North American Data Documentation Initiative (NADDI)

2018 Annual Conference

Bureau of Labor Statistics, Washington, D.C.
April 3 - 6, 2018

PROGRAM

Tuesday, April 3th, 2018

	WELCOME TO TOWN
7:00pm	Informal get-together at; TBD

Wednesday, April 4th, 2018

		TRAINING SESSIONS BLS Janet Norwood Conference and Training Center
8:45am	9:00am	Registration & Packet Pick-up
9:00am	12:00pm	MORNING WORKSHOP
		Longitudinal research center in a box: Using DDI to enhance the mission of the UC Davis Alzheimer's Disease Center and the Midlife in the U.S. (MIDUS) study This seminar provides a gentle introduction to the Data Documentation Initiative (DDI) metadata standards for biomedical research data. Two NIA funded program projects, UCD ADC and MIDUS, applied DDI center-wide to document linked data and organize complex distributed data collection processes. We review why/how the application of DDI enhances research productivity and provide real-world examples how these research projects benefit from a technological standard that provides a basis for richly-structured metadata. We will also discuss some of the tools each center has created to implement DDI in a longitudinal health research context.

		Instructor: David K. Johnson, UC Davis Alzheimer Disease Center, Barry Radler, University of Wisconsin
12:00pm	1:30pm	Lunch - On Your Own – See options listed in your information packet
1:30pm	4:30pm	AFTERNOON WORKSHOP
		Document Questionnaires and Datasets with DDI: A Hands-On Introduction with Colectica This workshop offers a hands-on, practical approach to creating and documenting both surveys and datasets with DDI and Colectica. Participants will build and field a DDI-driven survey using their own questions or samples provided in the workshop. They will then ingest, annotate, and publish DDI dataset descriptions using the collected survey data. The course will cover the following DDI content areas: Questionnaire Design • Survey Instruments • Questions • Concepts and Universes • Question banks Dataset Documentation • Datasets and dataset layouts • Summary Statistics • Code Lists and Categories • Data concordance with RepresentedVariables and ConceptualVariables Attendees may optionally bring their own Windows laptops to participate in the hands-on exercises. *Instructors: Jeremy Iverson and Dan Smith, Colectica*
7:00pm		Informal get-together dinner at; TBD

Thursday, April 5th, 2018

		MAIN CONFERENCE - DAY 1 BLS Janet Norwood Conference and Training Center
8:00am	4:00pm	Registration, Packet Pick-up & Information
8:00am	9:00am	Hosted Continental Breakfast and Networking
		WELCOME AND KEYNOTE PANEL
9:00am	9:05am	**Welcome Remarks and Introduction of Keynote Speakers:** *Jared Lyle, Director, DDI Alliance*
9:05am	10:30am	**Keynote Panel:** *John M. Abowd, Associate Director for Research and Methodology & Chief Scientist, U.S. Census Bureau* *Edmund Ezra Day Professor of Economics, Cornell University* *Robert M. Groves, Executive Vice President and Provost,Georgetown University* *Director, U.S. Census Bureau, 2009-2012* *John H. Thompson, Executive Director, Council of Professional Associations on Federal Statistics (COPAFS)* *Director, U.S. Census Bureau, 2013-2017* *CEO, NORC at the University of Chicago, 2008-2013* *Maggie Levenstein, Director, ICPSR (Moderator)*
		MORNING SESSIONS
10:30am	11:00am	Break – beverages and snacks provided
11:00am	12:30pm	**Implementing DDI to Document the Consumer Expenditure Surveys** *Reginald Noël, Bureau of Labor Statistics* **Using DDI to drive data governance in the 21st century** *Claire Stent, Statistics New Zealand* **Moving from Compliance to Reproducibility: Metadata for Supplementary Research Collections** *Courtney Butler and Brett Currier, Federal Reserve Bank of Kansas City* **Leveraging metadata, DDI + other standards to implement data as a service at Statistics Canada** *Kathryn Stevenson, Statistics Canada* *Session Chair: TBD*
12:30pm	1:30pm	Hosted lunch and Poster Session - BLS Conference Center

		AFTERNOON SESSIONS
1:30pm	3:00pm	**Panel: Work to develop common standards for metadata across federal statistical agencies** *Warren Brown, Cornell Institute for Social and Economic Research, Cornell University*
3:00pm	3:15pm	Break – beverages and snacks
3:15pm	4:15pm	**Documenting and Publishing Statistical Data with Colectica and DDI** *Jeremy Iverson, Colectica* **Blaise and Colectica - Building on Metadata Standards** *Dan Smith, Colectica* **Using DDI to build open source solutions for curation and dissemination of microdata: World Bank Tools and experience** *Matthew Welch, The World Bank* *Session Chair: TBD*
4:30pm	5:00pm	**Optional Session: Review of Strategic Plan for DDI Alliance 2018 - 2021** *Jared Lyle, DDI Alliance*
7:00pm		Hosted Dinner - TBD

Friday, April 6th, 2018

		MAIN CONFERENCE - DAY 2 **BLS Janet Norwood Conference and Training Center**
8:00am	9:00am	Registration, Packet Pick-up and Information Continental Breakfast and Networking
		MORNING SESSIONS
9:00am	10:30am	**Panel: Perspectives on the Value of Metadata and DDI from the Perspective of Federal Funders** *James McNally, ICPSR, University of Michigan*

10:30am	11:00am	Coffee Break
11:00am	12:30pm	**Improving Roper@Cornell: DDI as a foundation** *Kathleen Weldon, Roper Center for Public Opinion Research* **Enhancing ICPSR metadata with DDI-Lifecycle** *Jared Lyle, ICPSR, University of Michigan* **A role for DDI in management of data as a record** *Claire Stent, Statistics New Zealand* **C2Metadata: Continuous Capture of Metadata** *Jeremy Iverson, Colectica* *Jared Lyle, ICPSR, University of Michigan* *Session Chair: TBD*
12:30pm	1:30pm	Lunch
1:30pm	2:30pm	**OpenCBA - a step towards management metadata and paradata** *Ingo Barkow, University of Applied Sciences HTW Chur* **Ricochet: Developing standards around biomedical reproducibility** *Cynthia Vitale, Ripeta* **Discovering and using administrative data** *Claire Stent, Statistics New Zealand*
2:30pm	2:45pm	Break
2:45pm	3:45pm	**DDI: Where we've been, where we're going** *Wendy Thomas, Minnesota Population Center* **A Sample Codebook in DDI4 XML** *Larry Hoyle, University of Kansas* **The DDI4 Collections Pattern** *Dan Gillman, U.S. Bureau of Labor Statistics (BLS)*
3:45pm	4:15pm	**DDI Driven Evaluation** *Barry Radler, University of Wisconsin* **Wrap Up**

Thank you for coming and safe travels!

ABSTRACTS

Wednesday, April 4, 2018

9:00 – 12:00pm: Workshop: Longitudinal research center in a box: Using DDI to enhance the mission of the UC Davis Alzheimer's Disease Center and the Midlife in the U.S. (MIDUS) study

Instructor: David K. Johnson, UC Davis Alzheimer Disease Center, Barry Radler, University of Wisconsin

This seminar provides a gentle introduction to the Data Documentation Initiative (DDI) metadata standards for biomedical research data. Two NIA funded program projects, UCD ADC and MIDUS, applied DDI center-wide to document linked data and organize complex distributed data collection processes. We review why/how the application of DDI enhances research productivity and provide real-world examples how these research projects benefit from a technological standard that provides a basis for richly-structured metadata. We will also discuss some of the tools each center has created to implement DDI in a longitudinal health research context.

1:00 – 4:00pm: Workshop: Document Questionnaires and Datasets with DDI: a Hands-on Introduction with Colectica

Jeremy Iverson and Dan Smith, Colectica

This workshop offers a hands-on, practical approach to creating and documenting both surveys and datasets with DDI and Colectica.

Participants will build and field a DDI-driven survey using their own questions or samples provided in the workshop. They will then ingest, annotate, and publish DDI dataset descriptions using the collected survey data. The course will cover the following DDI content areas:

- Questionnaire Design
    - Survey Instruments
    - Questions
    - Concepts and Universes
    - Question banks
- Dataset Documentation
    - Datasets and dataset layouts
    - Summary Statistics
    - Code Lists and Categories
    - Data concordance with RepresentedVariables and ConceptualVariables

Attendees may optionally bring their own Windows laptops to participate in the hands-on exercises.

Thursday, April 5, 2018

9:00am – 10:30am: Keynote

NADDI 2018 will sit a distinguished panel of guests who will open the conference on April 5th with a plenary panel on the importance of open standards in federal statistics and research.

Margaret Levenstein, Director of the Inter-university Consortium of Political and Social Research at the University of Michigan, will moderate the panel.

John M. Abowd is Associate Director for Research and Methodology & Chief Scientist at the U.S. Census Bureau, and Edmund Ezra Day Professor of Economics at Cornell University.

Robert M. Groves is Executive Vice President and Provost at Georgetown University, and was Director of the U.S. Census Bureau from 2009-2012.

John H. Thompson is the Executive Director at the Council of Professional Associations on Federal Statistics (COPAFS). He was the Director of the U.S. Census Bureau from 2013-2017, and CEO of NORC at the University of Chicago from 2008-2013.

11:00am – 12:30pm: DDI and Official Statistics

Implementing DDI to Document the Consumer Expenditure Surveys

Reginald Noël, Bureau of Labor Statistics
Daniel Gillman, Evan Hubener, Bryan Rigg, Arcenis Rojas, Lucilla Tan, Taylor Wilson (all BLS)

Documenting survey metadata for large-scale household surveys presents unique challenges that require a formalized approach. The Consumer Expenditure Survey Program (CE) has a three dimensional nature which adds an additional layer of complexity. For these surveys, major changes occur biannually and affect the processing subsystems and survey instruments. DDI provides a way to describe the multi-dimensional nature of the CE and allows the survey program to track these changes in a consistent, reusable, and replicable way. Implementing DDI will drive the CE toward the use of an international metadata standard and provide increased interoperability. Mapping CE metadata to DDI elements represents an ongoing challenge to developing the system. We propose a potential solution and explore the process of its implementation. The goals of DDI implementation include improving the administrative efficiency of the survey, facilitating centralized access to information about the current state of the survey, and providing a way for both survey staff and the public to easily access information about survey changes across time. This paper outlines the process and the challenges with implementing DDI into the CE framework, as well as the work that has been done, and the work that is left for full implementation.

Using DDI to drive data governance in the 21st century

Claire Stent, Statistics New Zealand

Managing the flow of data in National Statistical Offices (NSOs) and other organizations is becoming more complex. There can be multiple sources of data for a dataset, with multiple owners and differing levels of quality. Datasets can move through organizations in multiple ways. There is pressure to release data earlier than traditionally. This is too complex for the traditional data custodian model.

A Data Governance Framework which is being developed for New Zealand manages this new world by: instilling data governance capabilities for everyone and by capturing quality decision points throughout the data life cycle. DDI quality standards and statements can be used to record decision points throughout the data life cycle. Using DDI quality objects enables an organization to see which measures are being used by which steady states and for which datasets. It enables governance by requiring approval to move on to the next steady state. Identifying steady states also enables datasets to be shared at the most optimum steady state. DDI provides a practical way to implement this new approach to data governance throughout the data life cycle.

Moving from Compliance to Reproducibility: Metadata for Supplementary Research Collections

Courtney Butler, Federal Reserve Bank of Kansas City
Brett Currier, Federal Reserve Bank of Kansas City

The research community has begun making supplementary materials (such as research data and software code) more open and accessible in recent years. Journals have started to require that authors submit their supplementary materials along with their papers, and many authors make these items available for compliance purposes with little thought given to how those materials might be reused or what metadata might be needed to facilitate such reuse. However, data reuse has continued to become more prominent, and the growing popularity of newer research methodologies like systematic reviews, meta-analyses, and verification studies reflect that. As a result, more authors are starting to voluntarily make their data available, and researchers expect to be able to find and build upon existing datasets. Before, supplementary materials often took on the metadata of the papers they accompanied. Now they must stand on their own with their own descriptions to facilitate reuse. Data repositories and initiatives have started to spring up and provide some guidance (e.g., DDI). However, a pervasive standard has not fully emerged. This presentation will describe how metadata for traditional research publications differs from metadata for supplementary materials that is intended for primary discovery and reuse.

Leveraging metadata, DDI + other standards to implement data as a service at Statistics Canada

Kathryn Stevenson, Statistics Canada

Building on Statistics Canada's metadata-driven architectural principle and metadata strategy themes: drive, make available, structure and manage, the Picasso project is this NSI's solution for statistical data and metadata management. Automated business rules will ensure metadata is gathered uniformly, adhering to common architecture, governance and policy instruments.

Picasso, a three-year project, will be deployed into production in Summer 2018. It replaces a variety of local tools and processes with a enterprise hub for managing metadata for surveys, administrative files and record linkage projects; a data service centre function for 'fit for use' data files; and enterprise search and discovery using metadata to facilitate reuse of information. New tools and components include a metadata designer with an entity lifecycle management and registration process.

The solution architecture is based on a hybrid relational/semantic graph (RDF) core registry and repository with a data model driven by standard vocabularies, e.g. SKOS/XKOS, PROV-O, and reference models, e.g. GSIM, DDI 4 and SDMX. Picasso component and external systems interact with the RDF core via a Data Access Layer and Entity Services to access metadata entities via Common Information Exchange Models. DDI enables efficient sharing of metadata stored in Picasso with external users through the cross-country network of Research Data Centres.

12:30pm – 1:30pm: Lunch and Posters

DDIPy: A Python package to work with DDI file

Guinsly Mondésir, University of Ottawa

DDIpy is a Python package to work with DDI file. This package is a implementation of concept found in one paper* and a previous DDI poster titled DDIR: An R Package for Handling DDI Files. This package is an Open Source project. We believe that giving the developpers tools to help them creating open source related project will help the DDI communities further by providing maybe new concept, new software or web applications. The main function will be explained such as the DDIDataFrame(), the DDIParser() and DDItoSQL.

Citations:

Amin, Alerk; Barkow, Ingo; Kramer, Stefan; Schiller, David; Williams, Jeremy (2012) : Representing and utilizing DDI in relational databases, Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 191

Nakano Y. EDDI16 – 8th Annual European DDI User Conference, (2016), DDIR: An R Package for Handling DDI Files

GUILD - Graphical User Interface for Legislative Data

Valentin Pentchev, Indiana University Network Science Initiative
Matthew Hutchinson, Scott McCaulay, Patricia Mabry

PURPOSE -This poster describes the initial stages of developing GUILD, a database designed to enhance the utility of the Indiana General Assembly (IGA) data made public by the state's Legislative Services Agency (LSA). GUILD's enhancements will enable users to rapidly query and study the network structures that reflect how public policy is made in Indiana, and ultimately in other states in the Midwest and beyond. To our knowledge, this is a first-of-its-kind resource, unique in its representation of the legislative process at a highly granular level and in a graphical (network) format.

PROJECT DESCRIPTION - The raw IGA data is available from the state of Indiana containing information on Legislators, Committees, and Bills for the past three legislative sessions, IUNI-IT acquired the IGA from the LSA, and has parsed into a graph database information for every state legislator, every bill filed and every committee convened; including how these entities relate to one another. The network can be queried and filtered based on a range of criteria; returning results based on party affiliation, ranking position within a committee, key words etc. Outside the database, we have archived the full text of every bill, version, amendment and fiscal impact study filed as well as the Indiana legal code

CISER

Forthcoming

Colectica

Colectica is software used to document and publish statistical data using open standards. The software is used by national statistical organizations and major longitudinal studies worldwide. In the poster session, Colectica staff will be available for questions and demonstrations.

DDI Alliance

The Data Documentation Initiative (DDI) is an international standard for describing the data produced by surveys and other observational methods in the social, behavioral, economic, and health sciences. DDI is a free standard that can document and manage different stages in the research data lifecycle, such as conceptualization, collection, processing, distribution, discovery, and archiving. Documenting data with DDI facilitates understanding, interpretation, and use -- by people, software systems, and computer networks. Use DDI to Document, Discover, and Interoperate!

Controlled Vocabularies

This poster presentation will focus on updating the audience on a less-known product of the DDI Alliance, the Controlled Vocabularies, with a view to increase the visibility and usage of this valuable metadata resource among data users and curators.

Controlled vocabularies are structured lists of terms, or concepts that maybe used to standardize metadata content and thus enhance both resource discovery and metadata interoperability. A clear advantage presented by the DDI Alliance Controlled Vocabularies is that they are published independently of the DDI specification, and therefore may be used in

conjunction with any version of the DDI standard, but also with other metadata standards that may have a different structure and need not be expressed in an XML language.

Our poster presentation will include a brief review of the published vocabularies and our plans for the future, will familiarize the audience with their Web presentation and download, will discuss translations and the possibility of other agencies contributing to the vocabularies development, with the main goal of encouraging the vocabularies' widespread usage.

1:30pm – 3:00pm: Panel: Work to develop common standards for metadata across federal statistical agencies

Organizer: Warren Brown, Cornell Institute for Social and Economic Research, Cornell University

Moderator: Lars Vilhuber, Executive Director, Labor Dynamics Institute, Cornell University

Panelists: Barbara Downs, Director of the Federal Statistical Research Data Center (FSRDC) Program, Center for Economic Studies, U.S. Census Bureau; Maggie Levenstein and Jared Lyle – ICPSR; Catherine Fitch – Associate Director, Michigan Population Center, University of Minnesota; Representative of another federal statistical agency on FSRDC

This is a panel on work to develop common standards for metadata across federal statistical agencies. Barbara downs will present on current objectives and longer range goals of the FSRDC Technical Working Group to harmonize procedures of federal statistical agencies for researcher access to confidential federal statistical data in an RDC. Maggie Levenstein and Jared Lyle will present the new Census metadata portal that ICPSR is hosting. Catherine Fitch will present IPUMS work with the Census Bureau as a model of collaboration to be extended to other federal statistical agencies. Representative from a federal agency involved in FSRDC will present on efforts to harmonize procedures and improve researcher access to metadata. This panel is organized by Warren Brown of CISER and moderated by Lars Vilhuber of the Labor Dynamics Institute, Cornell University

3:15pm – 4:15pm: DDI Software and Tools

Documenting and Publishing Statistical Data with Colectica and DDI

Jeremy Iverson, Colectica

Colectica is software used to document and publish statistical data using open standards. The software is used by national statistical organizations and major longitudinal studies worldwide.

Colectica provides several tools:

- Colectica Questionnaires for specifying surveys in a standard way
- Colectica Datasets for documenting SAS, SPSS, Stata, and other statistical datasets
- Colectica Designer for documenting the entire data lifecycle
- Colectica Portal for publishing richly-documented data on the Web, with full variable-level lineage and concordance across time

This presentation will provide an overview of the tools, show how they are used in production at various statistical agencies and research projects, and will highlight new functionality available in 2018.

Blaise and Colectica - Building on Metadata Standards

Dan Smith, Colectica

Colectica and Statistics Netherlands announced a long-term partnership to build software linking Blaise, Colectica, and the DDI Lifecycle standard. This partnership has resulted in Blaise Colectica Questionnaires, a software system that allows survey researchers to build surveys faster, to leverage the DDI metadata standard, and to generate rich documentation and reports.

The first tool offers an intuitive survey design surface and questionnaire palette, allowing survey designers to build questionnaires without learning a domain specific language. Questions, blocks, and logic can be created within the program or reused from question bank powered by DDI. Reusing standardized questions assists in creating more comparable data and quicker survey development.

The software stores questionnaire specifications using the open DDI and GSIM standards, and can connect to metadata repositories and question banks powered by Colectica software. Data descriptions can be linked with source questions, creating harmonized data and showing data lineages.

Surveys designed with this tool can be fielded using Blaise 5 on the desktop, on the Web, and on mobile devices. The tool converts the DDI metadata into a Blaise project and source code. Changes to surveys tool can be published and executed within the Blaise environment, allowing rapid iteration while developing surveys.

Using DDI to build open source solutions for curation and dissemination of microdata: World Bank Tools and experience

Matthew Welch, The World Bank
Olivier Dupriez, The World Bank
Mehmood Asghar, The World Bank

The World Bank Data Group and the International Household Survey Network (IHSN), which it coordinates, are providing data discovery and management tools to producers of microdata in over 80 countries. These tools cover all phases of survey implementation, from survey design to data dissemination. Our tools include open source DDI (Codebook) compliant data curation tools, comprising: A Metadata Editor, Data Deposit Application and Data Dissemination Application. Our tools are being used in National Statistics Offices, the main data producers in developing countries, as well as increasingly by Universities and International Development Agencies. This presentation will cover the latest version of our data dissemination application, our data deposit application and our new Multi-Standard Metadata editor. Usage examples will include those at a large International Development Agency and National Statistics Offices.

4:30pm – 5:00pm: Review of Strategic Plan for DDI Alliance 2018 - 2021 - solicit feedback

The DDI Alliance is developing a new strategic plan for 2018-2021, which will succeed the 2014-2017 strategic plan. In this optional session, you will hear about the proposed new plan and have the opportunity to provide feedback.

Friday, April 6, 2018

9:00am – 10:30am: Panel: Perspectives on the Value of Metadata and DDI from the Perspective of Federal Funders

Moderator: James McNally
Panelists: representatives from NIA, NICHD and OBSSR

The use of metadata and DDI to effectively share information represents an important part of federal funding research. The National Institutes of Health (NIH) see well-developed metadata as an important tool to encourage data use in secondary analysis and in disseminating data to the user community. As DDI has become a production standard, NIH Project Scientists have become more knowledgeable in its use in data development. Three Project Scientists from the NIH will discuss ways in which DDI and, metadata more generally, are impacting research and playing a role in funding decisions. Having a federal perspective on DDI and the creation of metadata tools to support the use of secondary data is important as Project Scientists are playing an active role in discussions regarding best practices and future directions for data repositories. The presentation will provide DDI users and researchers developing tools to carry DDI into the next decade with an overview of NIH perspectives on high priority directions that will help us better meet NIH's mission and increase our ability to serve the community. Representatives from the OBSSR, NIA, and NICHD will offer perspectives on the value of metadata and its practical applications for funded research.

11:00am – 12:30pm: DDI and Data Management

Improving Roper@Cornell: DDI as a foundation

Kathleen Weldon, Roper Center for Public Opinion Research

The Roper Center for Public Opinion Research is the oldest social science archive, and the world's largest archive devoted exclusively to public opinion survey research data. Roper@Cornell holds over 23,000 datasets and offers iPOLL, a question bank with over 700,000 entries. After the Center moved to Cornell University in late 2015, a major rebuild of the archival structure and data model was undertaken to map the archive's metadata to DDI standards.

In this presentation, we will describe how we are using more comprehensive, normalized, and granular DDI metadata to drive improvements at Roper@Cornell. We will also explain and define how our efforts align the Center with the American Association of Public Opinion Research (AAPOR) Transparency Initiative. Our presentation will include a particular emphasis on historical data—polling data from the 1930s-1960s—and how the Roper Center is providing new tools, facilitating new analysis, and helping researchers develop better understanding of these important collections. We will show how Roper is using DDI to describe early and transitional periods of poll sampling, and how that effort can inform description of new polling methods.

Enhancing ICPSR metadata with DDI-Lifecycle

Jared Lyle, ICPSR
Sanda Ionescu, ICPSR

As the host institution for the DDI Alliance, the Inter-university Consortium for Political and Social Research (ICPSR) has invested heavily in DDI since the 1990s. One example of the investment may be seen in the ICPSR Social Science Variables Database, which uses structured DDI metadata to enable ICPSR users to examine and compare variables and questions across studies or series. The majority of ICPSR's data collections currently are described using DDI Codebook (DDI 2.5), which is intended primarily to document simple survey data. ICPSR is now taking steps to document selected collections using DDI-Lifecycle (DDI 3), which is especially useful for helping data users understand the relationships among waves of longitudinal data. This presentation will highlight the process of moving to DDI-Lifecycle for one pilot collection, including benefits and lessons learned.

A role for DDI in management of data as a record

Claire Stent, Statistics New Zealand

Datasets are records and require plans which detail how they will be managed throughout their life cycle. There is pressure on organizations to demonstrate stewardship of the data in their care. DDI records include all the information required to produce a Data Management Plan:

- Descriptive information for discovery (title, abstract, purpose, subject, concepts)
- Statistical information for discovery and re-use (variables, methodology, population)
- Retention, preservation and disposal information (disposal class, time period for the disposal action, access to the data)

Having this information in DDI enables data management plans to be updated as data moves through its life cycle and a plan to be created as a document at any time.

C2Metadata: Continuous Capture of Metadata

Jeremy Iverson, Colectica
George Alter, Pascal Heus, Jared Lyle, Ørnulf Risnes, Dan Smith

Accurate and complete metadata is essential for data sharing and for interoperability across different data types. However, the process of describing and documenting scientific data has remained a tedious, manual process even when data collection is fully automated. Researchers are often reluctant to share data even with close colleagues, because creating documentation takes so much time.

This presentation will describe a project to greatly reduce the cost and increase the completeness of metadata by creating tools to capture data transformations from general purpose statistical analysis packages. Researchers in many fields use the main statistics packages (SPSS®, SAS®, Stata®, R) for data management as well as analysis, but these packages lack tools for documenting variable transformations in the manner of a workflow system or even a database. At best the operations performed by the statistical package are described in a script, which more often than not is unavailable to future data users.

Our project is developing new tools that will work with common statistical packages to automate the capture of metadata at the granularity of individual data transformations. Software-independent data transformation descriptions will be added to metadata in two internationally accepted standards, the Data Documentation Initiative (DDI) and Ecological Markup Language (EML). These tools will create efficiencies and reduce the costs of data collection, preparation, and re-use. Our project targets research communities with strong metadata standards and heavy reliance on statistical analysis software (social and behavioral sciences and earth observation sciences), but it is generalizable to other domains, such as biomedical research.

1:30pm – 2:30pm: Describing Educational, Medical, and Administrative Data

OpenCBA - a step towards management metadata and paradata

Ingo Barkow, University of Applied Sciences HTW Chur

Data collection of large scale studies present a variety of different problems. Currently most vendors tackle problems in delivery while the institutions involved in data collection have massive management problems (e.g. interviewer control, proceedings in the household, sampling, weighting, quality control). OpenCBA is a new project for a metadata-based computer-based assessment platform derived from the remains of Rogatus, but heading more into handling the processes around data collection while using standardized process metadata or paradata.

Ricochet: Developing standards around biomedical reproducibility

Cynthia Vitale, Ripeta
Anthony Juehne, Leslie McIntosh

Metadata standards for multiple data types and formats have existed for a number of years, yet similar metadata standards for reproducibility are only just developing. Reproducibility metadata standards seek to ensure not only that the data are properly described and linked, but all components of the research, including any code, software, workflows, and more, are richly annotated, and FAIR.

This presentation will highlight the development of a reproducibility framework and software automation tool. The framework contains over one hundred metadata, initially selected and validated within the biomedical field. We will also describe and briefly demonstrate the software application that automates the detection of the framework variables, thus programmatically assessing the reproducibility of research within specific biomedical fields. Presenters will highlight future work on the reproducibility framework, new software extensions, and current adoption.

Discovering and using administrative data

Claire Stent, Statistics New Zealand

Data is not always collected by a survey. It can be a by-product of an administrative process. Administrative data can be stored in a “data lake”, which can make the data more difficult to describe within the DDI framework. Administrative data may not be accessible to researchers without a formal process to ensure identifiable information is safe. However, researchers need the metadata to know if it worth developing a research proposal.

In the example presented here, the data in the Integrated Data Infrastructure (IDI) has been grouped by subject. Using subjects as collections in DDI, enables the datasets to be seen as part of a group within a subject area or as individual datasets. Metadata can now be searched across the “data lake”, enabling researchers to find relevant datasets. Updating the metadata only requires changing the DDI.

Once researchers find the datasets, many prefer a data dictionary they can print off and use to develop their proposal. If they are working in our secure Datalab, they will need a print data dictionary as the internet is not available. Using DDI, users can print off a data dictionary with Stats NZ branding, date and copyright information. The dictionary provides information on the abstract, purpose, data collection and methodology, significant events affecting the dataset, concepts and variables. Data Dictionaries can be printed off for Series, Studies and Datasets and will be automatically updated when the DDI is changed.

2:45pm – 3:45pm: DDI Past, Present, and Future

DDI: Where we've been, where we're going

Wendy Thomas, Minnesota Population Center

In the 1990's DDI grew out of the need of archives to exchange metadata in a consistent format and support storage, discovery, and access systems. By 2000 we were looking at metadata driven data creation systems and reusing metadata to support comparison, quality control, and consistency. DDI is now addressing issues of Big Data, Data Lakes, and data derived from activities rather than planned collection. In this ever-evolving world of data, why is DDI special? What do we do that others do not? Is it important and where might it lead? A view of DDI from 10,000 feet.

A Sample Codebook in DDI4 XML

Larry Hoyle, Institute for Policy & Social Research, University of Kansas

This presentation will be a first look at a sample codebook serialized in DDI4 XML. The sample codebook is for a subset of variables from the Australian Election Study, 2013 - au.edu.anu.ada.ddi.01259 from the Australian Data Archive (ADA). The subset was first developed at the DDI4 Dagstuhl week 2 Sprint 2016 as an example of physical formatting of data. A complete codebook for the subset was begun at the DDI4 Dagstuhl Week 1 Sprint 2017 as a test case for the simple codebook functional view. It has proved useful in showing what can be represented in the DDI4 model, as well as showing the style of the resulting XML.

The presentation will include a walk-through of the original codebook from the ADA showing how each piece of information is represented in DDI4 XML.

The DDI4 Collections Pattern

Dan Gillman, U.S. Bureau of Labor Statistics (BLS)

One of the innovations in DDI4 is the introduction of the notion of a pattern, a set of abstract classes that can be realized by other classes. A class that realizes a class in a pattern has all of the properties and relationships of the pattern class, but may have additional ones as its use in a particular area of the model demands. This practice ensures that classes that have similar functionality have a similar structure, which should, in turn, make DDI4 easier to use.

The collections pattern allows for the description of simple unordered and ordered sequences and lists as well as complex orders including hierarchies, cyclical networks and directed acyclic graphs

The Collections Pattern has many realizations. It gives structure to workflows, variable groups, concept systems, code lists and classifications, processes, data records, and more.

This presentation will describe the Collections Pattern and how it has evolved over the last year. We will describe some of the pattern's realizations and will include examples of applications like the description of processes, statistical classification, and concept networks.

3:45pm – 4:15pm: Wrap Up and Evaluation

The wrap up session is used to demonstrate a DDI-driven conference evaluation. DDI metadata will be used to describe the evaluation form design, the data collection from attendees, and connect the collected data back to its source questions.