UK Longitudinal Linkage Collaboration
Population Health Sciences
Bristol Medical School
Canynge Hall
39 Whatley Road
Bristol BS8 2PS

UK Longitudinal Linkage Collaboration (UK LLC)

NAMING OF PROJECTS AND DATA-RELATED OUTPUTS POLICY

PUBLIC

Version 1.1

8 November 2022

Policy number:	POL-DAT-006	**Version:**	V1.1
Author:	Katharine Evans, Governance & Policy Manager; Sammy Berman, Data Manager	**Date:**	15/04/2022
Authorised by:	Andy Boyd, Director	**Date:**	26/04/2022
Date published:	26/04/2022	**Date to review:**	08/11/2023
Permission to edit this policy must be provided by:	Director; Senior Research Manager		

Review History

Version:	Review date:	Reviewed by:	Section(s) amended:	Authorised by:
1.1	08/11/2022	Katharine Evans, Governance & Policy Manager	Review in preparation for DEA audit	Emma Turner, Senior Data Manager Acquisitions

Contents – UK LLC Naming of Projects and Data-related Outputs Policy

1. INTRODUCTION

1.1 Background

The UK Longitudinal Linkage Collaboration (UK LLC) organisation manages the collation, curation and access to de-identified data about Longitudinal Population Study (LPS) participants held in the UK LLC Trusted Research Environment (TRE). The UK LLC is led by the University of Bristol (UoB) and operated in collaboration with the University of Edinburgh (UoE).

1.2 Purpose

This policy explains how researchers working in the UK LLC TRE should name their projects and data-related outputs (e.g. syntax and code lists).

This policy will help the UK LLC to be transparent in its operations and enable reproducible and efficient research for public benefit.

A formalised naming convention for research projects is important both in terms of: (i) project management during the lifetime of the project; and (ii) longer-term provenance and documentation purposes. For similar reasons, a naming convention is also pertinent to new resources produced through projects, such as derived datasets and related syntax/code.

This naming convention – particularly the clear and consistent use of UK LLC project numbers across the lifecycle of the research activity – will also support the UK LLC's objectives to (i) operate an ISO 27001 and Digital Economy Act (DEA) accredited Information Security Management System; (ii) be transparent in our operations; and (iii) enable reproducible and efficient research using the 'Team Data Science' model.

NOTE: To understand how to apply to access the UK LLC TRE and the terms and conditions under which approved researchers access data, please refer to the UK LLC Data Access and Acceptable Use Policy.

This policy will be reviewed to respond to any changes in the UK LLC risk assessment or risk treatment plan and at least annually.

2. SCOPE

This policy applies to **all UK LLC staff** involved in the management of researchers' applications and projects and **all approved researchers** who access the UK LLC TRE for approved projects.

3. DEFINITIONS

Application – all research use within the UK LLC TRE is embedded within an approved application. As such, a consistent identifier is required from the point at which an application is submitted to the UK LLC for consideration, through to when it becomes an approved **project** in the UK LLC TRE. See the UK LLC Data Access and Acceptable Use Policy for further details.

Project – an application is referred to as a project at the point data are prepared for provisioning within the UK LLC TRE. A project is defined as a specific research activity addressing a pre-defined purpose. A project may consist of activities investigating a range of research questions, but these must all be specified and related to the project theme. Projects may be applied research, methodological research or specific resource enhancements (e.g. the production of a new derived dataset).

Data-related output – this is defined as any research output comprising derived data or documentation (e.g. data processing/analysis syntax or user guide material), that would be necessary

for other users to understand and reproduce the research; and/or which may be useful to others for reuse within their own project; and/or which are necessary for external stakeholders (including participants and the public), to understand how data are being used in the UK LLC TRE.

The **inter-relationship** between **applications, projects** and **data-related outputs** is defined in figure 1.

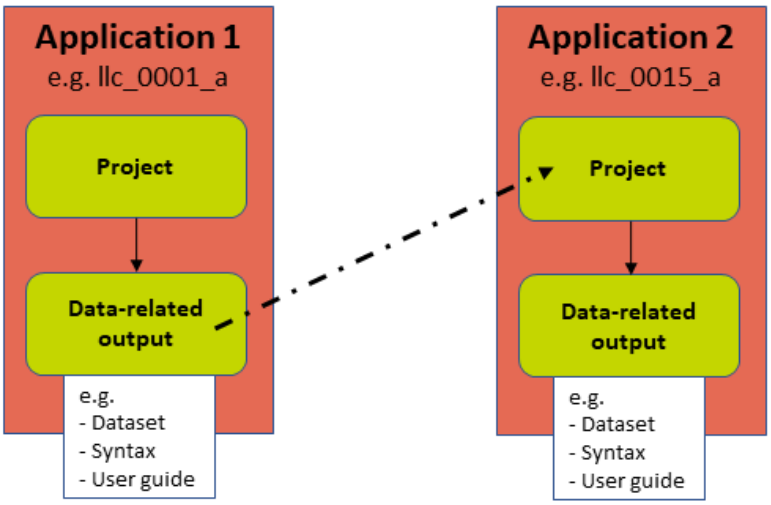

Figure 1 The relationship between an application, a project and data-related outputs in the UK LLC TRE
Note: the black dashed line illustrates the re-use of data-related outputs from one project to another (with prior approval(s) in the case of derived data objects)

4. NAMING CONVENTIONS

4.1 General rules

Some software packages (e.g. Stata and R) are case-sensitive in how they handle object names. Although converting to lower-case variable names is a straightforward task in Stata[1] and in base R[2], adopting and following a standard removes the need for additional checks and conversion routines. Moreover, a lower-case naming convention has the advantage of uniformity and minimises risk of namespace conflict and **lowercase is consequently enforced in the UK LLC file naming convention.**

Snake case is a specific variation of the lowercase naming convention in which words are separated by underscores. It is used to address challenges in embedding code to interact with databases within software applications and is a widely adopted convention amongst users of Python[3,4] and R[5,6,7], as well as other programming languages[8]. **Snake case is consequently enforced in the UK LLC naming convention.**

[1] https://www.icpsr.umich.edu/icpsrweb/content/shared/ICPSR/faqs/stata-variable-names.html
[2] https://stackoverflow.com/questions/13258020/change-letter-case-of-column-names
[3] https://www.python.org/dev/peps/pep-0008/#descriptive-naming-styles
[4] https://google.github.io/styleguide/pyguide.html#316-naming
[5] https://stackoverflow.com/questions/159720/what-is-the-naming-convention-in-python-for-variable-and-function-names
[6] http://adv-r.had.co.nz/Style.html
[7] https://style.tidyverse.org/syntax.html#object-names
[8] https://en.wikipedia.org/wiki/Snake_case

4.2 Project names

Projects in the UK LLC TRE should be named as follows:

<llc>_<project_number>_<amendment>

where

- **<llc>** corresponds to the UK LLC
- **<project_number>** corresponds to a positive integer of 4-digit length, assigned by the UK LLC Applications Team when an application is formally submitted for consideration
- **<amendment>** corresponds to a character, ranging in sequence from 'a' through to 'z'.

Table 1 Examples of the project naming convention

Example	Interpretation
llc_0001_a	The 1st application submitted to the UK LLC and no amendments have been made
llc_0001_b	The 1st application submitted to the UK LLC and one amendment has been made
llc_0012_a	The 12th application submitted to the UK LLC and no amendments have been made

4.3 Data-related output names

The UK LLC will maintain UK LLC Git repositories where, for each project, users will make their data-related outputs (syntax, codelists, protocols, methods and derived variables) accessible and understandable to future users. This will take the form of an internal Git Lab within the UK LLC TRE for active projects and an external, publicly accessible UK LLC Git Hub repository for completed projects.

Data-related outputs must follow a prescribed naming convention. To ensure provenance is retained, all names commence with **<llc>_<project_number>**, e.g. llc_0001. The <amendment> suffix is omitted to avoid conflict with versioning of the data-related outputs, which will likely be distinct from project versioning.

As data processing/analysis syntax can produce multiple data outputs, each data output should be given a unique name pertinent to the data. Multiple data outputs produced by one syntax file should not increment the version number. Version numbers should only increment when there has been a change in analyses or data pipeline, resulting in a data output with the same name being produced.

Where multiple syntax files are used in sequence/combination to produce a data output, a master syntax file should be maintained and versioned in accordance to the versioning of the data output.

Data-related outputs should therefore be named as follows:

<llc>_<project_number>_<output_name>_<output_type>_<version>
where

- **<llc>_<project_number>** – as described above
- **<output_name>** corresponds to a concise descriptor of the output of maximum 10 characters length which will be constrained within the UK LLC to novel values (to avoid duplication risk)
- **<output_type>** corresponds to a signifier of the type of content in the output, constrained currently to a narrow set of options ('data', 'doc', 'syntax')
- **<version>** corresponds to an integer of variable length, sequenced in order of versioning (decimal values are not permitted).

Table 2 Examples of the data-related output naming convention

Example	Interpretation
llc_0001_sesdemog_syntax_v1	The syntax used to create the sociodemographic dataset 'llc_0001_sesdemog_data_v1'
llc_0001_sesdemog_data_v1	The data derived by project llc_0001 researchers about sociodemographic measures using the syntax in the syntax file above
llc_0001_sesdemog_doc_v1	Documentation summarising descriptive sociodemographic information about the participants included in the dataset file above

4.4 Digital Object Identifiers (TBC)

Digital Object Identifiers (DOIs) are an internationally recognised standard for persistent and unique identification of any digital objects. They serve an important role in bibliographic citation and consequently, data-related outputs produced in the UK LLC will also require assignment of a DOI.

DataCite[9] have proposed a metadata schema for the citation of research data and related research outputs and offer useful guidance on the specification of DOI values in this context. Such values comprise both an organisational identifier (prefix) and a unique output identifier (suffix), separated by a forward slash. DataCite caps suffix values at 255 characters in length, but advises that an optimal length is 8-10 characters. While such suffix values can comprise any string term within those length constraints, the inclusion of semantic information is not advised[10]. As a consequence, the UK LLC will generate distinct suffix values for the data-related outputs as the basis for a DOI name that supplements the more informative detail captured by the primary naming convention as proposed above.

5. RELATED DOCUMENTS

- UK LLC Data Access and Acceptable Use Policy (POL-ISM-003)
- UK LLC Publication Policy (POL-ISM-007)
- UK LLC TRE User Guide (DOC-DAT-040)

[9] https://support.datacite.org/v1.2/docs/getting-started
[10] https://support.datacite.org/docs/doi-basics