DDI Longterm Infrastructure Manifesto

1. Overview (Jared, Katja) -- motivation, purpose, benefits, what the document is/isn’t

Founded in 1995, the Data Documentation Initiative (DDI) has been used primarily by the international social science research community to document and describe data. In recent years, new user groups, including the official statistics and medical research communities, are exploring and using DDI. Our vision is for a large-scale distributed infrastructure for all of empirical research based on DDI. Embedding DDI documentation into the research lifecycle will lower costs of re-use (or use of DDI?) while increasing DDI users.

[placeholder for paragraph on infrastructure vision]

Behold! By outlining our overall vision, we hope to inspire the DDI community to build DDI tools and services encompassing the processes of the entire research lifecycle. Development typically comes in stages; this document provides an overall vision and connects the individual dots of development: support discovery, analysis, preservation, harmonization, and reuse to more easily enable future research. A later, although hopefully not too distant goal, is for this manifesto to inspire the larger research community to embrace and implement data documentation.

[placeholder here for a paragraph explicitly mentioning the stakeholders and briefly outlining needs]

This vision provides numerous benefits for the DDI community specifically, but also to the overall research community. One of the most important benefits is making it easier to use DDI by reducing implementation and coordination costs across the data lifecycle, thereby increasing overall usage of DDI. Even the small-scale producers will be able to use DDI by having tools that are easy to use that don’t require investment and overhead. By automating metadata capture across the data lifecycle through tools and services, metadata are improved and more complete.

Our vision supports the research process from the beginning of data collection to reuse (meta)data. Designing and implementing of the new data collections with metadata from conceptualization and questionnaire production to collecting data will be aided making it faster and more efficient. Also the data processing phase will be supported by metadata capture. Having the metadata from the beginning of the data collection phase will facilitate the long-term preservation and reuse of data. It will also increase the research transparency and reproducibility through audit trails.

Our vision is to ease data harmonization, comparison and combination and encourage interoperability and comparability across studies and even between different domains/disciplines. The increased amount of improved metadata will give an opportunity for new ways of data discovery. As mentioned previously, DDI community is international and covers variety of domains/disciplines. By utilizing related standards with DDI and building also on existing tools we will be able to implement envisioned multi-lingual infrastructure. We also wish to give credit to all of them who are producing items to this infrastructure.

Please note that while this document provides long-term vision for DDI Alliance work, we do not provide specific details about requirements or implementation, nor do take a position on which organizations or individuals should undertake particular aspects of the work. Rather, our hope is that a broad range of actors within the international (social) science data community--ranging from individuals to companies to large organizations--will be inspired by our vision and take up particular aspects of the outlined work.

We also do not feel that our vision is complete. The international (social) science data community should consider this vision a work in progress and subject it to further consideration, criticism, and extension where appropriate. Our hope is that our professional colleagues will take our statement of vision and build upon it--run with it, so to speak--in ways that advance the capabilities of data across the research lifecycle.

We lay out the building blocks and ideas for knitting together a total documentation package.

1. Vision for long term infrastructure for the social sciences (George)
   1. Example use case: Survey design
2. Integrating DDI into the Data Lifecycle
   1. Element registry
   2. Instrument Design
   3. Data transformation
   4. Archiving
   5. Data Discovery
   6. Data extraction
   7. Data merge
   8. Data analysis

An integrated metadata-based life cycle for survey data is illustrated in Figures 1, 2, and 3. We describe a workflow extending from survey design and ending with publication in which all of the metadata is seamlessly transmitted to the next stage by automated tools. The survey design example is particularly useful, because the workflow of data collection is already automated. Most surveys today are conducted with Computer Assisted Interview (CAI) software that captures responses directly. Although questions from the survey can be represented in a document as they would appear on paper, the actual questionnaire is digital. Data are transmitted directly from the CAI system to processing and eventually to analysis. However, if survey data are born digital, the metadata needed to understand them are dead on arrival. Today, metadata are created by humans, and the same metadata are often recreated several times at different stages of the data lifecycle. The metadata infrastructure of the future will eliminate these redundancies, produce more and better metadata, and offer new capabilities to each of the participants in the data lifecycle.

The research workflow of the future will have new tools and capabilities at each stage:

Study design:

* Survey Design Tool will provide templates and process descriptions to make it easy to produce standardized design documents.
* The Instrument Design Tool will create new questionnaires draw on the Element Registry, a repository of questions and response schemas. Questions will be discoverable by browsing Concept ontologies and by searching for similar items.
* Each question in the Element Registry will be linked to data, paradata, and publications from previous surveys, so that the designer can learn from previous research.

CAI implementation

* Output in DDI format from the Survey Design and Instrument Design Tools will be passed electronically to the CAI applications, which assures that the instrument in the field matches the design criteria.
* Modifications of the CAI application will be recorded in the DDI metadata as they occur.
* The CAI application will export all data, metadata, and paradata in standard formats.
* Paradata will become available for analysis and reuse with their own metadata in DDI.

Data preparation

* When data are modified, the Metadata Capture Tool will add data transformation metadata to an updated version of the DDI. The Metadata Capture Tool will parse scripts used for major software applications, such as SPSS and R, so that the metadata record exactly reflects the current state of the data.
* The Workflow Analyzer Tool will generate audit trails for every variable on demand to validate the data preparation process.

Data archiving

* Data repositories will receive data and metadata packages that are ready for preservation and distribution with minimal need for curation.
* Researchers will be able to search for datasets at the study and variable level across data repositories. Discovery services will use the Element Registry to enable browsing by concept and searching for similar questions.
* Electronic Codebooks will provide variable descriptions and provenance displays including workflow descriptions and variable-level audit trails.

Data analysis

* Researchers will use the Data Shape Changer to design new data objects that combining data from multiple studies. For example, a time series of opinion on the death penalty can be created by extracting identical questions from hundreds of surveys into a single dataset.
* The Response Harmonizer Tool will access Response Schema Mappings in the Element Registry to harmonize studies that coded responses differently.
* The Metadata Capture Tool will maintain updated metadata throughout the data analysis process. Electronic codebooks will be available on demand.

Publication

* Authors will deposit data and DDI metadata files to accompany their publications.
* Electronic publications will link readers to an Electronic Codebook, where they will find workflows and audit trails describing data transformations, and Online Analysis tools for reproducing published results.

This quick tour of the future data lifecycle introduces future software applications which rely on new public resources: the Element Registry, the Metadata Pond, and the Datum Pond.

Element Registry

The Element Registry will be a curated repository of data elements stored in DDI metadata. We use “element” to refer to any item in a dataset, including questions (“How old are you today?”), measurements (blood pressure), and other attributes. The Element Registry adds important features to the “question banks,” which already exist in several places.[[1]](#footnote-0)

First, the Element Registry assigns a unique persistent identifier (PID) to every element. PIDs provide assurance that the questions found in different datasets are in fact exactly the same.

Second, elements are linked to the concepts that they represent. Concepts can be linked to existing ontologies, and new ontologies can be built on registered concepts.

Third, the Element Registry includes a repository of “response schemas,” which are also assigned PIDs. It is not uncommon for different surveys to code responses to the same question in different ways, such as “1=yes, 2=no” versus “Y=yes, N=no”.

Fourth, the Element Registry includes a repository of “response mappings” that allow a machine to automatically recode two datasets to the same codes.

Finally, the Element Registry may include one or more “similarity indexes” that assign scores to the differences between elements. Similarity indexes may be based on text comparisons, and we anticipate that multiple indexes will be created as semantic technologies evolve. For our purposes, translations of questions into different languages can also be considered similarity indexes. These indexes will be incorporated into data discovery tools and can be used to develop new ontologies.

Registration of persistent identifiers (PIDs) for elements plays a central role in our model. PIDs provide assurance that the elements in different data collections measure the same thing. They also provide a convenient way to communicate between software applications. The Survey Design Tool can use element PIDs to describe a survey to a CAI application, because all of the metadata associated with each element can be obtained from the Element Registry.

The Element Registry will offer a number of services to the community. Data creators will be able to submit new elements, concepts, and response schemas to the registry, which will be validated and curated by Element Registry staff. Each of the component repositories in the Element Registry will offer discovery services.

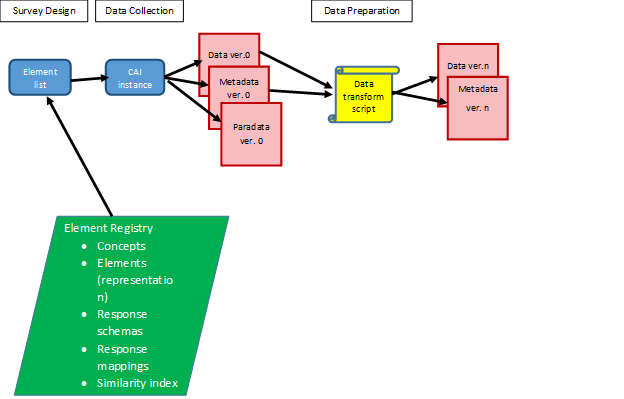
Metadata Pond

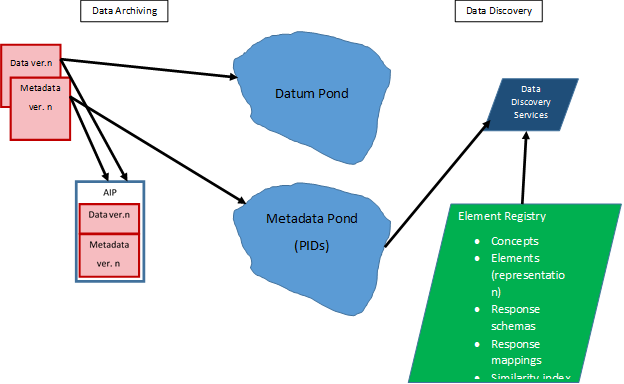
“Metadata Pond” is our metaphor for services that data repositories provide for automated searching of metadata describing their holdings. Users will want to search both the data collection- (study-) level metadata and element-level metadata. The Element Registry should offer a basic search capability in which a PID will elicit responses from all compliant data repositories. However, we expect that independent search applications will be developed that take advantage of the services of the Element Registry.

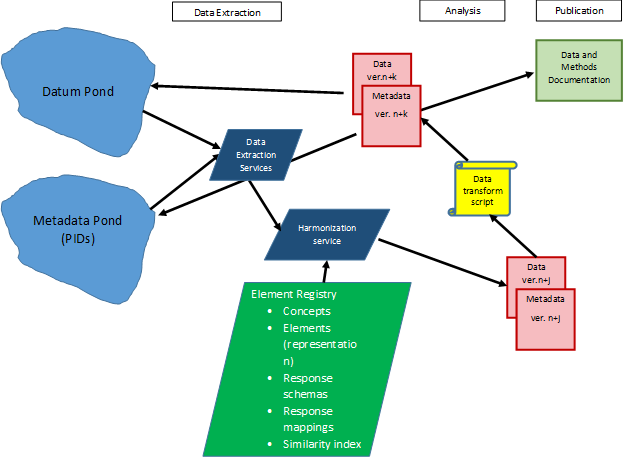
Datum Pond

The “Datum Pond” refers to services offered by data repositories that allow researchers to combine elements from multiple data collections into new datasets for analysis.

For example, suppose that one wants to study the relationship between attitudes toward the death penalty and inequality over time and internationally. Attitudes toward the death penalty are measured by questions (“Do you favor or oppose the death penalty for persons convicted of murder?”) answered by individuals. Inequality is an aggregate measure of the distribution of income in a population. We envision tools that will extract the relevant data from multiple locations and combine them into a dataset structured by individual or by country/time period, depending on the needs of the researcher. Standardized services provided by data repositories will allow researchers to request data identified by element PIDs, which will be accompanied by the necessary metadata describing the process that produced them. The Response Mapping repository is also essential to this process, because responses to the identical questions may be coded differently.



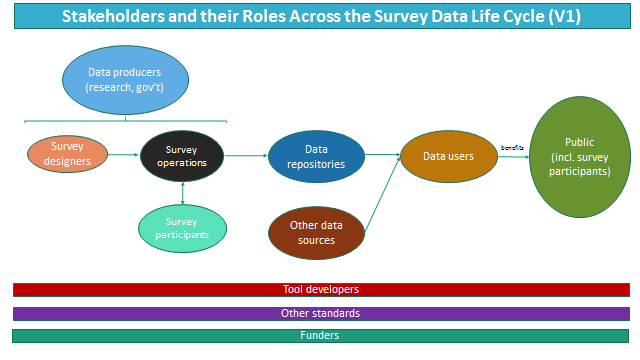


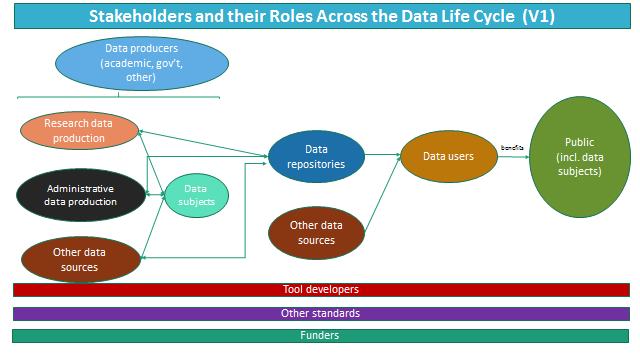


1. Lifecycles of other data types
   1. Administrative data (Katy?)
   2. Other data with existing metadata standards (e.g. EHR)
   3. Other object types: images, text, video, etc.

### Stakeholders

Stakeholder diagrams:





The data life cycle at its various stages involves numerous stakeholders who serve in complementary, interdependent, and sometimes overlapping roles. Moreover, an actor/stakeholder can play multiple roles (e.g., a researcher can be both a data producer and a data consumer; a government agency can produce both surveys and administrative data). A robust long-term [data? metadata? DDI?] infrastructure would facilitate and improve the work of all of these stakeholders and help meet a broad set of needs.

#### Data producers:

At the beginning of the life cycle are data producers. They are housed in a variety of settings (e.g., academic, government, NGO, or private sector) and collect data for distinct purposes (research, administration, business, or [placeholder for term to represent things like social media whereby people are putting out information about themselves]).

There are distinct differences in processes and infrastructure used between large- and small-scale data producers. Large-scale data producers (e.g., producers of large, long-standing surveys) are much more likely to use formalized systems for data collection and processing (e.g., CATI/CAPI software) and work on such a scale that they are able to invest time and resources into establishing and maintaining formal workflows, including metadata management. Many of these producers (as well as other types of stakeholders) already have been investing in the use and development of DDI and its related tools. Moreover, large-scale producers are much more likely to have distributed actors playing different roles: different organizations may conduct survey design vs. survey operations, and even within a survey design group one might have different staff members deciding upon the concepts to be covered by a survey from those who design the survey instrument. In those cases, it is of great benefit to be able to ease the transmission of information (requirements, data, and metadata) between stakeholders in the survey production workflow.

On the other hand, infrastructure is much less robust for small-scale data producers. Given the scale of their work, they are much less likely to reap the benefits of, and thus invest in, adopting DDI-related tools (which to date often have a significant entry cost). In addition, they often will be engaged in a diversity of data collection methods over time (as opposed to, say, being focused on a particular long-standing survey) and may be likely to collaborate with shifting set of collaborators. They would significantly benefit from an improved ability to document their research/create metadata as they do their work with minimal additional cost of using DDI (i.e., by integrating DDI into existing tools they use).

Data producers specifically conducting surveys need to be able to:

* Re-use existing survey components
* Design new survey components

However, many needs are common among most all types of data producers:

* Ability to transmit data and metadata among different roles within data production workflow
* Enable changes in measurement while maintaining comparability over time
* Ability to integrate various types of measures (quantitative, qualitative, biometric, open-ended responses, etc.)
* Transparency in the data collection process, including the ability to track and reproduce or replicate their work
* Greater efficiency in doing activities related to data collection and early-life-cycle data management
* Discoverability of their data
* Demonstrating use and impact of the data they produce

Looking at ways in which improved infrastructure can better meet the needs of producers has potential benefits in quality and efficiency of stakeholders further on in the process.

#### Data repositories:

Repositories which provide access to the data are the desired next step in the workflow. Some repositories are specific to a particular domain (subject and/or format) whereas others are more general. And repositories vary greatly in their levels of curation they provide. \**This infrastructure plan is designed around repositories focused on social science data which have professional staff dedicated to curation activities.* Such repositories take data from a range of producers. The most common format historically collected has been the survey, repositories are taking an increasing diversity of data; therefore, this proposal envisions infrastructure for the survey in particular but also begins to address requirements for a wider set of data types.

Common infrastructure needs among data repositories include:

* Provide access to data at various levels—e.g., open, safeguarded, and secured--given issues of licensing and confidentiality.
* Enable discovery and analysis by end-users, both within systems they maintain (e.g., repository catalogs) and external systems (e.g., systems which harvest metadata--e.g., a virtual metadata pond--as well as link data and publications)
* Preservation (for some)
* Richer metadata accompanying deposits, as well as more to be added/tracked throughout their workflows (which will improve ease/speed of processing pipeline and provide a higher-quality product)
* Demonstrate use and impact of the data they make available
* Ability to build upon/learn from the work/systems/tools that others have done to better leverage resources
  1. Data users:: Secondary data users (note: includes academia and beyond (public sector, non- and for-profits, etc.) and people with various levels of skill and expertise
     + N: Discovery (known-item and by features (topic, time, geography, etc.)) across multiple data sources and repositories; includes discovery at the variable level
     + N: Analysis (remote and on desktop) (given varied skill levels)
     + N: Linking and combining datasets from different sources (esp. in new ways)
     + N: Comparing change (over time, geography, etc.)
     + N: Reproducibility/transparency (incl. ability to track their work)
     + N: Using data as part of teaching research methods
     + [placeholder for referencing data producers as data users as well]
  2. Funding agencies
     + Large-scale national agencies
       1. National Science Foundation (NSF)
       2. National Institutes of Health (NIH)
       3. Deutsche Forschungsgemeinschaft (DFG)
       4. Economic and Social Research Council (ESRC), and potentially others, within the broader umbrella for Research Councils (Department for Business, Energy and Industrial Strategy)
       5. Swiss National Science Foundation (SNF)
       6. Swiss Commission for Technology and Innovation (CTI)
       7. Fill in others here; especially international examples
     + Large-scale international agencies
       1. European Commission (EC) - especially Horizon 2020 and Eurostars program
       2. Organisation for Economic Co-operation and Development (OECD)
       3. World Bank
     + Ministries
     + Private Research Foundations
       1. Alfred P. Sloan Foundation
       2. Wellcome Trust (esp. as relates to biomedical)
       3. Fill in others here
     + Other Government
       1. Institute of Museum and Library Services
       2. Other
     + Universities
  3. Data subjects/survey participants
     + N: confidentiality
     + N: demonstration of public benefit resulting from their participation
  4. Other standards:
     + N: having DDI complement their standard by documenting aspects they don’t cover
  5. (DDI) Tool/service developers (often would live in one of the aforementioned stakeholder organizations):
     + N: Having their tools widely known and adopted
     + N: Awareness of other tools upon which they can build
     + N: Finding collaborators to co-create tools when a need is shared by multiple organizations
     + N: Long-term hosting and maintenance of tools
  6. Members of the public:
     + N: to gain a benefit for society from data gathered and used
  7. Other communities

1. Services and Tools (Ingo)

Metadata standards are usually only successfully implemented if there are sufficient services and tools available which build up on them. A key factor for the success of this vision is therefore to motivate organizations in embedding these elements into new or already existing software products or to provide tools for converting DDI elements into their own proprietary formats. The following software packages or tools have been identified to play a key role in the overall survey workflow.

* 1. Survey Design Tool

The Survey Design Tool should support research organizations in setting up the organizational structure of the survey like describing the concepts, the expected sample sizes, the different institutions involved, the quality measures, the milestones of the project and internal workflows. Ideally it can be used to report the progress of the survey back to the funding agencies. In lifecycle models like the Generic Statistical Business Process Model (GSBPM) or Generic Longitudinal Business Process Model (GLBPM) these processes fall into the “Evaluate” category at the very beginning of a survey project.

* 1. Instrument Design Tool

The Instrument Design Tool should enable researchers to create instruments ideally via a graphical user interface by re-using elements like questions, response domains or controlled vocabularies from the element registry. Researchers can use search functionalities to check in the element registry if a suitable component for their design is already available. The output from this tool can be used in Computer-Aided Interview (CAI) systems which are used to capture data from questionnaires in the field. CAI systems are very well established in data collection agencies and normally consists of the following components:

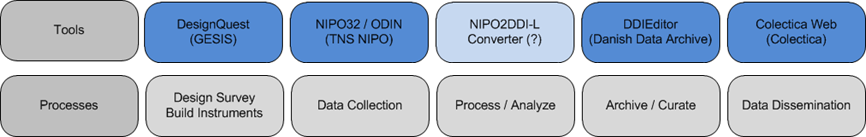
* + - Questionnaire Design Language or a basic Graphical Questionnaire Designer (e.g. does not include loops and cannot access an Element registry)
    - Survey Management System / Case Management System (e.g. to handle disposition codes, case assignment, sample management, interviewer assignment, interviewer tracking, synchronization mechanisms)
    - Logging mechanism / Audit trail
    - Reporting / Field Monitoring
    - Export of the results (data and paradata) into formats of statistical packages (e.g. SPSS, Stata, R, MPlus)

The use of an advanced Instrument Design Tool is the basis for reusing items or harmonization between different studies and conveys potential for huge cost-saving effects. As the instrument designer can also preview the questionnaire as it can be rendered to look like a survey instrument directly from the elements in the standard the time for instrument development is also shortened. Instead of transferring questions e.g. as Word documents or Excel tables from instrument designer to questionnaire programmer (which is the common workflow for systems like Blaise or MMIC) a less complex workflow can be established where the researcher is in full control of the instrument. Ideally the questionnaire programmer becomes obsolete and will only be needed for complex scenarios like rotating questions or loops within loops. This also means there is no need anymore to learn CAI-specific questionnaire description languages as the output of the Instrument Design Tool can be imported into the CAI system or in the long run there might be CAI systems which build upon the standard itself (an example for this is Rogatus Survey - a prototypical open source CAI system which is using DDI Lifecycle 3.2 in the backend). In the latter case the Instrument Design Tool is a module of the overall DDI-based CAI platform.

The Instrument Design Tool should also support different CAI modes which are the following:

* + - Paper and Pencil Interview (PAPI) – paper questionnaire conducted in house by an interviewer
    - Computer-Assisted Personal Interview (CAPI) – computer-based questionnaire conducted in house by an interviewer (Examples: Blaise, MMIC, TNS Nipo, SPSS Dimensions, CASES, Rogatus Survey)
    - Computer-Assisted Web Interview (CAWI) – web survey filled out by the participants themselves (Examples: Limesurvey, Surveymonkey, Redcap, Google Forms)
    - Computer-Assisted Self Interview (CASI) – computer-based questionnaire filled out by participants in a facility, sometimes observed by audio or video
    - Computer-Assisted Telephone Interview (CATI) – computer-based questionnaire conducted by an interviewer via phone (Example: Voxco)

This also means that the Element registry should contain elements for all CAI modes including paper&pencil interviews. Ideally the CAI system should seamlessly integrate with the Instrument Design Tool. This can be reached if the CAI system is built on top of the standard or is able to import or export the DDI format. Alternatively there can be tools like DDI-to-CAI and CAI-to-DDI converters which transform the questionnaire in DDI into the proprietary format of the CAI system and converts the output (metadata, paradata and data) back into DDI (metadata) and statistical package formats like SPSS, Stata, R or MPlus (paradata, data). The standard thus enables the creation of metadata toolchains between different tools like the following:



Unfortunately these and other tools do not support standards or are lacking interfaces to exchange metadata and data. Very often this leads to scenarios where the same data has to be curated in multiple systems with elaborate conversion procedures or manual work.

* 1. Data creation
  2. Documentation tools
  3. Need to define format for Datum Pond
  4. Need to define format for Metadata Pond (PIDs)

1. Strategies for realizing the vision

We anticipate a variety of strategies will be instrumental to realizing our vision. First, given the amount and the complexity of work to be completed, it will be important to build upon existing tools and projects whenever possible. Indeed, much of our vision consists of work that needs to occur *between* already-existing capabilities, entities, and functionality. For example, when fully realized, our vision expects a robust DDI-based exchange between CAI-supported data collection efforts and the data processing stage of the data lifecycle. In some specific instances such functionality already exists--see, for example, Colectica’s open source Blaise-to-DDI metadata converter.[[2]](#footnote-1) In the vast majority of instances where CAI applications are used today, however, an easy exchange from CAI to DDI (and back again) does not yet exist and needs to be created.

Second, our vision will be most fully realized by utilizing related standards whenever possible. While “standards” in this context is broadly understood (meaning it extends beyond metadata standards to software standards, good practice standards, etc.,), a good example of DDI working cooperatively with another metadata standard is shown in the work of SDMX (Statistical Data and Metadata eXchange). SDMX is an international effort to standardize and modernize the mechanisms of exchanging statistical data and metadata. DDI and SDMX have long been understood as being complementary, not competing standards.[[3]](#footnote-2)

and [Building on existing tools and projects

* 1. Utilizing related standards
  2. Research and demonstration projects
  3. How to engage stakeholders

1. Next steps
   1. Grant application components
   2. Publications
   3. Coordination

* Tools needed for Element Register
  + Submission
  + Curation
  + Discovery
  + Survey Design

1. See ICPSR’s Social Science Variables Database (<http://www.icpsr.umich.edu/icpsrweb/ICPSR/ssvd/index.jsp>) and CESSDA’s Euro Question Bank (<http://cessda.net/About-us/2016-Work-Plan/Euro-Question-Bank>). [↑](#footnote-ref-0)
2. http://www.colectica.com/news/Open-Source-DDI-Converter-Project [↑](#footnote-ref-1)
3. See, for example, the paper by Arofan Gregory and Pascal Heus on this topic, available at: <http://www.opendatafoundation.org/papers/DDI_and_SDMX.pdf>, as well as the curated UNECE wiki page <http://www1.unece.org/stat/platform/display/metis/Existing+resources+related+to+the+relationship+between+SDMX+and+DDI>

   [↑](#footnote-ref-2)