Workspaces Architecture

Use cases descriptions

These use case descriptions are related to the workspace architecture which can be found at:
https://docs.google.com/presentation/d/1L17rYK1UXCDfbR5pPw7qXy2CgHghfhJxizmsnDBfKa8/edit#slide=id.p

0a) Workspace

<<<start van workspace beschrijven ...datavraag>>>>

A workspace is a specific study that can be accessed via the DRE (Digital Research Environment). The workspace is there to process data. The data capturing resides outside the workspace.

Ideally, the workspace can be viewed as a specialized virtual machine that can be access from any device, any place and anytime with the right authorization.

It is not known, nor expected, that all studies should or can be processed within a specialized virtual machine. A physical machine that acts as a workspace should provide the required functionality as is depicted in the architecture workspace and is required by the specific study.

0b) Research Playground

This area is, depending on the freedom that the workspace owner / user has (see Config Management) for the workspace owner / user to work in.

1) Data is safely made available in the Workspace environment

This about how to make data available for a workspace from environment X

- From data owners
    - Pseudonymised via TTP or alternatives which makes it possible to:
        - Correctly merge / enricht the same subjects from data sources from different institutes (as opposed the birthday-hack)
            - Prevents faulty mergers / enrichments
            - Prevents data doubling
        - Source control by the monitor
        - For an auditable data trail
        - Report accidental findings
        - Via min=max[1] principle providing access to data elements, when more is needed, this can be relatively easy be added at a later stage (also useful for studies that build on a previous study)
    - Via a safe and for users easy to use protocol
    - Data can be made available both virtual as non-virtual in a workspace
        - virtual: e.g. a specific set of images is stored near a suitable HPC
- From different data sources
    - Users can add data via a safe upload data (e.g. from stick) in a self-served way that supports the user to:
        - Upload non-pseudonymised data using TTP or alternative
        - Ingestion of the dataset to map the data

[1] MIN = MAX, the minimum what is needed to realize the objective, is the maximum to do

        - With the Connector API data can be pushed from an outside source into the Data Landing zone of a workspace

2a) Data landing zone

The zone where the data owner makes the data available and the conditions under which the data can be used. The data in this zone cannot be modified by the workspace owner/user.

Ideally the data owner maintains control which workspace can use the data; data can be added or data access can be removed. The latter provides proof the data owner can meet its obligation for being a data custodian (e.g. accidental providing wrong data).

The data landing zone consists out of 2 parts:

- The data and conditions
    - Data = data & meta data
    - conditions = the conditions for usage
        - Usually confined by the research question / objective
        - Possible conditions relating to re-use
        - Possible condition regarding usage (e.g. the data is not allowed to leave that workspace)
    - Data can be removed by the workspace owner based on governance
    - Data access can be removed by the data owner
    - Disclosure risk assessment; if that data turns up in the wrong place, how likely can it be traced back to individual patients
- Log
    - Who made when what data available to this workspace
    - The log cannot be removed, important for:
        - Data audit trail
        - Establish who is data owner (e.g. for data citation)
- Connector API
    - A secure way to push data from an outside source directly into the Data landing zone of a work space

***Discussion / future**: block chaining data sets, this provides audit trails across studies and supports data citation. With big data, data creation and maintenance becomes relatively expensive that can earn itself back through proper citation and licensing.*

2b) Data curation zone

The data curation zone provides the tooling to correct source data in a controlled way; e.g. adding missing values, modifying values. Changed, but non approved / acknowledged are flagged. Who curates when which data when is logged. Also is logged who approved / acknowledged which modification when. The latter is supporting the monitor function in an auditable way.

Depends on license agreements ….

2c) Data sharing zone

A place where a workspace user can put data to be shared; this includes archiving of the study. The data in the Data sharing zone cannot be modified, only be deleted by a workspace owner. The data sharing zone is based on FAIR-principles (Findable, Accessible, Interoperable, and Re-usable).The Data sharing zone contains three parts: Meta, log and Data & Conditions.
The difference with a data landing zone, is who is in control of the data (=can modify): data landing zone = supplier/owner of the data; data sharing zone = workspace owner.

The data sharing zones support three use cases and provides TTP where required/needed[2]:

- Publication
    - Connectors to CRIS-systems
    - Supporting the publication of the study
    - Making the study findable (google like, via 'crawlers')
        - mandatory, fixed entries and system driven tagging
        - free format items and user driven tagging
        - Automatic generation of meta information on the data used (type, size, etc)
- Archiving
    - Restoring (e.g. for audit)
    - Duplication (conducting the same / updated version of the study) with and without data
- Other, like:
    - Providing access to data / developed tooling for other specified workspaces (e.g. follow-up studies)
    - Creation of applications (e.g. (semi) clinical applications)
    - Providing data to registries
    - Download of data (e.g. on a data carrier for usage outside the workspace / DRE; all downloads will be logged: who, when, what & reason and the downloader has to confirm his/her responsibilities)
    - Disclosure risk assessment; if that data turns up in the wrong place, how likely can it be traced back to individual patients

3) Compute & Data Staging

A self-managed solution to make use of different compute solutions. Where need be a self-managed easy to configure data staging to cost-effective buffer high compute calculations.

[2] TTP provides the possibility to:

4) Storage

A self-managed flexible storage solution to provide cost-effective storage during the study. Ideally after a configured amount of time, untouched data will be moved to lower tiers.

Different storage possibilities in the interaction between local and cloud/central storage. Local can be more than one instance.

- “Dropbox++ ~ syncing local and cloud/central”
- read/write local - read cloud/central
- Read local - read/write cloud/central
- ….

5) Configuration Management

Depending on the configuration of the DRE for that particular workspace owner / user, the owner/user has the freedom to self-manage the:

- The authorization / access to the workspace
- Roles
- The type of OS
- What kind of programmes / applications to install and use (especially paid versions)
- The amount and type of storage
- The amount and type of compute

The configuration management should be able to indicate the implications on cost (planning) and register the usage (with ideally remaining budget).

6) Reporting & Search

Reporting

This provides the ability for departmental heads and higher to report over their area of responsibility. E.g. # studies in which phase, burn rate per item, who used what data source. It is a system that is in the periphery of the workspace, the workspace generates/provides input for these systems.

Search

This provides the ability to find data within the workspace, in known data sources (RDP’s) and/or other studies. Search makes use of the meta data in the data sharing zone.

<<verder uitwerken, idealiter; haalbaarheid midden termijn is nog onbekend, aansluiting WP3>>

7) Vault

This provides the possibility to store versions of data workflows, algorithms and if need be (subsets) of data.

- Versions can be saved / retrieved
- Meta data is also stored
- Logging of who saved / retrieved what when

8a) Sharing: Other

- Other, like:
    - Providing access to data / developed tooling for other specified workspaces (e.g. follow-up studies)
    - Creation of applications (e.g. (semi) clinical applications)
    - Providing data to registries

8b) Sharing: Publish

- Publication
    - Connectors to CRIS-systems
    - Supporting the publication of the study
    - Making the study findable (google like, via ‘crawlers’)
        - mandatory, fixed entries and system driven tagging
        - free format items and user driven tagging
        - Automatic generation of meta information on the data used (type, size, etc)

8c) Sharing: Archive

- Archiving
    - Restoring (e.g. for audit)
    - Duplication (conducting the same / updated version of the study) with and without data

9) Data discovery

Being able to find data in various sources and (request) access to external data

- From open to the workspace sources, the user can self-serviced move data in his landing zone.
- For on-request data sources
    - Having access to the meta data (e.g. size, demographics)
    - Being able to view and ‘import’ the metadata and sample data

<<verder uitwerken, idealiter; haalbaarheid midden termijn is nog onbekend, aansluiting WP3, ook governance: governance op de vraag>>

10) Analytics software

11) External services

Access to interface with the a larger world than the DRE is providing for.

12) Analytics to the data

Within the workspace analytics is created. The analytics is send to one or more federated stored data that resides outside the workspace (could be a different workspace) and returns aggregated outcomes to the workspace.

There is no (direct) human access to the targeted data.

Explaining the Architecture of the Workspaces

Setting the scene

In choosing any solution the key question is: will the solution suit my needs.

One thing is for sure, Research is and will always remain a diversifying business. A key difference between BI questions and Research questions, is that for BI in 80+% of the cases both the required data and the questions is known. This makes for instance a data lake a very suitable solution for BI. BI questions can be considered as a longitudinal study. For Research, longitudinal studies is just one type. Other types do require different than data lake kind of solutions.

But also other considerations might apply for a specific study: can I collaborate with others, how much effort do I need to put into making the environment compliant with rules and regulations, what do I need to do to make it auditable, what if I want to develop analytics for (semi) clinical applications, what if I need to upload my own data, access other data sources, how can I manage the cost, how easily do I have access to HPC, how can I version my scripts? And plenty of other considerations that might or might not apply for your specific needs.

The architecture of the workspaces is not an ICT solution, nor a Executive Management solution, nor a Privacy solution, nor a Security solution ...not even a Researcher solution. The architecture of the workspaces is a Research Community solution.

The Research Community has many stakeholders. Each of these stakeholders has questions, concerns and ambitions flowing from their role, responsibilities and activities. It is our believe that any solution can only be viable if it serves the community as a whole. In the discussions with stakeholders we see evidence that seemingly conflicting interests can be bridged and make the solution stronger and better for the whole community.

Of course as the community grows and develops, new questions and concerns will arise that will impact the architecture of the workspaces. We not only recognize this, we embrace the change.

For instance, the future seems to be working in the cloud and on remote/virtual machines. However, we do not exclude a local physical install; be it for legacy products, or for whatever reason.

If there is one thing key in the architecture of the workspaces, it is the word: trust.

The start

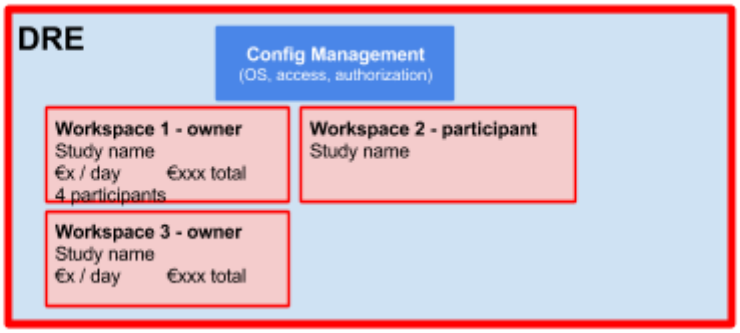
You have a research question; be it already very formalized, or more of the Friday-afternoon / Monday-morning kind of question.

What you then can do is log into the Digital Research Environment (DRE). The DRE is a portal. A portal where you can access all the workspaces; those you created yourself or were invited to.

A workspace is basically a secure environment for a specific research question. If you create the workspace, you are the workspace owner. You can choose from a menu for instance the type of OS, type and amount of memory, storage and compute. This can be adjusted according to the needs during the research. But, a for you standard pre configured workspace can also be selected; saving a lot of non-added value time.

If you want, you can (temporarily) invite colleagues; be it from the same faculty or another one from the other side of the world.

From the DRE portal you can configure self-serviced the workspace and all your workspaces and the ones you participate in.



The level of self-service and configuration possibilities can be limited by the organization or department you work for. It is after all a Research Community solution and the environment you work in must reflect the organization/department you work for.

It is our belief that the hard implementation of organizational and departmental policies should be limited as much as possible. The mechanism that will support this, is logging: who did what when, who had access to what when. Not only is a hard implementation of policies difficult and expensive to maintain, it invites ‘creativity’ for the right or wrong reasons. If you can’t trust people, you should not allow them to work within the system. It is easier to trust people if everybody knows that all relevant activities are recorded.

For instance, if you are a PI and want to have some external support on some analytics, you don’t want to send the analytics with some data to the outside world. Nor do you want to get

a service ticket from ICT to allow somebody access to your workspace. What you want is to provide temporary access to your workspace to help you out. Of course you trust this external support, but the both of you knowing that all activities will be logged (i.e. data downloads or uploads outside your workspace), makes it easier to trust.

When reading the use cases below, bear in mind the following:

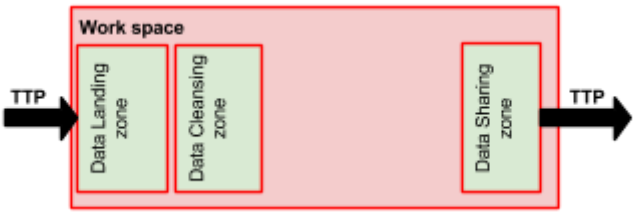
- It is about architecture
  Architecture provides direction (for the future) and is less concerned about current feasibility
- That said, this architecture is checked with people from the industry and aims easy implementation of solutions
- Research is very diverse, and no document can contain all the use cases for Research
  If you find something that you need but seem to miss, please contact us

Getting data

Assuming you are not generating data from within the workspace, which is also a possibility, you need to 'import' if not find data.

Before moving into the possibilities, it is important to know that a workspace has three Data Zones:

- Data Landing Zone
- Data Cleansing Zone
- Data Sharing Zone

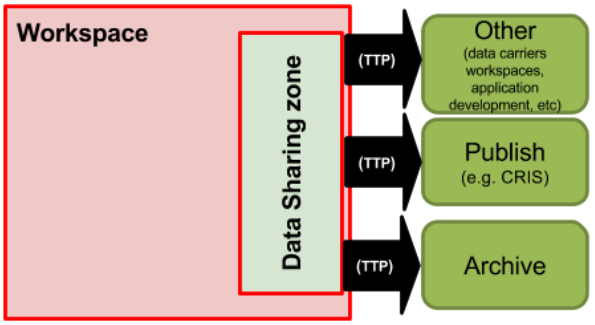


All data that is imported into your workspace will land in the Data Landing Zone. Not just the data, but when available and applicable also the meta-data, the licence agreement, the source, who provided the data and everything date stamped. You as a workspace owner can only remove a complete data set (for instance to free space/reduce cost of wrong data), but not the logging. Nobody can modify the data that resides in the Landing Zone. This will ensure a clear data audit trail provided by the system.

The Data Cleansing Zone is there to clean up the data. Because logging takes place who cleaned up what data when and optionally why, again a clear data audit trail exists. It is also the place where a monitor can verify if the data source was correct and any such cleaning

that might take place. Also the findings of the monitor are logged in the form of: who, when and what.

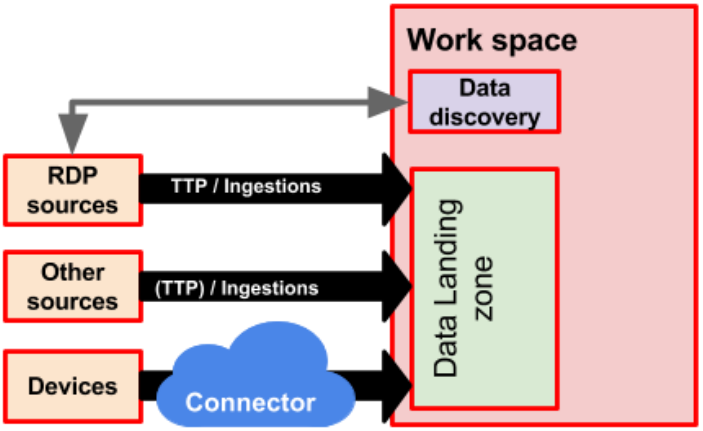
The Data Sharing Zone is there to provide access to your data and/or analytics to other workspaces, for publication, etc. This includes the available metadata and licence agreements applicable to the shared data and/or analytics. It provides the ability to get data citations and analytics citations.



For both the Data Landing Zone and Data Sharing Zone, revoking data by the owner is possible. We know it is not a popular feature for many, but think of it this way: trust is required to provide access to data/analytics, the level of trust that is required can be much lower when you know that you can revoke. This idea is applicable at data donor/patient level, but also on organizational all the way to workspace owner level.

There are multiple ways to get data.

- From a data carrier like a stick, drive, CD.
  You can import it and a wizard will help to ingest the data and metadata into your Data Landing Zone.
- From a device.
  You can create a database in the Data Landing Zone and provide credentials for to access an API so that one or devices can securely upload data directly into the database.
- From prepared clinical sources and disclosed studies
- From other sources, like municipal data



Before, or after all this, you might also want to see what data is available where through Data Discovery. With Data Discovery data and metadata can be found and explored at aggregated level over all connected organizations. Data Discovery supports Federated Search.

Some sources require a formal request and review before access might be given for your study. This workflow can be triggered for instance via the Data Discovery.

Getting data might require a specialist like a Data Scientist. Via the configuration management in the DRE (temporary) access can be given to such a specialist in order to get the data inside your workspace.

One of the challenges is sometimes merging different datasets to enrich the data. Not uncommon a birthday-gender hack is used. With larger datasets this can become more and more of an issue. Especially when data needs to be pulled from different data sources. Using TTP each record be coded uniquely identifiable by the TTP provider. Meaning re-identification is possible for future enrichment or usage of the data, and reporting accidental findings without the researcher having to know the relevant person.

When you share data, all data will be repseudonymised via the TTP. This meets regulations while it still enables enrichment and allows the reporting of accidental findings.

Analyzing and processing the data

To analyze and process data programs, apps, scripts, statistical software, etc is required. There are no restrictions from the workspace what you can use. A reason why, a, that part of the workspace is called Research Playground.

Analytics is available through a webstore where vendors/owners can provide their analytics that can be used in the workspace. The conditions and licence models can vary.

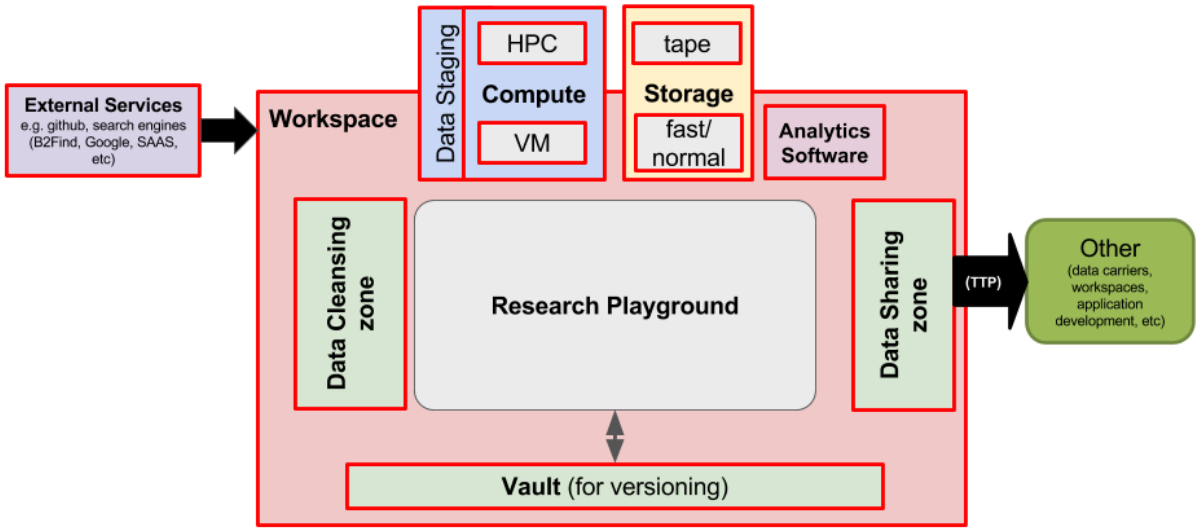
If SaaS solutions are provided and available through the webstore for your workspace, you can assume they are compliant with your organization's policies. If a SaaS solution is not

available in the webstore, a proof of compliance to your organization's policies must be obtained. This will not block using the SaaS solution, but the due diligence then becomes your responsibility.

Solutions exist to make use of (local) HPC if it is required. As well as beefing up your workspace for temporary number crushing. This can be done using self-service in the Configurator in the DRE.

Data of course includes documents and drafts that are relevant to the study.

Versioning can be an important asset in your research and for this there is the Vault. It provides the possibility to version data, metadata and analytics.



For sure there will be use cases where the data has to leave the workspace. Downloading data to a portable data carrier, uploading data to for instance an HPC is supported, moving data to and back to cheap external storage is supported.. Who downloads or uploads what data when will be logged, the due diligence is yours.

Of course as a workspace owner you can limit who can upload/download data. However, this is our premise, we're dealing with some very clever people. Rather than play into their creativity driven by frustration, when the option exist, we prefer the logging over restricting.

Managing costs

For some type of research, managing cost of storage, compute and memory might be very important. This can be done through the Configuration Management at the DRE. From putting 'hybernating' the workspace to the cheapest storage level, to having access to a lot of compute, fast and plenty of storage and a lot of memory.

Dealing with exceptions

For instance the analytics web store will contain a growing collection of analytics. However, not necessarily (yet) what is required for your study. In that case, if it is available on the web, you can find and install it through the External Services. Here you might want to use a Github like solution for your versioning.

Reporting & Search

You, the departmental head or even the organization as a whole need to report. Usually the required data is available in more than one place and yet very hard to collect. The Reporting & Search module stores relevant pieces of (meta)data regarding the studies that takes place in the workspace.

Through Federated Search others, provided they have clearance, can obtain that information. But also you can ‘crawl’ through all this data. This might be relevant for finding similar (themed) studies, or datasets.

Reporting & Search is as much about you being able to find studies, as others can find you.