DATA ACCESS PROTOCOLS

Deliverable D1.4

December 2022

Project Acronym	Grant Agreement #	Project title
ATHLETE	874583	Advancing Tools for Human Early Lifecourse Exposome Research and Translation

Nature	Dissemination level	Leading institution
Report	Public	UMCG

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 874583

DELIVERABLE REFERENCE NUMBER AND TITLE	**D 1.4** Data Access Protocols V1.0
LEADING BENEFICIARY	UMCG
NATURE	REPORT
DISSEMINATION LEVEL	PUBLIC

AUTHORS

Eleanor Hyde UMCG	Serena Fossati ISGlobal	Mariona Bustamante ISGlobal
Morris Swertz UMCG		

REVISION HISTORY

REVISION	DATE	AUTHOR	ORGANISATION	DESCRIPTION
V.1	14.12.2022	E. Hyde S. Fossati M. Bustamante M. Swertz	UMCG, ISGlobal	*Final version*

Contents

1. Abstract

This deliverable is a protocol, or technical explanation / 'manual' for implementation and maintenance of the federated data access system which has been put in place by WP1. The intended audience are members of technical teams within institutions taking -or wanting to take- part in the Athlete research network. Installation and administration of the elements of the network using the protocols (manual) outlined below are required in order to allow authorised researchers technical access to the data across the network.

The partners of WP1 have implemented a federated data access system, whereby the data remains locally stored and analyses are sent to the data rather than vice versa. The privacy-preserving protocol implemented to this end is called DataSHIELD. Each clinical study has therefore installed or will install a DataSHIELD server allowing a clinical study to take part in the Athlete research network (a DataSHIELD node') using the open source MOLGENIS/Armadillo[1] or OBIBA/Opal software[2], and we created a central analysis server that can be used by researchers to analyse data in these nodes. The Athlete analysis network now consists of 18 cohorts across Europe, and most are now ready to receive harmonised data and to start pooled data analyses of all these data. This Athlete deliverable will provide the protocol to access the Athlete cohorts as a written report.

2. Introduction

The aim of Task 1.2 was to provide **data access systems** for the 18 ATHLETE cohort owners to make their datasets accessible to **members inside and outside the consortium in a secure and controlled manner**. For this, we have implemented **'federated'** (data stays on local servers and is analysed remotely) and **'centralised'** (data sent from cohort to central analysis site and analysed centrally by analyst) systems. Each cohort stores and updates their **own harmonised data on a local site server** and has implemented or is in the process of implementing a '**data access node'**. For many of the exposome analyses we plan to deploy **'DataSHIELD'** (WP3, task 3.4) as one of the federated access protocols, which enables access from the open-source 'R' statistical environment using either **MOLGENIS (Armadillo) or Opal software**. The federated system overcomes governance restrictions that prohibit the release or sharing of some of the required data, or render data access slow. Next to the federated approach, and because not all exposome analyses can be done through DataSHIELD or similar protocols, the local data access nodes **also enable cohorts and database owners to submit their data centrally**, where data is then **analysed centrally on a trusted facility** with strict data access policies (managed by the project steering committee). In all cases, the cohort and data owners remain in full control of data access.

The data controllers can install a DataSHIELD node using the open source **MOLGENIS/Armadillo or OBIBA/Opal** software. The central analysis server uses **Jupyter Notebooks** and **RStudio.**

[1] https://github.com/molgenis/molgenis-service-armadillo
[2] https://opaldoc.obiba.org/

We have joined the data access network developed in the H2020 LifeCycle[3] and EUCAN-Connect[4] projects that provide solutions for federated data analysis of cohorts, and agreed with the European Human Exposome Network (EHEN)[5] project LongITools[6], to follow the same data access protocols to achieve economies of scale and scope.

For **ATHLETE** we have extended the network with the studies included in the ATHLETE project. Each partner not already equipped has set up or is setting up the **site-specific data server and will link this with the central DataSHIELD analysis platform**. See Table 1 below. Data on the site server can now be uploaded and updated where appropriate during the duration of the project. In all cases the **cohort and data controllers are in full control of data access**. To ensure trustworthiness of users in this network we will implement the EU-wide 'federated' authentication/authorization systems for **BBMRI-ERIC/ELIXIR** known as 'life science authentication and authorisation infrastructure' (LifeSciences AAI[7]).

In this document we first summarise the state specific to the ATHLETE project. Subsequently, we provide a draft of the operational handbook for the distributed data protocols (Annex 1). All components are open source, to ensure that other projects in need of distributed access protocols can also freely implement this infrastructure (or join the current installation, in case of overlapping needs).

---

3. Progress

3.1 'Federated' systems

The ATHLETE network now consists of 18 cohorts. Each cohort (data controller) has installed or is in the process of installing Armadillo* or Opal* locally in order to take advantage of the federated data analysis network which has been provided by this task. Progress is being made, and the current state of affairs is set out below in Table 1.

* Please see the appendix for detailed explanation of these technologies. Both Armadillo and Opal provide the standard DataSHIELD interface, and cohorts only need to install one of them. The decision to install one or the other depends on operational details and performance requirements.

[3] https://lifecycle-project.eu/
[4] https://eucanconnect.com/
[5] https://www.humanexposome.eu/
[6] https://longitools.org/
[7] https://lifescience-ri.eu/ls-login.html

Table 1: Progress of cohorts in installation of Armadillo/Opal and DataSHIELD

Cohort	Institute	Armadillo / Opal version installed	DataSHIELD version installed
BiSC	ISGLOBAL	opal-3.0.9	6.2.0
BiB	BTHFT	armadillo-2.0.0	6.2.0
CELSPAC-TNG	RECETOX	*armadillo-0.0.17*	*6.1.0*
DNBC	UCPH	opal-4.5.2	6.1.0
EDEN	INSERM	armadillo-2.0.0	6.2.0
ELSPAC	RECETOX	*armadillo-0.0.17*	*6.1.0*
EnvironAGE	Hasselt University	*armadillo-0.0.17*	*6.1.0*
Generation R	ERASMUSMC	*armadillo-2.0.0* opal-3.0.9	*6.2.0* 6.1.0
Generation R Next	ERASMUSMC	*armadillo-2.0.0* opal-3.0.9	*6.2.0* 6.1.0
Generation XXI	ISPUP	opal-4.2.8	6.2.0
INMA	ISGLOBAL	opal-4.4.9	6.2.0
KANC	Vytauto Didziojo Universitetas (VDU)	armadillo-2.0.0	6.1.0
MOBA	NIPH	opal-4.5.2	6.1.0
NINFEA	UNITO	opal-3.0.2 armadillo-0.0.17	6.1.0 6.2.0

Cohort	Institute	Armadillo / Opal version installed	DataSHIELD version installed
PELAGIE	INSERM (Rennes)	armadillo-2.0.0	6.1.0
Piccolipiù	UNITO	*armadillo-2.0.0*	*6.1.0*
RHEA	UoC	opal-3.0.3	6.1.0
SEPAGES	INSERM	armadillo-2.0.0	6.1.0

Legend:

armadillo-2.0.0 installed
armadillo-2.0.0 installation / upgrade in progress

3.2 'Centralised' systems

In ATHLETE new data are being generated in Task 1.5 as part of the new follow up of the HELIX subcohort. This data will be centralised at ISGlobal and added to an existing database, created within the FP7-funded HELIX project, and already stored at ISGlobal and available for central analysis at ISGlobal, and for transfer of analysis datasets to analysts both internal and external to the project. By the end of ATHLETE this database will also be made available through the federated platform and DataShield access will be enabled. Current access procedures for external researchers are described here: https://www.projecthelix.eu/index.php/es/data-inventory

3.3. Omics data

Omics datasets are held locally at the cohorts and, just as all other data, are subject to cohort-specific access procedures. Technically speaking, each cohort requires a server with >=100GB storage capacity and 8 GB RAM in order to process the large volumes of data contained in omics datasets when using DataShield federated analysis.

Omics data were obtained using different methods and thus interoperability protocols are required to combine data. For DNA methylation most of the cohorts obtained data using the same array, and this facilitates the combination of the data from different cohorts. The same is true for the GWAS, that through an imputation process they can be combined easily. Things are more complex for the other omics, specifically for metabolomics and metagenomics. ICL is making a dictionary of metabolites available in each cohort which will help the harmonization of the data (on-going). D4.1 Inventory and protocols multi-omics data_v1.pdf covers the harmonisation protocols currently in place for omics data.

The complex nature of omics data has meant that a stepped approach has been applied to the use of DataSHIELD for analysis of omics data. Cohorts are limited for differing reasons, whether it be server

capacity, knowledge, or personnel, and to combat this we have started with methylation data, with the aim of moving on later to GWAS, transcriptomics and then metabolomics and other omics which are more complex. DNA methylation data is currently available on DataSHIELD in INMA and EDEN. BiB will also have DNA methylation data available once server capacity issues have been resolved.

4. Conclusion

This document reports a summary of the implementation of the task 1.2 of the ATHLETE project:

"Developing a data access network for rapid controlled data sharing [M6-M36]". The teams involved have performed the task according to the plan described in the document of action. To date, we have implemented a federated data access system, including data access nodes employed locally to facilitate this federated data access, a central analysis server which researchers can use to analyse data from multiple studies in one, integrated analysis, and ensured the DataSHIELD method of distributed access analysis can be utilised for that analysis.

We have joined the data access network developed in the H2020 LifeCycle and EUCAN-Connect projects that provide solutions for federated data analysis of cohorts, and agreed with the European Human Exposome Network (EHEN) project LongITools to follow the same data access protocols to achieve economies of scale and scope.

The task has led to the implementation of a data access network, supporting the data management plan of ATHLETE (and collaborative projects within the EHEN) and a protocol to guide the users (see annex 1).

The work has been disseminated (for example) at the DataSHIELD conference October 2022, the EHEN Network Days in March 2021 and May 2022, Research Data Alliance April 2021, Health-RI October 2022 and EUCAN-connect general assemblies, and will lead to at least one peer review article (in collaboration with EHEN partners).

To conclude, the work has been performed without delay and has reached its expectations allowing us to deploy the system for ATHLETE, continue implementing WP1, as well as to increase further interoperability for the EHEN data working group.

Annex 1: DataSHIELD Armadillo Network Handbook

This document provides an overview of operational aspects of the distributed data access protocols, linking to detailed manuals that can be used by the ATHLETE participants during installation and operations of the network.

Background

We use the DataSHIELD protocol, a unique computer interface addressing key governance and regulatory challenges faced by researchers when working with sensitive individual-level data. Initially developed to support biomedical and social science research teams, DataSHIELD can be used in scenarios where individual-level data must be analysed but where data sharing with anyone other than the originating data controller entails legal and financial challenges. DataSHIELD provides a researcher with a set of "R" methods through which analysis of data can be requested. The results of such analysis can be either stored on the server on which the sensitive data resides, or if the results are non-disclosive passed back to the researcher. The "R" methods provided by DataSHIELD include code whose purpose is to identify and block disclosure risks specific to that function. When the research involves multiple studies, DataSHIELD supports their simultaneous analysis by a single request from the researcher, and the individual results from the studies will be combined by DataSHIELD into an overall result.

For the invocation of DataSHIELD methods we have used the MOLGENIS Armadillo system. Before the Armadillo starts the DataSHIELD methods to perform the required analysis, Armadillo performs authentication to validate the identity of the requester, then authorization on that identity to ensure they are permitted access to the indicated sensitive data, and that the analysis methods are permitted to be used by that individual.

DataSHIELD web-site: https://www.datashield.org/

Molgenis' Armadillo Data Managers API: https://github.com/molgenis/molgenis-r-armadillo/

Molgenis' DataSHIELD Interface Driver: https://github.com/molgenis/molgenis-r-datashield/

Which people have which roles to play?

Data:

- Data managers of the institute: working with delegated responsibility from the owner of the data, import data and grant permissions
- Data managers Molgenis-support: help with import and/or data questions if needed
- Researchers: analysis of the data

Infrastructure management:

- IT department: manages the machine, the Operating System (OS) and patching

- Molgenis Operations team: managing DataSHIELD Armadillo suite if needed and updating the DataSHIELD Armadillo suite

Architecture and systems overview

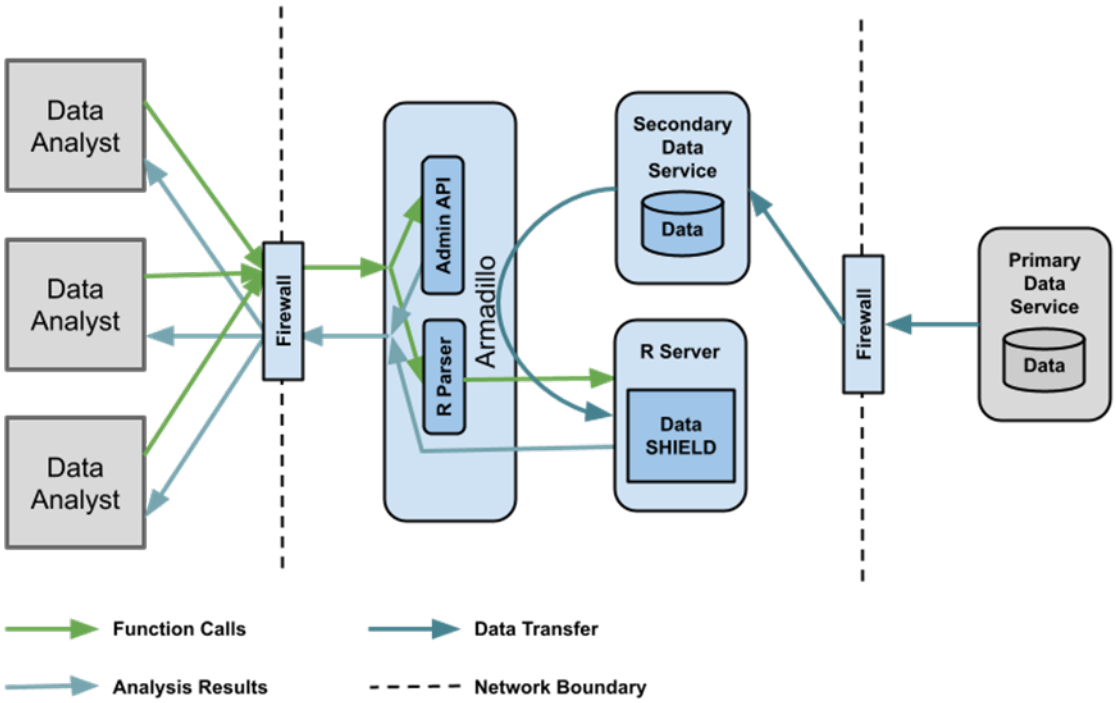

Fig 1. Architecture and systems overview

The deployment shown in figure 1 represents an environment which is split over three network regions connected by firewall and reverse proxies (not shown), which restrict traffic flows between the network regions. The firewall between the data analysts and the Armadillo Server will be configured to only allow 'https' network traffic from data analysts nodes to a single port on the Server, response will be routed back to the data analysts via the firewall. Another firewall between the primary data service and the secondary data service is used to ensure that access cannot be gained to the network region holding the primary data service from the network region containing the secondary data service. The purpose of this firewall is to permit updates to the primary data service to be copied to the secondary data service.

The purposes of the components and interactions within the reference deployment are as follows:

- Data Analyst: the node from which data analysts make analysis requests to the Server, and receive analysis results;
- External Firewall: the firewall between the data analysis and Server;
- Internal Firewall: the firewall between primary and secondary data service;
- Analysis Requests: the request sent by the data analysis to DataSHIELD, via the Server;

- Analysis Results: the non-disclosive results sent from DataSHIELD to the data analysis, via the Server;
- Armadillo Server: the Server manages authentication and authorization of sessions and demultiplexing of requests to the R software or Web Portal. User authentication, by the Server, can be performed via password or certificate, and supports 2-factor authentication;
- Armadillo Admin API: within the Server provides a standard computer interface ('REST') based administrative interface to Server;
- R Parser: the R parser checks the analysis requests only contain valid requests, and in particular only permitted functions;
- Primary Data Service: this data service contains the primary copy of the data to be analysed;
- Secondary Data Service: the data service, which could be provided by standard file and database software systems (MinIO, MySQL or mongodb), contains the secondary copy of the data to be analysed. This version of the data will be copied into the R server to be analysed;
- R Server: the R server is an execution environment, one per user, within which analysis requests are executed;
- DataSHIELD: an R package, within the R server, which provides data analytic functions.

Most of the components in the Armadillo suite are open-source, available on GitHub and maintained by Molgenis Support. Table 2 gives an overview of the components described earlier.

Table 2. Overview of components

Component	Description
DataSHIELD	DataSHIELD is a software ecosystem which allows researchers to analyse data in distributed locations in an actively disclosure-protected manner.
Armadillo Service molgenis/molgenis-service-armadillo	The Armadillo Service is the central component that manages the execution of analysis requests and the feedback of their results. It acts as the security layer between the outside world and the data service.

Component	Description
DataSHIELD Armadillo Driver molgenis/molgenis-r-datashield	This R client is the Armadillo implementation of the DataSHIELD interface. It is the tool Data Analysts use to analyse the data they have access to.
Armadillo R Client molgenis/molgenis-r-armadillo	The Armadillo client interfaces with the Admin API of Armadillo and is used by Data Managers to move and manage data (or subsets of the data) in the secondary storage (MinIO).
Permission Manager molgenis/molgenis-js-auth	The Permission Manager provides a user interface that Data Managers use to grant Data Analysts access to specific data within an application. This UI was built as an extension of FusionAuth, which does not have this feature.
Custom Rserve molgenis/Rserve	Armadillo uses a custom Rserve implementation that has a broader and predictable port range. This fixes the server becoming unresponsive under heavy loads.
Opal	While most cohorts operated by ATHLETE WP1 staff are using Armadillo, cohorts are free to use Opal instead. Armadillo and Opal collaborate on the DataSHIELD 'driver' interface.
Jupyter notebook	The Jupyter notebook is an image that is started on the analysis server. This image contains an environment in which R scripts can be created and in which analysis using those R scripts can be performed. An RStudio environment is also included in the image. The Jupyter notebook can be started anywhere with a sufficiently stable internet connection.

Component	Description
RStudio	RStudio is an application with which to create, test and run R scripts. It can be installed locally on the user's own environment or it can be run in the cloud via Jupyter notebooks.

One Armadillo service can manage many R servers. Each of these servers may have packages installed that are specific for certain types of research questions. In DataSHIELD, this is the concept of profiles. Armadillo implements these profiles by having a Docker container for every R server instance, with the packages belonging to that profile pre-installed. For example, the "omics"-profile will contain *dsBase* and *dsOmics*. The scripts for building the Docker images for these profiles are also open-source and available on GitHub: datashield/docker-armadillo-rserver-base and isglobal-brge/docker-armadillo-rserver. The actual Docker images are stored on DockerHub: organisation datashield and user brgelab.

What data managers need to do to upload data

Data managers will upload harmonised data into a local Armadillo (or Opal) instance in order to make it available for users to run analyses with DataSHIELD. To ensure secure access, the Armadillo works with a central authentication service. This means that in order to work with the Armadillo, you need to have an account on the central authentication service.

There are two phases to uploading data to the Armadillo. The initial upload transforms your source data to the correct format for analysis. Besides this you can perform some data manipulation on the initially uploaded data. The initial upload can be done with the dsUpload.

After the initial upload, subsets of the data can be created for specific projects or research questions. Permissions can be set for each subset. In this way it can be very specifically defined who can access certain parts of the data. To manipulate the data after the initial upload the MolgenisArmadillo client can be used. Check the documentation to create subsets.

What data managers need to do to provide access permissions

After the correct authentication and authorisation steps have been set up, researchers will be able to analyse the data via DataSHIELD. We use the authentication service to give people permission to analyse the data. Data managers perform the following steps to give people access:

- create a role
- register a user
- give a user a role

Permissions can be provided to complete datasets or subsets. First, the user must register via the central analysis server (https://lifecycle.analysis.molgenis.org/) using their institutional accounts.

Now the data manager can select the user on the authentication server. Here, the data manager can assign roles to the user, as shown in the example below (Fig. 2). Each role corresponds with a project and (sub)set of the data.

Email	First Name	Last Name	Roles
t.kennedy@inserm.fr	Terry	Kennedy	ALSPAC_RESEARCHER × GECKO_RESEARCHER × MLMALSPAC_RESEARCHER × MLMBCG_RESEARCHER × MLMBIB_RESEARCHER × MLMCHS_RESEARCHER × MLMPROBIT_RESEARCHER ×
t.brown@umcg.nl	Tabitha	Brown	ALSPAC_RESEARCHER × DNBCTEST_RESEARCHER × GECKOSTUDY30_RESEARCHER × SU × gecko4wcr_researcher × ninfeaasd_researcher ×
o.sullivan@isglobal.org	Omar	O'Sullivan	SU ×
t.hernandez@isglobal.org	Tanya	Hernandez	SU ×
r.connor@isglobal.org	Robert	O'Connor	ALSPAC_RESEARCHER × GECKO_RESEARCHER × MLMALSPAC_RESEARCHER × MLMBCG_RESEARCHER × MLMBIB_RESEARCHER × MLMCHS_RESEARCHER × MLMPROBIT_RESEARCHER ×

Fig. 2 an example of authorisation assignment using roles (fictitious names and email addresses)

What researchers need to do to start a multi-center analysis

Researchers will use the DataSHIELD client (R package *dsBaseClient*) to connect with one or more DataSHIELD servers. These can be Armadillo or Opal servers, which both require their own driver package.

Step 1. Include the packages needed to connect to Armadillo and Opal.

```
library("dsBaseClient")

library("DSMolgenisArmadillo")

library("DSOpal")
```

Step 2. Armadillo requires online authentication and will open a browser.

```
armadillo_url <- "https//armadillo.example.org"

token <- armadillo.get_token(armadillo_url)
```

Step 3. Create all login information using the login builder from the *dsBaseClient*. The researcher needs to supply the server and (optionally) a table they require access to.

```
builder <- DSI::newDSLoginBuilder()

builder$append(server = "armadillo",

  url = armadillo_url,

token = token,

table = "gecko/2_1-core-1_0/nonrep",

driver = "ArmadilloDriver")

builder$append(server="server1", url="https://opal.example.org",

user="dsuser", password="password")

logindata <- builder$build()
```

Step 4. Login to all specified servers.

```
connections <- datashield.login(logins=logindata)
```

Now the researcher can use *dsBaseClient* to analyse data on both servers. A complete Armadillo guide can be found here. An overview of the available analysis-methods in *dsBaseClient* can be found here.

How to install Armadillo for system administrators

https://galaxy.ansible.com/molgenis/armadillo

https://github.com/molgenis/molgenis-ops-galaxy/tree/main/armadillo1

DataSHIELD Armadillo instances require network traffic to and from the central analysis server's IP address. When security concerns or firewalls complicate this, either a separate authentication server has to be deployed, or specific traffic will need to be allowed access. A server with its own authentication server however cannot connect to other studies.

In order for the DataSHIELD Armadillo server to be reached by different users, port 80/tcp and port 443/tcp need to be opened. It is possible for the instance to be behind a firewall or virtual private network (VPN), however in order to allow other study sites to work on the instance, their IP addresses need to be allowed access. It is possible to install additional packages on separate servers. To allow the servers to communicate, access to and from port 6311/tcp needs to be possible.

By default the DataSHIELD Armadillo suite listens to port 80 (HTTP) instead of port 443 because every study/institute (data controller) has their own method of installing SSL certificates. SSL certificates can be provided, if requested, using Let's Encrypt. Other certificates need to be manually installed and the NGINX webserver configuration needs to be adjusted.

To simplify the installation process, ansible is used to set up and install all requirements. Ansible is a free and open-source automation platform in which tasks required for installation can be defined in scripts called playbooks. Users can install ansible and run playbooks to setup and install all requirements. For the Armadillo playbook, certain parts need to be configured before running it, specifically usernames, passwords and domain names. After these gaps are filled, the playbook should be handled with care.