EDDI
EUROPEAN
DDI USER CONFERENCE

13th Annual European DDI User Conference

Virtual (Paris) Conference, 30 November – 1 December

Twitter: #EDDI2021

Sponsors

SciencesPo
CENTER FOR SOCIO-POLITICAL DATA

gesis
Leibniz-Institut
für Sozialwissenschaften

I Z A Institute
of Labor Economics

IDSC
INTERNATIONAL
DATA SERVICE CENTER

DDI

About

EDDI21 is organized jointly by CDSP, Center for Socio-Political Data, GESIS, Leibniz Institute for the Social Sciences, and IDSC of IZA, International Data Service Center of the Institute for the Study of Labor. The DDI Alliance kindly supports EDDI 2021.

The conference will be held virtually, the program includes links to the sessions.

Recording of sessions

Sessions will be recorded, the presentations and slides will be made available after the event on the DDI YouTube channel. The discussions will NOT be made publicly available.

Conference Co-Chairs

Jon Johnson, CLOSER, UCL, Social Research Institute, United Kingdom

Mari Kleemola, FSD – Finnish Social Science Data Archive, Finland

Program Committee

Alina Danciu, Center of Socio-Political Data, Sciences Po Paris, France

Uwe Jensen, GESIS – Leibniz Institute for the Social Sciences, Germany

Jon Johnson, CLOSER, UCL, Social Research Institute, United Kingdom

Mari Kleemola, FSD – Finnish Social Science Data Archive, Finland

Nicolas Sauger, Center of Socio-Political Data, Sciences Po Paris, France

Knut Wenzig, German Institute for Economic Research (DIW Berlin) / German Socio-Economic Panel (SOEP), Germany

Wolfgang Zenk-Möltgen, GESIS – Leibniz Institute for the Social Sciences, Germany

Organisation Committee

Nikos Askitas, IDSC, IZA – Institute of Labor Economics, Germany

Alina Danciu, Center of Socio-Political Data, Sciences Po Paris, France

Marie Darcimoles, Center of Socio-Political Data, Sciences Po Paris, France

Uwe Jensen, GESIS – Leibniz Institute for the Social Sciences, Germany

Jon Johnson, CLOSER, UCL, Social Research Institute, United Kingdom

Mari Kleemola, FSD – Finnish Social Science Data Archive, Finland

Nicolas Sauger, Center of Socio-Political Data, Sciences Po Paris, France

Amelie Vairelles, Center of Socio-Political Data, Sciences Po Paris, France

Timetable

All times are CET

Day	Time	Track One	Track Two
Tues 30 Nov	09:30 - 09:50	Welcome to EDDI 2021 - https://sciencespo.zoom.us/j/93426390826	
	10:00 - 11:00	Enhancing Metadata (1) https://sciencespo.zoom.us/j/98997713068	Discovery (1) https://sciencespo.zoom.us/j/99778568227
	12:00 - 13:00	Enhancing Metadata (2) https://sciencespo.zoom.us/j/92088621800	Discovery (2) https://sciencespo.zoom.us/j/95021305482
	14:00 - 15:00	Infrastructure https://sciencespo.zoom.us/j/91337079798	Discovery (3) https://sciencespo.zoom.us/j/92394021642
Weds 1 Dec	10:00 - 11:00	Enhancing Metadata (3) https://sciencespo.zoom.us/j/95495750926	Interoperability (1) https://sciencespo.zoom.us/j/92725767670
	12:00 - 13:15	Workflows https://sciencespo.zoom.us/j/96931984565	Interoperability (2) https://sciencespo.zoom.us/j/97000444531
	14:00 - 15:00	User Needs https://sciencespo.zoom.us/j/94852706141	Interoperability (3) https://sciencespo.zoom.us/j/92669920880
	15:30 - 16:00	Plenary & Announcement of EDDI 2022 - https://sciencespo.zoom.us/j/99373030588	

Friday Nov. 26 & Monday Nov. 29

CODATA & EDDI Training FAIR

Details of registration for this event are available at the EDDI Website at:

https://www.eddi-conferences.eu/eddi-2021/eddi-2021-training-fair-26-29-11/

Tuesday November 30

Welcome (9:30 CET)

Chair: Mari Kleemola & Jon Johnson

Zoom Link: https://sciencespo.zoom.us/j/93426390826

Enhancing Metadata 1 (10:00 CET)

Chair: Hilde Orten - Norwegian Centre for Research Data

Zoom Link: https://sciencespo.zoom.us/j/98997713068

Engineering a Machine Learning Pipeline for Automating Metadata Extraction from Longitudinal Survey Questionnaires

Suparna De[1], Harry Moss[2], Jon Johnson[3], Jenny Li[3], Haeron Pereira[1], Sanaz Jabbari [2]
[1] University of Surrey – United Kingdom
[2] Research IT Services, UCL – United Kingdom
[3] CLOSER, UCL Institute of Education – United Kingdom

Data Documentation Initiative-Lifecycle (DDI-L) introduced a robust metadata model to support the capture of questionnaire content and flow, and encouraged through support for versioning and provenancing, objects such as BasedOn for the reuse of existing question items. However, the dearth of questionnaire banks including both question text and response domains has meant that an ecosystem to support the development of DDI ready CAI tools has been limited. Archives hold the information in PDFs associated with surveys, but extracting that in an efficient manner into DDI-Lifecycle is a significant challenge.

While CLOSER Discovery has been championing the provision of high-quality questionnaire metadata in DDI-Lifecycle, this has primarily been done manually. More automated methods need to be explored to ensure scalable metadata annotation and uplift.

This paper presents initial results in engineering a machine learning (ML) pipeline to automate the extraction of questions from survey questionnaires as PDFs. Using CLOSER Discovery as a 'training dataset', a number of machine learning approaches have been explored to classify parsed text from questionnaires to be output as valid DDI items for inclusion in a DDI-L com- pliant repository.

The developed ML pipeline adopts a continuous build and integrate approach, with processes in place to keep track of various combinations of the structured DDI-L input metadata, ML models and model parameters against the defined evaluation metrics, thus enabling reproducibility and comparative analysis of the experiments. Tangible outputs include a map of the various meta- data and model parameters with the corresponding evaluation metrics' values, which enable model tuning as well as transparent management of data and experiments.

Transform legacy study, question, and variable documentation into DDI 3.2 LifeCycle

Claus-Peter Klas[1], **Knut Wenzig**[2]
1 GESIS – Leibniz Institute for the Social Sciences – Germany
2 German Institute for Economic Research – Germany

Many service providers and institutions in the social sciences document survey data including study information, questions, and variables in self-developed legacy software in proprietary databases or other forms according to the specific needs and use cases. We introducean easy-to-use REST API to create a DDI LC 3.2 documentation based on a use case of the question/variable documentation by SOEP (DIW). DIW has documented their questions and variables in CSV files using paneldata.org. The REST API is attached into the GESIS questionnaire editor.

The resulting documentation is a self-contained DDI XML file including the study unit documentation, instruments, questions, answers, and connected variables. The documentation can be visually explored, evaluated or even further documented in the GESIS questionnaire editor, but also downloaded as DDI XML file or Word/PDF report. The created documentation can also be directly re-used for question search, e.g., based on the CESSDA European Question Bank. In addition, REST API is capable to import multiple languages and translations.

Discovery 1 (10:00 CET)

Chair: Maja Dolinar - Slovenian Social Science Data Archives

Zoom Link: https://sciencespo.zoom.us/j/99778568227

Developing a DDI-based Online DataCatalogue Using the Software NADA

Julie Lenoir [1], **Julie Baron**[1], **Arianna Caporali***[2]

[1] Institut national d´tudes d´emographiques - DataLab – INED – France
[2] Institut national d´tudes d´emographiques - DataLab – INED – France

In the past year, the DataLab of the French Institute for Demographic Studies (INED) has been evaluating different software solutions to develop an online data catalogue based on DDI-Codebook. This was necessary because Nesstar, the software we used so far, is being discontinued. The new tool should meet INED-specific requirements, provide a user-friendly access to the metadata we manage and promote INED socio-demographic data among expert and non-expert users.

In this presentation, we explain why, after a thorough investigation, we decided to implement NADA Microdata Cataloging Tool rather than Dataverse. We then illustrate the work carried out to develop the new data catalogue. This includes updating our metadata in order to comply with the CESSDA metadata model and upgrade to DDI-Codebook 2.5. We present our metadata updating process and our procedure to migrate from Nesstar to NADA. We conclude with showing our new data catalogue under development. This paper is the follow-up of a study presented at the 2020 European DDI User Conference by one of the authors (https://doi.org/10.5281/zenodo.4326739).

Providing data at the National Archive of Official Statistics (ADISP)

Elodie Petorin[1]

[1] Archives de donn´ees issues de la statistique publique (ADISP-ProGeDo) – CentreNationalde la Recherche Scientifique - CNRS, PROGEDO – France

The National Archive of Data from Official Statistics, part of the Large Research Infrastructure ProGeDo, has been providing access to social sciences data from more than 20 years.

Our mission is to disseminate studies and database produced by Official Statistics to the whole scientific community in order to foster the use of statistical data in social sciences.

In the context of the FAIR principle, to facilitate the discovery and re-use of data by humans and information systems, an indexation and contextualization of the data is necessary. In other word, we need to question our documentation processes to disseminate data that are findable, accessible, interoperable and reusable under the DDI standard.

Therefore, this presentation will highlight the processes and questions for providing DDI-compliant data in adequation with the FAIR principle.

Enhancing Metadata 2 (12:00 CET)

Chair: Benjamin Beuster - Norwegian Centre for Research Data

Zoom Link: https://sciencespo.zoom.us/j/92088621800

Metadata-based synthetic data generation

Erik-Jan Van Kesteren[1,2]
[1] Utrecht University – Netherlands
[2] ODISSEI (Open Data Infrastructure for Social Science and Economic Innovations) – Netherlands

The Dutch Central Bureau of Statistics (CBS) possesses extremely valuable sensitive datasets with information on the entire population of the Netherlands. Researchers in the Netherlands may gain secure access to these datasets after submitting a detailed proposal with information on which datasets are required to answer their research question. However, it is difficult, time- consuming, and costly to manually comb through the metadata to understand whether and how the available data may answer their question of interest. To help researchers in this initial exploration step, we are developing a software program that generates synthetic example data (public use files) based on variable-level information from public metadata – e.g., means, standard deviations, range, variable names and types, and dataset size. Using this synthetic data, researchers can define their requirements and even write analysis code before formally requesting access.

In this talk, I will show off a prototype app (https://github.com/sodascience/ddi-synth), discuss the implications, and talk about how this concept can grow in the future. We intend for this software to integrate with the Dataverse instance in development at ODISSEI, the the Dutch Social Science research infrastructure, and we are considering the DDI metadata format for this task.

Open, metadata enriched, non-proprietary data format for data dissemination

Claudia Saalbach[1], **Xiaoyao Han**[1]
[1] DIW Berlin – Germany

At the moment, social scientists are using different and sometimes proprietary software which processes metadata in different ways to analyze their data. Different data formats of statistical software packages, which are only partially compatible, represent an obstacle for replication studies. In particular, proprietary data formats endanger the requirement of interoperability anchored in the FAIR principles. Our project aimed to address this problem by introducing a metadata-enriched open data format that can be easily accessible, readable and interoperable in various statistical software.

As a first milestone, we would like to present the conceptual model of an open data format, where DDI might come into play, and a minimal data example derived from survey data. Furthermore, we would like to discuss our preliminary work on importing the open-data format into (proprietary) statistical programs and converting data of a proprietary format into the open format.

Discovery 2 (12:00 CET)

Chair: Iris Alfredson – Swedish National Data Service

Zoom Link: https://sciencespo.zoom.us/j/95021305482

DDI Search: The missing link between researcher and repository

Thomas Krämer[1], Esra Akdeniz[1], Claus-Peter Klas[2], Alexander Mühlbauer[2], Oliver Hopt[2], Martin Friedrichs[1]
[1] GESIS – Leibniz Institute for the Social Sciences – Germany
[2] GESIS – Leibniz Institute for the Social Sciences – Germany

Once research institutions manage to document studies at question and variable level, an appropriate solution is required to make the data findable (as in FAIR) at a fine granular level. DDI Search is a set of microservices and a standalone user interface, that provide responsive search and browse functionality and personalization such as creating question collections for re-use.

In conjunction with the GESIS FlatDB, it is a strong, tested, open source and DDI compli- ant technology stack for Social Sciences research institutes to increase findability, accessibility, interoperability and re-usability at question and variable level.

User tests with the CESSDA EQB community were the basis for the UI development. With an open-source licensing policy, institutions are safe from the risk of vendor lock-ins often unavoidable with commercial solutions like NESSTAR or Colectica. Repositories can choose smooth, step-by-step implementation paths, unlike with monolithic architectures such as DataVerse.

We describe the technical components and data pipeline from institutional repositories to the search index, the functionalities of the search user interface and give directions for research institutes and data repositories how to employ the DDI Search for their purpose.

From Documentation to Search for Questions for Small Service Provider in DDI

Claus-Peter Klas[1], Oliver Hopt[1], Thomas Krämer[1], Sigit Nugraha[1]
[1] GESIS – Leibniz Institute for the Social Sciences – Germany

Many smaller service providers and institutions in the social sciences document survey data only on the level of studies. The reasons are multifold, but one prominent reason is missing soft- ware or tool support to document questions and variables. In addition, such tools are usually self-developed and lack functionality and do not provide DDI as metadata standard. Current tools to document studies and questions are DataVerse, NESSTAR and the commercial tool suite Colectica.

In conjunction with the GESIS FlatDB, it is a strong, tested, open-source and DDI compliant technology stack for Social Sciences research institutes to increase findability, accessibility, interoperability and re-usability at question and variable level we will provide updates about the GESIS questionnaire editor as open-source service for documentation and translation.
The documented information can directly be published and used for searching e.g., in the GESIS DDI search tool (see presentation: DDI Search: The missing link between researcher and repository).

Infrastructure (14:00 CET)

Chair: John Shepherdson - CESSDA

Zoom Link: https://sciencespo.zoom.us/j/91337079798

Rich Data Services and DDI

Pascal Heus[1], **Andrew Decarlo**[1]

[1] Metadata Technology North America Inc. – United States

Launched in 2020, MTNA's Rich Data Services (RDS) (https://www.richdataservices.com) offers a comprehensive solution for the concurrent publication of data and metadata as a service through a modern REST API. The platform comes equipped with three web applications to enable the ability to manage content, interactively access record level data, and rapidly create aggregated tables.

The RDS API provides the foundation needed to build rich web or desktop based data applications, access data directly from analytical and development environments like R, Python, SAS, Stata, C++, or Java, or for machine learning purposes. Open source libraries and multiple forms of documentation are available to developers and data scientists to rapidly leverage the API.

The RDS Open Data Packaging service further caters to the needs of users who prefer to work with the (meta)data offline. This provides the option to download the data in various text formats, alongside with scripts/syntax files, PDF documentation, or DDI metadata, enabling immediate reuse and minimizing wrangling.

From its inception, RDS leveraged DDI for loading metadata and as an export option. The latest version adds DDI-C and schema.org endpoints on each data product, providing easy access to metadata in standard formats. These latest features further facilitate alignment on GOFAIR recommendations, linked data integration, metadata harvesting, or discovery through Google Data Search.

Finally, RDS is being increasingly recognized as a potential alternative to Nesstar Server. Pop- ular amongst the DDI community, Nesstar has for over two decades been instrumental in the establishment of the standard, but unfortunately is no longer being actively maintained. Migration from Nesstar to RDS can mostly be automated and is a smooth and natural transition.

This short presentation will provide a refresher on what RDS can do for you, highlight the platform DDI aspects, and briefly demo the latest features and capabilities.

What's New in Colectica 6.2 and 7.0

Kevin Mcginnis[1], **Benjamin Adams**[1]

[1] Colectica – United States

Colectica is announcing the availability of two new releases, Colectica 6.2 and Colectica 7.0. 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 of Colectica 6.2 and Colectica 7.0, including:

- Calculation and display of weighted summary statistics
- SKOS Controlled Vocabulary import and usage enhancements
- Native storage for DDI OtherMaterial files
- Customizable facets and enhanced ElasticSearch indexer

- Linux support for Repository+Portal and Colectica Workflow
- Individually managed contributors and roles
- Qualtrics import to DDI
- A new difference viewer between DDI items or versions
- Nesstar and DDI 2 import improvements

Discovery 3 (14:00 CET)

Chair: Olof Olsson – Swedish National Data Service

Zoom Link: https://sciencespo.zoom.us/j/92394021642

Controlled Vocabularies in Colectica

Jeremy Iverson * [1]
[1] Colectica – United States

A controlled vocabulary is an organized set of terms. Controlled vocabularies make it easier to make information consistent. They also improve indexing and searching for information. DDI Lifecycle supports using controlled vocabularies for many of its metadata fields.

Colectica 6.2 adds support for controlled vocabularies in its desktop and Web tools. Administrators can configure which vocabulary to use for specific content fields. Users can choose terms from the appropriate vocabulary when editing content.

Vocabularies are specified as DDI Lifecycle code lists. DDI code lists allow specifying the values and terms of a vocabulary in one or more languages. Organizations can register these code lists in a metadata repository for persistence and revision tracking.

Colectica can import vocabularies from the Simple Knowledge Organization System (SKOS). The Consortium of European Social Science Data Archives (CESSDA) runs a vocabulary ser- vice, providing 28 vocabularies in more than 10 languages. These vocabularies are downloadable in SKOS format. This presentation will show how all these vocabularies can integrate with DDI Lifecycle and the Colectica software.

Making metadata FAIR: combining DDI

Flavio Rizzolo 1, Farrah Sanjari1
[1] Statistics Canada – Canada

Metadata in statistical production is ubiquitous: from concepts, classifications and variables to retention and provenance information, metadata is created, used and shared across all phases of the data lifecycle. Unfortunately, metadata management is sometimes silo-based and tool- specific, which impairs all four FAIR principles (Findable, Accessible, Interoperable, Reusable).

Statistics Canada is in the early stages of implementing a virtual metadata integration platform, the Metadata Hub, that integrates a collection of metadata repositories, e.g. Colectica, Aria, SDMX Istat Toolkit, OpenLink Virtuoso and CKAN across a number of standards, e.g. DDI, SDMX, XKOS, DCAT, and RDF/OWL, among others. In this presentation we describe our experience so far and the way forward

Wednesday December 1

Enhancing Metadata 3 (10:00 CET)

Chair: Kerrin Borschewski - GESIS

Zoom Link: https://sciencespo.zoom.us/j/95495750926

Exploring upgrade options for the CESSDA Data Catalogue

Dolinar Maja[1], Katja Moilanen[2], Markus Tuominen[2], Benjamin Beuster[3]
[1] Slovenian Social Science Data Archives (ADP), Faculty of Social Sciences, University of Ljubljana – Slovenia
[2] Finnish Social Science Data Archive (FSD), Tampere University – Finland,
[3] Norwegian Centre for Research Data - NSD – Norway

The CESSDA Data Catalogue (CDC) contains metadata of studies from CESSDA service providers (SPs) and serves as an entry point for search and discovery of European social science data. The CDC has been present in the European research space for several years, but its development has been hampered by metadata harmonization and technical support issues.

The goal of the CDC Upgrade Task group under Agenda 21-24 is to develop a list of recommended features and updates for the next version of CDC, including search and discovery of variable-level information. Based on desk research of available resources, a survey of SPs about needed improvements of the CDC, and an analysis of their variable-level metadata examples, we have developed a list of recommendations for the next CDC release.

The challenge is to either build the CDC on the SPs' existing metadata (i.e., primarily DDI Codebook) or to advocate for a metadata upgrade (i.e., DDI Lifecycle, currently used by only a handful of SPs). The result could be limited functionality for end-users who want to search and find variables documented in different languages. The goal of this presentation is to give a first insight into possible future developments of the catalogue.

Improving Metadata Management in the CSO, Ireland

Ciara Cummins[1]
[1] Central Statistics Office – Ireland

Quality is often defined in terms of being 'fit for purpose' for the user and a key element of this involves providing the metadata to gain clarity and to capture the reality of the data. Our goal in the Central Statistics Office (CSO), is to move towards the systematic storage and production of survey information and away from multiple standalone approaches used by Statistical areas. The CSO has recently begun a journey towards standardization and using Colectica products to utilize the international DDI (Data Documentation Initiative) metadata standard to integrate the transfer of statistical metadata through its statistical production environment.

This presentation will explain how we first began using the product as a Questionnaire Design tool to enable reusability, standardization and to help improve the quality and processes across statistical outputs. This presentation will also address our current progress incorporating DDI and Colectica into our Statistical production processes, the benefits, lessons learnt along with some challenges from our own experiences and our future objectives for metadata management in the CSO.

Interoperability 1 (10:00 CET)

Chair: Oliver Hopt - GESIS

Zoom Link: https://sciencespo.zoom.us/j/92725767670

Harvesting DDI for an integrated catalogue

Mari Kleemola[1], Claudia Martens[2], Anna-Lena Flügel[2]
[1] Finnish Social Science Data Archive (FSD), Tampere University – Finland
[2] Deutsches Klimarechenzentrum (DKRZ) – Germany

Integrated metadata catalogues and sharing metadata across all disciplines are needed to al- low scientists and machines to collaborate across borders and disciplines in finding research data. The FAIRsFAIR project has been working on metadata integration pilot. The pilot includes integration of social science metadata into the generic research discovery system B2FIND which is part of the EUDAT CDI and therefore a central indexing tool for EOSC-hub.

At the same time, CESSDA ERIC has been developing an open OAI-PMH API endpoint (called CESSDA Metadata Aggregator) to expose the metadata from the CESSDA Data Catalogue. Therefore, the decision was made to harvest CESSDA DDI Codebook 2.5 metadata records into B2FIND using the beta version of the Aggregator.

Our presentation will describe the work done, the selections made, and the findings from this pilot. We will also discuss the benefits and challenges of harvesting metadata records from various sources, experiences on DDI, and the possible ways forward.

Best practice for publishing classifications as linked data with XKOS

Franck Cotton[1]
[1] INSEE – Institut national de la statistique et des études économiques (INSEE) – France

XKOS is an RDF vocabulary extending SKOS and dedicated to the representation of statistical classifications. Though XKOS recently celebrated its tenth anniversary, it was officially published in 2019 by the DDI Alliance and the XKOS Working Group was formally created in 2021.

Also in 2021, the production of a “Best Practice” document was started. This guide intends to provide advice on how to use XKOS for maximum interoperability and reusability, and will cover topics like the description of classifications (labels, explanatory notes, levels), the different types of correspondences, how to represent evolution over time of the different elements, etc.

A number of specific topics will also be addressed, for example how to tackle multiple languages, how to publish statistical classifications as XKOS, or what kind of descriptive metadata to attach in order to maximize findability.

Where appropriate, best practice rules will be formalized in the SHACL validation language, so that publishers can check if their classifications conform to the guidance. In addition, the guide will propose a list of useful tools for the production and publication of XKOS statistical classifications.

The presentation will provide an overview of the XKOS specification, then detail the progress made regarding the best practice guide.

Workflows (12:00 CET)

Chair: Henri Ala-Lahti – Finnish Social Science Data Archive

Zoom Link: https://sciencespo.zoom.us/j/96931984565

CESSDA’s approach to bulk DDI study-level metadata validation and feedback

John Shepherdson[1], Matthew Morris[1]

[1] CESSDA – Norway

The CESSDA Data Catalogue (CDC) provides researchers with a single point of reference for the data holdings of CESSDA’s Service Providers. It harvests study-level metadata in DDI XML format from numerous OAI-PMH endpoints. A high degree of metadata standardization is required, in order to support sophisticated search and browsing techniques and provide re- searchers with relevant results.

The CESSDA Metadata Validator (CMV) has been developed to allow both data publishers and consumers to check metadata quality against published standards. DDI Profiles (formal, machine-actionable documents that specify additional constraints on the content of a DDI XML document, over and above those specified by the document’s associated XSD schema) are used to define the standards that must be met by study-level metadata. The constraints are assigned to validation gates that build on one another and thus allow different levels of compliance to be specified and validated.

CMV has been incorporated into CDC and is used to check harvested records for compliance with specific DDI profile/quality gate combinations. Non-compliant records may be automatically excluded from the catalogue, or just flagged as such. A dashboard is available so publishers can view and locate any constraint violations flagged against their metadata records.

Integrating DDI validation into daily workflow at FSD

Emil Rekola[1], Oskari Niskanen[1]

[1] Finnish Social Science Data Archive (FSD) – Finland

DDI users across the world face the problem of producing DDI Codebook compliant descriptions of datasets in the form of XML-files. Errors in the structure of the file are common if the validation of the file is done only by a human inspector. Therefore, many kinds of DDI/XML- validators have been developed to ensure data repositories can produce valid DDI descriptions that follow DDI Codebook structure

As a data archive, we too have faced the problem of producing compliant DDI descriptions. Therefore, we began developing our own DDI/XML-validator to solve the problem, and at the same, ease the workload of our employees that process the data. The validator that we developed is integrated into our operational database platform, which guides and aids us in our daily workflow.

The reason why we chose to develop our own validator instead of utilizing existing solutions, was because we wanted to have a tailor-made solution that fits our platform flawlessly. Our validator consists of a back-end micro-service and a user interface, which checks if the file is well-formed as well as follows DDI’s and FSD’s specifications.

In our presentation we’ll discuss the background information related to DDI/XML validation and showcase our newly built software.

Customizing Dataverse to workflows of ADP

Dolinar Maja[1], Gregor Žibert[1]
[1] Slovenian Social Science Data Archives (ADP), Faculty of Social Sciences, University of
Ljubljana –Slovenia

The Slovenian Social Science Data Archives (ADP) has identified the Dataverse application as the best new option for distributing its curated catalogue, which is currently hosted on its website and on Nesstar. In addition, the application was also seen as the best way to provide additional services for self-archiving and basically as an ingest point for all studies added to the catalogue (e.g. curated and self-archived). However, it soon became apparent that the default version of Dataverse lacked many features that were essential to our workflow and the services we wanted.

We have invested many resources over the last few years in customizing the default version and have tested several plugins that could be crucial (e.g. Data Curation plugin, Data Explorer plugin, Two Ravens). These include a multi-language user interface, implementation of the CMM CESSDA Metadata Model (and some ADP custom metadata fields), including DDI Controlled Vocabularies and ELSST keywords, and customization of deposit and access features (e.g., inclusion of online deposit agreements and terms of use).

The paper will present our experience in customizing the Dataverse software application to our needs, and identify issues and workarounds that repositories should be aware of when customizing the application to their workflows.

Interoperability 2 (12:00 CET)

Chair: Claus-Peter Klas - GESIS

Zoom Link: https://sciencespo.zoom.us/j/97000444531

StatConverter: statistical conversion tool using DDIwR

Adrian Dusa[1], Emilian Adrian Hossu[1]

[1] RODA – Romania

StatConverter is an open source alternative to the well-known, commercial software Stat- Transfer. It is built as a graphical user interface on top of R, as an ElectronJS application under Node.js and it is available from the list of software tools provided by the Romanian Social Data Archive (RODA).

This is a cross-platform tool which needs R to be installed on the local computer, and a series of R packages among which the dedicated DDIwR (DDI with R). This package uses the well-known package Haven written by Hadley Wickham at RStudio, which in turn uses the open source C library ReadStat by Evan Miller. Apart from being a wrapper around haven, the package DDIwR offers additional functionality by using DDI as a conversion platform between the various statistical software.

One other unique feature of DDIwR that no other software offers (including the commercial ones) is the treatment of missing values, converting those in the native format for each software. This is an automatic procedure scanning all variables' metadata for information about missing values, and creates a dictionary for consistently converting missing values with the same codes.

CESSDA Metadata Aggregator

Toni Sissala[1], **Matti Heinonen**[1]

[1] Finnish Social Science Data Archive (FSD) – Finland

Sharing machine actionable metadata increases the discoverability of scientific studies and promotes archival work. Metadata harvesting is a way to share such metadata on a large scale and to a wide group of interested parties and is typically achieved automatically with minimal human interaction.

CESSDA is developing a metadata aggregator service which will expose the metadata records in CESSDA Data Catalogue for re-harvesting via OAI-PMH. Together with the Catalogue, the aggregator will provide a comprehensive discovery solution for CESSDA Service Providers' metadata records. The aggregator will make records available in DDI-C, OAI-DC and Datacite. Initial targets are B2Find, OpenAIRE and Triple catalogues. The aggregator features OAI-sets to group records based on their originating archive and provides provenance information to track the origin of each record.

CESSDA and FSD partnered up to develop the aggregator system. FSD had prior knowledge of OAI-PMH and machine actionable metadata standards, in fact the aggregator software is based on FSD's Kuha2. While FSD is responsible of providing the development work, CESSDA is committed to operating the software and maintaining its runtime environment.

The presentation will outline the purpose of the software, main functionality, status of the project and immediate plans.

User Needs (14:00 CET)

Chair: Adrian Dusa - RODA

Zoom Link: https://sciencespo.zoom.us/j/94852706141

The DDI Alliance Training Group – Checkout what's new!

Jane Fry[1]

[1] Carleton University – Canada

The DDI Alliance Training Group has had a successful year and we want to tell you all about it. There are new Training webpages and new Zenodo communities with links to different DDI training materials and past DDI presentations. These resources are a result of work which has been done by our working groups.
One exciting development is the Zenodo community for DDI training materials: it is now available to use when you are conducting DDI training.

These materials have been created by DDI experts at a Schloss Dagstuhl Workshop in 2018 and in the DDI Training Group. The materials are intended to be reused, and they cover many different and relevant DDI topics. Other exciting new developments include a series of sprints and regular workshops hosted by CODATA. So come and be part of this presentation as you are taken on a tour of the different training resources available for you. It will be worth your while!

Listening to our data users: a qualitative study to better know their perception of our data and metadata

Quentin Gallis[1]

[1] Centre de données socio-politiques de Sciences Po – Sciences Po, Centre National de la Recherche Scientifique : UMS828 – France

In the open science "era", data centers are on the frontline of data sharing. Sharing data collected by itself and by and others, the French Center for Socio-Political Data (CDSP) faces multiple problematics concerning the usability of its services, especially those related to the FAIR principles. The CDSP has been using DDI to document its data since its creation in 2006. If the standard provides a rich documentation, up to the level of the variable, what are the other elements to be taken into consideration for a "good" data user experience?

Does DDI answer to all of the four letters of FAIR? In order to get a better grasp of the secondary data users' point of view, we have undertaken a qualitative study: we conducted semi-directive interviews with people having downloading data at least once in the last year from the CDSP. This study has taught us that the documentation of data and metadata is perceived as a great strength of our data, permitting optimal reuse.

However, if this fact is encouraging in regard to the academic world's efforts to open science, a practical limit emerges. This showed us that even if standards are a huge step to data dissemination, we still have efforts to do to make our data more accessible.

This paper will present the outcomes of our survey, as well as our next steps to make our DDI metadata and our data more findable and discoverable.

Interoperability 3 (14:00 CET)

Chair: Jon Johnson - CLOSER

Zoom Link: https://sciencespo.zoom.us/j/92669920880

DDI Registry upgrades and new HTTP based resolution

Dan Smith[1]
[1] Colectica – United States

The DDI Registry is a service run by the DDI Alliance that supports the registration and resolution of DDI agency ids. Agency ids are used in the creation of a DDI URN, and provide a unique namespace for all organizations using the DDI standards.

Since 2013 the DDI Registry has supported agency registration and DNS SRV based agency id resolution. There are many environments where DNS resolution services cannot be accessed, such as from within a browser using Javascript.

This presentation will outline the new HTTP based resolution feature of the DDI Registry for agency id service resolution. It will describe the service location options for resolution, give an overview of the new agency service discovery document, and provide a demonstration of the DDI URN to website resolution for web browser users.

Dissemination of contextualized statistical information: achieving interoperability between DDI-CDI and NGSI-LD

Franck Cotton[1], Arofan Gregory[2]
[1] INSEE – Institut national de la statistique et des ´etudes ´economiques (INSEE) – France
[2] Consultant – United States

INTERSTAT, a project funded under the European Connecting Europe Facility (CEF) initiative [1], aims at developing an open framework allowing the interoperability between national statistical portals and the European Data portal in order to build cross-border data services.

One way to reach this interoperability is to allow statistical open data to be made available through a key building block of the CEF infrastructure known as the Context Broker [2]. This would allow to disseminate statistical data in a way that can be contextualized in time and space, interoperable with other sources, and available via simple and standard web services for consumption on various platforms.

The core specification for the Context Broker is NGSI-LD [3], an ETSI [4] standard for context information (IoT, Smart Cities, etc.) which combines an information model and an API. To disseminate statistical information through the Context Broker, it is necessary to map statistical information models to NGSI-LD, in particular the DDI Cross-domain integration model and other foundational information models used in the statistical community.

INTERSTAT is working at building this cross-standard data model and a software implementation of this bridge capability. The project includes three concrete pilots that will allow to experiment and validate the results of this activity.

References:

[1] https://ec.europa.eu/inea/en/connecting-europe-facility
[2] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/Context+Broker
[3] https://www.etsi.org/deliver/etsi_gs/CIM/001_099/009/01.01.01_60/gs_cim009v010101p.pdf
[4] https://www.etsi.org

Plenary & EDDI 2022 Annoucement (15:30 CET)

Chair: Mari Kleemola & Jon Johnson

Zoom Link: https://sciencespo.zoom.us/j/99373030588

Author Index

Organisation Index