Workshop report

closer
The home of longitudinal research

Understanding metadata management in longitudinal studies

Hayley Mills & Jon Johnson

CLOSER

April 2020

UKRI Economic and Social Research Council

Table of contents

1. Key insights

- More training and online resources are needed in research data management, metadata and the Data Documentation Initiative (DDI) standard.
- For training to be effective and to ensure high uptake of metadata standards, best practice guidelines with case studies and examples are required.
- Communication with and buy-in from all stakeholders, including Principal Investigators (PIs) and funders, is key to ensuring resources for data management activities.

2. Introduction

The workshop explored how structured metadata and open standards can be used to actively manage longitudinal study data and resources.

The first day provided delegates with an overview of different structured metadata concepts, as well as use-cases for data management, question management, versioning and data sharing.

The second day included presentations from an international perspective: Kerrin Borschewski (Leibniz Institute for the Social Sciences (GESIS), Cologne, Germany) presented on the CESSDA Metadata Office Project; Alina Danciu (Center for Socio-Political Data, SciencesPo (CDSP) Paris, France) provided a case study in how to standardise metadata; and Bodil Agasøster (Norwegian Centre for Research Data (NSD), Bergen, Norway) presented a new survey design tool to be used by the European Social Survey (ESS).

UK perspectives were presented by: Aida Sanchez (Centre for Longitudinal Studies (CLS), UK) presented on the CLS birth cohorts and areas where CLOSER could help with providing guidance and improving CLOSER Discovery for longitudinal studies; Catherine Yuen (Institute of Economic Research (ISER), Colchester, UK) provided a case study of moving to DDI and Colectica software for surveys; and Jon Johnson (CLOSER, UK), provided a cross-study and institutional perspective.

There was a group discussion about future needs for metadata and data management, some of which were highlighted in Aida Sanchez's presentation, as well as a discussion identifying potential barriers to making progress in these areas. The following question was also put to delegates for discussion: *Where might some co-ordinated effort make a difference for improving longitudinal metadata management?*

3. Summary of themes arising from discussions

Potential advantages of using DDI

As part of the first day of training, delegates were asked to think about how DDI could be used in their current metadata management processes, as well as the potential advantages.

Several advantages related to improving data quality were identified. The production of summary statistics was suggested as a way to improve data quality, consistency and re-use. The possibility of cross-study validation of variables in cases of odd covariation could be used to establish if there is consistency in the data. Questions can be used as a mechanism to constrain the data for quality assurance e.g. if the data reflects the question and the routing. There are also options for recording any changes made to data and versioning, which will improve data quality and transparency.

The creation and understanding of relationships were also discussed as having several advantages, as it provides a way to bring together inputs from different sources, and is useful for documenting multi-agency data relationships and comparability. For example, the DDI framework offers a way of working with different agencies who use different terminology. DDI also allows relationships between questions to be created, which means questions can be linked across sweeps and across studies, and questions which are based on other questions can also be indicated.

Lastly, designing questionnaires and managing the metadata was seen as a potential area where DDI could be utilised within longitudinal studies. This was identified in Aida Sanchez's presentation and was later discussed by Catherine Yuen and Bodil Agasøster, where examples of how NSD and ISER are doing this were provided. The ability to produce a codebook more easily and in a more automated way was also seen as an advantage. The general consensus was that utilising standards and improving metadata practices should save time in data management in the long run.

Potential barriers of using DDI

Delegates indicated several barriers which would have an impact of the use of the DDI standard. Firstly, the structure of DDI can be complex, especially when working with very complicated routing in questionnaires. It would therefore involve a level of training and learning process to understand this complexity.

The second barrier was related to best practice and the different ways that DDI could be used. This was discussed later in the workshop by Kerrin Borschewski and Alina Danciu. Kerrin provided an overview of the CESSDA Metadata Model and how this was being used, and Alina provided examples of how inconsistencies in the use of the standard can be introduced, and how they could be resolved. Questionnaire design can benefit from the use of DDI, however, it is very important for those using it to know the protocols and good practice, for example, policies for version control.

Additionally, the time and resources needed to get setup was identified as a key barrier. Many of these discussions centred on buy-in, communication and community. It was thought that a lack of buy-in for metadata management from senior management could result in a lack of funding. Barriers were also identified which related to existing cultures, processes and internal politics, as well as resistance to change.

Technical requirements

Producing guidelines for best practice, particularly for the technical aspects, were identified as key for improving data management in longitudinal studies. Specific guidance related to advice about: preventing and preparing for software mortality, the advantages and disadvantages of flat files versus relational databases, off-the-shelf versus bespoke software, different programming languages, and data security. In addition, case studies which are of different scales and cover different timeframes were identified as being valuable examples to follow for planning improvements to current systems.

The software tools required was an area identified as needing coordinated effort. It was agreed that more specific tools are needed e.g. to document metadata quicker, to automate the topic mapping between controlled vocabularies and internal topics, and for developing Computer Aided Personal Interviews (CAPIs). Further to this, technical advice on existing DDI tools, for example questionnaire design, as well as a network for users of off-the-shelf products was thought to be worthwhile pursuing. For those that had created or were in the process of developing new tools, a method of sharing this information and knowledge, which could take the form of a community of developers and collaborators in order to develop stronger tools, was recommended. In addition, it was thought that communication and promotion of the value of adopting new technology and software so that users (internally and externally) are less resistant to change was important to pursue. Lastly, having more metadata available and in different formats was discussed as being helpful for others to build upon.

Funding requirements

Continuity of funding in order to retain staff as well as longer terms of funding for key infrastructures, e.g. archives and metadata platforms, were identified as important for improving data management. Funders were seen as being in a position to influence stakeholders both internally and externally. The importance of (meta)data management plans when issuing grant proposals and projects were discussed, including mandatory minimum standards. This would ensure that resources are available to data management teams, and would also emphasise the importance of metadata and data in research. Whether funders can influence external stakeholders was also discussed, such as the potential to influence data providers or collectors in improving what metadata is produced and retained. More generally, lack of funding for infrastructure and the underestimation of the complexity of the data management process, was also raised as a concern and was discussed as part of the community and communication requirements - outlined below.

Training requirements

Training needs were identified at all levels: Researchers and Survey Managers, early career practitioners, undergraduates/students, as well as at an organisational level. Areas of training required included: the importance of metadata, DDI, specific data management software, terminology and research data management.

Understanding and selecting the best format for training was also discussed. Online formats were identified as being key, so they could be used as a resource. This included podcasts, videos, online materials and expanding the CLOSER Learning Hub. In addition, in-person training, including training days, tutorials, and cascade training (e.g. software carpentry) or train-the-trainers was considered important, particularly for more interactive training and creating buy-in.

Community and communication

Building a community of several studies brings benefits from scale, as the more studies, data handlers and data users, the greater the collaboration and knowledge sharing between studies and institutions. Promoting best practice and providing advice at all levels, including senior management, is an important part of the process for ensuring uptake of metadata standards. Generally, increasing awareness of the importance of metadata management, via open door days, information sessions and the media, was seen as key for both ensuring data quality and securing funding. Improving communication between disciplines, including understanding of differences in the language and background, and using common terminology would ensure better collaboration as well interoperability. Communication with Researchers and understanding their needs was seen as fundamental for ensuring that the data management meets requirements.

4. Priorities

Delegates agreed that priorities should be made, but the consensus was to determine these after the event. A short questionnaire was circulated to the delegates six months after the workshop to determine which of the requirements were their highest priorities.

Delegates were asked to select up to three priorities for each area: technology, communication, funders and training, and were then asked to select their top three priorities. See Appendix A for the list and the overall scores in each area. The overall priorities were identified as:

- Guidance on how to prevent/prepare for software mortality;
- A network of users for off-the-shelf products e.g. Colectica lists/forum;
- Training and advice for Principal Investigators (PIs);
- Training for early career practitioners apart from Researchers;
- Understanding Researchers' needs.

5. Feedback

Delegates were asked to complete an evaluation form at the end of the workshop either on paper or online. Twenty evaluation forms were received giving an average satisfaction rating of 10/10 and an average relevance rating of 10/10. The networking opportunity, the presentations and the discussions were seen as the best aspects of the workshop. Areas which needed to be improved upon included having more examples and/or case studies, as well as practical activities.

6. Appendix A

Technology

Identified priorities	**Votes**
Specific tools: • document metadata quicker • automation between CLOSER topics and internal topics	8
A community of developers and collaborators	7
Guidance on how to prevent/prepare for software mortality	6
Access to metadata directly and in different formats	4
Network of users for off-the-shell products e.g. Colectica lists/forum	3
Guidance on how to develop CAPIs	1
Cluster licenses	0
Information on the pros and cons of: • flat files vs relational databases • off-the-shelf vs bespoke software • programming languages	0
Guidance on data security e.g. theft and impact of such, leaks and hacks	0

Communication

Identified priorities	**Votes**
A community of Data Managers for collaboration and cooperation with other institutions	9
Increasing awareness of metadata/DDI e.g. Open Door days, Information Sessions	5
Improve communication (language and background and common terminology)	4
Promote best practice of DDI	4
Understanding Researchers' needs	3
Promoting the value of adopting new technology/software so users are less resistance to change	3

Funders

Identified priorities	**Votes**
Continuity of funding to retain staff with specialist skills	10
Funders to influence external stakeholders (e.g. data providers)	8
Longer terms from funders for Archives and metadata platforms	5
Introduce Metadata Management Plans when issuing the grant prospects and apply sanctions to those who do not comply	4

Training

Identified priorities	**Votes**
Training and advice for Principal Investigators	6
Online resources and tutorials	5
Guidance for best practices: • Standard practice of using DDI at different scales and timeframes • Use cases for successes	5
Training/guidance on the importance of metadata	3
Training for early career practitioners apart from Researchers	3
Training for Researchers and Survey Managers	2
Free and effective resources to learn new software properly and efficiently	2
DDI training (including videos, online materials)	1
Cascade training (e.g. software carpentry) or train the trainers for spreading the training	1
Expand the CLOSER Learning Hub	1
Research Data Management training	1
Podcasts	0
More training on terminology	0
Training undergraduate students	0

Correspondence:
closer@ucl.ac.uk

First published in 2020 by

CLOSER (Cohort and Longitudinal Studies Enhancement Resources)

UCL Institute of Education

20 Bedford Way

London WC1H 0AL

www.closer.ac.uk

About CLOSER

The UK is home to the world's largest and longest-running longitudinal studies. CLOSER aims to maximise their use, value and impact both at home and abroad. Bringing together eight leading studies, the British Library and the UK Data Service, CLOSER works to stimulate interdisciplinary research, develop shared resources, provide training, and share expertise. In this way, CLOSER is helping to build the body of knowledge on how life in the UK is changing – both across generations and in comparison to the rest of the world. CLOSER was funded by the Economic and Social Research Council (ESRC) and the Medical Research Council (MRC) from 2012-17, and by the ESRC from 2017 to present. Visit www.closer.ac.uk.

The views expressed in this work are those of the conference delegates and do not necessarily reflect the views of CLOSER, UCL, ESRC, MRC or the Wellcome Trust.

This document is available in alternative formats.

Please contact CLOSER
Tel: +44 (0)20 7612 6938
Email: closer@ucl.ac.uk