From Finding to Re-using Research Data

Research Data Exchange

Emma Schreurs[1], Frans Oort[1], Freek Dijkstra[2], Tim Kok[2], Iza Witkowska[2], Mike Kotsur[3]

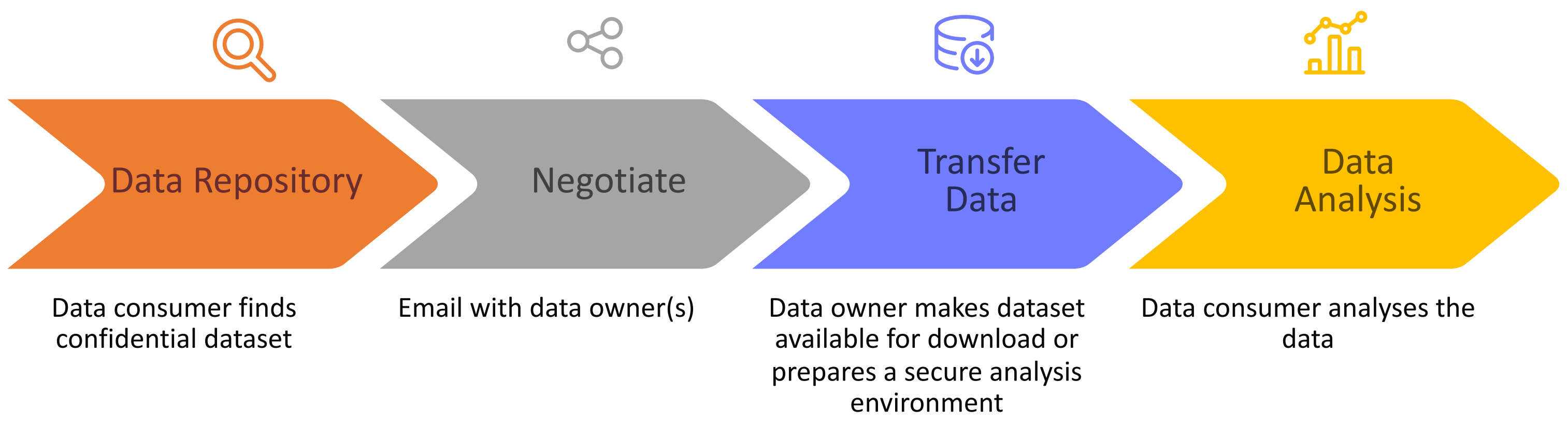
[1] Research Institute for Child Development and Education, Universiteit van Amsterdam; [2] Innovation lab, SURF; [3] Absolute Value

Research Data Exchange (RDX) allows researchers to share data in a controlled and secure manner, while also adhering to legal requirements and institutional policies. RDX is a prototype that integrates existing data repositories and algorithm-to-data solutions and is the next step in solving the Open Science Dilemma.

Current Situation & Problem Statement

There are analysis tools available that allow for the re-use of data while ensuring its security (e.g., confidentiality and knowledge safety) (see *Integration with Existing Tools*). However, the current process requires a manual effort each time a researcher wishes to utilize a dataset provided by someone else. In this process, the researcher, acting as the data consumer, must locate the dataset on existing data repositories (e.g., DANS or OSF). Unfortunately, the lack of a download option for datasets means that the only available course of action is to communicate via email with the data owner and hope that they are willing to either provide the dataset for download or make it accessible within a secure analysis environment. This method is not only tedious and time-consuming, but it also places a burden on the data owners.

Existing workflow — re-using data



Data consumer finds confidential dataset

Email with data owner(s)

Data owner makes dataset available for download or prepares a secure analysis environment

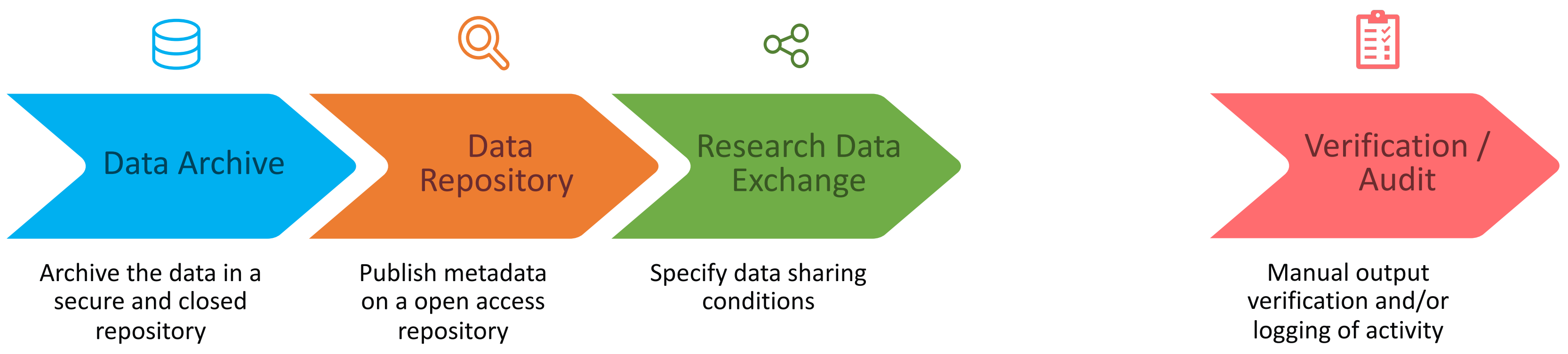
Data consumer analyses the data

Solution: Research Data Exchange (RDX)

Research Data Exchange is a prototype that automates the process of making data available for re-use. In contrast to the current situation (see *Current Situation & Problem Statement*), the workflow is divided into two: one for the data owners and one for the data consumers.

Publication workflow (for data owners)

With RDX, a data owner can **specify the data sharing conditions** for each dataset (see *Which Data Sharing Conditions?*) and makes the dataset available for re-use. This only needs to be done once, at the same time the (meta)data is published on a data repository to make it findable.



Archive the data in a secure and closed repository

Publish metadata on a open access repository

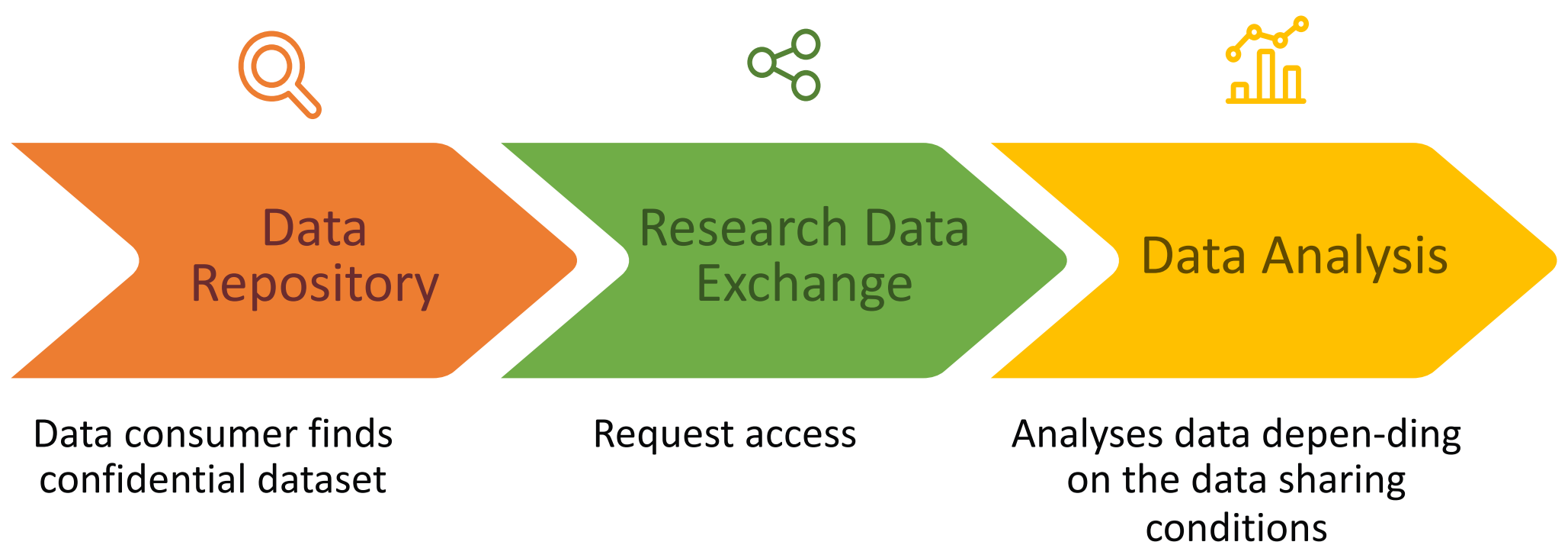
Specify data sharing conditions

Manual output verification and/or logging of activity

If desired by the data owner, there is an option to perform manual output verification for each analysis conducted by a data consumer, as well as access the records of previous analyses conducted on the dataset. This level of control is crucial as it empowers the data owner to maintain complete oversight of the dataset and its usage.

Re-use workflow (for data consumers):

When a researcher is interested in re-using a dataset, they can still locate the (meta)data on an existing repository. However, instead of engaging in negotiations with the data owner for access, the RDX prototype automatically enforces access permissions. Depending on the data sharing conditions, the data consumer must first prove its affiliation (e.g., be part of an existing research community) and agree to the designated sharing conditions (e.g., non-commercial use or citation requirements). Once these conditions are met, the data consumer can proceed to download the data or conduct analyses within a secure analysis environment. The specific actions allowed are contingent upon the data sharing conditions established by the data owner in the publication workflow.



Data consumer finds confidential dataset

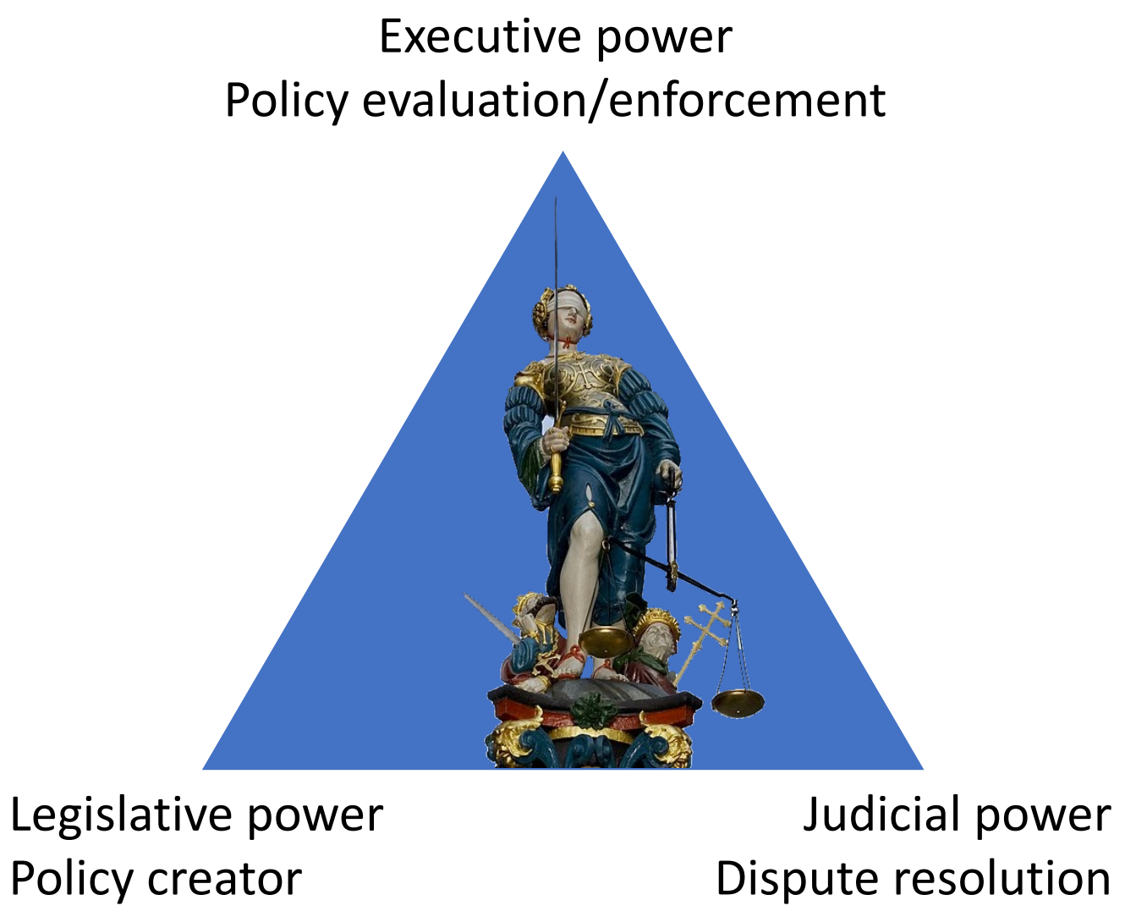
Request access

Analyses data depen-ding on the data sharing conditions

A *Trias Politica* for Data

Our prototype helps in establishing a "trias politica" for data, ensuring a clear separation of responsibilities. Policy creation rests with the data owner, while policy enforcement is carried out by the RDX prototype. This stands in contrast to traditional data lakes where infrastructure providers dictate data sharing conditions, even to the data owners themselves.

Ideally, the introduction of an independent entity would further enhance the framework. This entity would be responsible for resolving disputes that may arise between data owners and data consumers regarding adherence to the data sharing conditions. It's important to note that even the RDX can not enforce every condition. Some conditions cannot be verified beforehand (such as requirement to cite a dataset in a publication) or are not easy to verify (such as the non-commercial use only requirement).



Open Science Dilemma

RDX plays a crucial role in addressing the Open Science Dilemma by offering a solution that enable the publication of all types of data, including confidential data, while still ensuring data control.

On the one hand Open Science advocates for the dissemination of as much data as possible on an open platform to promote scientific progress, enable transparency, and allow for the replication of analyses. Additionally, given that scientific research is often funded with public money, it is important to make research results and data accessible to the public.

On the other hand, legal and sovereignty issues can limit the extent to which data can be openly shared. It is important to maintain control over your own data, which can entail legal issues such as ownership and copyright, confidentiality of personal data, restrictions on informed consent letters, purpose limitations, bans on dual use, and prohibitions on resale.

Integration with Existing Tools

The Research Data Exchange (RDX) enhances the functionality of current tools designed for the re-use of data by seamlessly integrating these analysis tools with data repositories, which host (meta)data. This integration effectively eliminates manual labour for both the data owner and the data consumer, streamlining the entire process.

Data Repositories

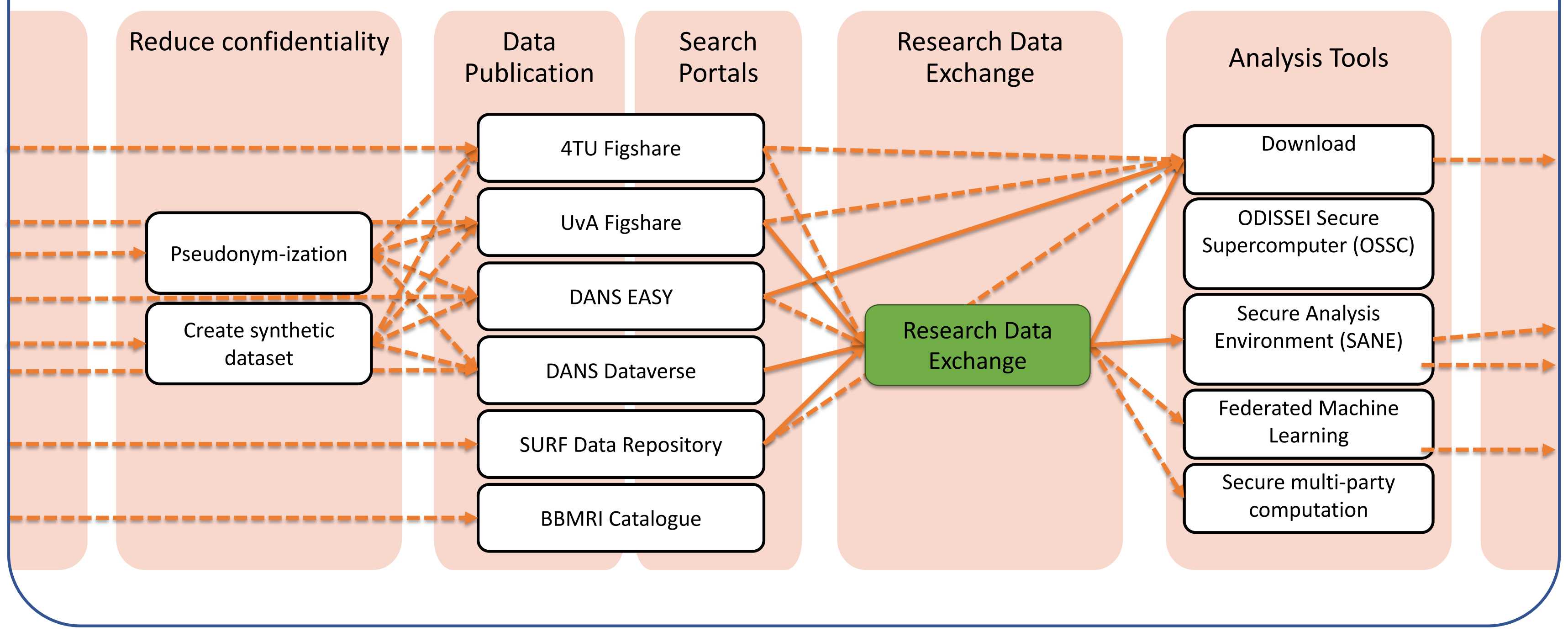
The RDX prototype currently interfaces with UvA Figshare data repository, but it will be no problem to integrate with other repositories. In our upcoming demonstrations, we anticipate showcasing the prototype's compatibility with ODISSEI, which is built on the Dataverse software. Whilst data repositories do not provide data directly to a data consumer, they must make it available to the analysis tools.

Analysis Tools

SURF currently offers two algorithm-to-data analysis tools that enable researchers to perform analyses on datasets without the need to make the data available for download. Both systems operate within secure environments and only provide output to the researcher once it has been verified that the output does not contain any confidential data.

- The first tool is the ODISSEI Secure Supercomputer (OSSC). It allows researchers to analyse CBS data on the Snellius supercomputer, ensuring a high level of security.
- The second tool is the Secure Analysis Environment (SANE), which is a prototype specifically designed for social sciences, economics, and humanities. Developed in collaboration with ODISSEI and Clariah, SANE offers two variants: "blind" and "tinker." The "blind" variant implements a job-submission system, while the "tinker" variant provides a remote desktop solution.

The RDX prototype currently interfaces with the Secure Analysis Environment (SANE).



Which Data Sharing Conditions?

Potential enforceable data sharing conditions

- Verify identity and affiliation
- Sign data sharing conditions
- Download dataset
- Analyse in a secure analysis environment
- Verification of the analysis output before releasing the output
- Only allow verified algorithms for analysis
- ...

Already implemented by the prototype

- Verify email address
- Sign data sharing conditions
- Download dataset
- Analyse in a secure analysis environment
- Verification of the analysis output before releasing the output

Objectives & Findings

The current prototype has two primary objectives: to showcase the technology and to gain a deeper understanding of the roles played by the data owner, particularly the distinction between the researcher who generated the dataset and the data steward at the providing research institute. One key aspect we aimed to explore was the verification of analysis outputs for data: whether it should be performed by the researcher, who knows the ins-and-outs of the dataset, or by the data steward, who has a better grasp of GDPR compliance.

Our findings indicate that both parties desire to retain control. However, for many datasets, it may not be necessary to verify the output of every analysis if robust logging and monitoring mechanisms are in place and there is the option to audit past analyses. While the implementation of an automated output check could potentially aid in prioritizing these audits, this topic warrants further consideration in future discussions.

Are You Interested?

If you are a data owner or data steward with datasets that require specific use conditions, we would greatly appreciate your insights on the viability of this approach. Specifically, we would like to hear your perspective on the data sharing conditions that would enable you to make these datasets available for re-use, the feasibility of enforcing such conditions, and whether the logging and monitoring features provided are adequate for cases where enforcement may not be possible. Your input will help us further refine and enhance the system.

Please email freek.dijkstra@surf.nl to reach out to us!

The Research Data Exchange (RDX) is a use-case in the Amsterdam Data Exchange (AMdEX) project.
The AMdEX project is sponsored the European Union and the Province of North-Holland.

SURF

AMdEX

European Union
European Regional Development Fund

Provincie Noord-Holland