15th European DDI Users Conference, Ljubljana

Monday, 27 November 2023 - Wednesday, 29 November 2023

Hotel Slon

Book of Proposals

Contents

Official Statistics / 4

Empowering society by reusing privately held data for official statistics - A European approach

Authors: Geta Mitrea[1]; Florent Diverchy[2]

[1] *University 'Stefan cel Mare' of Suceava, Romania*

[2] *Lecturer European Communication School in Brussels, Belgium & Marketing Intelligence Manager, Produpress, Belgium*

Corresponding Authors: mitrea.geta@gmail.com, florent.diverchy@mediaschool.education

The present paper is focused on the work of an expert group on 'Facilitating the use of new data sources for official statistics' created by Eurostat during March 2021 – May 2022. The main purpose of this expert group was to reflect on new opportunities riced once with our rapidly changing and increasingly data-driven society and make recommendations to enhance the reuse of private sector data in official statistics under the European strategy for data.

In this context the Data Documentation Initiative (DDI) standards could play their relevant role to reuse of data, transparency, data sharing respecting the legal framework and the principles described below.
The final report highlights the regulatory gaps, fragmentation of practices, and lack of clarity regarding businesses' rights and obligations. Voluntary partnerships between businesses and statistical authorities have recently been set up and rolled out, but they have rarely led to sustainable data reuse in the regular production of official statistics. To obtain the maximum benefit for society, the Expert Group proposes to prioritize action in four areas.

First, fair and effective partnerships between businesses and statistical authorities must be promoted on a systematic and regular basis. Such partnerships have to be based on a mutual recognition that it is legitimate for the different parties to have different roles and interests, based on trust, social responsibility.

Second, statistical authorities and private data holders should develop a partnership approach to maximize business incentives and minimize risk, based on mutually agreed operational modalities of data reuse.

Third, the legal framework should set out a clear set of requirements and safeguards for private data holders. Where it is necessary to collect data from private data holders, the data protection framework must be respected, and data subjects' rights must be ensured. The legal framework should also clarify businesses' rights and obligations regarding data sharing and reuse for statistical purposes, including issues related to ownership, intellectual property rights, and commercial confidentiality.

Fourth, the regulatory framework should enable statistical authorities to access privately held data for statistical purposes while ensuring the data protection framework is respected, data subjects' rights are ensured, and businesses' interests are protected. The report proposes a model based on the principles of transparency, accountability, proportionality, and risk management.

The expert group and the report have contributed in clarifying several issues related to the use of new data sources for official statistics. And, was used for the Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL amending Regulation (EC) No 223/2009 on European statistics. The proposal will be further discussed in the European Parliament and the Council following its journey in the EU legislative process.

Harmonisation / 6

Cross-Country Collaboration Efforts with Common Conceptual Groups - NACDA and CLOSER

Authors: Kathryn Lavender[1]; Hayley Mills[2]

Co-authors: Jon Johnson [3]; Jennifer Zeiger [1]; Sanda Ionescu [1]

[1] *ICPSR, NACDA*

[2] *CLOSER*

[3] *CLOSER, UCL*

Corresponding Authors: h.mills@ucl.ac.uk, kfrania@umich.edu

The National Archive of Computerized Data on Aging (NACDA) and CLOSER - the interdisciplinary partnership of leading UK social and biomedical longitudinal population studies (LPS), have been discussing ways to collaborate and create common conceptual groups across their social science data collections in their Colectica Portals.

Although NACDA and CLOSER have approached metadata organization efforts in different ways, and present the metadata differently in their portals, the use of common DDI-Lifecycle metadata structures simplifies the possibilities for collaboration, in particular the mapping of overlapping concepts which have been identified across both LPS data collections.

We plan to discuss how we have managed our collaborations thus far and include highlights from an in-person workshop we held together earlier this year. Further, we will explain some of the details of our respective approaches to creating conceptual variable groups and why these are important to our projects and to future reuse.

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. CLOSER is based at the University College London (UCL) Social Research Institute (SRI).

DDI & Other Standards / 7

Getting running with ARK persistable identifiers

Author: John Kunze[1]

Co-author: Donny Winston

[1] *ARK Alliance and Ronin Institute*

Corresponding Authors: donny@polyneme.xyz, jakkbl@gmail.com

Any DDI dataset, ancillary, or supporting file is a candidate for systematic, persistent identification with ARK identifiers. End users, especially researchers, rely on ARKs for long term access to the global scientific and cultural record. Since 2001 some 8.2 billion ARKs have been created by over 1100 organizations — libraries, data centers, archives, museums, publishers, government agencies, and vendors.

They are open, mainstream, non-paywalled, decentralized persistent identifiers that you can start creating in under 48 hours. They identify anything digital, physical, or abstract, and from any domain or discipline. This has been true since 2001; in contrast, the well-known service for dataset DOIs, DataCite, only incorporated in 2009. ARKs are similar to DOIs, URNs, and Handles. All of them:

- were introduced over 20 years ago,
- exist in large numbers (8.2 billion ARKs, 240 million DOIs, etc.),
- start with a string to identify the name assigning authority,
- require the active updating of URL redirects, and

- support research and scholarship, appearing in the Data Citation Index, Wikipedia, ORCiD.org profiles, etc.

In contrast, ARKs are cheaper, more flexible, and less centralized, letting you:

- create unlimited identifiers without paying for the right to do so,
- add any kind of metadata, including no metadata,
- append extensions and query strings during resolution,
- link directly to an article, image, or spreadsheet that is immediately usable by people and software without making them first stop at a landing page, and
- make millions of ARKs resolvable by managing just one ARK, via a mechanism called suffix passthrough.

As a low cost alternative to the DOI (for example), ARK adoption is rapidly accelerating in the global South and in any organization that cannot pay for large numbers. ARK organizations include 10 national libraries, 184 archives, 75 journals, and 145 universities.

We will cover:

- Why ARKs – non-paywalled, decentralized, flexible
- Use cases – Smithsonian, French National Library, Internet Archive
- Metadata for early and ongoing object development
- How to get started – fill out one form
- Minting and assigning ARK identifiers
- Resolvers, resolution, redirection Persistence considerations

Questionnaires / **9**

DDI 2.5 to EQB metadata transformation

Authors: Gregor Zibert[1]; Janez Stebe[1]

[1] *UL-FDV-ADP*

Corresponding Authors: janez.stebe@fdv.uni-lj.si, gregor.zibert@fdv.uni-lj.si

The Slovenian Social Science Data Archives (ADP) have developed an XSLT (eXtensible Stylesheet Language Transformations) stylesheet to align the data, variable, and question descriptions within a selection of current ADP DDI 2.5 codebook instances with the established CESSDA European Question Bank (EQB) DDI 2.5 harvesting profile. This effort was undertaken as part of the Social Sciences & Humanities Open Cloud (SSHOC) project.

This paper offers a concise overview of the transformation process, including insights into our local development environment setup, the systematic approach taken to create the XSLT transformation file , the resultant EQB DDI 2.5 XML files and potential avenues for future enhancements. Emphasis is on extracting information from existing variable description, and transforming and adding the missing default ingredients in order to pass the EQB profile validation. The functional XSLT can serve as a model for others aiming to adapt their local DDI 2.5 metadata to the common EQB profile, thus opening the metadata to further transformations including DDI3 family. The paper also discusses the challenges and potential pitfalls that were encountered during the preparation of the XSLT stylesheet.

ADP is presently engaged in the task of converting its existing metadata to DDI 2.5. Once this transition is complete, we will be equipped to convert the metadata into the EQB DDI 2.5 profile, thus enabling its accessibility for harvesting through the OAI-PMH endpoint.

DDI Standards / 10

State of the DDI Cloud

Authors: Knut Wenzig[1]; Xiaoyao Han[2]

[1] *DIW Berlin*

[2] *DIW Berlin / SOEP*

Corresponding Authors: kwenzig@diw.de, xhan@diw.de

An investigation was conducted to examine the extent to which metadata in different Data Documentation Initiative (DDI) standards is openly available and which elements of these standards are used. DDI is a set of international standards for describing and documenting data used in social, behavioural, economic, and health sciences research.

To identify the online repositories, where DDI metadata is available, re3data.org (a global registry of research data repositories that covers research repositories from different academic disciplines) and an enquiry on the DDI-users mailing list were used. We compare this with findings from 2017 and could add new repositories after a first presentation of this project in July 2023.

Then we tried to access and analyse the metadata, e.g. by using a standardised protocol like the Open Archive Initiative-Protocol for Metadata Harvesting (OAI-PMH). This makes it possible to show which elements are more commonly used than others.

The findings have implications for deploying DDI metadata and the further development of the standards. They could also inform users like researchers and data stewards, how the standards are used by the community. Overall, the investigation highlights the value of openly available metadata in supporting research to achieve the goals of the FAIR data movement.

Software / 11

Enhancing DDI Support in the Open Source Dataverse Repository Software

Author: Victoria Lubitch[1]

[1] *Scholars Portal*

Corresponding Author: victoria.lubitch@utoronto.ca

Recognizing the significance of DDI for use in data archives and research data management, this presentation aims to introduce the Dataverse Data Curation Tool, an open and integrated DDI application that supports editing dataset variable metadata in Dataverse. An overview of the current challenges and limitations of integration of DDI into Dataverse, including issues related to metadata transfer, compliance, reuse, and user accessibility will be discussed. Furthermore, an overview of proposed enhancements and recent developments in Dataverse related to DDI support, including expanded support for variable metadata and the Data Curation Tool, highlights new features, tools, and best practices for leveraging DDI in Dataverse. This presentation seeks to demonstrate the Data Curation Tool and DDI support within the Dataverse repository software for the DDI community as a whole.

An example of the Data Curation Tool used for a specific dataset.

Harmonisation / 12

Creating Longitudinal Data Series Comparisons - New Highlights from the NACDA Portal

Authors: Kathryn Lavender[1]; Jennifer Zeiger[1]

Co-author: Sanda Ionescu [1]

[1] *ICPSR, NACDA*

Corresponding Authors: zeiger@umich.edu, kfrania@umich.edu

In this presentation, we plan to describe NACDA's efforts over the last year to develop a multi-series comparison of longitudinal, nationally representative, National Institute on Aging funded data collections using DDI-Lifecycle, as well as share some lessons learned.

The National Archive of Computerized Data on Aging (NACDA, part of ICPSR) began working with DDI-Lifecycle in 2018. Since then, NACDA has made efforts to document some of our most established and frequently-used longitudinal data collections to DDI-L and display them in a Colectica Portal. In this presentation, we will discuss how our portal and the way we use it have evolved. Namely, we have built on our earliest efforts in comparing variables within a single series and across two series, to develop new methods that include the recent addition of a multi-series concordance to the portal.

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.

DDI & Other Standards / 13

IZRK Metadata Portal: Cataloguing the multidisciplinarity of karst (meta)data

Authors: Magdalena Năpăruş-Aljancic[1]; Žan Kafol[2]; Tanja Pipan[3]; Stanka Šebela[3]; Jasmina Čeligoj Biščak[3]; Nataša Ravbar[3]

[1] *Karst research Institute ZRC SAZU*

[2] *Karst Research Institute ZRC SAZU; KAFOL.NET,*

[3] *Karst Research Institute ZRC SAZU*

Corresponding Authors: zan@kafol.net, magdalena.aljancic@zrc-sazu.si

Karstology is a multidisciplinary science that encompasses a wide range of earth and life sciences: hydrology, geology, biology, geomorphology, ecology, microbiology, speleology, and history of karst science. The Karst Research Institute ZRC SAZU reinforced its long tradition of excellent research by becoming the national coordinator and headquarters of three major environmental European RIs in Slovenia: LifeWatch ERIC, EPOS ERIC and eLTER RI. Karst Research Institute Data Centre recently established the IZRK metadata portal (https://metadata.izrk.zrc-sazu.si).

Built on the GeoNetwork platform, the portal adheres to FAIR data principles, collecting and mapping various datasets, databases, research sites, equipment, virtual labs, models, and codes. GeoNetwork core standards do not appear to be interoperable with DDI standards. Due to the open source nature and broader use of such platforms in various research domains, researchers are constantly

confronted with these issues - which platform to use, which standards, so that they are able to satisfy multiple aggregation catalogues at European or global scale.

Here we present our practical approach to publishing metadata using different standards on one platform. We also propose theoretical and practical solutions to be platform independent and/or provide ad hoc conversion web services to transform and export metadata to other standards including DDI.

Questionnaires / 14

Semi-automating questionnaire metadata entry for increased job satisfaction

Authors: Becky Oldroyd[1]; Hayley Mills[1]; Jenny Li[1]

[1] *CLOSER / UCL*

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

CLOSER Discovery is the UK's most comprehensive research tool for longitudinal population studies, containing questionnaire and dataset metadata for 11 leading UK studies.

Creating questionnaire metadata can be a time-consuming and challenging task. Historically, CLOSER's Metadata Assistants (MAs) entered the questionnaire metadata into our in-house developed DDI questionnaire editor – Archivist – by manually entering them into the tool.

CLOSER are committed to creating enriching and fulfilling jobs, particularly for those who create the content that enables CLOSER Discovery to be an evolving and valuable resource. Subsequently, CLOSER's MA role has advanced from manual metadata *entry* to semi-automated metadata *editing* using GitLab parsers.

Gitlab is freely available for educational institutes and open-source software projects, and allows the automation of tasks through a simple interface.

These parsers use the available structured information from the studies (e.g., PDF, XML) so that questionnaire metadata can be loaded into Archivist, and then checked and edited. Consequently, our workflow is more efficient with reduced human error and, importantly, the MA role is more fulfilling and allows staff to focus on the aspects that are most engaging and creative.

This presentation will provide an overview of CLOSER's GitLab parsers, and explain how they have advanced CLOSER's MA role.

Software / 15

Implementing Colectica at the GESIS Data Archive

Author: Wolfgang Zenk-Möltgen[1]

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

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

Data collections at GESIS - Leibniz Institute for the Social Sciences are currently managed with a common metadata database to support the search and re-use of research data. This involves a variety of tools, e.g. for study-level documentation, PID registration, data file management, and question and variable documentation. Colectica is currently added to the available metadata management tools to improve the effectiveness and stability of procedures. This process holds both challenges and opportunities for the metadata workflow.

Current tools support the DDI-Codebook and DDI-Lifecycle standard, depending on publication needs. Given the existing metadata production and management ecosystem, the presentation will highlight the benefits and disadvantages of the new way of working with Colectica. Other presentation areas will include technical challenges, improvements in using DDI-Lifecycle features, and better search functions for research data at the GESIS website.

DDI & Other Standards / 16

Bridging Portals and Formats: Transforming Metadata from DDI-C 2.5 to OmicsDI in BY-COVID Project

Author: Markus Tuominen[1]

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

Corresponding Author: markus.tuominen@tuni.fi

One of the primary objectives of the BeYond-COVID project was the integration of social sciences & humanities studies' metadata into the COVID-19 Data Portal. This presentation focuses on the technical work done to achieve this for studies found in CESSDA Data Catalogue (CDC).

Focus of the presentation is the implementation of XML transformation from DDI-Codebook 2.5 to extended OmicsDI format using an XSL template (XSLT). Other parts of the overall process are also covered. XSLT transformations are ran with SaxonC which fully supports XSLT 2.0 and even 3.0, version 2.0 being the one used in this project. Other necessary steps consist of things like harvesting metadata from CDC's OAI-PMH endpoint, filtering and XML validation.

All steps are combined into one fully automated pipeline that creates an XML file with filtered subset of metadata in OmicsDI format for the COVID-19 Data Portal to harvest and use. First version has been in production since September 2022. However, improvements will be made until the end of the project in September 2024.

The presentation also covers the encountered challenges, lessons learned and what can still be improved.

DDI-CDI / 17

DDI-CDI: Current Status and Further Developments

Author: Arofan Gregory[1]

[1] *CODATA*

Corresponding Author: ilg21@yahoo.com

This presentation will summarize the capabilities of the DDI Coss-Domain Interoperability specification, and describe how early imple,entations have used the standard, both on its own and in combination with other DDI specifications. The focus of the working group in the near- to mid-term will be described, as will prospects for adoption through various FAIR initiatives. The range of available tools for working with the specification will be covered, as well as expected and supported implementation approaches in RDF as well as XML, including alignment with non-DDI standards such as SKOS, Schema.org, and DCAT.

Tutorial / 18

Applying the DDI Specifications to Organisational Challenges: An Introduction to the Standards

Authors: Arofan Gregory[1]; Maja Dolinar[2]; Adrian Dusa[3]; Christophe Dzikowski[4]

[1] *CODATA*

[2] *Slovenian Social Science Data Archives (ADP), Faculty of Social Sciences, University of Ljubljana*

[3] *University of Bucharest*

[4] *INSEE*

Corresponding Author: ilg21@yahoo.com

The DDI specifications cover a wide range of needs in data production, management, reuse, and dissemination. Deciding on which specification is best depends on the organisational challenges faced, and specific requirements. This tutorial is for an audience which is unfamiliar with the DDI standards. It introduces the different specifications, and considers the uses to which standard, machine-actionable metadata and associated tools can be put. It is not a technical introduction, but provides information about how the standards support data production, management, archiving, and dissemination.

I. Introduction: brief background and the evolution of DDI specifications in light of changing user needs and technology.

II. Data archiving: supporting secondary use of data with DDI Codebook.

III. Data production: how metadata management with DDI Lifecycle improves efficiency in collection and management, and enhances data quality/usability. Looks at description of questionnaires, multi-wave studies, data comparison, and metadata reuse.

IV. Widespread data reuse: implementation of the FAIR data principles with the DDI standards. Covers both domain-specific FAIR as well as cross-domain FAIR with DDI-CDI.

In each case, live examples/use cases and exercises will illustrate the topics, and provide hands-on interactions by participants.

This will be a half-day, in-person training event.

Software / 19

What's New in Colectica

Author: Jeremy Iverson[1]

[1] *Colectica*

Corresponding Author: jeremy@colectica.com

Colectica is delighted to launch Colectica 7.3 at EDDI. Colectica is software for creating, publishing, centralizing, and managing DDI metadata within and across organizations. It is used by national statistical organizations, university research groups, and data collection agencies to provide well-documented data to researchers and the public. Colectica is built on open standards like DDI and GSIM, ensuring that information can be presented in numerous formats and shared among different organizations and tools.

In this session we will give an overview of new features in Colectica 7.2 and 7.3, including:

- Colectica Repository: advanced search capabilities, health checks
- Web Portal: Improved concordance views, item comparison

- Colectica Workflow: new user interface, multiple state transitions to support ISO 11179 workflows
- Elasticsearch integration: additional configuration and data type support
- Improved DDI integration with Blaise, including direct use of the Blaise API for survey imports
- Expanded DDI support
- Hundreds of other enhancements, performance improvements, and fixes

We will also discuss forthcoming new products, including Colectica Datasets and web editors for DDI content.

Harmonisation / **20**

New Project and Tools To Aid In DDI-based Variable Concordance and Harmonization

Author: Dan Smith[1]

[1] *Colectica*

Corresponding Author: dan@colectica.com

The current research data environment provides many opportunities for linking similar topical datasets and harmonizing extant common variables. The DDI Lifecycle standard supports documenting these linkages, but few software tools are available to facilitate the actual performance of this resource-intensive task. This project uses a DDI based framework to assemble richly-described datasets that are mapped against DDI represented and conceptual variables to identify equivalent concepts and variables. The tools use machine learning and advanced text analysis algorithms to guide the creation of concorded databases (variable crosswalks) that support harmonization and discoverability, both within and across statistical datasets and studies. Specifically, the tools use several human-in-the-loop algorithms to operate as a "recommendation engine" to guide the concordance of potentially equivalent or similar variables among multiple datasets. The goal of this project is to significantly decrease the labor, time, and resources required to create accurate and standardized concorded databases and store the results using the DDI standard.

Questionnaires / **21**

DDI-L, a keystone for panel data harmonization

Authors: Malaury Lemaître-Salmon[1]; Lucie Marie[1]; Alexia Ricard[1]

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

Corresponding Author: lucie.marie2@sciencespo.fr

Panel survey data is often repeated to allow comparisons over time. However, questionnaires may be slightly adjusted over data collection waves - and therefore the datasets variables. Impacting the comparability, data harmonization may be required to maintain the panel data "mission".

Based on the ELIPSS panel use case, this talk will show opportunities of a centralized metadata management system project using DDI-Lifecyle.

We aim to highlight the value of collaborative work for creating a question bank and its usefulness for both input harmonization (at the conception stage) and post-harmonization of existing data (for a secondary user).

Official Statistics / 23

Historical Data Conversion and Archiving

Author: Chandra Shekhar Roy[1]

Co-author: Alamgir Hossain [2]

[1] *Bangladesh Bureau of Statistics.*

[2] *Labcom Technology*

Corresponding Author: csroy1@gmail.com

Making historical analog data Re-useable:
a successful outcome of Data Rescue/Conservation Discipline at Bangladesh Bureau of Statistics.
Short Title: Historical Data Conversion and Archiving.

Chandra Shekhar Roy1
Alamgir Hossain2
1Senior Maintenance Engineer-IT, Bangladesh Bureau of Statistics, Statistics & Informatics Division, Ministry of Planning, E27/A, Agargaon, Dhaka-1207, Bangladesh.
2Computer Lab Manager, Labcom Technology, House 17/1D, Road 28, Dhaka-1209,
Corresponding author E-mail: csroy1@gmail.com

Abstract: Bangladesh starts its journey focusing on ICT with a view to realizing the vision of 'Smart Bangladesh 2041'. Thus, in the long run, census and survey data can be used exhaustively in the planning process to transform into Smart Bangladesh by 2041. In Bangladesh, several types of official data are released under the Statistics Act. Bangladesh Bureau of Statistics (BBS) played a significant role in the field of historical Statistical data preservation. After the independence of Bangladesh in 1971, there was a rich repository of statistical microdata in IBM 360 to ES/9000 model mainframe tapes dating back to the late 1970s and early 2000s. Almost 8600 nine-track ½ inch spool tapes were used to preserve those data. Recently BBS has converted all those data from EBCDIC format to ASCII format. About 165 data sets have been recovered which have been declared as Digital Assets. The overarching objective is to strengthen the prevailing national statistical archiving system. BBS will be making available this large volume of converted data to the citizens of Bangladesh as well as globally so that academic and scholarly debates can take place taking cognizance of historical data. Since independence, 2,391 BBS surveys and census publications have been digitized and converted to e-book systems. The overarching objective is to strengthen the prevailing national statistical archiving system. By revisiting time series data, it is hoped that well-informed and meticulous policies can be designed and formulated in the future. Most fundamentally, the availability and easy accessibility to such a large volume of Big Data will inspire reassessing economic theories and indicators of development informing Bangladesh's position in global rankings like Sustainable Development Goals (SDG). An alternate backup has been established for data Preservation at a distance of 200 km from NSO headquarters.

Keywords. Historical data, Data preservation, Microdata, Statistics Act, SDG. IBM, Archiving.

24

DDI Trainining Group side meeting

Authors: Alina DANCIU[1]; Hayley Mills[2]; Kathryn Lavender [3]

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

[2] *CLOSER*

[3] *ICPSR*

Corresponding Authors: alina.danciu@sciencespo.fr, kfrania@umich.edu

The DDI Training Group (TG) expects to have several members attending the EDDI 2023 conference in person. With this in mind we plan to discuss notable EDDI presentations, debrief on the training workshop and discuss training plans for 2024, and possible target groups (e.g. researchers). We are considering opening the meeting to non-members of the training group.
This will be a half-day meeting.

Workshop / **25**

Metadata Uplift and Machine Learning - European Perspectives

Authors: Alina Danciu[1]; Jon Johnson[2]; András Micsik[3]; Christophe Dzikowski[4]; Claus-Peter Klas[5]; Knut Wenzig[6]; Judit Gárdos[7]

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

[2] *CLOSER, UCL*

[3] *MTA SZTAKI*

[4] *INSEE*

[5] *GESIS*

[6] *DIW Berlin*

[7] *Centre for Social Sciences – MTA Centre for Excellence*

Corresponding Author: jon.johnson@ucl.ac.uk

The background for this workshop are recent calls for proposals where participants had limited understanding of the scope of metadata holdings, local computer science expertise and knowledge of current work being done at institutions to be able to develop a focused and convincing proposal for funding.
The development of the European Question Bank, and the European Language Social Science Thesaurus (ELSST), in addition to other national and cross national metadata resources could be utilised to develop methods and training data in a collaborative way that support open science.
The purpose of this workshop is to bring together metadata content providers and infrastructures that are interested in co-ordinating such efforts to uplift existing and future metadata holdings with the aim to put together future research proposals.
Expected topics for the workshop will be, available metadata, in DDI (or otherwise), current and future plans for ML or AI activities, development of cross-language training datasets, and the types of collaborative project which are possible.

Variable Cascade / **26**

Managing repeated representation of variables in DDI Lifecycle

Authors: Christophe Dzikowski[1]; Hayley Mills[2]; Jon Johnson[3]; Frank Cotton[1]; Sophiane Kab[4]

[1] *INSEE*

[2] *CLOSER*

[3] *CLOSER, UCL*

[4] *INSERM*

Corresponding Authors: jon.johnson@ucl.ac.uk, christophe.dzikowski@insee.fr, sofiane.kab@inserm.fr

DDI-Lifecycle utilises the variable cascade to organise and describe data from conception to collection.

The organisation of conceptual variables and conceptual variable groups allows comparison of data at different time points, universes, representations and many other dimensions through concordance tables and is well suited to iterative data collected as panels, cohorts, repeated surveys and periodic administrative files.
Within a single study design, such comparisons are straight-forward to describe and comprehend. Whilst it is possible to extend the comparison between studies or from other sources such as administrative data, by creating groups of conceptual variables, conveying to the user (human or machine) the data collection dimensions will likely require the alignment of this metadata also.
This is potentially a significant challenge for multi-study metadata infrastructures or national statistical agencies which hold data from diverse sources.
The presentation will discuss the outputs from a workshop organised between CLOSER, INSEE and Constance that seeks to address these issues

Posters / 27

Python Library for Colectica Portal API

Authors: Jenny Li[1]; Jon Johnson[1]

[1] *CLOSER, UCL*

Corresponding Author: jenny.li@ucl.ac.uk

through display as web pages and as downloads in a number of formats.
The API is documented in the swagger documentation that is supplied with Colectica Portal, but for those unfamiliar with either API programming or DDI-Lifecycle, using the API can be a significant barrier.
The colectica-api wrapper allows a user with little experience of using APIs or DDI-Lifecycle a straight-forward entry point to those with python programming experience to access and discovery DDI-LIfecycle metadata. The colectica-api wrapper is open sourced and available as a pip install package for ease of use, and provides example code to get a python developer up and running. The wrapper takes advantage of the introduction of JSON format in the recent versions of Colectica which is preferred over the previous XML format.

28

DDI Alliance Technical Committee

Authors: Jon Johnson[1]; Wendy Thomas[2]

[1] *CLOSER, UCL*

[2] *Minnesota Population Center*

Corresponding Authors: wlt@umn.edu, jon.johnson@ucl.ac.uk

The Technical Committee meet face-to-face on a yearly basis to review progress and plan work over the year, in addition to resolving issues which need focused discussion which is more suited to a face-to-face environments

FAIR / 29

FAIRs not fair for metadata aggregators?

Authors: John Shepherdson[1]; Joshua Tetteh Ocansey[1]; Kostas Papagiannopoulos[2]; Matthew Morris[1]

[1] *CESSDA ERIC*

[2] *National Centre for Social Research (EKKE)*

Corresponding Authors: john.shepherdson@cessda.eu, joshua.ocansey@cessda.eu, matthew.morris@cessda.eu, papagiannopoulos.konstantinos@outlook.com

The domain agnostic metrics adopted by FAIR data assessment tools tend to penalise metadata aggregators, like the CESSDA Data Catalogue (CDC). This became apparent during the work done for the 'Bulk FAIR assessment of the CESSDA Data Catalogue using the F-UJI API' (as presented at EDDI2022).

Building on that work, FAIR scores were generated by the F-UJI and FAIR EVA tools for some of the records held in the CDC and the same records at source. Then they were analysed to determine the relative differences and underlying causes on a criteria by criteria basis.

In order to do this, the script previously used to run the bulk assessments was modified to allow it to run different tools against different sources of metadata. The results were stored as ElasticSearch indices, and a Kibana dashboard provided a visual representation of the relative differences in the scores.

The results show that the lower scores for aggregated metadata records are due to the definition of the FAIR criteria rather than their implementation by the tools. Which begs the question, 'is it OK that aggregated metadata is, by definition, less FAIR, or should the definition be changed to level the playing field?'

FAIR / 30

FAIRness assessment and the role of DDI in the Generations and Gender Programme

Authors: Arianna Caporali[1]; Olga Grünwald[2]

[1] *French Institute for Demographic Studies (INED)*

[2] *Netherlands Interdisciplinary Demographic Institute (NIDI)*

Corresponding Authors: grunwald@nidi.nl, arianna.caporali@ined.fr

The Generations and Gender Programme (GGP) is a cross-national longitudinal panel survey on life-course and family dynamics launched in the year 2000. It comprises two rounds of data collection: Generations and Gender Survey-I (GGS-I), covering 19 countries, and GGS-II, initiated in 2017 and currently ongoing, with data available so far for 10 countries. The GGP is on the European Strategy Forum on Research Infrastructures (ESFRI) Roadmap and aims to become a European Research Infrastructure Consortium (ERIC).
Since its inception, the GGP has been committed to high quality data documentation and has chosen the Data Documentation Initiative (DDI) standard to document its surveys. It has implemented DDI-Codebook, before opting for DDI-Lifecycle with the start of GGS-II. Today users can browse the datasets in the GGP Colectica Portal (https://ggp.colectica.org/). As part of the work to enhance its technical excellence, GGP takes part in a support action promoted by FAIR-Impact (https://fair-impact.eu/) to assess its level of FAIRness.
This paper presents the results of this assessment and discusses the contribution of DDI to making GGS datasets FAIR. It ends with consideration of the actions needed for GGP to increase FAIRness and the role that compliance with DDI will play to implement them.

DDI-CDI / 31

Implementing the DDI-CDI Process Model for Describing Data

Integration: Insights from the EOSC Future Science Project "Climate Neutral and Smart Cities"

Authors: Benjamin Beuster[1]; Joachim Wackerow[None]

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

Corresponding Authors: benjamin.beuster@sikt.no, joachim.wackerow@posteo.de

The "Climate Neutral and Smart Cities" Science Project contributes significantly to the European Open Science Cloud (EOSC) Future Project by showcasing cross-domain data integration using the new DDI-CDI metadata specification. This presentation demonstrates the use of the DDI-CDI Process Model, which offers a standardized approach to comprehensively describe data lineage and integration processes. This helps researchers understand the complexity of the integrated data and the resulting variables.

The presentation focuses on a specific facet of the DDI-CDI Process Model, breaking down individual process steps with designated purposes and distinct input/output parameters. This approach emphasizes the process sequence rather than the entire DDI-CDI process model.

Furthermore, we introduce a prototype tool that not only demonstrates the use for researchers but also facilitates universal understanding of the DDI-CDI process description. Additionally, we address limitations of the DDI-CDI model identified during the process tool implementation, which prompted updates by the DDI-CDI working group.

User Needs / **32**

The Swiss Virtual Educational Observatory and DDI

Authors: David Schiller[1]; Marcel Hanselmann[1]

[1] *Swiss Institute for Information Science*

Corresponding Author: marcel.hanselmann@fhgr.ch

The Swiss Virtual Educational Observatory (VEO) project is funded by the Swiss National Science Foundation. It has the goal to link and to visual sources of research data about education and learning. While concentrating on research data itself data documentation and sources of open research data are relevant as well.
The first step to find research data is looking for open data and data documentations. This is where DDI can play is crucial role. But sadly, it does not.
Still there are sources not documented at all or documented in Excel or Word. Institutions using a machine-readable documentation have developed solutions individual to their institution. Provider of smaller data sources struggle with having enough resources for a good data documentation. All this is making data discovery a hard job. Accordingly, hindering research.
Why is this still the case and how can it be overcome?
Switch, a Swiss service provider for infrastructures, has developed a tool for open data sources that collected data form different standards and variations of those standards. Is this the way to go? Or need data providers to be forced to use a standardized standard?
The talk will give examples and provide thesis about how to go ahead in Switzerland.

FAIR / **34**

Qualitative Data and DDI: Chances to move forward

Author: Noemi Betancort[1]

Co-author: Kati Mozygemba [2]

[1] *RDC Qualiservice, University of Bremen / State and University Library Bremen*

[2] *RDC Qualiservice, University of Bremen*

Corresponding Author: noemi.betancort@suub.uni-bremen.de

The QualidataNetwork (QualidataNet) links several research data centers that archive and provide access to sensitive qualitative research data. It is part of the Consortium for Social, Behavioural, Educational and Economic Sciences (KonsortSWD) at the National Research Data Infrastructure (NFDI) in Germany.

In 2022, we presented the project, our goals and how we plan to move forward. Many of you have supported us in the meantime. For example, we are currently in contact with the DDI Controlled Vocabularies Working Group to develop the *QualiTerm* controlled vocabulary to improve the findability of qualitative data.

We would like to take these efforts a step further and use the feedback and comments from the last EDDI conference and throughout the year as an opportunity to discuss the creation (or *revival*) of a Qualitative Data Working Group within the DDI Alliance. We would like to propose what we consider to be important aspects of applying the FAIR Data Principles to qualitative data, and to explore with you the idea of international collaboration at the DDI level. We hope to reach out to all interested persons to create a forum for discussion of issues related to the management, infrastructure and reuse of qualitative data.

DDI - New Directions / 35

Documenting and validating administrative data with DDI

Authors: Romain Tailhurat[1]; Thomas Dubois[1]

[1] *Insee*

Corresponding Author: romain.tailhurat@insee.fr

Official statistics increasingly rely on external sources, particularly administrative data, to produce statistics. This requires further industrialisation of the data integration before the downstream steps leading to dissemination.
In 2021, INSEE has launched a project named Resil with the objective of centralising administrative data ingestion for further processing of social statistics. Raw data received must be delivered to statisticians with documentation, in particular structural metadata.

DDI is used for describing the structure of the data stored. The data acquisition tool takes this formal description as input and generates the model of ingested data. So, DDI is used in an active way, since the DDI documentation is required to process the data received.

In order to assess the compliance between data and associated structural metadata (e.g., an integer value must be between limits, a code value must be present in a code list…), INSEE has developed a tool which generates VTL scripts (Validation and Transformation Language is a standard, part of SDMX) from structural metadata in DDI. The validation is then performed using Trevas, an open source VTL engine for big data distributed environments.

The presentation will focus on DDI as the pivot model for documenting and validating administrative data in the context of Resil.

DDI-CDI / 36

The role of structured metadata in the EOSC Future Science Project 'Climate Neutral and Smart Cities'

Author: Hilde Orten[1]

Co-authors: Arofan Gregory [2]; Benjamin Beuster [3]; Joachim Wackerow

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

[2] *CODATA*

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

Corresponding Author: hilde.orten@sikt.no

The objective of the 'Climate Neutral and Smart Cities' Science Project of EOSC Future is to demonstrate that relevant environmental data and data on citizens' values, attitudes, behavior and involvement can be combined in a meaningful way for social, political and scientific analysis.
The Science Project rests on three pillars: Indicator production and integration of data from three different research domains, structured metadata for interdisciplinary use, and dissemination of data and research outcomes.
The presentation gives an introduction to the project and describes needs and usages of structured metadata in the project, with main focus on DDI-Cross Domain Integration (DDI-CDI), and how DDI-CDI and DDI-Lifecycle are used together in the project.

Variable Cascade / **37**

Variables and their Value Domains

Author: Dan Gillman[1]

[1] *US Bureau of Labor Statistics*

Corresponding Author: gillman.daniel@bls.gov

A value domain is the description of the values a variable is allowed to take. The idea originated with the ISO/IEC 11179 series of standards. DDI-CDI makes use of them explicitly, and DDI-L uses the idea as well. However, metadata reuse depends on which metadata are assigned to a variable, and which are assigned to a value domain. We address this here.

We can illustrate the problem with an example: Suppose we have a variable for household income measured in 2023 USD. The applicable values are non-negative scaled numbers to 2 decimal places, such as $75,000.00 (in the American format). Applying the variable to households in developed countries might mean (ten million dollars), but the same variable used in a developing country might only need a maximum of $100,000.00 (one hundred thousand dollars). However, the value domain ought to be the same: non-negative monetary amount in USD.

In both cases, the constraints makes the two applications different. Now the question arises as to where we specify the constraints, the maxima? Placing them on either the variable or the value domain reduces reuse and interoperability, as they make the variable or the value domain different in each usage.

Mostly, it is the value domain where people tend to place the constraints, It makes sense from a strict perspective. The constraints affect the allowable values. We argue here that the more relevant choice for the constraints is on the variable since it is the application of the variable where the constraints are needed. Luckily, the variable cascade defined in both DDI-CDI and DDI-L can account for the differences and increase reuse and interoperability.

We describe an application of these ideas from work done at the Office of Data Governance at the US Department of Labor.

Controlled Vocabularies / 38

Enhancing FAIR compliance: A controlled vocabulary for mapping Social Sciences survey variables

Authors: Janete Saldanha Bach[1]; Claus-Peter Klas[1]

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

Corresponding Author: janete.saldanhabach@gesis.org

In Social Sciences surveys, the dynamic relationship among survey instruments and study entities like questionnaires, variables, questions, and response formats evolve. When reusing variables, researchers may need to modify variable attributes such as labels or names, question-wording, or response scales. Therefore, explaining these relations across different waves and studies is necessary to track how variables relate to each other. Although standards like Data Documentation Initiative – Lifecycle (DDI-LC) and DataCite model these relationships, these frameworks fall short of capturing the complexity of variable relationships. The DDI Alliance Controlled Vocabulary for Commonality Type employs codes—such as 'identical,' 'some,' and 'none'—to outline shifts in entities like variables; however, this approach is insufficient for disambiguating these relationships since they do not differentiate the variable attributes subject to change. To bridge this gap, we introduce the GESIS Controlled Vocabulary (CV) for Variables in Social Sciences Research Data. This CV is specifically designed to enhance semantic interoperability across various organizations and systems. By establishing explicit relationships, it not only facilitates harmonization across different study waves but also enriches data reuse. This enhancement supports advanced search and browse functionalities. The CV, published via the CESSDA vocabulary manager, seeks to forge a semantically rich, interconnected knowledge graph specifically tailored for Social Science Research. This endeavour aligns with the FAIR data principles, aiming to foster a more integrated and accessible research landscape.

Posters / 39

CLOSER Discovery update: user-driven redesign

Author: Hayley Mills[None]

Corresponding Author: h.mills@ucl.ac.uk

CLOSER Discovery is the UK's most comprehensive research tool for longitudinal population studies (LPS). DDI-Lifecycle is used to document questionnaire, question, dataset and variable metadata from 11 (and counting) leading UK studies. CLOSER Discovery is powered by the Colectica software stack including a customised Colectica Portal which sits on top of Colectica Repository, and provides access to item level metadata through display as web pages.

Since the launch in 2017, the CLOSER Discovery portal has had only minor visual updates within the bounds of the underlying software. In the latest update, launched in September 2023, we have worked with Colectica to update the Portal software, enabling the implementation of the design created by Bravand, a digital design and build company with experience of complex websites.

CLOSER Discovery has been redesigned using detailed feedback from one-to-one interviews with users, focussing on the home, search and explore pages. The poster will detail the new design features and the feedback it is based on.

Controlled Vocabularies / 40

Showcasing Progress of the CESSDA's Controlled Vocabulary Service (CVS)

Author: Maja Dolinar[1]

[1] *Slovenian Social Science Data Archives, Faculty of Arts, University of Ljubljana*

Corresponding Author: maja.dolinar@fdv.uni-lj.si

CESSDA has launched version 3 of its Vocabulary Service (CVS), a significant upgrade from version 2. Accessible at https://vocabularies.cessda.eu, CVS offers users the capability to explore and download multilingual controlled vocabularies in formats like SKOS, HTML, and PDF. The Editor component empowers authorized individuals to manage and translate vocabularies. Many vocabularies have been shaped by the DDI Alliance and its Controlled Vocabularies Working Group, with multilingual content enriched by CESSDA members. Such vocabularies are instrumental in harmonizing data descriptions, as showcased in the CESSDA Data Catalogue. Primarily, the tool benefits administrators, translators, data repositories, and researchers. Administrators can benefit from using an easy and robust vocabulary management system to host their in-house vocabularies and access those provided by other agencies, for use in their data catalogues. Translators can use it to reduce the complexity of their processes. Data repositories and metadata catalogue maintainers can use the REST API to consume controlled vocabularies in machine readable format.
The aim of this presentation is to introduce the CVS and highlight its latest advancements. Version 3 introduces key features such as term deprecation, a 3-digit versioning system, and enhanced workflows. Updates to user guides illuminate these changes, optimizing the platform for users aiming for precision and efficient vocabulary management.

DDI Standards / **41**

The Art of DDI!? - The Model-Driven Approach of DDI-CDI

Author: Joachim Wackerow[1]

[1] *Independent expert*

Corresponding Author: joachim.wackerow@posteo.de

This presentation describes the model-driven approach of DDI-CDI. In this way, it is possible to generate related syntax representations (such as XML Schema and RDF) and field-level documentation of the UML model. A subset mechanism allows targeted generation for specific use cases such as the process description. Experimental work will also be shown using visualization and sonification methods to represent complex data such as the network represented by the DDI-CDI model. The talk will highlight model- and data-driven approaches.

DDI Alliance Plenary / **42**

Envisioning the Future of DDI in the Metadata Landscape

Authors: Hilde Orten[1]; Jared Lyle[2]; Jon Johnson[3]

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

[2] *ICPSR, University of Michigan*

[3] *CLOSER, UCL*

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

Join the chairs of the DDI Executive Board and the DDI Scientific Board, as well as the DDI Executive Director, to discuss Alliance priorities and plans, especially with an eye toward user and member needs. Engage in thoughtful conversations and Q&A, especially as we envision DDI's strategic role in the exciting future of metadata.

User Needs / **43**

How DDI supports data management, archiving and secondary reuse at the French Center for Socio-Political Data

Author: Alina DANCIU[1]

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

Corresponding Author: alina.danciu@sciencespo.fr

The French Center for Socio-Political Data is jointly operated and financed by Sciences Po, one of the leading SSH French universities, and the French Center for National Research (CNRS). One of the center's main missions has been to serve the French and international SSH communities by facilitating the reuse of surveys and data, both quantitative and qualitative, in the fields of sociology and political science. The main themes of CDSP's data repository are: political attitudes and behaviour, gender, family, immigration, school, health, cultural practices, new technologies, etc. The CDSP is one of the authoritative data sources concerning French elections, providing ballot results 1958 to 2012, making them available both to the research community and to the general public.
The CDSP has been using DDI for the last 15 years and has recently been certified with the CoreTrustSeal. In this presentation, we'll discuss the role DDI has in our data management and archiving processes and how it helped us be eligible for the CoreTrustSeal.

Variable Cascade / **44**

Mixed-mode survey creation as a key scenario for tool support of the DDI variable cascade

Authors: Claus-peter Klas[1]; Oliver Hopt[None]

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

Corresponding Author: oliver.hopt@gesis.org

Mixed mode surveys are not new but have become quite popular or necessary recently due to a variety of reasons such as enhanced response rates, access to diverse populations or flexibility.
When it comes to creation, documentation, or management of mixed-mode survey they become complex to handle, then having one master questionnaire. More complexity is followed by more resources, making mixed-mode surveys more expensive. To reduce the complexity and costs we propose to provide the right tools and base them on the DDI variable cascade model. We want to discuss potential workflows and key requirements focusing on creation and documentation of mixed-mode questionnaires and surveys. The support should also include the seamless integration in lifecycle management for further reduction effort.

Small mixed mode surveys are a key scenario for tooling because they cover the same needs as large projects with the resource restrictions of single studies. By handling the needs of small mixed mode studies, the approach would also offer a raised standard of documentation for single studies and reduce the effort of metadata management within the large survey programs.

Posters / **46**

The Path to Open Access for Restricted Data with the Data Product Builder

Authors: Deirdre Lungley[None]; Thomas Gilders[1]

[1] *University of Essex*

Corresponding Author: dmlung@essex.ac.uk

With the Data Product Builder (DPB) the UK Data Service (UKDS) aims to allow researchers access to on-demand linked subsets of data, dynamically assessed for emergent disclosure risk in real time. Such a system depends both on sophisticated upstream metadata and powerful downstream computation. We detail the pipeline components required to take original curated data in traditional dissemination formats, e.g. SPSS, and make it available as a secure RDF linked data resource. These components encompass:

• Transforming the binary files into a highly granular structural representation in DDI-CDI
• Using machine learning models to aid automation of metadata enhancement, e.g. determining 'key' variables for Disclosure Risk Analysis
• Aligning study variable representation with aggregate census variable representation to allow us to benchmark 'risk' using population statistics
• Transforming the user determined data product into RDF triples, as well as permitting further deserialization into multiple binary formats - SPSS, STATA, Excel etc., - if required for download and/or desktop analysis.

This pipeline results in a new type of digital resource, where data products are dynamically built by the researcher based on their individual research need, augmented by real-time disclosure risk mitigations, unblocking their access to data which has traditionally been difficult and time-consuming to procure.

47

Large Scale HPC Infrastructures: Societal Impact and the Role of Training for Data-driven Professionals, Scientists, and Innovators

Author: Antonino Rotolo[1]

[1] *Alma Mater Studiorum - Università di Bologna*

Corresponding Author: antonino.rotolo@unibo.it

High performance computing (HPC) is key to Europe's future prosperity, digital transformation and resilience. This has been acknowledged by the EU strategy and investments. One recent initiative in this context is the establishment of the Italian National Centre for HPC, Big Data and Quantum Computing. This centre, which is funded under the National Recovery and Resilience Plan (NextGenerationEU), conducts R&D, nationally and internationally, for innovation in high-performance computing, simulations, and big data analytics. The aim is pursued through a world class research infrastructure for high-performance computing and big data management, which leverages existing resources and integrates emerging technologies.
This workshop aims at discussing two dimensions:
1. The need to elaborate disruptive methods for assessing the societal impact of this type of infrastructures, encompassing the following spectrum of implications: on the scientific side, generating high-quality knowledge, enhancing human capital in research and innovation, and promoting Open Science and data-driven research methods; on the social and ethical front, boosting efficiency, awareness, participation, equity, fairness, trust, sustainability, and transparency while addressing discrimination and improving education access; economically, striving for efficiency, innovation, and new business models, with potential effects on employment, public funding reliance, fintech, and inclusive economic growth; legally, grappling with fundamental rights, intellectual property, privacy, liability, cybersecurity, fairness, and compliance; politically, working on international control, decision-making, e-democracy, abuse, and digital sovereignty.
2. The need of training in order to attain systemic objectives such as promoting "culture" and competencies of data-driven research and innovation, improving the competencies of young researchers through the access to large HPC infrastructure, creating a knowledge ecosystem for a tailored access modes for industry, supporting the whole research lifecycle through integrated and federated HPC and Big Data resources, ensuring that more scientific communities have access to state-of-the-art

HPC and Big Data services.
The workshop is part of the RItrainPlus Community of Practice workshops (see https://ritrainplus.eu/join-the-events-of-ritrainplus-community-of-practice/).

DDI Standards / **48**

DDI: From a simple Codebook to an integrated suite of products

Author: Wendy Thomas[None]

Corresponding Author: wlt@umn.edu

The vision of DDI has grown over the years from a simple codebook for microdata and aggregate data files, to a model that supported metadata-driven data systems, to a suite of products that supports a variety of applications in a broad area of coverage across several related disciplines. This shift in focus reflects changes in the data environment, technology, and user needs over time. It raises the need for a clear road-map for individual products as well as the overall integration of DDI products and support mechanisms. The Technical Committee of the DDI is responsible for overseeing the coordinated development, publication, and maintenance of the suite of DDI products. As such, the overall roadmap needs to reflect the intended purpose and means of ensuring the development and maintenance of a well-integrated suite of products as well as products to support product development and the smooth transfer of metadata between products to support different applications and uses over time.

This presentation will address activities and plans to address issues concerning:

- The integration and interplay of products in the DDI Suite
- Common ontology/common objects used across DDI products
- Movement of content between products
- Product alignment in terms of coverage and applied usage
- Consistency f new coverage areas in multiple products
- Tooling to support input to the development process, transfer of content between products, and decision support

Keynote (Day 1) / **49**

Keynote: Open Access roadmaps in SloveniaSlovenia's efforts to align with the open science policies in the European Research Area

Miro Pušnik the Director of the Central Technical Library at the In 2021, Slovenia signed up to the Pact for Research and Innovation, a commitment that establishes common values and principles for research and innovation in the European Research Area (ERA). Slovenia has actively responded to the need to harmonise research and innovation activities by adopting comprehensive legislation and implementing strategic measures to guide these activities. In these official documents, open science, adherence to the FAIR principles for the sharing of research results and an urgent changes in the assessment of research work are identified as key points for the advancement of science in Slovenia. Therefore, the purpose of this paper is to provide a concise overview of the measures taken by Slovenia to align with the ERA open science policies, together with the underlying motivations for doing so.

DDI - New Directions / 50

Survey Variables Classification with Hierarchical Machine Learning

Author: Ivan Evdokimov[1]

[1] *University of Essex*

Corresponding Author: ie20391@essex.ac.uk

Recent developments in Machine Learning (ML) show robust performance in the area of Natural Language Processing (NLP) tasks, such as sentiment analysis and document classification. Our ML task is one of short text classification, specifically we are endeavouring to annotate variables using the variable name, label, question text and representation. Our task is one of multi-class classification, where accuracy is known to be sensitive to the number of labels in the dictionary. For this particular ML task we are bounding the annotation to purely learn 'key variables' - socio-demographic indicators which feed our Disclosure Risk Analysis (DRA).

In this presentation, we present a Hierarchical Machine Learning (HML) approach to recognising variable concepts from studies available at the UKDS. Specifically, we decompose the task into first learning the broad group to which the variable belongs, e.g. Education and then the concept within that group, e.g. Highest Educational Qualification.

At a high level, we use a mix of shallow and deep learning models to minimize the number of target class labels and boost the overall performance of each individual algorithm. We present details of the cost function simplification process and each hierarchy metrics and benchmarks.

Conference Opening / 51

EDDI 2023 Welcome and Chair's Opening Remarks

Corresponding Authors: jon.johnson@ucl.ac.uk, mari.kleemola@tuni.fi

Conference Opening / 52

Welcome from the Host Institution

Iztok Prezel, Dean of the Faculty of Social Sciences

Keynote (Day 2) / 53

How to ensure semantic interoperability between FAIR digital objects

FAIR Digital Objects (FDOs) are datasets, publications, software, services, workflows, lab notebooks, and other digital results of research that are easily discoverable by humans and machines. FDOs are accessible to anyone with permission, interoperable with other digital objects regardless of their format or software, and reusable without modification for the same or different purposes. The semantic interoperability of FDOs is achieved by using common vocabularies and ontologies to define

the meaning of data elements. These vocabularies and ontologies provide a common understanding of the data, even if it is stored in different systems or formats. We will first outline the minimum set of metadata required for FDOs. The lecture will focus on the National Open Science Infrastructure, what processes we have put in place, and how we have achieved semantic interoperability of FDOs within this infrastructure. Finally, we will outline the European Genomic Data Infrastructure (GDI) project, where we are working to establish a federated, sustainable, and secure infrastructure that will provide access to genomic and related phenotypic and clinical data across Europe, with controlled access for clinicians, public and private sector researchers, and health policymakers. Non-sensitive and aggregated data will be openly discoverable through the federated query system.

Closing Session / **54**

Closing Remarks

Corresponding Authors: jon.johnson@ucl.ac.uk, mari.kleemola@tuni.fi

Closing Session / **55**

Announcement of EDDI 2024

Conference Opening / **56**

Welcome from Local Organisation Committee

Author: Irena Vipavc Brvar[1]

[1] *ADP*

Corresponding Author: irena.vipavc@fdv.uni-lj.si

Welcome from Local Organisation Committee