Tools and Resources for FAIR Data

Anca Vlad
Research Data Publishing Officer,
Data Repository Administrator

18 May 2020

UK Data Service

SSHOC
social sciences & humanities open cloud

Summary

- Short introduction of the EMM Survey Registry
- Overview of FAIR data publishing standards
- Tools and resources for producing and publishing good quality FAIR data
- QAMyData, open-source tool – about and demo
- Questions

UK Data Service

UK Data Service(UKDS): Who are we?

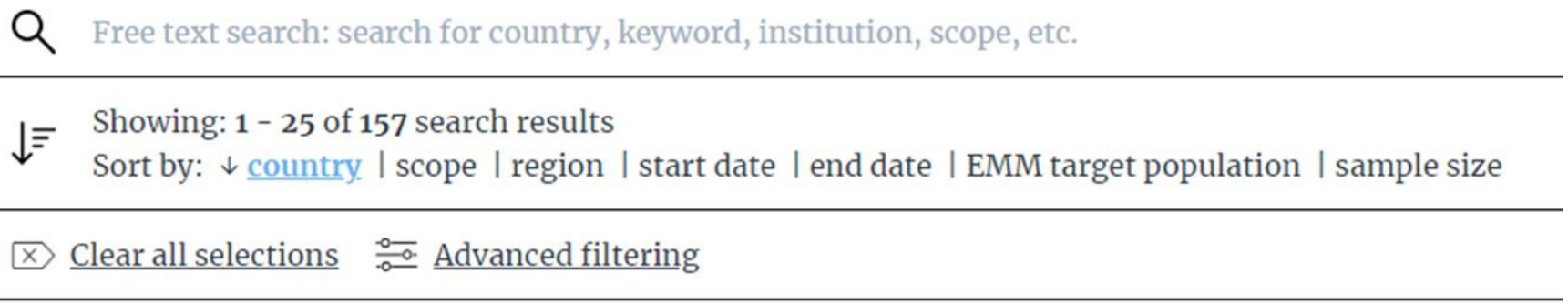
- UK's largest collection of UK and international social, economic and population data:
- Over 7,900 studies in our collection: https://www.ukdataservice.ac.uk/
- About 25,000 users
- Collection includes major UK government-sponsored surveys, cross-national surveys, longitudinal studies, UK census data, international aggregate, business data, and qualitative data

UK Data Service

EMM Survey Registry - About

The Ethnic and Migrant Minority (EMM) Survey Registry is a free online tool that allows users to search for and learn about existing quantitative surveys to EMM populations through the compiled survey-level metadata.

The EMM Survey Registry can be accessed here.



UK Data Service

Make your data FAIR?

What makes data good for sharing and reuse?

Other researchers can understand and reuse the data if:

- ✓ high quality
- ✓ accurate
- ✓ well organised
- ✓ easily accessible
- ✓ well documented
- ✓ long-term validity

or **FAIR** data:

Findable, **A**ccessible, **I**nteroperable, **R**eusable

Publish your data FAIRly

FAIR are guiding principles meant to improve the Findability, Accessibility, Interoperability, and Reuse of digital assets.

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

UK Data Service

FAIR in a nutshell

Findable – assigned a persistent identifier and described by detailed metadata

Accessible –clear access conditions

Interoperable – standardised formats and vocabulary

Reusable – well-defined license and ready to be reused in future research and processed using computational methods

UK Data Service

Is data Findable?

- The dataset should have a unique, permanent “address” that would allow it to be found using discovery portals. This is referred to as a persistent identifier and is normally part of that dataset’s citation.

Example: **Kovacheva, Siyka** and **Demireva, Neli** (2018). *The lived experiences of migration 1996-2017.* [Data Collection]. Colchester, Essex: UK Data Service. 10.5255/UKDA-SN-853333

- Another aspect that can make a dataset more findable is rich, machine readable metadata that describes the dataset and is used to index datasets in data catalogues and allow them to be found efficiently through discovery portals. Example: UKDS Data Catalogue, CESSDA Data Catalogue,

UK Data Service

Is data Findable?

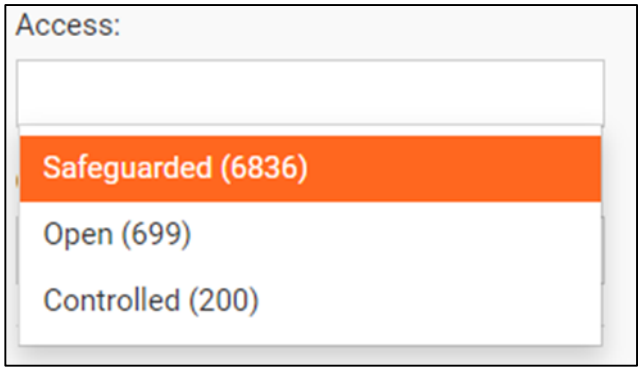
Migration and HIV risk in India: Places of origin study

About: A cross-sectional behavioural survey was conducted among non-migrants, returned migrants (with a history of migration), and active (current) migrants in rural areas across two districts with high levels of male outmigration.

UK Data Service

Is the data Accessible?

- Can be open but not necessarily.
- Some data needs to be placed under access restrictions due to privacy concerns, consent agreements, disclosure risk or commercial interests.
- Key: access should be implemented using a standardised protocol ->
- - > terms and conditions governing access and reuse should be clear, standardised and transparent.



UK Data Service

Is the data Accessible?

- Access is implemented using a standardised protocol.

Unravelling the Mediterranean migration crisis: The MEDMIG project journey data

Details | Access data

Details

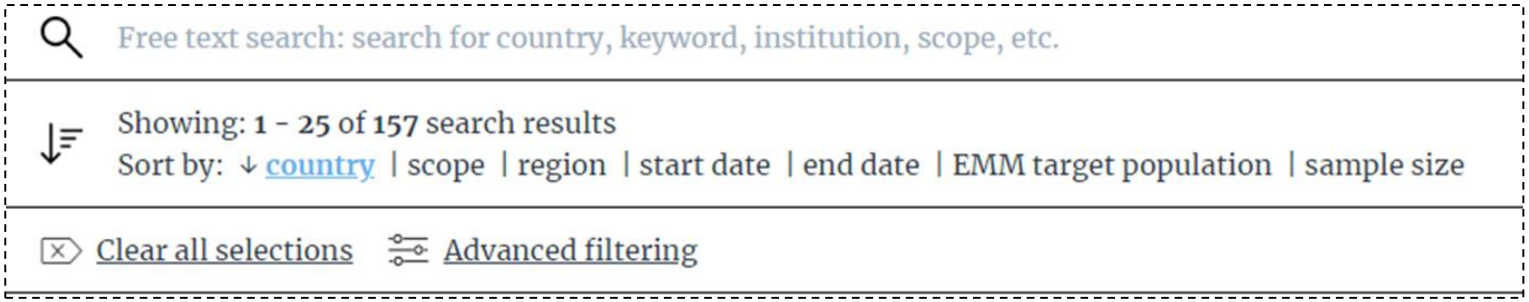
Title:	Unravelling the Mediterranean migration crisis: The MEDMIG project journey data
Study number (SN):	852674
Access:	These data are safeguarded
Persistent identifier:	10.5255/UKDA-SN-852674
Principal investigator(s):	Crawley, H, Coventry University Duvell, F, University of Oxford Sigona, N, University of Birmingham

UK Data Service

Is data Interoperable?

- Datasets are Interoperable if they are machine readable (metadata) and they are in specific formats, language and vocabularies and/or ontologies.
- *Digital data is software dependent, so endangered by obsolescence of software/hardware.*
- Formats used should be:
    - community agreed (vary across disciplines)
    - open (as opposed to proprietary)
    - unencrypted
    - suitable for long-term preservation

The metadata will also need to use a community agreed standards and vocabularies (such as the DDI Schema), and contain links to related information using persistent identifiers.



UK Data Service

Is the data Reusable?

- Does it have accurate and relevant attributes and provenance information? (machine readable metadata or text format)
- Does it meet domain-relevant community standards to allow it to be reused?

- Licencing - Stating clear re-use rights is like having a warm 'Welcome' on the doormat of your dataset.

- To make re-use as likely as possible, choose a licence which:
  - makes data available to the widest audience possible
  - makes the widest range of uses possible

UK Data Service

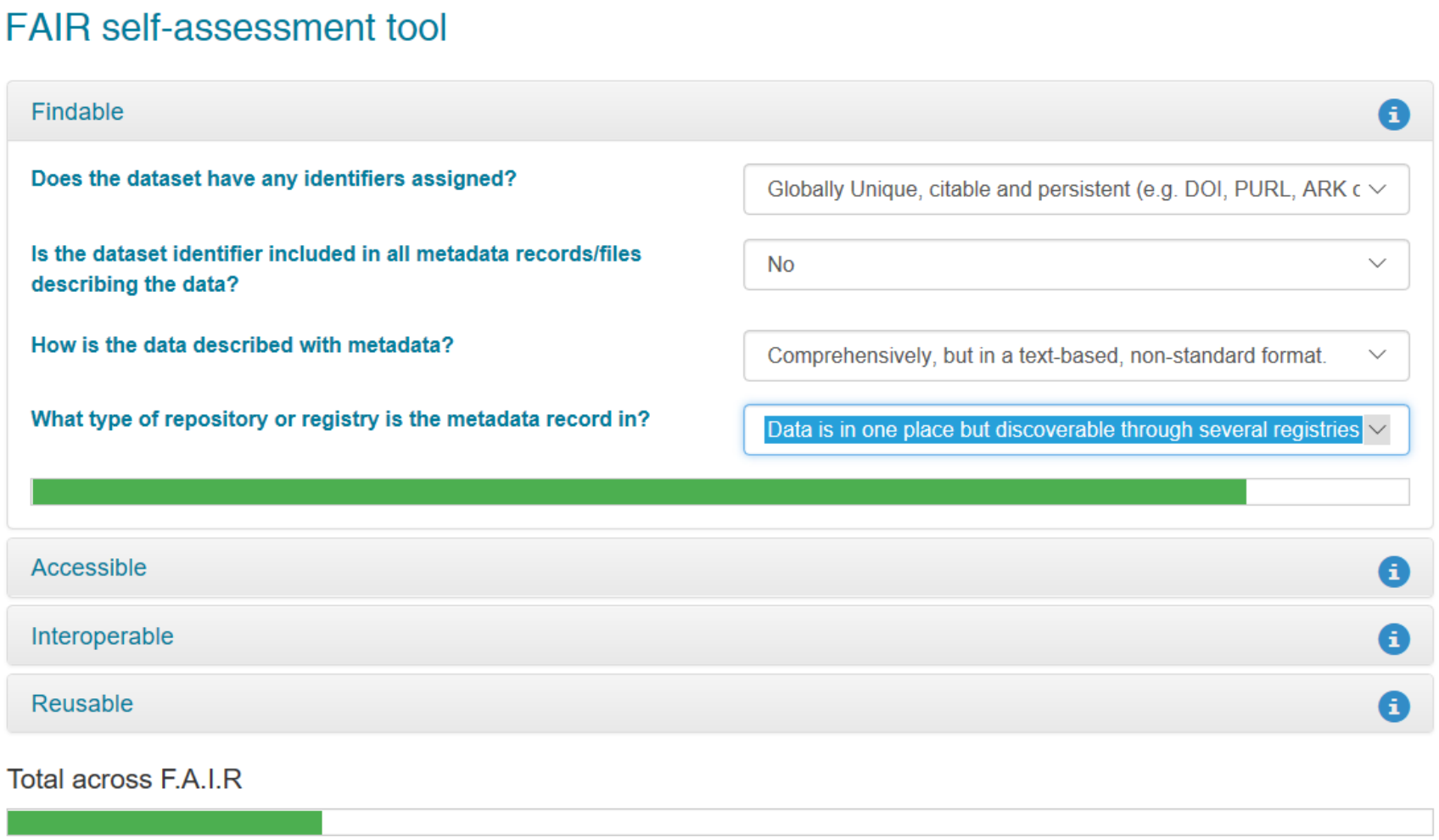
Tools and resources

FAIR data is the ultimate goal, still long way to go!

- FAIR self-assessment tool
- Data management plan
- CESSDA's Data Management Expert Guide
- SSH Open Marketplace
- Go FAIR Starter kit
- QAMyData

UK Data Service

FAIR assessment tool

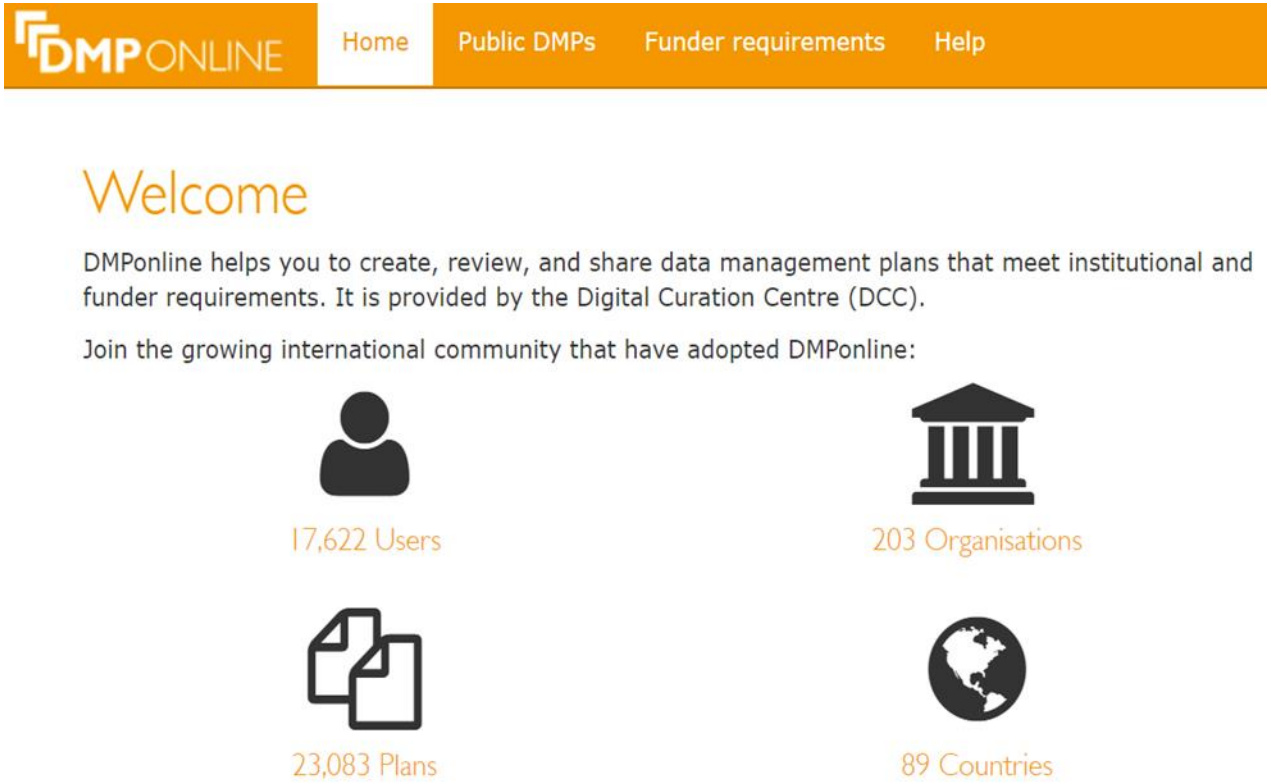


ARDC (2018) FAIR self-assessment tool, Australian Research Data Commons: https://www.ands-nectar-rds.org.au/fair-tool

UK Data Service

DMP Online

- A data management plan is a great tool to have along, as it helps decide how research data will be managed throughout the research cycle and how best to prepare for publishing.



UK Data Service

CESSDA's Data Management Expert Guide

- guide designed by European experts to help social science researchers make their research data FAIR
- entire research data lifecycle from planning, organising, documenting, processing, storing and protecting your data to sharing and publishing.

Includes:
- ✓ expert tips,
- ✓ European diversity in data management & data protection
- ✓ in depth tips for specific types of data
- ✓ training events

UK Data Service

SSH Open Marketplace

Want to make your data FAIR?

The SSH Open Marketplace can be used to:

- ✓ find relevant software and services
- ✓ describe how it aligns with standards and open science principles
- ✓ link to tutorials and other training material
- ✓ access a forum where other SSH peers would have commented on tools/software

UK Data Service

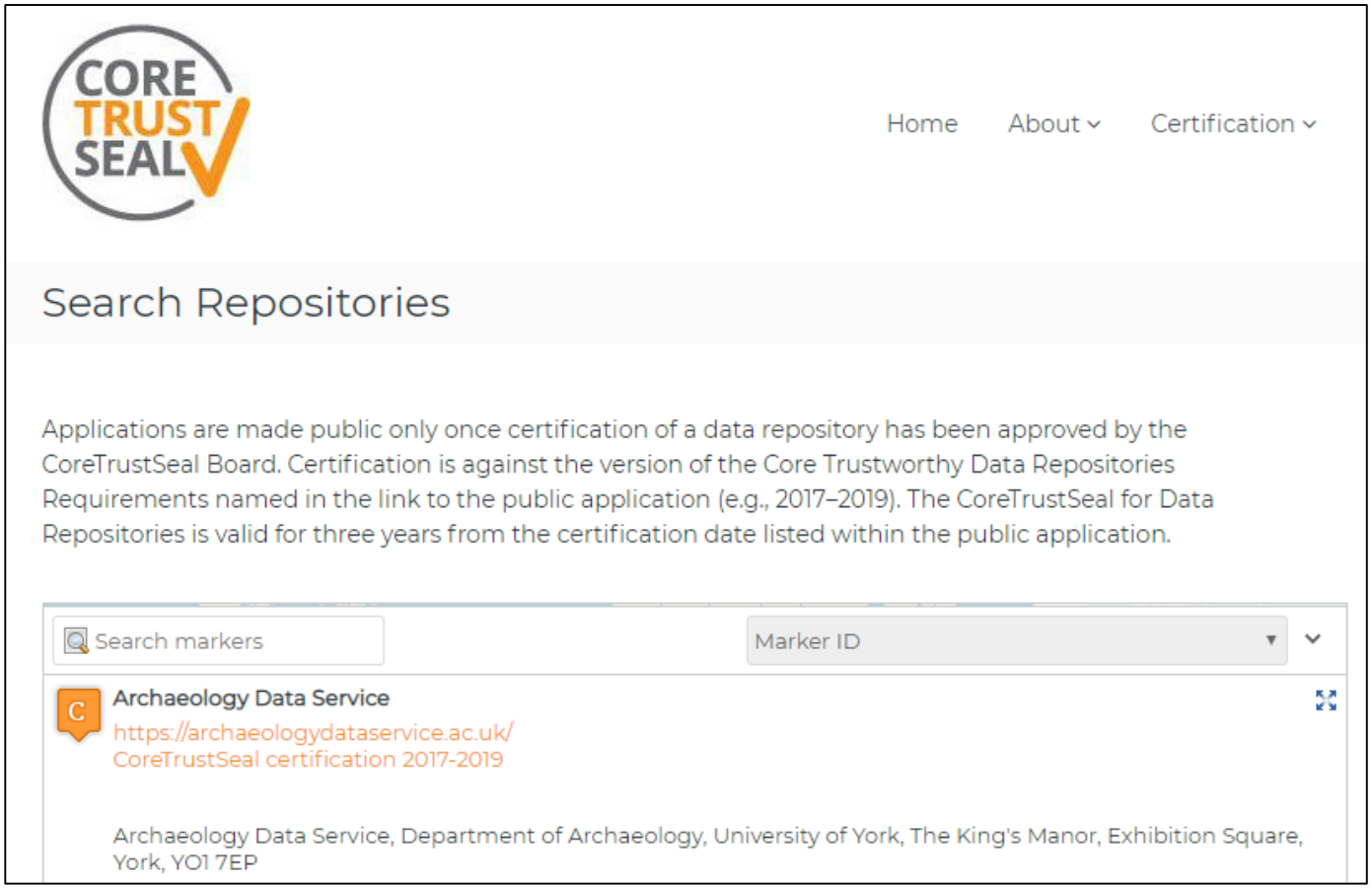
Go FAIR Starter kit

- ✓ Where to find a suitable research data repository
- ✓ RDM support
- ✓ Research Data Management Plans
- ✓ Information about Persistent Identifiers (PID)
- ✓ Information about licences
- ✓ Helpful webpages and presentations for Research Data Management

UK Data Service

Choose a trustworthy data repository

- For FAIR data – choose a trustworthy data repository.



UK Data Service

QAMyData

The UK Data Service developed a free easy-to-use open source tool known as **QAMyData** that provides a **health check for numeric data**.

The tool uses automated methods to detect and report on some of the most common problems in survey or numeric data.

Tool can:

- check structure, find issues
- incorrect, missing, inconsistent values
- check for unanticipated/accidental disclosure risk
- flag issues: enable a machine or human to resolve the problems
- be deployed as a service for self deposit repository, eg DataVerse for a submission health check
- be deployed into data publishing pipelines

UK Data Service

QAMyData

The tool offers a number of configurable tests that have been categorised into four types: file, metadata, data integrity, and identifiers.

Can be run on popular file formats, including SPSS, Stata, SAS and CSV.

The software creates a '**data health check**' that details errors and issues as both a summary and detailed report, providing a location of the failed test. New tests can easily be added.

The **QAMyData** software is easily downloaded to a laptop or server and can be quickly used and integrated into data cleaning and processing pipelines.

It is available to download from the UK Data Service Github page under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

UK Data Service

QAMyData: Metadata checks

Report on number of cases and variables	Always run
Count of grouping variables	
Missing variable labels	Must be set to true . If set to false the test will not run
No label for user defined missing values e.g. - 9 not labelled	SPSS only
'Odd' characters in variable names and labels	User specifies the characters
'Odd' characters in value labels	User specifies the characters
Maximum length of variable labels, e.g. >79 characters	User specifies the length
Maximum length of value labels, e.g. >39 characters	User specifies the length
Spelling mistakes (non-dictionary words) in variable labels using a dictionary file	User specifies a dictionary file.
Spelling mistakes (non-dictionary words) in value labels using a dictionary file	User specifies a dictionary file

QAMyData: Data integrity checks

Report number of numeric and string variables	
Check for duplicate IDs	User specifies the variables. Multiple variables can be added on new lines e.g. - Caseno - AnotherVariableHere
'Odd' characters in string data	User specifies the characters
Spelling mistakes (non-dictionary words) in string data using a dictionary file (can check if date format set correctly!)	User specifies a dictionary file
Percentage of values missing ('Sys miss' and undefined missing)	User sets the threshold, e.g. more than 25%

QAMyData: Disclosure control checks

Identifying disclosure risk from unique values or low thresholds (frequencies of categorical vars or minimum values)	User sets the threshold value, e.g. 5
Direct identifiers using a RegEx pattern search	User runs separately for postcodes, telephone numbers etc. Advise tests are separately as may be resource intensive
Direct identifiers/named entities in string data using a dictionary file (to be added)	Specify a dictionary file containing lists of stop words or named entities e.g. for places, names etc. Advise tests are separately as may be resource intensive

QAMyData Demo

Information and installation files can be found here. These include user guide, installation guide, training exercise and a purposely-erroneous dataset.

Outputs:

- Overview presentation (PDF)
- Blog
- Table of tests included (PDF)
- Tool and installation documentation
- Config file
- QAMyData Guide: How to install and run (PDF)
- Teaching exercise one: Identifying issues (PDF)
- Teaching exercise two: Using the QAMyData tool (PDF)
- Dummy Dataset (ZIP)

UK Data Service

Contact

Enquiries/ Help Desk:

http://ukdataservice.ac.uk/help/get-in-touch.aspx

help@ukdataservice.ac.uk

Follow us on:

https://twitter.com/UKDataService

https://www.facebook.com/UKDataService

https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=UKDATASERVICE