16th European DDI Users Conference, Chur

Monday, 2 December 2024 - Friday, 6 December 2024

University of Applied Sciences of the Grisons (Pulvermühlestrasse 57)

Book of Proposals

Contents

Documenting Variables / 3

Creating a Custom Concept System to Document Longitudinal Studies

Author: Jennifer Zeiger[1]

Co-authors: Kathryn Lavender [1]; Sanda Ionescu [1]

[1] *ICPSR*

Corresponding Author: zeiger@umich.edu

The National Archive of Computerized Data on Aging (NACDA) began working with DDI-Lifecycle in 2018. Since then, NACDA has made efforts to document in DDI-L some of our most established and frequently-used longitudinal data collections and display them on a Colectica Portal. In this presentation, I will discuss how the system of topical groups and subgroups we use to organize the conceptual variables in these data collections improves the interoperability and findability of the metadata on our portal, how the system has evolved over time, and our plans for its improvement and use in the future.

NACDA is part of the Inter-university Consortium for Political and Social Research (ICPSR) and based at the Institute for Social Research (ISR) at the University of Michigan.

Sustainable, Ethical and Economical Controlled Vocabularies / 4

DDI RDF for the masses

Author: Darren Bell[1]

[1] *UK Data Service*

Corresponding Author: dbell@essex.ac.uk

While the semantic web has been around for over twenty years, practical and sustainable implementations have been thin on the ground. During 2024, DDI Controlled Vocabularies and the DDI-CDI ontology have been made available, for the first time, as persistently resolvable linked open data. This presentation digs into the underlying cloud infrastructure, the rationale for creating it and some practical guidance on how to get the most out of the resources now available. We will detail how to use a triple store like Apache Fuseki as the primary storage layer for making RDF data available, and why you should avoid trying to do this with SQL. As well as technical challenges, we'll also explore some of the organisational challenges we faced creating an RDF data pipeline from CESSDA (where the CVs are created and edited), to the DDI Alliance platform, where both HTML and RDF data are made available, and how to achieve this seamlessly using HTTP content negotiation.

Official Statistics / 5

Advancing Statistical Metadata at StatsCanada: The Role of DDI 3.3

Author: Philippe Bisson[1]

[1] *Statistics Canada*

Corresponding Author: philippe.bisson@statcan.gc.ca

This presentation will explore the strategic implementation and evolution of the DDI 3.3 standard at Statistics Canada, focusing on how it has enhanced metadata management practices within the organization. Under the guidance of StatCan's Information Governance Committee, DDI 3.3 has been mandated as a standard for microdata description, integrated into the agency's policy suite alongside SDMX. Its application is enforced through a Minimum Metadata Requirement (MMR), which is part of the directive on statistical standards. The presentation will discuss the strategic im-portance of DDI 3.3 in ensuring consistent, accurate, and comprehensive metadata documentation.

It will also cover real-world examples of DDI 3.3 implementation, particularly in managing structural and reference metadata through the MMR. Additionally, the presentation will address the role of the GSIM in defining metadata fields and the use of extension pairs to support policy and administrative functions. Finally, it will highlight enforcement strategies (XPATHs) to ensure compliance with metadata standards, contributing to a robust metadata governance structure. In conclusion, this presentation will offer valuable insights into the lessons learned from StatCan's DDI 3.3 implementation and provide a forward-looking perspective on how the agency plans to further leverage DDI to advance its metadata journey.

Interoperability.3 / 6

Integrating DDI 3.3 with ModernStats Models at StatCan

Author: Flavio Rizzolo[1]

Co-author: Philippe Bisson [1]

[1] *Statistics Canada*

Corresponding Author: philippe.bisson@statcan.gc.ca

The ModernStats models, including GSBPM and GSIM, alongside reference architectures like CSPA and CSDA, provide a robust framework for understanding statistical production, guiding business decisions, and designing reusable software components. However, bridging the gap between these conceptual models and their implementation standards—particularly DDI 3.3 and SDMX—has become a focal point of discussion within the international community.

This presentation details Statistics Canada's recent efforts to bridge this gap by deploying a suite of standard-based data and metadata management tools (e.g., Colectica, Aria, Fusion Metadata Registry, Data Lifecycle Manager) aligned with the ModernStat models. The challenge lies in making these tools, developed by different communities, interoperate at a semantic level. Achieving this requires mapping between the standards and developing customized micro-utilities that act as "connective tissue," enabling the creation of data and metadata pipelines across the phases of the GSBPM.

We will present key use cases that highlight our progress, describe the challenges we have encountered, and outline the road ahead. This session aims to provide a deep dive into how DDI 3.3 is being leveraged to enhance metadata-driven processes at StatCan, offering valuable insights for those working with or planning to implement similar standards-based approaches.

DDI & Policy / 7

Data policy to save decades

Author: Noé Nessel[1]

[1] *Asociación Trabajadores del Estado (ATE)*

Corresponding Author: noenessel@hotmail.com

In Buenos Aires, the data for importing, exporting and correct interoperability of academic years and subjects approved in other South American countries to be validated in the Argentine Republic are not harmonized.
For this reason, administrative processes have been accelerated in order not to stagnate the development of individuals from various Latin communities. Bolivians, Peruvians, Venezuelans; generally, they must study at least 4 years to enter the University. Most of them speak Spanish/Castilian perfectly.
The majority of ministerial officials do not activate or perfect the metadata system to validate the academic qualifications of young people and adults. From my position, pressure is exerted for the automation of international educational procedures taking into account the DDI standard.

User Needs 1 / 8

New Study and New Features: A Colectica Portal for the Wisconsin Longitudinal Study

Author: Barry Radler[1]

[1] *University of Wisconsin-Madison*

Corresponding Author: bradler@wisc.edu

This presentation will introduce the latest longitudinal study to use DDI Lifecycle and Colectica software. The Wisconsin Longitudinal Study (WLS) is a long-term study of a random sample of 10,317 men and women who graduated from the Wisconsin stat (U.S.) high schools in 1957. Since then, there have been six rounds of survey data collected from the original respondents and the sample has been expanded to include parents, siblings, and spouses of the graduates. The current round of data collection focuses on Alzheimer's Disease and Related Dementias (ADRD) by capturing detailed measures of cognitive change and collecting blood-based biomarkers of Alzheimer's Disease.
After first experimenting with DDI using the Nesstar suite of products, WLS began producing online codebooks in 2004 using Perl scripts to assemble DDI Codebook XML from a wide array of input sources. A successful National Institute on Aging (NIA) grant in 2023 included funding to migrate WLS codebooks to DDI Lifecycle and create a Colectica Portal much like other longitudinal studies (MIDUS, NHATS, etc.) have done. The WLS Portal (https://wls.portal.ssc.wisc.edu/) went live in June, 2024 and this presentation will provide a metadata administrator's perspective on the upgrade process, with a focus on two themes:
• Comparing and contrasting the electronic metadata before and after the upgrade, highlighting the pros and cons of using scripts vs. tools , and describing new features developed by Colectica software.
• Describe how updating the metadata production pipelines prompted a careful review of the actual microdata and documentation, leading to improvements in variable representations and streamlining how metadata are assembled .
The new Portal also presents the potential for cross-study harmonization using Lifecycle and Colectica; possible integration with NACDA's cross-series comparisons project will be reviewed.

User Needs 2 / 9

PROGEDO's new and improved data repository: shining a spotlight on metadata

Authors: Ami SAJI[1]; Nicolas Sauger[2]

[1] *CNRS (PROGEDO)*

[2] *PROGEDO / Sciences Po*

Corresponding Authors: ami-katherine.saji@cnrs.fr, nicolas.sauger@cnrs.fr

PROGEDO's data repository, Quetelet-Progedo-Diffusion (https://data.progedo.fr/), holds more than 1,600 datasets produced by the social sciences and humanities community in France. In spring 2024, PROGEDO upgraded this repository to offer a single entry point for discovering and requesting access to the available datasets.

This upgrade inevitably brought to the forefront the rich and well-structured metadata, in DDI Codebook, provided for each of the datasets. This is because users must exploit this metadata to not only browse the data collection, but also locate datasets of interest.

This presentation therefore examines how the new and improved Quetelet-Progedo-Diffusion facilitates data re-use via the metadata documented in DDI Codebook. It also explores how this heightened focus on metadata helps to showcase the benefits and critical role of metadata in disseminating and sharing data to the research community and beyond.

Interoperability 1 / 10

Publishing Fine-Grained DDI Metadata: Learnings from Efforts in Three Different RDCs

Authors: Knut Wenzig[1]; Andreas Daniel[2]; Dominique Hansen[1]; Tobias Koberg[3]; Mihaela Tudose[3]

[1] *DIW Berlin/SOEP*

[2] *DZHW*

[3] *LIfBi*

Corresponding Author: kwenzig@diw.de

As the availability of DDI metadata at the variable level is quite low, three German research data centers (DIW Berlin/SOEP, LIfBi, DZHW) collaborated in a KonsortSWD-financed project to make progress in this area.

All partners have a metadata system in place that is based on partly multilingual structured metadata at the variable level, including, for example, variable labels, categories, question texts, and keywords. Based on this, they wrote an initial export to DDI-Codebook 2.5. A comparison of these exports led to a suggestion on how the elements within <fileDscr> and <dataDscr> can be used to describe studies with multiple datasets at the variable level using the preexisting metadata.

While the exported DDI metadata in XML format could be easily published on a webpage, the project partners also evaluated the options for using the OAI-PMH protocol and available server software. The presentation will discuss the conceptual and technical questions that arise.

11

DDI Training Working Group - Train the Trainer Collaboration Workshop, EDDI 2024

Authors: Alina DANCIU[1]; Catherine Yuen[2]; Jennifer Zeiger[3]

[1] *Sciences Po, Center for Socio-Political Data (CDSP)*

[2] *Understanding Society*

[3] *ICPSR*

Corresponding Authors: catherine.yuen@essex.ac.uk, alina.danciu@sciencespo.fr, zeiger@umich.edu

The DDI Training Working Group (TG) expects to have several members attending the EDDI 2024 conference in person. With this in mind, we propose having a side in-person workshop after the EDDI conference. We plan to engage attendees on the topics of DDI training standards, requirements, and resources. Attendees are expected to be a part of the DDI Alliance, with some knowledge and/or interest in DDI and community training demands and needs. We also expect to provide attendees with examples of past training - noting successes and improvement areas, and an overview of the currently available training slide decks. Lastly, we will plan to discuss targeted audiences, such as statistical agencies, and brainstorm together the best approaches to training that audience.

Poster Session / 12

Use of Artificial Intelligence in Metadata Creation in the Social Science Japan Data Archive: Metadata Extraction in Social Surveys Using Large Language Models

Authors: Koichi Iriyama[None]; Takenori Konaka[None]; Yukihiro Nishimura[None]

Corresponding Author: konaka@iss.u-tokyo.ac.jp

To improve the effectiveness of metadata production in Japanese for social surveys, the Social Science Japan Data Archive (SSJDA) at the University of Tokyo's Institute of Social Science has built a metadata extraction method using OpenAI's API. Since April 1998, SSJDA has been steadily providing data by making metadata manually. However, with the increase in deposited data, we need to improve the efficiency of metadata production. The system demonstrated considerable improvements in processing time and cost efficiency compared to the manual approach, while also showing a high level of concordance with expert-created metadata. In the future SSJDA will implement LLMs in local environments and develop a user-friendly GUI, which is expected to yield further enhancements and to support self-archiving.

Interoperability 1 / 13

Harvesting DDI Codebook Metadata at Scale: Challenges and Experiences using OAI-PMH

Author: Xiaoyao Han[None]

Corresponding Author: xhan@diw.de

As the DDI community continues to grow and its user base expands, the demand for sharing and accessing metadata in this format is steadily increasing. This presentation shares our experiences in collecting DDI codebook metadata records using the OAI-PMH protocol from diverse repositories. We will demonstrate the tools and processes employed to access and extract DDI metadata elements and the technical challenges encountered during the data acquisition process, such as prefix normalization issues and the current state of data provision will be shared. In addition, we hope to discuss how to facilitate data collection standardization through OAI-PMH and foster the wider sharing and utilization of DDI metadata for academic research.

Poster Session / 14

Using the Association of Religion Data Archives (ARDA) to Strengthen the Religion Research Community

Author: Tom Clark[1]

[1] *University of Sheffield*

Corresponding Author: t.clark@sheffield.ac.uk

Founded in 1997, the Association of Religion Data Archives strives to democratize access to the best data on religion. The ARDA includes American and international data collection and a host of free resources. Data included in the ARDA are submitted by the foremost religion scholars and research centers in the world. This poster will introduce three major initiatives that are building upon and complementing existing ARDA features and services. The initiatives include: a) building on the strengths of the current Data Archive; b.) new tools for discovering data and measures; and c.) building a religion research community that promotes open science principles. These initiatives all leverage the ARDA's newly rebuilt website architecture and will include: a new self-submission portal; a collaborative programme of work with journal editors in the field of religion; and, a religion research hub that will incentivize the submission of replication data to the ARDA.

Machine Learning and Media / 15

Metadata for Social Media and Web Tracking Data

Authors: Wolfgang Zenk-Möltgen[1]; Kokila Jamwal[2]; Libby Bishop[2]

[1] *GESIS - Leibniz Institute for the Social Sciences*

[2] *GESIS*

Corresponding Author: wolfgang.zenk-moeltgen@gesis.org

At GESIS, we plan to collect more digital behavioral data, e.g. social media data and web tracking data. Data sources are currently X/Twitter and tracking data collected by GESIS. The GESIS Web Tracking software works via a browser plugin on desktop devices. To document these data sources for archiving, additional new information is needed beyond the usual survey metadata.

Challenges are the different sources, types of data collection methods, and new data preparation and selection procedures. Another challenge is the sensitive character of some of the collected data that needs some preprocessing before it can be shared. Further consideration needs to be given to legal aspects of data collection and data sharing.

The presentation will report the proposals to capture existing and additional types of information for these new social science datasets and what an implementation could look like. We use Colectica as a tool for documentation and will highlight possible usages and extensions of the current DDI-Lifecycle standard for the purpose of documenting social media and web tracking data.

Software 2 / 16

Enhancing metadata interoperability: The journey to DDI Lifecycle implementation in the CESSDA Data Catalogue

Author: Matthew Morris[1]

[1] *CESSDA*

Corresponding Author: matthew.morris@cessda.eu

The CESSDA Data Catalogue (CDC) has long supported metadata about Social Science studies in DDI Codebook 1.2.2 and 2.5 forms. Historically, however, metadata in DDI Lifecycle formats has not been supported. This was a pain point for CESSDA's service providers who work with this format.

The CESSDA Metadata Office has created mapping from DDI 3 metadata to CDC UI elements which was the basis for this work.

In order for the CDC to ingest Lifecycle metadata CESSDA MO has implemented a parser. Implementing Lifecycle is significantly more complex than Codebook. Simple parsing techniques like linearly reading through the document are not sufficient for Lifecycle. This is because DDI 3 elements can reference other parts of the document. These references need to be resolved.

To implement this, the behaviour of the parser needed to be defined programmatically at the XPath. This is different from what the parser did previously and required significant rewrites to introduce the required flexibility. Extensive use of lambdas was used to associate XPaths with parsing behaviour.

Next steps will be looking at performance and memory optimisations could be reduced when parsing large XML source files. Could this be accomplished with a streaming XML parser?

Documenting Variables / 17

Mapping PhysicalInstance and DCAT

Author: Christophe Dzikowski[1]

[1] *INSEE*

Corresponding Author: christophe.dzikowski@insee.fr

INSEE, the French national statistical office, has launched an overhaul of its dissemination processes in order to improve efficiency and services to producers. Part of this project involves developing a data dissemination platform where data can be discovered and consistency between data and metadata verified. In terms of metadata, this boils down to describing the structure of the data and creating a catalogue to document the data.

In practice, and in the context of INSEE's RMéS metadata repository, this means using different standards and different parts of the repository. On the one hand, we need to describe the micro-data with DDI L, in particular with PhysicalInstance objects, and on the other hand, we need to use the DCAT RDF model which is convenient for publishing data semantically on the web.

The questions of this project are therefore the correspondence between the PhysicalInstance DDI-L objects and the DCAT standard. We will present the extent to which these two standards can be used together, and we will highlight the issues and difficulties of this mapping which are linked to the limited possibility of defining the data structure at the level of the RepresentedVariables in DDI-L.

Practical considerations will be presented regarding the management and requirements of this type of multi-repository metadata description, and recommandations could be made to streamline the relation between these standards.

DDI-CDI / 18

DDI-Cross Domain Integration: Features, Tools, and Early Adoption

Author: Arofan Gregory[1]

[1] *CODATA*

Corresponding Authors: benjamin.beuster@sikt.no, dbell@essex.ac.uk, ilg21@yahoo.com

This is a full session featuring three presentations and a panel discussion. The goal is to present the capabilities of DDI-CDI as an overview, to describe those organisations and efforts which are already using the standard and planning to do so in the near term, and to show the active tools development which is taking place.

1. **DDI-CDI: A General Introduction** - describes the features of the specification and gives an update on status of the specification and documentation. Emphasises the way in which DDI-CDI works with other standards and compleents DDI Codebook and DDI Lifecycle.
2. **DDI-CDI: Early Adopters and New Projects** - A presentation on who is using the standard and for what applications. This would include UKDA, the Cross Domain Interoperability Framework (CDIF), some of the EOSC andd EU projects implementing the standard (such as CLIMATE ADAPT, the ESS Labs, and the X-Ray Absorption Spectroscopy OSCAR project), and possibly others.
3. **DDI-CDI : Tools and Services** - this presentation gives a general overview of those who are implementing support for DDI-CDI (the Nectar work from the DDI Developer's Group, the "high-Value Data Services" tools Pascal Heus is developing at Postman, the SPSS/SAS converter which came out of WorldFAIR, the Process browser from ESS, implementation in Dataverse, the UN SDG Indicators SDMX transformation service, and others). In addition to the general overview, an in-depth look at one area will be provided - Benjamin Beuster will demonstrate current developments at Sikt around SPSS/Stata transformation and working with SQL databases.
4. **Panel Discussion** - presenters and implementers will discuss various aspects of DDI-CDI and will respond to questions from the audience.

Sustainable, Ethical and Economical Controlled Vocabularies / 19

"You are what you eat": Deploying a method for creating an accurate ML model for variable tagging using ELSST vocabularies

Authors: Jieun Jeong[1]; Lucie MARIE[1]; Mathieu Olivier[1]

[1] *Centre for socio-political data, Sciences Po, CNRS*

Corresponding Authors: jieun.jeong@sciencespo.fr, lucie.marie2@sciencespo.fr

With more than 400 DDI documented datasets, Center for Socio-Political Data's catalogue (CDSP) counts ten of thousands of variables - mainly quantitative survey data collected from structured questionnaires.
With the final goal to produce accurate and consistent data training material for a machine learning model (camemBERT), the CDSP's engineers launched a working group for variable tagging using the French version of the European Language Social Science Thesaurus (ELSST) keywords.
Experimenting with machine learning for classifying data at the variable level, this paper evaluates the machines' capabilities to process and classify large datasets, while emphasising the accuracy and contextual understanding that human experts provide.
This presentation aims to provide feedback on the methodology developed for the human tagging process in order to minimise bias and provide a harmonised classification.

20

DDI Marketing Group

Authors: Barry Radler[1]; Jon Johnson[2]

[1] *University of Wisconsin-Madison*

[2] *CLOSER, UCL*

Corresponding Authors: jon.johnson@ucl.ac.uk, bradler@wisc.edu

The DDI Marketing Group has recently reformed and will meet to work on developing the Marketing Strategy, messaging and the development of Case Studies

Official Statistics / 21

DDI-Lifecycle 3.0 in the production of official statistics: views on a preliminary experience

Authors: Alicia Nieto[1]; Carmen Elena Guaza Picallo[None]; Clara Marín Cuadros[None]; Sandra Barragán Andrés[None]

Co-authors: David Salgado Fernández ; David Sánchez Hernández

[1] *Statistics Spain*

Corresponding Author: alicia.nieto.ramos@ine.es

In mid-2022, Statistics Spain (INE) decided to design and set the strategic guidelines for a new technological infrastructure that would provide our statistical office with the necessary hardware and software to automate, standardize, and modernize the entire statistical production process, encompassing improvements in terms of storage, virtualization, ingestion, analysis, processing, and dissemination of information.

This ambitious, innovative, and promising project was soon envisioned as a strategic axis of the organization, and one that could not be conceptualized without an exhaustive, necessary, sufficient, and standardized metadata ecosystem.

For this reason, a multidisciplinary group of experts from differerent units was tasked with analyzing current metadata international standards, giving special priority and emphasis to the set of structural metadata for microdata, its advantages and disadvantages, as well as the potential a priori materialization of these standards in the institution.

This paper summarizes the development of a first proof-of-concept study carried out by Statistics Spain between 2022 and 2024, in which, using the microdata from the Retail Trade Index Survey for January 2021, DDI-L 3.0 has been tested as a potential microdata standard in the production of official statistics. We share our main preliminary findings in this experience to document microdata in a statistical production process.

User Needs 2 / 22

Optimising Metadata Quality in LIFEOBS: The Role of DDI Standards in Harmonisation Across Diverse Data Documentation Teams

Authors: Anna Sidorets[1]; Lucas Bourcier[2]

[1] *CNRS - PROGEDO*

[2] *INED*

Corresponding Authors: lucas.bourcier@ined.fr, anna.sidorets@cnrs.fr

Launched in 2020, the LIFEOBS project spans seven major French national surveys covering all life stages, including three integrated into European research infrastructures: the Generations and Gender Programme (GGP2020), SHARE, and GUIDE-EuroCohort. This project is conducted by key French institutions, including but not limited to INED and PROGEDO. To improve international visibility and usability, the project uses DDI standards for the documentation of collected data, ensuring that documentation is optimised to align with global data practices and support accessibility for researchers worldwide.
Given the project's scale, we focus extensively on documentation and data distribution. This enables us to document new surveys and review previous survey waves, ensuring all documentation complies with DDI standards.
Our presentation highlights how DDI standards have been crucial in reinforcing collaboration and normalising practices between the separate teams at INED and PROGEDO, leading to harmonised, high-quality metadata across LIFEOBS. This process enhances the value of collected data by ensuring consistent and high-quality documentation, which is essential for the effectiveness of subsequent data processing.
This presentation is particularly relevant to DDI users and professionals. We aim to share practical insights on applying DDI standards in large, multi-site projects, offering useful takeaways for similar work globally.

User Needs 2 / **23**

COORDINATE Project and CESSDA Tools: Empowering Child and Youth Wellbeing Research through a Thematic Metadata Portal

Author: Markus Tuominen[1]

[1] *Finnish Social Science Data Archive (FSD)*

Corresponding Author: markus.tuominen@tuni.fi

One objective of the COORDINATE project is to provide improved access to studies related to child and youth wellbeing. This has been achieved by utilizing CESSDA's software and metadata provided by CESSDA Service Providers in DDI-C and DDI-L formats.

In this presentation, the various components that enable the portal's functionality will be covered. This includes harvesting DDI metadata from OAI-PMH, indexing it into Elasticsearch, and ensuring that only studies related to child and youth wellbeing are included in the portal.

Development of the portal is ongoing and will continue at least until early 2025, when the portal is expected to be officially released. However, a demo version of the portal has been publicly available since late 2023, offering a preview of its capabilities. This demo will be showcased during the presentation to illustrate the current state of the portal, how thematic views can be created based on DDI metadata, and the portal's potential impact on research accessibility.

The presentation will also address the challenges encountered during development, the lessons learned, and areas identified for further improvement.

Software 1 / **24**

The Road Forward

Author: Wendy Thomas[None]

Corresponding Author: wlt@umn.edu

In 2022 the Technical Committee provided a roadmap for their work during the period 2023-2027. We are now about 1/3rd of the way through the initial roadmap and it's time to see what we have

accomplished, what needs adjustment, and what lies ahead in the 18-24 months.

This roadmap reflects the long-term priorities of the Technical Committee and identifies the specific tasks that are needed to accomplish these goals. We have made substantial progress on the original task list and have expanded our activities in developing a technical infrastructure that will broadly support the development and long-term management of the DDI Product Suite. Following a quick status report this presentation will focus on the overall technology and product goals, additions to activities in the roadmap, activity areas for the next 12-18 months, and new opportunities for DDI members and community to involve themselves in product development and maintenance.

25

Tools and processes for generating and manipulating DDI-XML files

Author: Julie Lenoir[1]

[1] *French Institute for Demographic Studies*

Corresponding Author: julie.lenoir@ined.fr

Location: Building A: Room A3.08

Producing, harmonising and updating metadata are tasks metadata curators have to do regularly in order to make sure catalogues and their contents stay up-to-standards.

Drawing from the French Institute for Demographic Studies (INED) experience, this workshop aims at sharing about new procedures implemented in the last two years in order to support the needs for generating and manipulating DDI-XML files. The goal is for attendees to understand these procedures and be able to adapt them to their specific needs.

The first block will focus on the process for study-level documentation implemented at INED to facilitate data collection from research teams and accelerate the documentation writing process. This consists of an R script drawing information from a form (in our case a Microsoft Word one) and exporting it to a DDI-Codebook 2.5 XML file. The goal will be both to explain how the script works and to touch on some ways it could be adapted to other institutions standards or needs in terms of documentation. I will also quickly showcase the script used for variable-level documentation.

The second block will present how to use the eXtensible Stylesheet Language (XSL) and more specifically how one can use XSL Transformations (XSLT) to transform DDI-XML files. The presentation will consist of a quick overview of XSL, and will then focus on providing working examples (notably add attributes, add, replace or reorder tags).

Intended audience: metadata producers and/or curators working with DDI-Codebook 2.5
Pre-requisites: good knowledge of XML and of DDI-Codebook schema, good knowledge of the R language (the script uses the following packages: xml2, XML and xslt).

Software 2 / 26

A Journey into (meta)data management with DDI

Authors: Julie Baron[1]; Julie Lenoir[1]

[1] *French Institute for Demographic Studies*

Corresponding Authors: julie.baron@ined.fr, julie.lenoir@ined.fr

Since 2020, the French Institute for Demographic Studies (INED) has undertaken a significant transformation in its data dissemination and metadata production processes. Following a thorough evaluation of alternatives to the Nesstar software to the implementation of the NADA Microdata Cataloging Tool and the metadata update and migration process, we are now two years into using NADA. We devised a refined documentation and publication routine, supported by a suite of tools that enable us to adhere to the best practices of DDI Codebook, FAIR principles and European standards.

The metadata production is carried out using a formatted Microsoft Word form, supplemented by R scripts. Information at the variable level (labels and attributes) is retrieved through the use of R scripts. Finally, the publication of metadata from related publications down to the variable level is made easy thanks to the user-friendly interface and straightforward administration of NADA.

In this presentation, we will focus on this new process and provide a hands-on overview of our documentation workflow from data acquisition to publication. Special attention will be given to catalog administration and the publication stage.

Interoperability 2 / **27**

Creating an ISO Standard for the DDI Suite

Author: Dan Gillman[1]

Co-author: Arofan Gregory [2]

[1] *Data Unchained LLC*

[2] *CODATA*

Corresponding Author: dgillman4909@gmail.com

Since DDI is a suite of standards and other semantic products, the idea of creating an ISO standard for them is a challenge. Do we pick one of the DDI products and create a standard for that? Which one? Why? Will the revision schedule for the product selected interfere with the creation and maintenance for the standard? Will the Alliance be able to provide the resources for an ambitious standardization process? And can we produce a standard that represents the work under DDI yet stands on its own and is not subject to the revision schedule for DDI products?

Rather than try to address these questions directly, the Scientific Board established a temporary working group (tWG) to look into the question. The tWG elected to copy the approach SDMX took for a similar project 20 years ago. The SDMX standard focused on the core strength of SDMX - exchange. So, the tWG endeavored to find similar strengths of DDI. What are the core strengths of the products in the DDI suite?

The tWG decided to focus on variables (the variable cascade in particular) and the data lifecycle. Each DDI standard addresses variables and can be situated in the data lifecycle. These are the core strengths. Creating a standard around these will satisfy all the conditions laid out above.

This paper considers the questions posed above and addresses the answers outlined. Finally, the path forward in the ISO community will be described.

User Needs 1 / **28**

Processing cross-national longitudinal panel surveys to document rich metadata using automation and open standards: the case of the Generations & Gender Programme

Author: Thibaud Ritzenthaler[None]

Corresponding Author: thibaud.ritzenthaler@ined.fr

High quality metadata is a prerequisite for any data producer in contemporary research. The Generations and Gender Programme has a strong interest in this matter. The GGP is a cross-national (Europe and beyond) longitudinal survey providing data on a variety of topics including partnerships, fertility, work-life balance, transition to adulthood and later life. The documentation of the latest round of the survey (GGS-II) is in constant evolution to follow the latest FAIR best-practices, supported by the Data Documentation Initiative (DDI) framework. The survey metadata are available in DDI-Lifecycle and hosted on Colectica portal.
The challenge is to publish high quality metadata, following international standards, with a good level of FAIRness, through a fast and easy procedure, especially as the GGP is on the roadmap of the European Strategy Forum on Research Infrastructures (ESFRI) and aims to become a European Research Infrastructure Consortium (ERIC). This can be achieved by having a strong pipeline of automation. Therefore, this presentation will focus on how to develop this type of process, choose the good technical stack and have a robust quality control pipeline in order to go from raw data to comprehensive metadata using DDI Lifecycle, appropriate CESSDA controlled vocabulary and future developments.

Metacurate-ML / 29

Metacurate-ML: Metadata Extraction from CAI

Authors: Suparna De[1]; Jon Johnson[2]

Co-authors: Zeqiang Wang [1]; Chandresh Pravin [1]; Deirdre Lungley ; Paul Bradshaw [3]

[1] *University of Surrey*

[2] *CLOSER, UCL*

[3] *Scottish Centre for Social Research (ScotCen)*

Corresponding Author: s.de@surrey.ac.uk

Extending the results of our work on pre-trained language models with recent developments in text-layout models and zero-shot techniques. Since relying solely on textual information makes it difficult to accurately classify and extract metadata, a combination of textual content and visual logic that incorporates vision transformers with optimisation techniques will be explored.
This will allow us to extract the specific items with questionnaires such as question texts, responses and routing to create a rich source of metadata which provenances'data collection methodology to the resultant data which can be transformed into DDI-Lifecycle. We will investigate the feasibility of document understanding multimodal models that employ masked language techniques and present the resulting challenges.

Metacurate-ML / 30

Metacurate-ML: Conceptual Comparison

Authors: Suparna De[1]; Zeqiang Wang[1]

Co-authors: Wing Yan (Justina) Li [1]; Deirdre Lungley ; Paul Bradshaw [2]; Jon Johnson [3]

[1] *University of Surrey*

[2] *Scottish Centre for Social Research (ScotCen)*

[3] *CLOSER, UCL*

Corresponding Author: s.de@surrey.ac.uk

Questions from the CLOSER DDI-Lifecycle repository will be used to assist in training a model that is capable of using questions and response domains from the metadata extraction workstream to create conceptually equivalent items from which data variables can be concorded. Approaches such as fine-tuned large language model (LLM)-based relevance scores model and vector retrieval-LLM reordering will be presented.
The session will present initial results in question concept tagging that feed into the conceptual comparison task, addressing challenges of long-tail distribution of the data, model memorisation and human annotation bias in the dataset. Higher-level machine learning (ML) limitations of identifying indeterminate tags and the notion of probability in model outputs will be explored.

Metacurate-ML / 31

Metacurate-ML: Enhanced Data Curation - Automation of Disclosure Control Assessment

Authors: Deirdre Lungley[None]; Ivan Evdokimov[1]; Jon Johnson[2]; Paul Bradshaw[3]; Suparna De[4]

[1] *University of Essex*

[2] *CLOSER, UCL*

[3] *Scottish Centre for Social Research (ScotCen)*

[4] *University of Surrey*

Corresponding Author: dmlung@essex.ac.uk

Conceptual annotations and provenance can provide contextual information to inform a range of data processing activities. In this workstream we will be utilising the metadata generated in the earlier workstreams –the questions and response domains from the metadata extraction phase and the concorded variables from the conceptual comparison phase –to identify key variables, those that although are not sensitive in of themselves, have the potential to be disclosive if used in combination. This identification will be achieved using state-of-the-art text classification methods, which we will also use to identify such metadata as identifiers and weight variables. Rule-based classifiers will further interrogate the variable metadata to determine its classification hierarchy and level, e.g., a socio-economic variable may be coded using the ONS NS-SeC classification hierarchy at the 8-class analytic level.

This enhanced metadata can then be combined with the data itself to provide an enhanced curation platform –one which allows our data curators to evaluate and mitigate the disclosure risk of a dataset with relative ease. The resulting platform will be powered by metadata and microdata stored using the DDI-CDI schema, utilising such aspects as its variable cascade.

32

DDI Scientific Board Meeting

Author: Ingo Barkow[1]

[1] *FHGR*

Corresponding Author: ingo.barkow@fhgr.ch

Meeting of the DDI Alliance Scientific Board

Software 1 / 33

Colectica in Action: Real-World Applications of DDI in Europe across the Data Lifecycle

Author: Jeremy Iverson[1]

[1] *Colectica*

Corresponding Author: jeremy@colectica.com

This presentation will showcase public, in-production use cases of Colectica software, highlighting its role across various stages of the data lifecycle. Colectica, built on the Data Documentation Initiative (DDI) standard, serves as a powerful tool for data documentation and management. Specific use cases will include projects that leverage the software and the DDI standard in the following areas:

- Question banks
- Survey design and implementation
- Longitudinal study data documentation
- Data catalogs for cross-national, longitudinal datasets
- Eurostat quality reporting
- Classification management and publication

These examples will illustrate how DDI and Colectica enable enhanced data discovery, promote interoperability, and improve accessibility for European national statistical offices, research projects, and data archives.

Interoperability.3 / 34

Metadata Interoperability with RDF and JSON-LD in DDI Lifecycle 4 and the Colectica Portal

Author: Dan Smith[1]

[1] *Colectica*

Corresponding Author: dan@colectica.com

This presentation explores the newly implemented RDF (Resource Description Framework) support in DDI Lifecycle Version 4 and its significant impact on enhancing metadata interoperability. With the growing use of linked data and semantic web technologies, RDF offers a standardized method for representing and exchanging metadata, enabling smoother integration across diverse metadata models and systems.

In DDI Lifecycle Version 4, RDF is introduced as a primary serialization format. This presentation examines how this RDF support enables metadata to be published in a structured, machine-readable format that adheres to semantic web standards. A key focus will be the Colectica Portal's utilization of DDI in JSON-LD format to embed metadata directly into web pages for studies, datasets and variables. By embedding JSON-LD, each dataset and variable page includes rich, machine-readable metadata that can be easily interpreted by search engines and external systems. These new linked data capabilities improve data discovery, accessibility, and interoperability across platforms.

Additionally, this presentation discusses how these advancements align with the FAIR (Findable, Accessible, Interoperable, Reusable) data principles. By providing seamless access to well-structured metadata directly from the web, this integration promotes more effective data sharing and reuse within the research community.

35

Colectica Datasets Unveiled: Software for Data Viewing, Curation, and Publication

Author: Jeremy Iverson[1]

[1] *Colectica*

Corresponding Author: jeremy@colectica.com

This workshop introduces Colectica Datasets, a powerful new application designed for viewing, improving, and publishing data files. Running on both Windows and macOS, Colectica Datasets supports a wide range of dataset formats, including Parquet, SPSS, SAS, Stata, and CSV. Participants will learn to:

- **View**: Learn how to inspect, visualize, and analyze datasets to quickly understand the data.
- **Curate**: Discover techniques for curating, cleaning, and annotating datasets and their metadata, ensuring high-quality documentation and preparation for analysis.
- **Publish**: Explore methods to share, export, convert, and archive datasets in a variety of formats, making data dissemination and long-term preservation seamless.

Through hands-on activities, attendees will experience the full functionality of Colectica Datasets. This workshop is ideal for researchers, data managers, and anyone looking to enhance their data management skills.

Documenting Variables / **36**

Guidelines for handling variables in repeated contexts

Authors: Hayley Mills[None]; Romain Tailhurat[1]

[1] *Making Sense*

Corresponding Author: romain.tailhurat@making-sense.info

DDI standards offer an extremely valuable solution to metadata management throughout the whole data acquisition, processing and dissemination phases. However, when documenting variables in repeated context - one of the most fundamental entities of any data process - challenges may arise to find the ideal way of using DDI. There are several contexts in which repeated variables are created including longitudinal surveys and question reuse over different studies, which have been captured by use cases.

An informal working group has been working on the subject since 2023 and presented some first results at the EDDI23 conference. This presentation will provide an update since then and will include an outline of guidelines driven by real life use cases, as well as pending issues, and ways to help contribute to this topic.

Sustainable, Ethical and Economical Controlled Vocabularies / **38**

The KDK Thesaurus –sustainable thematic metadata?

Authors: Júlia Egyed-Gergely[1]; Judit Gárdos[1]; Enikő Meisztercics[1]

[1] *Hungarian Research Network Centre for Social Sciences, Research Documentation Centre (KDK)*

Corresponding Authors: egyed-gergely.julia@tk.hu, meiszterics.eniko@tk.hu, gardos.judit@tk.hu

This paper is on the development of the KDK Thesaurus, a CV partly based on ELSST, used for the topical discovery of interview materials. The presentation discusses the workload for such a project, its sustainability and future perspectives of similar projects, incl. ONTOLISST and touch on the issue of economical and ethical considerations of metadata curation on smaller levels of datasets like parts of interviews and variables. The paper also examines how social science data archives may operate in economically scarce environments and how big of a burden the costs of the introduction of elaborate metadata standards like DDI or the use of different controlled vocabularies might entail.

Machine Learning and Media / **39**

Streamlining Media File Conversion to DDI-Lifecycle ”Other Material” Using AI

Author: Benjamin Beuster[1]

[1] *Sikt - Norwegian agency for shared services in education and research*

Corresponding Author: benjamin.beuster@sikt.no

Social research increasingly includes media formats like audio and video, which are often poorly documented and inaccessible. While archives handle traditional survey data well, media files are mostly limited to minimally annotated zip files due to the complexity of proper documentation. Recent advancements in AI, including the Whisper model, along with the use of Pydantic models and structured output, now allow for rapid metadata extraction, transcription, and reliable, structured summaries.

This presentation demonstrates how AI can streamline the documentation and conversion of media files into the DDI-Lifecycle “Other Material” format, significantly improving accessibility and usability for researchers and data archives.

Poster Session / **40**

DDI-CDI Converter Prototype: Generating Wide Tables for Stata & SPSS

Author: Benjamin Beuster[1]

[1] *Sikt - Norwegian agency for shared services in education and research*

Corresponding Author: benjamin.beuster@sikt.no

The DDI-CDI Converter Prototype is a Python-based web application designed to convert proprietary statistical files from Stata and SPSS into the open DDI-CDI format. This tool addresses the growing need for data interoperability and sharing by transforming closed data formats into a standardized, machine-readable structure. By converting both data and metadata from Stata and SPSS, including variable information such as types, labels, and missing values, the tool generates a comprehensive DDI-CDI XML file. The prototype also offers a detailed overview of the 25 DDI-CDI model classes used, linking to online documentation to support practical implementation and training within the DDI community.

Poster Session / 41

Metadata training as a foundation for DDI implementation

Authors: Becky Oldroyd[1]; Jon Johnson[1]; Hayley Mills[1]

[1] *CLOSER, UCL*

Corresponding Author: r.oldroyd@ucl.ac.uk

CLOSER is the interdisciplinary partnership of leading social and biomedical longitudinal population studies (LPS), the UK Data Service and The British Library. One of our areas of focus is training and capacity building. We currently offer free, online educational resources on our Learning Hub on a range of LPS-related topics, including Understanding metadata, which provides a high-level overview of metadata and its importance.

For researchers and data managers to recognise the value of the DDI standard, and integrate DDI tools into their workflows, a foundational knowledge and understanding of metadata as a first-class citizen in the use and management of data is required.

At CLOSER, we are building on our current online training offer by developing an in-person training course which covers the basics of data, metadata, and the FAIR principles. The course includes several exercises and actionable steps to help participants implement these concepts in their work, providing a solid entry point for DDI training and implementation.

In September 2024, we piloted our training course at two conferences: MethodsCon and the CLOSER conference. In this poster we will reflect on these events, highlighting successful elements, sharing user feedback, and describing how these insights will shape our future training offerings.

Machine Learning and Media / 42

AI for Data / Data for AI: Augmentation and Improved Discoverability of DDI Metadata Using LLMs

Author: Aivin Solatorio[1]

Co-author: Olivier Dupriez [1]

[1] *The World Bank*

Corresponding Author: asolatorio@worldbank.org

Microdata provides tremendous value in socioeconomic analysis. However, these data may not be easily discoverable when metadata are not as rich, structured, and optimized as they could be. In the case of microdata, an issue is the semantic discoverability of information contained in the variable-level metadata (the data dictionary). This paper presents an unsupervised framework that leverages large language models (LLMs) to generate variable groups in DDI and thematic description of these groups automatically from microdata's data dictionary. The framework leverages natural language processing (NLP) methods to improve the context accessible to the LLM, and self-consistent prompting is proposed to automate the validation of the generated themes. The framework also implements an AI agent to assess the self-consistency of the LLM's output for automating the quality assurance (QA) process. The automatically generated thematic descriptions of the variables serve as input for lexical search, for generating embeddings for semantic searchability, and recommendations for microdata.

Interoperability 1 / 43

Developing and testing harmonisation workflows for comparative survey data using DDI –a WorldFAIR case study

Author: Steven McEachern[1]

Co-authors: Hilde Orten [2]; Kristina Strand [3]; Ryan Perry [4]

[1] *UK Data Service, University of Essex*

[2] *Sikt - Norwegian Agency for Shared Services in Education and Research*

[3] *Sikt*

[4] *Australian Data Archive*

Corresponding Author: sm24412@essex.ac.uk

DDI Lifecycle and DDI-CDI provide significant capabilities for the integration and harmonisation of content across datasets. As part of the recently completed WorldFAIR project lead by CODATA, a team from the Australian Data Archive (ADA) and Sikt lead a work package to examine ways for improvement of FAIR practices in the management of harmonised content in cross-national social surveys.

This work was completed in three stages –a review of comparative survey data management practices at Sikt and ADA; development of a human and machine-actionable workflow for harmonisation of social surveys (the Cross-Cultural Survey Harmonisation workflow –CCSH) that leverages DDI and other standards; and a proof-of-concept test of the CCSH workflows leveraging services available at ADA and Sikt through their respective Colectica registries.

Overall, the pilot demonstrated that the CCSH workflow forms a viable foundation for standardising and progressively automating the process of survey data harmonisation. However the pilot also showed that there is still a significant degree of human manual input required –and thus has more work to do to be truly FAIR. We thus provide recommendations for data managers and the Alliance as to how more integration and automation might be achieved in future.

Official Statistics / 44

New Data Documentation Initiative (DDI) Insights acquired from assessing Openness of Air Quality Data in Smart Cities

Author: Hugo Wai Leung MAK[1]

[1] *The Chinese University of Hong Kong & The Hong Kong University of Science and Technology*

Corresponding Author: hwlmak@ust.hk

Using top 50 "smart"city governments identified by Eden Strategy Institute and ONG&ONG Pte Ltd. (2018), we first establish a data analytic and statistical scoring framework to assess and investigate their existing open data policies, and review respective performance in releasing environmental and air quality attributes and information to public. The framework considers data availability, data accessibility and visualization perspectives, which explains how metadata related to air quality and meteorology could be accessed, interpreted and utilized by general public, professional experts and city officers. Some of these datasets are highly sensitive and subjected to quality control, thus local and national governments have responsibilities in achieving systematic data documentation and integration among data types, provision and management of centralized big data information system related to air quality attributes and health-related factors, and implement DDI to obtain prescribed formatting for data release and usage, at the same time protect the rights of relevant air quality data providers. All these require the engagement of community and joint efforts of government and citizens, in creating a safe and useful documentation of open scientific database for enhancing scientific innovation and steering modern smart city development forward, as a result promoting DDI in the long run.

Software 1 / 45

Metadata Editor for DDI Codebook

Author: Mehmood Asghar[1]

[1] *World Bank Group*

Corresponding Author: masghar@worldbank.org

The Metadata Editor is an open-source web application developed by the World Bank, fully compliant with the DDI Codebook standard. It supports study, data file, and variable-level elements from DDI, and for variable-level metadata, it allows importing and exporting data and data dictionaries from various versions of SPSS and Stata. The editor features a robust templating system that helps users manage metadata efficiently by allowing them to customize fields to fit specific project needs. This simplifies documentation by focusing on the necessary metadata fields, reducing clutter, and ensuring consistency across projects.

In addition to the user-friendly web interface, the Metadata Editor provides an extensive REST API enabling programmatic access and automation. The editor offers built-in support for publishing directly to NADA data catalogs, and for other DDI-compliant catalogs, the API can be utilized for seamless automation. Additional features include user management, the ability to organize projects into collections, project sharing for collaboration, and support for translations, making it a comprehensive tool for metadata management and publishing.

Interoperability.3 / 47

Why GESIS dropped the DDI-FlatDB

Author: Oliver Hopt[1]

[1] *GESIS*

Corresponding Author: oliver.hopt@gesis.org

This talk will not contain grieve about the end of software projects. Instead, we will give some insight into the reasons for switching from a homegrown development towards bying a commercial sollution, Colectica in this case.
Besides the DDI-FlatDB we also discontinued the development of our questionnaire editor, our publishing pipeline and some other tooling around DDI metadata. Therefore this talk will also cover further planning on supporting lifecycle support for end users besides Colectica Designer.
And finally we will give an overview on the features being lost by discontinueing the FlatDB ecosystem in comparisson to the features we get from Colectica. This part will contain some thoughts on occuring gaps and how to close them.

Software 2 / 48

Community driven data documentation tool - Nectar Publisher

Author: Olof Olsson[1]

Co-authors: Deirdre Lungley ; Julie Lenoir [2]; Marc Iten [3]; Oliver Hopt [4]

[1] *Swedish National Data Service (SND)*

[2] *French Institute for Demographic Studies*

[3] *FHGR - Fachhochschule Graubünden*

[4] *GESIS*

Corresponding Author: olof.olsson@snd.gu.se

In the DDI Developers group we have been discussing the need for a simple data documentation tool to do basic documentation of a dataset. The goal is to develop a simple client side only application to document datasets with the possibility of building integration with repositories to load data and metadata. During 2024, the first basic concept of this tool was developed during the DDI Hackathon and resulted in a prototype with basic functionalities. This presentation will present the current goal and approach to developing a simplistic tool and the plans for new features.
https://github.com/ddi-developers/nectar-publisher

Poster Session / **49**

Build software in the DDI Developers community

Authors: Ingo Barkow[1]; Johan Fihn Marberg[2]; Oliver Hopt[None]; Olof Olsson[3]

[1] *FHGR*

[2] *Swedish National Data Service*

[3] *Swedish National Data Service (SND)*

Corresponding Authors: johan.fihn@snd.gu.se, olof.olsson@snd.gu.se, oliver.hopt@gesis.org, ingo.barkow@fhgr.ch

This is an informal group for developers of software implementations based on DDI. The group gets together periodically to discuss their implementations of the DDI specification. Join us at our poster to discuss developing software with DDI or if you are interested in the DDI Hackathons.
https://github.com/ddi-developers/

Interoperability 2 / **50**

Researchdata.se - A portal to Swedish research data, but why not DDI?

Author: Johan Fihn Marberg[1]

Co-author: Olof Olsson [2]

[1] *Swedish National Data Service*

[2] *Swedish National Data Service (SND)*

Corresponding Authors: johan.fihn@snd.gu.se, olof.olsson@snd.gu.se

Early 2025 SND and eight other Swedish research infrastructures will launch a new joint data portal targeting researchers interested in finding Swedish research data and support pages for research data management. Although considered, DDI was not selected as a format to be used to harvest metadata from the participating research data repositories. This presentation will focus on how we work with FAIR metadata on the resource landing pages in the portal and where we see potential for improvement to make DDI more machine-readable. We will also delve into the metadata flows of the portal and how we support exports to DDI for the harvested metadata, although not used as a source for harvesting.

51

ONTOLISST kick-off

Authors: Alina Danciu[1]; Judit Gardos[2]; Mari Kleemola[3]

[1] *CDSP*

[2] *HUN-REN*

[3] *Finnish Social Science Data Archive, Tampere University*

Corresponding Authors: mari.kleemola@tuni.fi, gardos.judit@tk.hu, alina.danciu@sciencespo.fr

The new 2-year ONTOLISST project starts in December 2024. The project idea is a result of discussions in the EDDI2023 conference. ONTOLISST will develop a simplified multilingual ontology and research whether and how NLP tools can help with (semi)automated (meta)data curation. The work will build on social scientific metadata in DDI format in different languages and from varius sources. The project is funded by the first OSCARS Cascading Grant call.

52

Side meeting of the DDI-CDI WG (including the Non-Quantitative Subgroup)

Authors: Arofan Gregory[1]; Noemi Betancort[2]

[1] *CODATA*

[2] *RDC Qualiservice, University of Bremen*

Corresponding Author: ilg21@yahoo.com

This meeting will carry forward the work from the Dagstuhl Workshop in October. Topics will include syntax representation of DDI-CDI, tools, documentation, and the integration of the non-quantitative model. Participants who are not already members of the DDI-CDI WG (or the non-quantitative subgroup) are welcome, especially members of the DDI Developer's Group. The meeting will provide an opportunity for people to learn what is going on, as well as being a working meeting on the topics in hand.

User Needs 1 / 53

Automating Survey Processes: Metadata Based Improvements to the NEPS Survey Life Cycle

Author: Simon Dickopf[1]

Co-author: Daniel Bela [1]

[1] *Leibniz Institute for Educational Trajectories (LIfBi)*

Corresponding Author: simon.dickopf@lifbi.de

Beyond documentation, machine-actionable survey metadata offer a wide range of possibilities for more efficient and less error-prone survey management.
We want to illustrate how the German National Educational Panel Study (NEPS) made use of metadata throughout the survey life cycle in its newly recruited Starting Cohort 8, a panel sample of 5th graders which started in 2022. With this new cohort, the NEPS consortium decided to implement a new technical backbone to its survey infrastructure, enabling us to take advantage of the centralized

metadata storage at the Leibniz Institute for Educational Trajectories (LIfBi). We developed a fully automated process of generating survey instruments based on their reference metadata, eliminating the need for (manually produced) programming templates as well as the manual programming of the instruments itself. Putting survey's metadata in the center of that infrastructure, it furthermore accelerates the survey testing procedures and ensures the coherence of the stored metadata with the content of the instruments and, ultimately, the disseminated data products.
Future developments aim to extend these workflows beyond the scope of in-house software environments, so that NEPS' metadata can be exchanged by a DDI compliant transfer format that can be used e. g. by contracted field institutes.

Poster Session / 54

Engaging with the DDI Strategic and Scientific Work Plans

Authors: Darren Bell[1]; Hilde Orten[2]; Jared Lyle[3]; Jon Johnson[4]; Steven McEachern[5]

[1] *UK Data Service*

[2] *Sikt - Norwegian Agency for Shared Services in Education and Research*

[3] *ICPSR, University of Michigan*

[4] *CLOSER, UCL*

[5] *UK Data Service, University of Essex*

Corresponding Author: lyle@umich.edu

Earlier this year, the DDI Alliance Executive Board and the Scientific Board announced a new DDI Strategic Plan, 2024-2027 and the complementary new DDI Scientific Work Plan, 2024-2026. These documents represent the culmination of collaborative efforts and thoughtful input from our community, and they are now ready for implementation and use.

This poster will highlight the focal points of the plans, especially for this upcoming year. It will also highlight how individual community members and working groups can participate and contribute, as well as opportunities for community members to provide feedback to the boards.

Birds of a Feather / 55

Shaping the Future of DDI Together

Authors: Darren Bell[1]; Hilde Orten[2]; Jared Lyle[3]; Jon Johnson[4]; Steven McEachern[5]

[1] *UK Data Service*

[2] *Sikt - Norwegian Agency for Shared Services in Education and Research*

[3] *ICPSR, University of Michigan*

[4] *CLOSER, UCL*

[5] *UK Data Service, University of Essex*

Corresponding Authors: sm24412@essex.ac.uk, hilde.orten@sikt.no, lyle@umich.edu, jon.johnson@ucl.ac.uk, dbell@essex.ac.uk

Earlier this year, the DDI Alliance Executive Board and Scientific Board launched a new DDI Strategic Plan, 2024-2027 and the complementary new DDI Scientific Work Plan, 2024-2026. These plans reflect the collective vision and contributions of our community and are poised to drive meaningful change.

This Birds of a Feather session led by leaders of the Executive and Scientific Boards invites you to join an open, dynamic discussion on how we can collectively support and implement the priorities

of these plans —particularly those set for the coming year Whether you're part of a working group, an individual contributor, or simply passionate about DDI's future, this is your chance to share ideas, explore opportunities for collaboration, and help shape the path ahead.

Birds of a Feather / 56

Questionnaires with DDI-Lifecycle: community feedback for future improvements

Authors: Hayley Mills[None]; Romain Tailhurat[1]

[1] *Making Sense*

Corresponding Author: romain.tailhurat@making-sense.info

Do you currently, or do you plan to document questionnaires with DDI-Lifecycle? The DDI Questions and Questionnaire Working Group invites you to participate in this birds of a feather to discuss how the DDI Alliance can make use of the standard easier for you.

We will discuss what you wish DDI-Lifecycle could do, and what aspects are most challenging when documenting questionnaires. The outputs will be used to; 1) improve the DDI standard 2) provide insights into where improved guidance is needed 3) feed into the DDI Developers Working Group for potential new light weight tooling.

57

DDI Alliance Questionnaires Working Group

Authors: Hayley Mills[None]; Romain Tailhurat[1]

[1] *Making Sense*

Corresponding Author: romain.tailhurat@making-sense.info

Do you document questionnaires with DDI-Lifecycle? The DDI Questions and Questionnaire Working Group invites you to participate in a workshop to discuss how the DDI Alliance can make use of the standard easier for you.

The session will include discussions on what you wish DDI-Lifecycle could do, and what aspects are most challenging when documenting questionnaires. The outputs will be used to; 1) improve the DDI standard 2) provide insights into where improved guidance is needed 3) feed into the DDI Developers Working Group for potential new light weight tooling.

The intended audience is anyone currently or is planning on working with DDI-Lifecycle and questionnaires. There are no prerequisites.

Poster Session / 58

Populating a DDI Codebook using R

Author: Adrian Dusa[1]

[1] *University of Bucharest*

Corresponding Author: dusa.adrian@unibuc.ro

As more and more data producers and research institutions are interested in documenting their data, following the often mandatory Data Management Plan, they find themselves in need for ready to use software tools that are capable of creating and modifying a Codebook, in the vein of Nesstar Publisher.
Until such a software will appear, it is possible to fully populate a DDI Codebook using the R package DDIwR, which is capable of reading, writing, and updating a DDI Codebook versions 2.5 and 2.6.
This presentation will demonstrate how to use the command line, interactively or using an automated script, to populate the verious sections of the DDI Codebook.

59

Populating a DDI Codebook using R

Author: Adrian Dusa[1]

[1] *University of Bucharest*

Corresponding Author: dusa.adrian@unibuc.ro

As more and more data producers and research institutions are interested in documenting their data, following the often mandatory Data Management Plan, they find themselves in need for ready to use software tools that are capable of creating and modifying a Codebook, in the vein of Nesstar Publisher.
Until such a software will appear, it is possible to fully populate a DDI Codebook using the R package DDIwR, which is capable of reading, writing, and updating a DDI Codebook versions 2.5 and 2.6.
This presentation will demonstrate how to use the command line, interactively or using an automated script, to populate the verious sections of the DDI Codebook.

Sustainable, Ethical and Economical Controlled Vocabularies / 60

The ONTOLISST project on DDI metadata, vocabularies and NLP

Authors: Alina DANCIU[1]; Judit Gárdos[2]; Mari Kleemola[3]

[1] *Sciences Po, Center for Socio-Political Data (CDSP)*

[2] *Hungarian Research Network Centre for Social Sciences, Research Documentation Centre (KDK)*

[3] *Finnish Social Science Data Archive, Tampere University*

Corresponding Authors: mari.kleemola@tuni.fi, gardos.judit@tk.hu, alina.danciu@sciencespo.fr

The talk introduces the new 2-year ONTOLISST project starting in December 2024, funded by the first OSCARS Cascading grant call. The project will develop a simplified multilingual ontology (LiSST) to describe social science research data, create a corpus of social science metadata, and research whether and how NLP tools can help with (semi)automated (meta)data curation. The aim is to better understand how social science archives assign thematic metadata to their datasets in order to describe their contents and how data curation practices shape social scientific understanding. ONTOLISST will build on metadata in DDI format in different languages from various sources and using different CVs. The presentation outlines the project tasks, expected outputs and relationships with existing standards and tools. It also discusses how AI could help to accelerate the tedious, resource-intensive but important work of metadata and data curation and improve (meta)data interoperability across languages and disciplinary barriers.

Keynote / 61

Navigating Political and Practical Barriers to Open Science in Europe and Switzerland

Author: Georg Lutz[1]

[1] *FORS, University of Lausanne*

Corresponding Author: georg.lutz@fors.unil.ch

Many stakeholders in science policy have demonstrated a strong commitment to open science over the past decade. However, this commitment contrasts with the reality, where a comprehensive and effective open science framework is still lacking. This talk focuses on political and practical challenges faced by open science in Europe, and how they are (not) met in Switzerland.

While the benefits of open science are widely acknowledged, implementation is hampered by varying national policies, coordination gaps, and limited integration of open science into research assessments. Realizing open science and open data requires a coherent and efficient interplay of platforms, services, standards, and policies. Increasingly, discussions also highlight concerns over data protection and national interests, which may limit openness and require striking a balance between openness and limiting access.

Furthermore, disparities in funding and resources across institutions lead to uneven adoption of open science. Scientific developments are often based on bottom-up initiatives and mechanisms to steer such bottom-up prioritization and funding are well developed in many countries and at the European level. For putting a conclusive open science framework in place, a bottom-up approach is not sufficient, top-down steering is also required. However, such top-down steering mechanisms and even more harmonizing such steering with allowing for community driven initiatives are even less established.

Keynote / 62

Enhancing FAIR Principles via Metadata: The Dublin Core Journey

Author: Sam Oh[1]

[1] *Executive Director, Dublin Core Metadata Initiative (DCMI)*

Corresponding Author: samoh@g.skku.edu

The FAIR principles—Findability, Accessibility, Interoperability, and Reusability—are essential for maximizing the value of resources in today's data-driven world. Metadata serves as the cornerstone in achieving these principles, providing structure and meaning to data.

This keynote will trace the evolution of metadata, from its origins in ancient civilizations using clay tablets to its critical role in contemporary large-scale AI systems. It will highlight the Dublin Core's contributions to advancing FAIR principles, demonstrating its reliability as a metadata standard.

The talk will also introduce the latest initiative from DCMI: the Dublin Core Tabular Application Profiles (DCTAP). Through practical examples, it will showcase how DCTAP extends beyond data integration and interoperability to align with the goals of the Data Documentation Initiative (DDI), offering a forward-thinking approach to metadata and resource management.

63

City Tour

Meeting point: In the hall of the Chur main train station; in front of the "Flying Tiger"Store (https://www.sbb.ch/de/reiseinforma
finden/bahnhof-chur/geschaefte/shop-detail.html/geo-flying-tiger-copenhagen-eff9). Participants will
be guided to the conference-opening after the tour.

64

Registration