ODAP Overview

Version 0.1

Table of contents

1 ODAP Overview

This is a shared description for the internal team of the core purpose, scope, operational components, and governance structure of the ODAP. It provides an introduction, plan and a reference.

1.1 Introduction

ODAP is long term, UK wide, 4 nations outbreak response infrastructure providing the unique and scalable capabilities of the academic sector to tackle public health challenges. ODAP creates a UK-wide capability by curating and linking outbreak relevant data from clinical records, research studies and audit data. It brings together key initiatives and leadership across the UK including ISARIC, COG-UK, MRC CLIMB and GenOMICC. The platform combines a national Trusted Research Environment (TRE) infrastructure collocated with >£100M of world-class computational and data science capacity including the UK National Supercomputer, with a UK-wide governance framework.

1.2 What is the Outbreak Data Analysis Platform (ODAP)?

The ODAP is the overarching term for a range of computers supporting research within a specific scope. These include the ODAP TRE (formerly called: flexible compute space; ultra; ultra2), and various project spaces in the National Safe Haven.

It has two components:

1. The ODAP Core. *Getting fast answers to important public health research questions.* ODAP Core is a complicated network of overlapping, linked data from research studies and national datasets, created through a range of independent research activities, sitting in multiple computer systems. These datasets are linked through the ISARIC Clinical Characterisation Protocol (CCP).
    - Purpose: to prepare for and support urgent public health research.
    - Access: data are openly accessible to appropriately-qualified researchers studying questions within the scope of the platform by individual agreement with the data providers. Much of the data is under embargo and publication of results requires consent from the data providers. Put simply, access is open but not easy.
    - Funding: The core ODAP is funded by the Baillie Gifford Pandemic Science Hub.
2. ODAP Open - *Making data access easy for approved researchers.* ODAP Open will be a robust, fully-staffed data access governance system to lower the barrier to entry by taking over data controllership from data contributors, so that a single approval is required for data access.
    - Funding: not in place. Funding from HDR UK infrastructure team is anticipated.
    - Leadership: open call for a senior appointment when funding is available.
    - Access: data are openly accessible to appropriately-qualified researchers studying questions within the scope of the platform following a single, streamlined application

The foundation for the ODAP is the International Severe Acute Respiratory Infection Consortium (ISARIC) Clinical Characterisation Protocol (CCP).

2 ODAP Core

2.1 Scope of the ODAP

The purpose of the ODAP is to facilitate biomedical research to advance understanding of severe infectious disease* and other exposures of public health interest. Research within the ODAP is strictly limited to this purpose.**

* Severe infectious disease - this term describes all severe infectious agents, including new, re-emerging or therapy-resistant forms of existing infectious agents.

** Other exposures of public health interest: this term describes new or unexplained poisoning, or exposure to harmful energy sources such as electromagnetic radiation.

The scope of ODAP is designed to carefully match the scope of the ISARIC-CCP.

Examples of in-scope research: - Understanding the evolution and biology of SARS-CoV-2. - Understanding the epidemiology and transmission of SARS-CoV-2. - Understanding COVID-19 disease risk, severity and outcomes. - Monitoring and understanding the impact of non-pharmacological interventions against SARS-CoV-2 transmissions and COVID-19 disease. - Monitoring, understanding, and assessing the impact of treatments, vaccines and prior infections in COVID-19 disease. - Analysing or modelling SARS-CoV-2 and COVID-19 data for future pandemic preparedness.

Examples of out of scope activities - Their research questions primarily focusses on a non-infectious disease area with incidental involvement of infectious disease (Focussed on a pathogen with no potential public health concern)

2.2 ISARIC Clinical Characterisation Protocol (CCP)

The scope of the ODAP mirrors the objective of the CCP, an ethically-approved research study in the UK (Joint Chief Investigators: Calum Semple(Liverpool, Oxford) and Kenneth Baillie (Edinburgh, Oxford)). A broad range of scientists with relevant expertise have come together to form a UK-wide group: the ISARIC Comprehensive Clinical Characterisation Collaboration. Membership of this collaboration is by invitation and is extended to researchers performing high-quality biomedical research to advance understanding of severe infectious disease and other exposures of public health interest.

2.2.1 ISARIC Spine

Some studies, addressing aims that are within the scope of the CCP, have chosen to contribute data into the ODAP for linkage to CCP and other datasets. Because these studies are addressing questions that are directly within the scope of the CCP, where appropriate, participants in these studies are eligible to be enrolled under the CCP.

Under our existing protocol and approvals for the ISARIC CCP, we have recruited over 300,000 Covid patients without explicit consent, into an observational study, with clear and specific aims. We are now linking this data to other studies, including consented studies such as GenOMICC and RECOVERY, and non-consented studies such as COG-UK viral sequencing.

Applying the established governance and data handling procedures for the ISARIC CCP to the additional linked data from these studies will enable cross-cutting research, all within the specific aims of the CCP. To achieve this, we have agreed to recruit all patients in the following studies into the ISARIC CCP, if they are known to meet the inclusion criteria for the CCP:

- GenOMICC
- RECOVERY
- PHOSP
- COG-UK
- HEAL-COVID

In each case, we will only recruit patients who meet the inclusion criteria for CCP. In our view, there is no ethical or information governance difference between recruiting a patient remotely from a hospital, or recruiting them remotely from within our trusted research environment. This has been agreed with the ISARIC CCP study sponsor and is explicitly stated in CAG and PBPP applications.

2.2.1.1 Examples The ISARIC4C consortium will incorporate participants from other studies into ISARIC4C study, creating a superset of participants in a single “spine”. Doing so would allow them to match to NHS data under existing ISARIC4C data agreements.

To do this, we need to have an ID and an NHS number, or CHI number, as a minimum standard. Some participants in the other studies may already be in ISARIC4C. These need to be identified and tagged as being

the same individual. This will be done by comparison of NHS or CHI numbers. The spine will contain all the joins between projects and participants. Those participants that are not already in the ISARIC4C system will have an ISARIC4C ID allocated to them.

The data from the spine will be used for several purposes:

- it will feed into the NHS Digital data matching system. This system requires ID, NHS number, date of birth and postcode (the latter two are optional). This will allow the ongoing monthly matching of study participants to English & Welsh Health Data for the superset of ISARIC4C participants.
- it will allow study participants to be identified against a cohort of NHS or CHI numbers for individual research studies, and as requested by Trusted Research Environments (TREs).

Currently the data for participants in these studies are held in a series of REDCap databases. These are the source of truth for the data and update on a regular basis. We will therefore need a process to update the spine from these sources.

2.3 Data access

2.4 Data under embargo

The default status for new data entering the ODAP is to be `under embargo`. This means that either:

1. the data contributor controls who has access to the data, and what analyses and reports or publications can arise from it.
2. the data contributor chooses to link data within the ISARIC4C study (using the Spine), wherupon all members of the ISARIC4C consortium have access to the data and can run exploratory analyses under the consortium rules. This enables the data contributor and others in the consortium to make use of existing contracts and approvals for linkage to other datasets.

2.4.1 ISARIC4C Consortium data access rules

1. Consortium members are recorded on the consortium membership list
2. Membership is extended to researchers with skills and techniques necessary to answer important research questions within the scope of the ISARIC Clinical Characterisation Protocol, interpreted by the Consortium Leadership in consultation with the broader consortium.
3. Consortium members have access to all data within the ISARIC4C consortium space within the ODAP.

4. All members agree to:
    - inform data contributors of all substantive analyses *in advance* of undertaking them (this does not apply to preliminary analyses)
    - obtain express agreement from all data contributors before writing a manuscript

2.5 Non-embargo data

Some data providers choose to delegate approval of data access to the ODAP team. The diagram in Figure 1 describes the five safes application form and light-touch review process. These processes are overseen by the Data Access Governance Committee (DAGC).

The data access review processes are described in detail in the following documents:

- Data Access Review Process
- DAGC Terms of Reference
- PRP Terms of Reference
- PSG Termos of Reference

2.6 Computer architecture

The ODAP consists of two main compute areas.

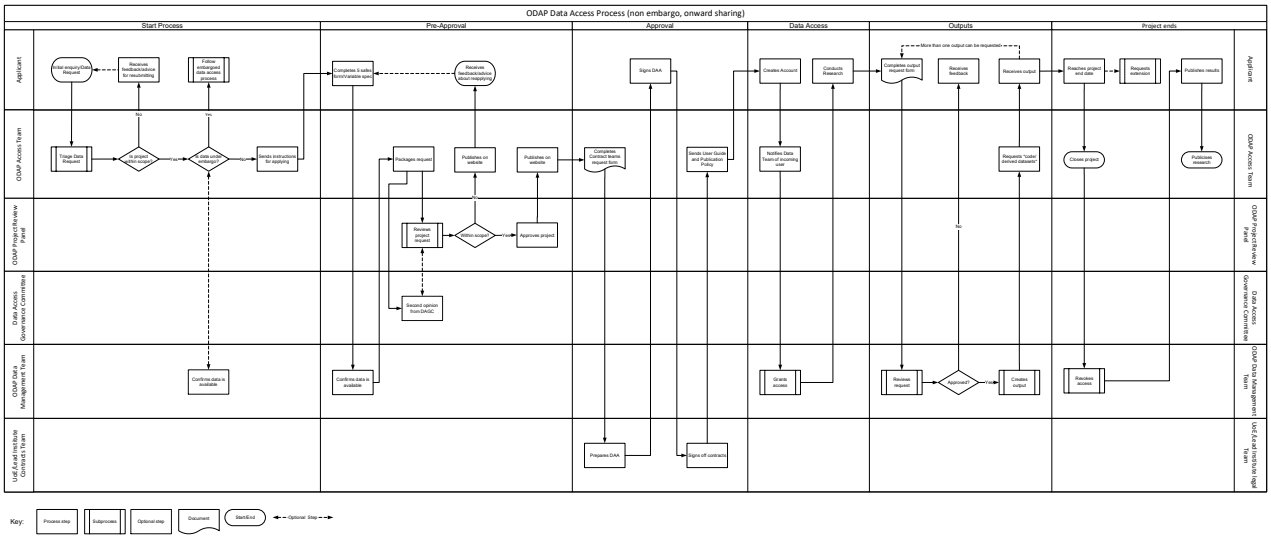

Figure 1: Data access flows

2.6.1 Flexible Compute Space

The Flexible Compute Space (FCS) is a Trusted Research Environment with large scale compute capacity and access to a range of software tools for data analysis and machine learning. Data within the FCS can be either under embargo, or not under embargo.

Access to the FCS is provided to researchers through *private project zones*

2.6.2 PHS National Safe Haven

Some data remains within the PHS National Safe Haven but will transfer to the FCS on completion of the Systems Security Policy.

2.7 Legal architecture

The legal entity hosting ODAP is the University of Edinburgh. A data processing agreement UoE(controller)-PHS(processor) directs PHS to (1) host and (2) index data from multiple data contributors within the ODAP. The ISARIC Clinical Characterisation Protocol (CCP) has ethical approval across the four nations of the UK and is sponsored by Oxford University. Both joint Chief Investigators (Calum Semple and Kenneth Baillie) have honorary contracts at Oxford Unversity to facilitate this. Data linkage for ISARIC CCP Covid-19 data is authorised under the COPI notice (regulation 3).

2.8 Data ingress

Data are linked into the ISARIC Spine using the patient's NHS numbers by PHS before transfer to the FCS as shown in Figure 2.

2.9 API

There will be a secure API in place from the Flexible Compute Space to External TREs and UK Public Health Agencies. This is to accelerate research by enabling data linkage across multiple data sets using unique identifiers to facilitate secure transfer of specified data fields e.g., viral sequencing data. The future purpose of the API is more in line with providing a mechanism for quick pandemic updates to national health services.

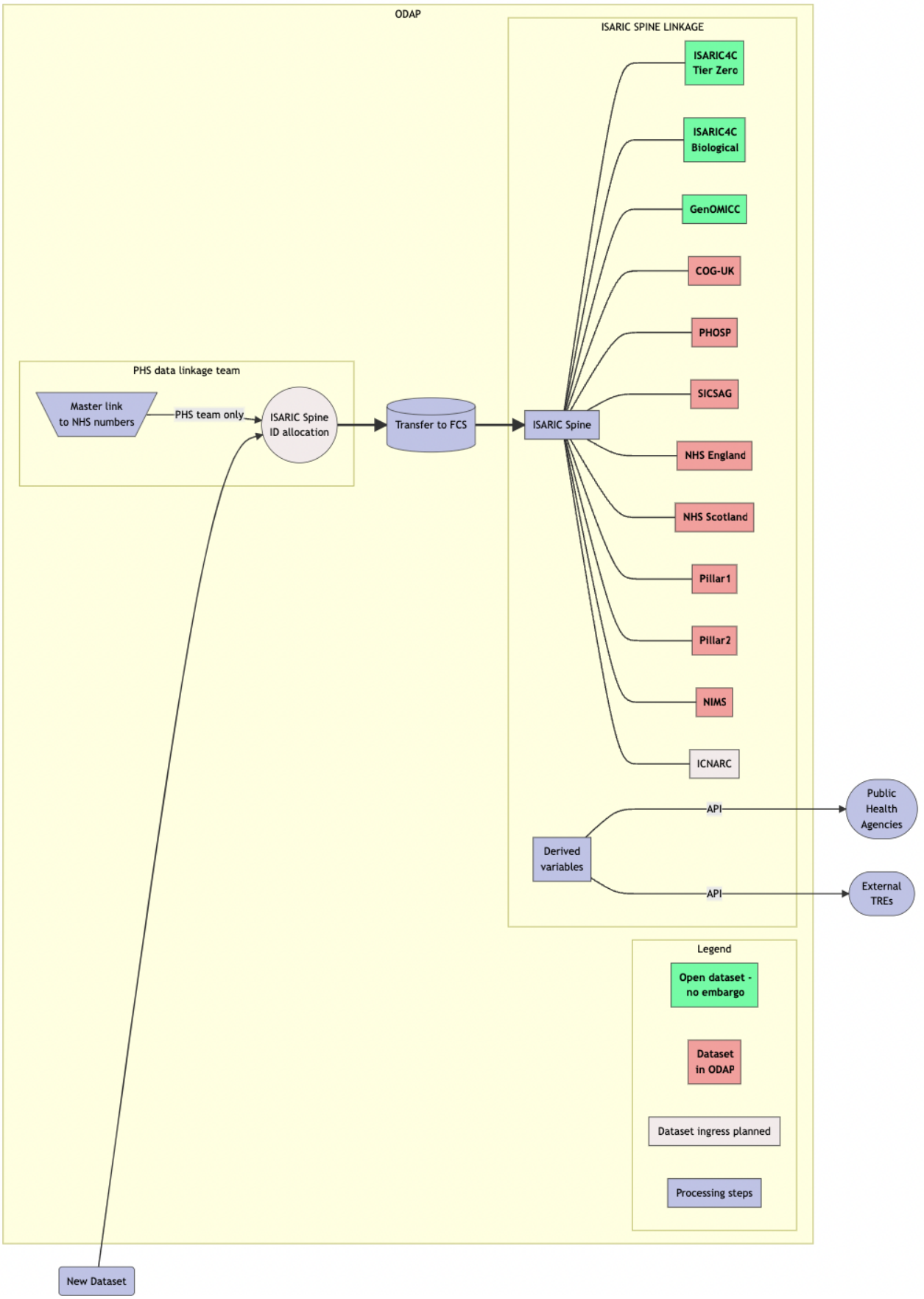

Figure 2: Flowchart showing data linkage and accessibility within ODAP

2.10 Data Contributors

Each data contributor will explicitly consent to the lifting of any embargo on their data. A permanent record of these instructions will be held in a shared file.

2.11 Dataset catalogue

The ODAP dataset catalogue is a flexible central record of the datasets held within the platform, approval requirements for access, embargo status, and contents. Each row of the catalogue provides information about an `atomic dataset`.

An `atomic dataset` should be the smallest unit of data that a user would apply for access to. All data elements in an `atomic dataset` will have, by definition, - the same version number - the same data controllers - the same data processors - the same embargo status - the same legal agreements - the same users with data access

Any legal agreement that applies to part of an `atomic dataset` will apply to all of the data within it. Any user who has access to part of an `atomic dataset` has access to all of the data within it. Where possible an `atomic dataset` will be in tidy format.

A `compound dataset` is a group of `atomic datasets` that commonly go together. They are grouped together for convenience because researchers often need access to all of them at once.

When an `atomic dataset` is split, two new `atomic datasets` are created, and the name of the superceded `atomic dataset` is recorded as a `compound dataset` in both new `atomic datasets`.

3 OPAP Open

Once the embargo is lifted by a data contributor, or (in the case of public data) delegated access is approved by the data controller, data provision will be *FAIR* (Findable, Accessible, Interoperable, Reusable) and access will adhere to the *five safes* principles:

- Safe data: data is treated to protect any confidentiality concerns.
- Safe projects: research projects are approved by data owners for the public good.
- Safe people: researchers are trained and authorised to use data safely.
- Safe settings: a SecureLab environment prevents unauthorised use.
- Safe outputs: screened and approved outputs that are non-disclosive.