexity of data transfers (due to the need to set up different servers that can be deployed when access to a dataset is granted). The increase in security comes from the possibility of disabling outside connections from the server, which would disable the possibility of unlawful sharing of the data. It would also be possible to give access for only a specific period of time.

- **Limited access to data for distributed learning**
  This option is one that would only become feasible later in the platform implementation process but has the potential to greatly reduce data security concerns. It consists of a technique that allows training of predictive models on data without the need for retrieval of the data. Some examples include the Personal Health Train or Google's federated learning.

Advice

We consider access through a server to be the optimal final solution.

It is important to ensure the security of the data transactions so due to this and the shift of most business' data to cloud servers we consider this the optimal solution. Furthermore, by using a dedicated server for data transactions they could be more easily monitored and limited to specific types of analysis, and limited or temporary access.

In the future, using distributed learning could provide additional layers of security for the more sensitive datasets. Therefore this option could be added to a data sharing platform implementation in its later stages. However due to the complexity of such an implementation and how it limits the possible types of data transactions (only training models) it should only be implemented on later phases.

Resulting IP

How is ownership of resulting intellectual property distributed among the involved parties?

One of the main causes for disagreements when developing projects that involve data sharing is the distribution of ownership of the resulting IP. When data providers grant access to their data and a different user develops an innovative algorithm to use their data, who should own the resulting IP? The following options arise:

- **Appointing a third party that evaluates the estimated contribution of each of the members (researchers, data providers, etc.) and splits ownership of the IP accordingly**
  As with all third-party solutions this would require all parties to trust said party to be impartial.

- **Determining a priori together with the project proposals by the parties involved**
  They could equally split, make it proportional or even auction off ownership of the resulting IP. This option would lead to slower project proposals and data sharing due to the needed negotiations but would provide a clear and predetermined legal basis which would avoid future conflict. A possible issue is determining the nature of the potential IP results before actually completing the project.

- **Freely distributing the resulting IP among platform members**
  This option would create a high collaboration environment but is unlikely to be successful due to unbalanced contributions by the different stakeholders.

Advice

IP ownership should be determined by the involved members and included in the data sharing request (and data sharing agreement).

A big part of the value of a data sharing platform is the intellectual property resulting from combining previously separated datasets and expertise. Since the fair distribution of the results could lead to conflict among parties we concluded the best possible solution was for the involved members to determine the distribution before concluding the data sharing request. Some options could include either an even split, a proportional split according to contributions or all the resulting IP going to data providers

DSP Design: Summary

Prerequisites

Ownership of data	Ownership of the source data always stays with the original owner, unless the ownership of the source data is transferred to a different owner.
Permission to the platform	An invitation-only platform.
Access to metadata	Open sharing of metadata among platform users.

Platform parameters

Platform styles

Data storage	A centralized data store containing both unstructured and structured private data vaults	OP	MP	CP
Standardization	A glossary with clear and specific descriptions of the variables. Strictly PHC-related data.	OP	MP	CP
Data governance	Access to the platform granted through invitation, access to data granted by data providers. Governance of the standard by either a platform member or third-party.	OP	MP	CP
Data permission	Access to, and use of, data granted on a per-project basis.	OP	MP	CP
Data distribution	Access to data via a server with limited connections.	OP	MP	CP
Resulting IP	IP ownership should be determined by the involved members and included in the data sharing request (and data sharing agreement).	OP	MP	CP

Sharing is Caring?

Appendix A

Current implementations and frameworks

Appendix A

Existing Data Marketplace solutions

DX Network

Ocean Protocol

Existing Data Marketplace solutions

- They offer a technical framework for the creation of DMPs.
- Mostly blockchain-based solutions.
- Buying and selling through a cryptocurrency of its own.
- Tokenization of assets.

DX Network

Seems to not be active since the end of 2018. Right around what could be referred to as "The blockchain winter". Seems like together with many other projects that heavily relied on blockchain and cryptocurrencies.

According to the founder this project failed due to the decentralised approach with free floating cryptocurrencies.

Use cases are mostly related with business data, including data on food companies and crowdfunding platforms.

Ocean protocol

Seems very developed and with an active community. Open documents white papers and implementations. Supported by a Singapore based non-profit foundation.

Some use cases include chronic disease data and heart disease data (in collaboration with Roche Diagnostics).

Appendix A

Current data sharing platform implementations

Genomic data commons (GDC)

Clinical Study Data Request (CSDR)

Genomic data commons (GDC)

Stakeholders

National Cancer Institute (NCI): part of the National Institutes of Health (U.S. Department of health and Human services).

Scope

Offering cancer researchers a unified data repository enabling data sharing among cancer genomic studies for precision medicine. Mainly focused on giving access to government funded data to researchers.

Solution

Metadata	Metadata and part of the data is open for everyone
Data permission	Access to GDC controlled data through submission of a project request (only available for senior investigators)
Standardization	Submission of data possible after strict quality assessment and standardization procedure performed by individual NIH Institutes and Centers
Data distribution	Direct download of data or through API

Clinical Study Data Request (CSDR)

Stakeholders

CSDR is a consortium of clinical study sponsors/funders among which are:

- Roche
- Bayer
- GSK
- Cancer Research UK
- Medical Research Council.

Scope

A platform where researchers can obtain access to high quality patient-level data with the objective of facilitating innovative data-driven patient care. Data is high quality and offered by a range of clinical study Sponsors/Funders. The platform is research-friendly and includes an independent review of proposals and protection of data privacy.

Solution

Metadata	Data description is open for everyone. Possibility to ask questions about the data.
Data permission	Access to data available through research proposals. Three stages: first Welcome trust then Sponsors/Funders then Independent Review Panel.
Data distribution	Data is accessed through a secure and controlled data access system only after signing a Data Sharing Agreement (DSA).
Resulting IP	Publication of research results into CSDR is part of the DS.

Appendix A

Data Marketplace case studies

oneTRANSPORT

BDEX

Quandl

JoinData

Datapace

oneTRANSPORT

Transportation data

Permission to platform	Subscription-based service.
Data permission	Data providers get to determine the price of their data, oneTRANSPORT takes 10% of transactions.
Data distribution	Handles both B2B and G2B data distribution.
	Defines three roles: Data Publishers, Data Enhancers and Data Consumers.

BDEX

Consumer data

Permission to platform	BDEX as an entity determines who can sell in the platform.
Standardization	Data passes strict quality and standardization controls.
	Stronger focus on the Data Consumer side, very little information about the conditions under which Data providers share their data.

Quandl

Financial and economic data

Data permission	Possibility to buy and sell data through Quandl.
	Pricing set by the data providers, terms and conditions for access set by them too.
Standardization	Contains both market data and what they call “alternative” data. This alternative data consists of non-traditional data assets that can bring value to institutional clients.

JoinData

Agricultural data

Standardization	Dutch initiative to connect commercial enterprises, knowledge institutions and agrarian businesses. Their objective is to share, reuse and combine data.
	Mainly focused on smart sensor data to develop applications.
Data Governance and Data permission	They include free and licensed data. They also define the concept of aggregated data (anonymized data). This allows for two types of permissions, either by the data provider or the data subject.

Datapace

IoT data

Data storage	A decentralized marketplace based on Hyperledger.
Standardization	Focus on data streams of data coming from sensors (IoT).
Data Governance	Private and permissioned blockchain, so control possible by the stakeholders.
	Seems to be in the early stages of development.

Sharing is Caring?

Appendix B

Sources

Sources

Marketplace platforms

1. DX Network: Promotes itself as a real-time marketplace for structured data. Based on ethereum and blockchain technologies. Some real use cases here.
2. Ocean Protocol: A protocol for decentralized data exchange. Built upon blockchain technology. A potential case study here.
3. Hu-manity: They try to let users decide what companies get access to their data (starting with health data).
4. Dawex: For the creation of DMPs. More of a tool to create the system.

Marketplace implementations

1. OneTransport: IoT platform for mainly transportation and city planning data, both historical and real time. Subscription based service
2. Centiva: Blockchain based health information platform that connects users to companies.
3. DataSpace: IoT data marketplace from sensors. Mostly stream data. Based on a permissioned blockchain.
4. Openprise DMP: Purchasing third-party Marketing & Sales data. Strong focus on data standardization.
5. BDEX: Purchasing consumer data.
6. Quandl: financial data for investment professionals. Part monetized part free. Plenty of datasets.

Data exchanges/marketplaces implemented in the Netherlands

1. https://www.ishareworks.org/en: Logistics sector
2. https://amsterdamdatascience.nl/news/launch-of-the-amsterdam-data-exchange-amdex/: Data exchange within the Amsterdam region
3. https://www.join-data.nl/?lang=en: Data sharing in agriculture (Wageningen
4. https://www.medmij.nl/en/ Health data exchange standard in the Netherlands

Data exchanges/marketplaces in the medical field

1. https://gdc.cancer.gov/
2. https://projectdatasphere.org/projectdatasphere/html/home
3. https://www.cancercoreeurope.eu/data-sharing
4. https://icgc.org/
5. https://www.clinicalstudydatarequest.com/Default.aspx
6. https://www.ga4gh.org/
7. https://media.sitra.fi/2018/11/14144842/261018-ihan-blueprint-2.0.pdf

Legal, contractual and other interesting readings

1. https://ec.europa.eu/digital-single-market/en/guidance-private-sector-data-sharing
   a. https://ec.europa.eu/digital-single-market/en/news/staff-working-document-guidance-sharing-private-sector-data-european-data-economy: EU Guidelines on private sector data)
   b. https://ec.europa.eu/digital-single-market/en/news/sme-panel-consultation-b2b-data-sharing: results of questionnaire on b2b data sharing)
2. https://www.government.nl/documents/reports/2019/02/01/dutch-vision-on-data-sharing-between-businesses
3. https://www.ncbi.nlm.nih.gov/books/NBK91503/
4. https://ec.europa.eu/growth/industry/policy/digital-transformation/big-data-digital-platforms/b2b_en
5. https://www.imi.europa.eu/projects-results/project-factsheets/melloddy
6. https://www.smartindustry.nl/wiki-smart-industry/data-delen/: Legal guidelines for data sharing.
7. https://www.sciencedirect.com/science/article/pii/S0958166918301903: paper studying big data analytics for precision medicine, includes some of the biggest current initiatives)

Sharing is Caring?

Appendix C

Interviews

Interviews

As part of the validation of the preferred solutions, the following subject matter experts were interviewed:

Andre Dekker	Maastro clinic
Xander Verbeek	IKNL
Ron Herings	VU Amsterdam
Kees Stuurman	Considerati
Judith van Schie	Considerati
Marc van Lieshout	TNO

Sharing is Caring?

Appendix D

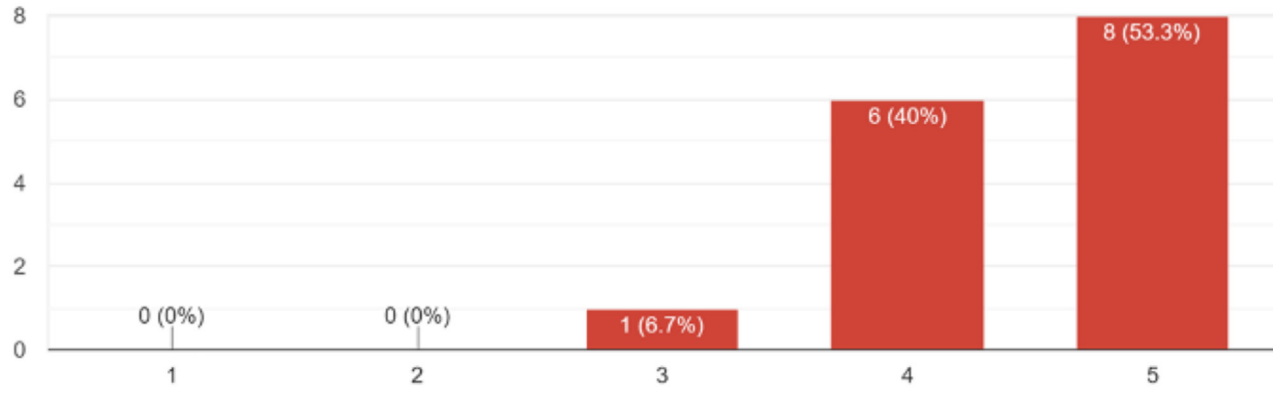
Survey results

Survey - Explained

The survey was sent to 40 people, of which 15 responded. The answers to the questions are depicted on the next pages. Some of the questions used a Likert scale of 1 to 5, where 1 means: low impact, and 5 means: high impact. The results of these questions are represented in bar charts. Other questions used multiple choice answers; results of these questions are represented in pie charts.

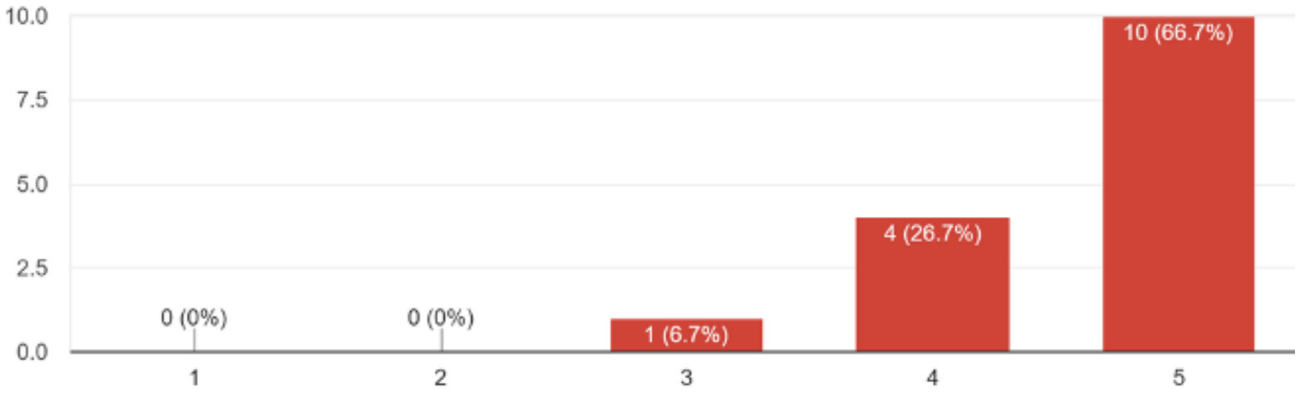
Ensure that data from different sources can be combined

15 responses



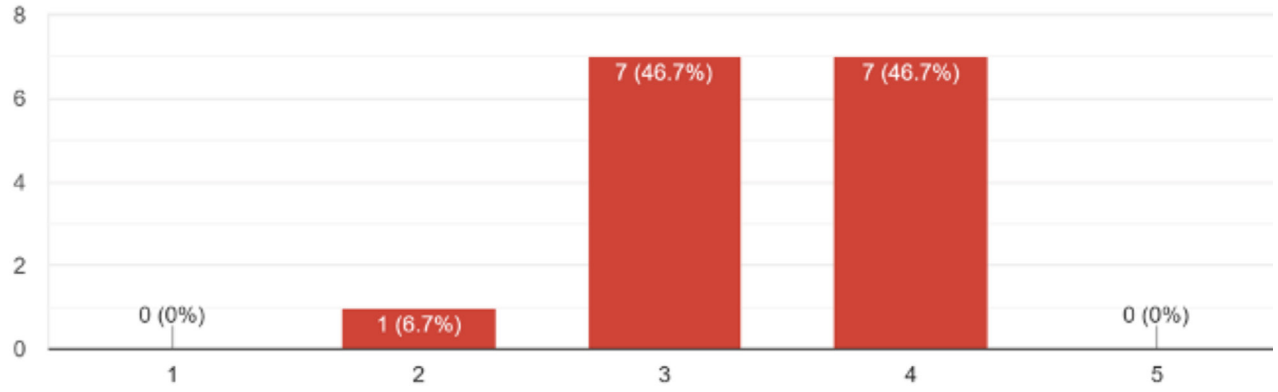
Make sure medical data and patient related data is shared within the boundaries of the legal

15 responses



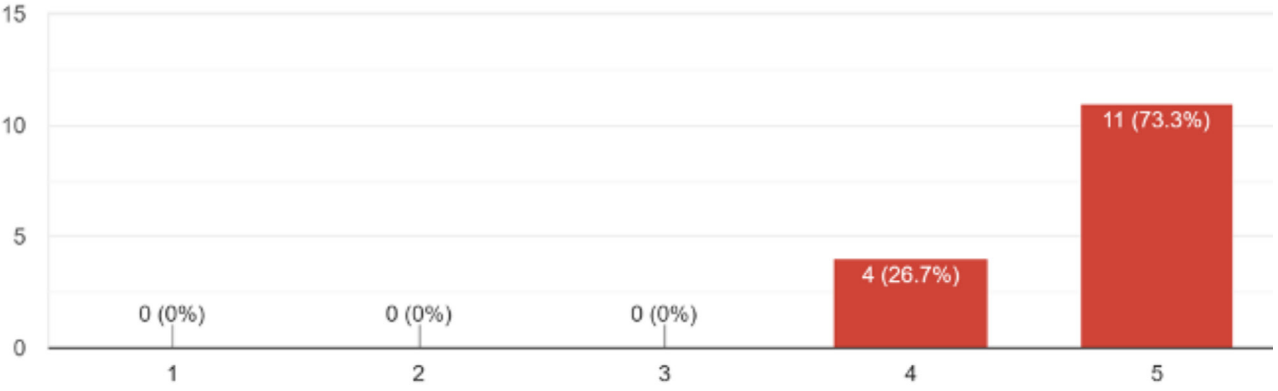
Value data as an asset

15 responses



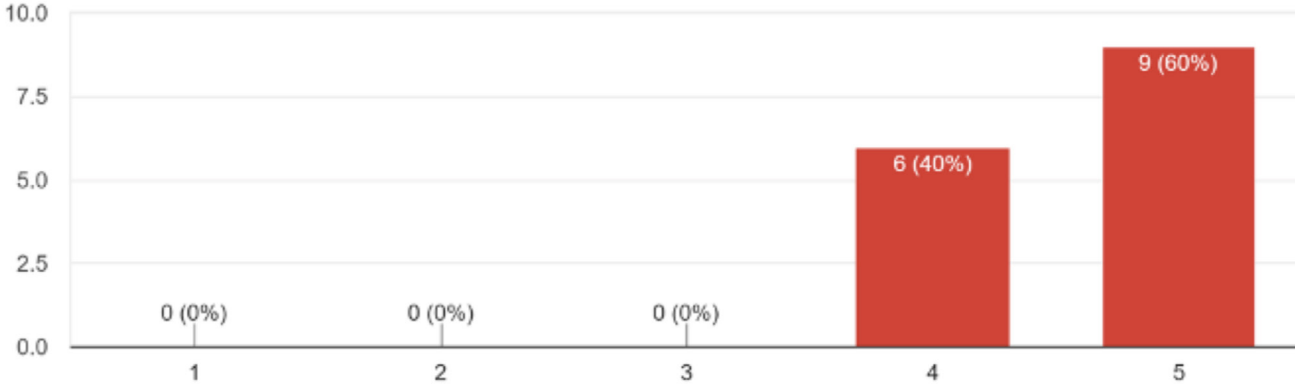
Develop trust between involved parties so they are willing to share data

15 responses



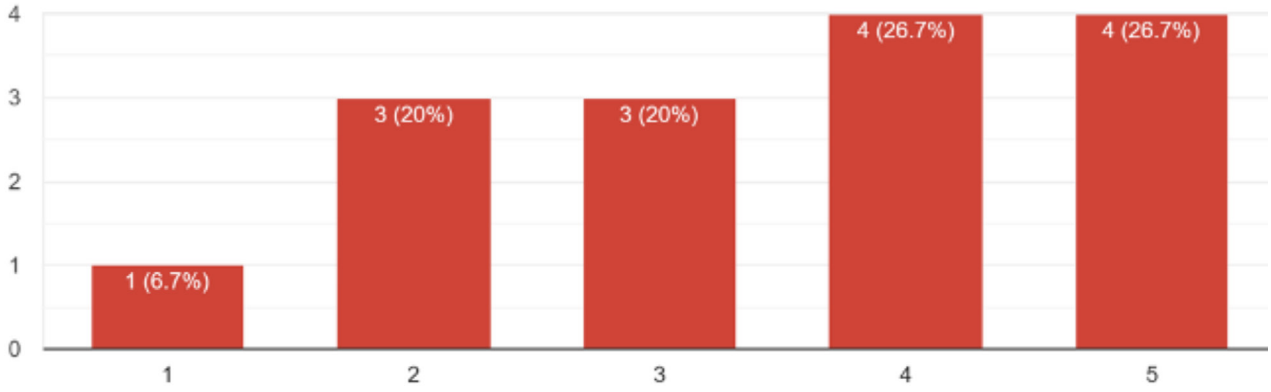
Develop culture in which sharing of data is given

15 responses



Develop the technologies needed for a DSP

15 responses



Who should be able to acces a description (metadata) of the data shared on the DSP?

15 responses

Where should data be stored?

15 responses

How should data be stored?

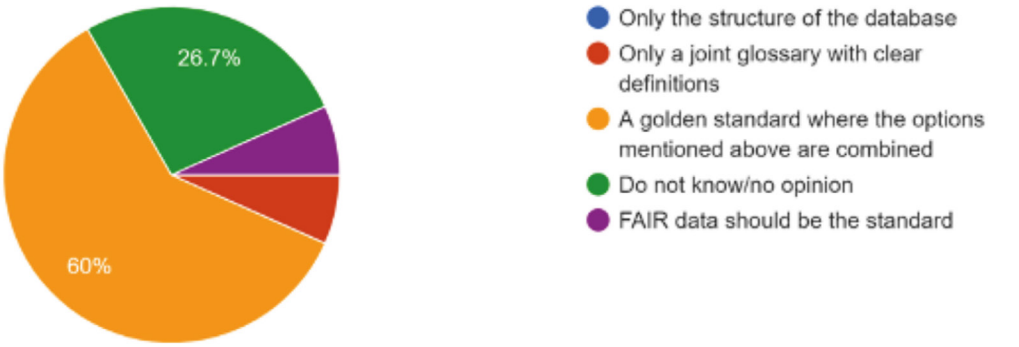
15 responses

What data should be accepted on the platform?

15 responses

What should be defined in a DSP standard?

15 responses



Who decides who gets acces to the platform?

15 responses

Who ensures adherence to the data standard?

15 responses

Who gives acces to datasets?

15 responses

How do users get acces to the data?

15 responses

How is the data distributed?

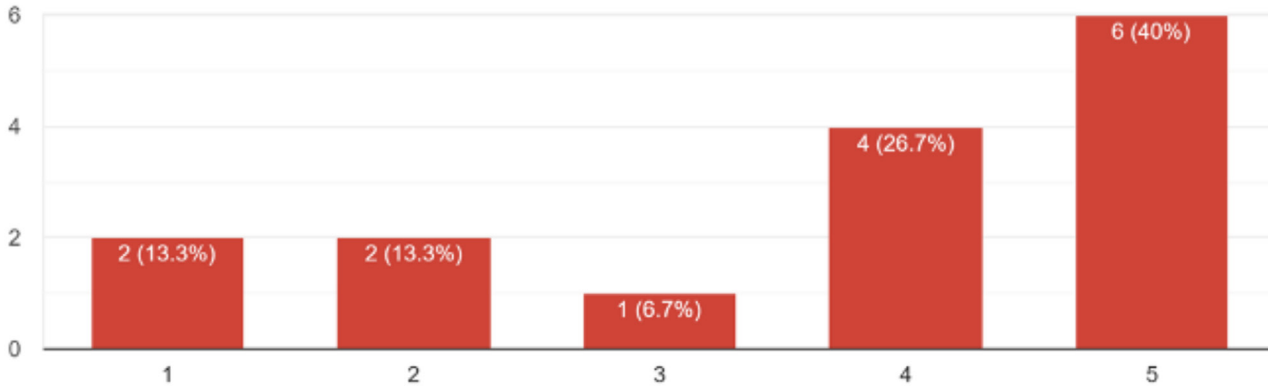
15 responses

How is ownership of resulting intellectual property distributed among the involved parties?

15 responses

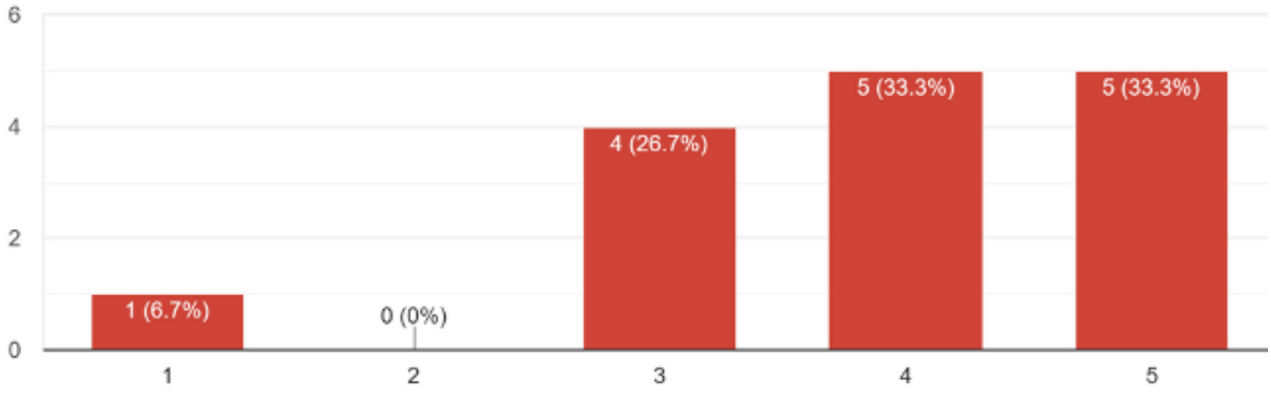
A DSP requires data sharing agreements for all data transactions

15 responses



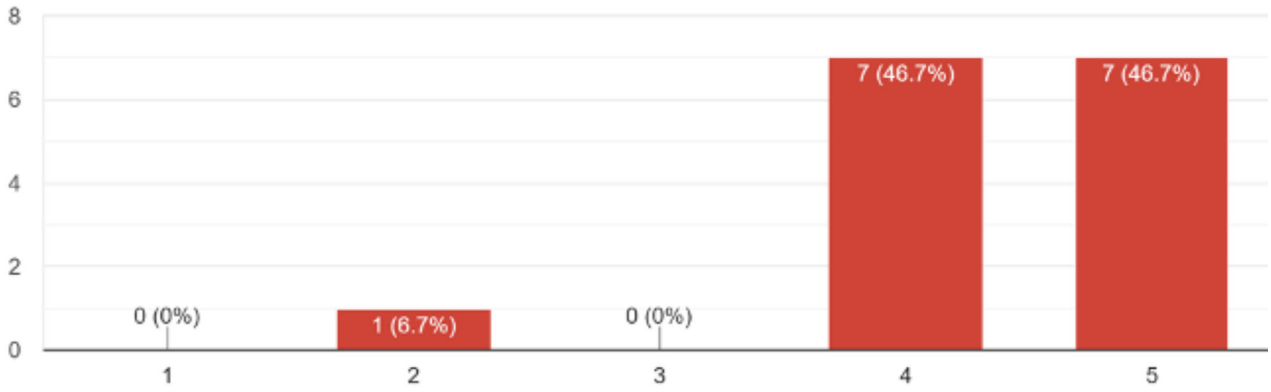
Metadata are accessible for all registered users of the platform

15 responses



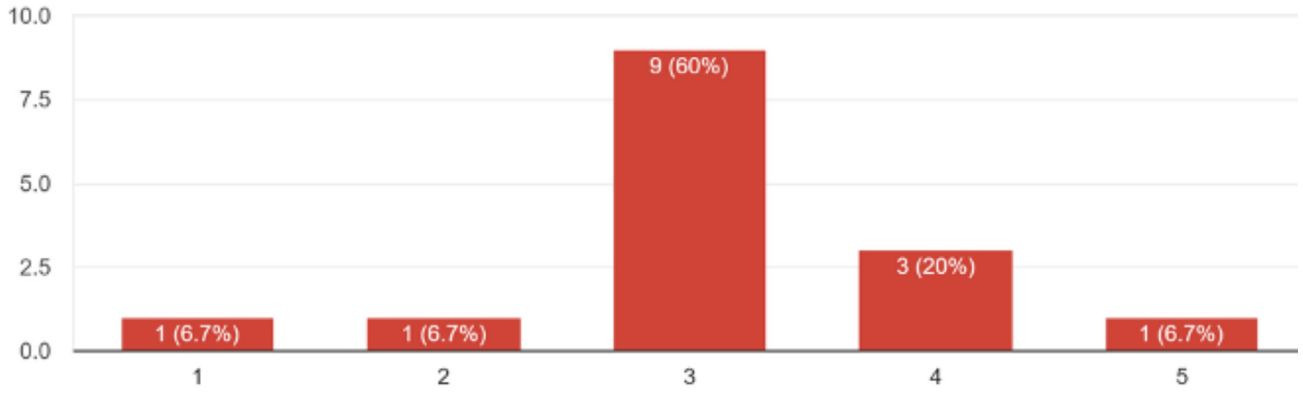
Data providers always maintain full control over the data they provide

15 responses



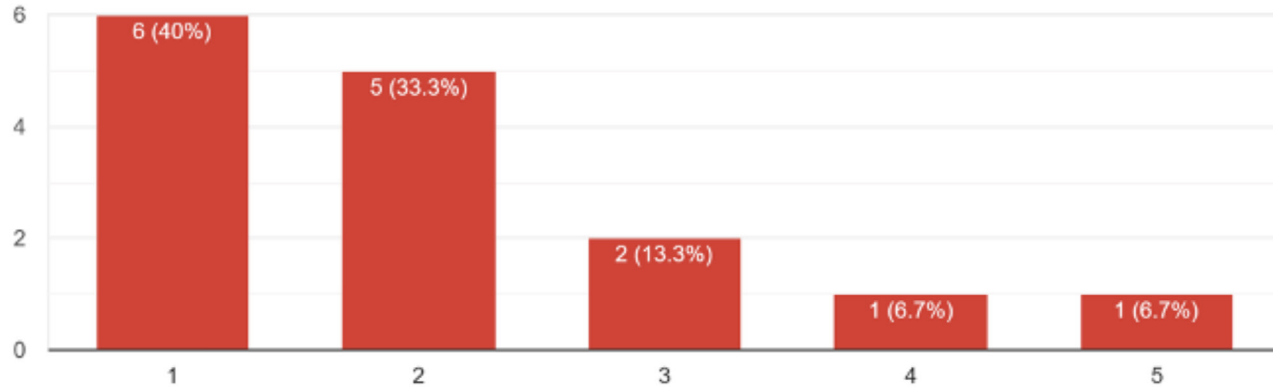
A data sharing platform enables economic transactions with data

15 responses



Data are stored on a central location for increased efficiency of distribution

15 responses



Only authorised users are able to acces the DSP

15 responses