DDI Long Term Vision[[1]](#footnote-0)

George Alter

Ingo Barkow

William Block

Jared Lyle

Steven McEachern

Katherine McNeill

Katja Moilanen

Joachim Wackerow

# Purpose

\*\*\*\*\*\*\*\*\* Introductory statement on purpose moved to context [document](https://docs.google.com/document/d/1ytCi8ttkC5eiQ3NHsIzDosQWO_jEN0_JgaswpSk0KmE/edit). \*\*\*\*\*\*\*\*\*

The DDI community has surpassed the two-decade mark in developing and maintaining DDI products. The first specification was for codebooks and addressed needs to document datasets employing XML that would represent printed codebooks, which traditionally was associated with a dataset at the conclusion of a project. XML permitted rendering this codebook in a variety of formats, including the capability to publish on the Web. This specification still satisfies the needs of many users, especially those who primarily need a way of sharing documented datasets.

The second specification introduced a lifecycle perspective, meeting requirements beyond the DDI-Codebook. It provides a way to describe the processes through which data move from planning, collection, processing, analysis, dissemination, and preservation. In addition to documenting a more comprehensive record for data, this descriptive content serves as valuable input driving further processes both within and outside the project in which it was produced. As a consequence, this specification broadens the reuse value of this metadata. Not only can the metadata be used to reexpress this information in different formats, it can serve as the input to machine-actionable tasks.

The third specification is DDI-Model and features new functionality enabling the description of data working from a model of concepts and relationships that can readily employ bindings to current technologies and new technologies in the future. Global digital research infrastructure is changing the way researchers conduct their work. Increasingly, a variety of online services are essential to conduct research and, as a consequence, there is a need to develop workflows moving across infrastructure services for which metadata play a critical role. DDI-Model will increase the capacity for technical interoperability of DDI products on digital research infrastructure.

The first two specifications have addressed user requirements over the years and remain for many as relevant today as in the past. However, each specification has addressed new needs while building on the previously one. In the context for DDI-Model, it has added functionality to support services operating on digital research infrastructure. This is especially important given the ecosystem in which data repositories exist and operate in silo-like services. Researchers are looking for easy ways to find and extract data from a network of data repositories without having to deal with each one separately. Data archivists are in need of exposing the data collections in their repositories to the global community of researchers and are in need of infrastructure services to facilitate this. Data producers require services that help them harmonise their data collection instruments at the point of design instead of a stage later in production. Publishers increasingly are requiring researchers to deposit their data with the submission of research articles and are in need of simplified ways to verify compliance with this condition for publication. Research councils and universities are concerned about the reproducibility of research findings and are in need of practices that support the deposit and ready access to publication-related datasets. All of these user groups require solutions within the global digital infrastructure supporting research. This is an important opportunity for DDI-Model to support key services providing solutions to the above requirements that are listed.

---- old stuff below here ----

will facilitate data reuse and enable new kinds of discovery, and analysis. A later, although hopefully not too distant goal, is for this manifesto to inspire the larger research community to embrace and implement data documentation.

The types of problems for which solutions are being sought often involve complexities that require multiple stakeholders working together. For example, assessing research impact requires aggregating indicators of researcher contributions to knowledge over a variety of scholarly communication channels and the subsequent use of such knowledge by other researchers. Some of the stakeholders engaged in the process of measuring impact include researchers, publishers, citation aggregators, libraries, funding councils, universities, and standards organisations. Researchers are looking to be rewarded for strong impact measures; publishers seek higher impact rankings for their journals; citation aggregators want higher usage of their indexes; libraries contribute supportive metadata; funding councils want a good return on their research investments; universities look to advance their research reputation; and standards organisations contribute ways to define, discover, and use appropriate information. Digital research infrastructure that provides the information technology platforms and tools needed to collect, manage, aggregate, and disseminate metadata and data has become the field on which stakeholders, collaborating with each other, are finding solutions to common problems.

There are various stakeholders in our vision who will make use of the envisioned infrastructure. The data producers as survey designers and survey operators will benefit most from the study design, CAI implementation and data preparation tools and aids. Methods for improved data archiving will benefit data repositories, while data discovery, analysis and publication tools will be an advantage for data users. And funders and governments will benefit from a more efficient research process and infrastructure and enhanced possibilities for analyzing the impacts of research funding.

This vision provides numerous benefits for the overall research community as well as the DDI 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 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 new data collections with metadata from conceptualization and questionnaire production to collecting data will be faster and more efficient. Also, the data processing phase will be supported by metadata capture. Retaining metadata from the beginning of the data collection phase will facilitate the long-term preservation and reuse of data. It will also increase research transparency and reproducibility through audit trails.

Our vision will ease data harmonization, comparison and combination and encourage interoperability and comparability across studies and even between different domains/disciplines and countries. More and better metadata will create opportunities for new kinds of analysis and data discovery. As mentioned previously, the DDI community is international and covers a variety of domains and disciplines. By coordinating with related standards and building on existing tools, DDI will be able to implement a multi-disciplinary, multi-lingual infrastructure. We also wish to give credit to all who are contributing 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 we 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 claim 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. We encourage our professional colleagues to take our vision and build upon it--run with it, so to speak--in ways that advance the capabilities of data across the research lifecycle.

# Functionality

In the future researchers should be able to find and access comparable data in different places, and they should be able to combine types of data collected in different ways. Integrating data across sources requires detailed metadata about the meaning as well as the location of data and the access rights. One cannot do comparisons across time or space if the data are not measuring the same thing. For example, if we want to know how attitudes about crime vary across countries and over time, we must be assured that we are measuring these attitudes in the same way in every place and time. In other words, merging and integrating data from disparate sources is only possible when the data are well described. The DDI Alliance is providing fundamental standards and tools infrastructure that will enable new kinds of research with distributed data, but we need new infrastructures to realize their potential. In particular, we need to embed DDI metadata at all stages of the data life cycle to assure the completeness and quality of metadata and to reduce the costs of data collection, archiving, and analysis.

What kinds of functionality are needed to realize the full potential of a data rich environment?

First, we need tools that make it easier and cheaper to document data. Metadata should be captured automatically as data are created and updated as the data are transformed. Most data are ‘born digital.’ They are created by computer assisted interviews, online surveys, administrative information systems, sensors, etc. Data are managed in other digital systems such as databases and statistical analysis packages. Metadata tends to be an afterthought that is created manually in text documents and spreadsheets that have no connection to the software used to create, transform, and manage the data. Consequently, metadata is incomplete, and documentation is often re-created several times for the same dataset at different stages of the data life cycle. We should integrate the capture of metadata into all data creation and management processes and transmit metadata in parallel with data. Since data creation and management is entirely digital, characteristics of the data, data transformations, and data management processes can be harvested directly from those tools. Automating the capture of metadata will produce better documentation, save time for data creators and researchers, and save money. [structured by a standard]

Second, we need a way to determine when two datasets are measuring the same thing. Small variations in questions can have big effects on responses. Surveys administered internationally are translated into multiple languages. How do we know that variables with the same or similar names were collected in the same way? Even if the same question was used, different responses might have been allowed, such as “Yes/No” versus “Yes/No/Maybe.” Harmonizing variables after data collection is expensive and often ineffective. Re-using established questions and response schemas ensures comparison (“harmonization by design”) and saves resources. International organizations that must produce comparable data devote significant resources to ensuring that measures collected in different countries can be compared. We propose services, “data element registries,” that will assure data harmonization by assigning unique persistent identifiers to variables. These registries will be community institutions that researchers can use to find and re-use existing data items and to register new ones for others to re-use. [interoperability needs to be based on a standard]

Third, datasets should not live in isolation. We should have tools that extract data from multiple sources and combine them into new customized datasets. We often want to add contextual data (population size, unemployment rate, income distribution, rainfall) to individual-level data, or to create time series of diverse measures for national, state, or local entities. If two sources of data use standard geographic codes, we should be able to link and combine them in an automated way. Experienced researchers and business analysts develop skills for extracting and combining data, but new tools would expand the ways that students and the general public can use data. Citizens should be able to use government data to answer questions, even when the data have been created by different government agencies.

-- role of standards

-- need for partners

-- role of community

 3. Use Cases

Data Design

Data collection instruments are not created in a vacuum.

Researchers beginning a new project should use DDI-based tools to identify variables and data relevant to the concepts that they wish to study. They should be able to search a catalog of alternative ways that previous researchers have attempted to measure these concepts, and the catalog should lead to datasets that will allow them to examine descriptive statistics and correlations with other variables. DDI metadata should be a way of documenting not only the variables that are collected but also the concepts that the researcher intended these variables to measure.

From Design to Implementation

When a researcher has selected variables, there should be an easy way to arrange them into an instrument and communicate the proposed instruments to a data collection agency. The data collection agency should not need to re-enter information about the variable into its own system or wonder about the wording of questions or acceptable values of responses. All information about the survey should flow electronically from the researcher into forms, websites, and automated interview systems.

Data collection agencies should have systems that automatically capture and record data collection protocols and changes in these protocols. Paradata about the data collection process should also be standardized and documented for management and future analysis. At the end of the data collection process the data should be accompanied by a full metadata record that includes information about the conceptual development of the project and the data collection process.

Documenting Data Transformations

“Raw” data almost always requires processing before it can be analyzed and shared. Identifiers and flags included for purely administrative purposes are dropped. Some variables are modified to reduce the risk of re-identifying subjects. Other variables are aggregated into scales and indexes. All of these transformations should be automatically captured by software and recorded in the metadata.

Data Discovery Tools

The processed dataset will be transferred directly to a data repository, where other researchers can find and reuse it.[[2]](#footnote-1) The data curation costs will be greatly reduced, because the data will come with complete and detailed metadata. Researchers looking for data will have a new tools for data discovery. They will be able discover all data including a specific version of a variable by searching for a persistent identifier uniquely assigned to that variable. This will encourage data re-use, meta-analysis, and relieve researchers of the heavy burden of harmonizing variables across datasets.

Data Extraction, Merging, and Analysis

Application developers will take advantage of DDI metadata to create new ways to combine and view data from several sources. Use Cases

### Multilinguality

In social sciences, cross-national research projects collect data in many countries and in different languages. From designing the study to reusing the data, the researchers are facing many data management challenges related to multilinguality. They will benefit from systems that help manage translations of research instruments, harmonize the variables and enable multilingual searches to retrieve information. Today, a great deal of this work is carried out manually using very basic tools like spreadsheets, which means a loss of efficiency, probably also loss of information, and waste of resources.

The researchers translate the research instrument into various languages, manage the language versions, and document the relationships between the question items: what are the different language versions of the same question item and what are the differences or the degree of similarity between language versions. In a perfect world, all this information is documented in a way that allows it to be carried into the next stages of the research process and data lifecycle and used for automated harmonization of questions or variables documented in different languages. (For example, education, marital status,..)

DDI provides Controlled Vocabularies in several languages to describe different aspects of a dataset across the data lifecycle. Using these vocabularies for metadata enables searching across collections and languages, and also enable automatic translation of contents.

Researchers (re-users?) will be searching for datasets containing specific concepts, measures, question items or variables. For example, suppose that one is planning a survey and would like to see how a certain concept has been measured earlier, what have been the questions used in the survey. If they can’t find what they are looking for in their preferred language, they should have ways to search questions documented in other languages. And if they find the question in another language, they should be able to translate that into their preferred language and link that back to the original question, with information about the degree of similarity. Another similar use case is when a user analysing a group of variables would like to know if there are similar datasets from other countries or in other languages that could be used for a comparative study.

Research data infrastructures working in multilingual environments, like CESSDA, need to create multilingual services to enable discovery of studies and datasets created in various countries and in various languages. For example, it is unreasonable to translate all variable descriptions into English so both automatic translation via controlled vocabularies and discovery of content in multilingual catalogue are essential. Another issue is that the same data (like Eurobarometer data or ISSP data) is available from multiple CESSDA Service Providers, and typically each has documented the data in their own language at least in study level. These relationships between studies need to be defined so that they can be shown to the researcher searching for data.

### Research Replication/Reproducible Research/Academic Publishers

The academic publishing enterprise is one use case that encompasses much of the data lifecycle and involves multiple stakeholders (including researchers, journal editors, and data archivists), all of whom have an interest in building efficient workflows enabling research to be replicated and reproduced with relative ease. Increasingly, journals are requiring authors of accepted articles to submit data and code that will allow future researchers to reproduce the reported results, and, indeed, in some cases the journals themselves are performing the replication work to make sure that the reported results are, in fact, reproducible as a final step before publication (see, for example, the case of the American Journal of Political Science: <https://ajps.org/guidelines-for-manuscripts/>).

The work involved in reproducing work can be daunting, especially for authors who have never faced this requirement in the publication process. The effort involved usually requires much iteration between the article’s author and a highly trained analyst familiar both with the software and analysis techniques commonly used in social science research. Hours of effort by the analyst and the author are usually required before articles are deemed fully reproducible.

From the perspective of the researcher the (relatively new) emphasis on reproducible research represents an additional required step along the path from scientific discovery to publication. From the perspective of the journal editor, however, the requirements add both effort and expense, because validating the reproducibility of findings is not free.

Thus in this use case, both parties--representing the beginning and end of the data cycle, have incentives to make reproducible research easier and to steamline the process of verifying the ability of files to replicate.

# Overview - [old introduction]

Founded in 1995, the Data Documentation Initiative (DDI) has been adopted by social science data repositories around the world to document and describe data in a machine-actionable form. In recent years, new user groups, including the official statistics and medical research communities, have been exploring and using DDI. We argue that realizing the full benefits of DDI requires a reconceptualization of the entire data life cycle. Metadata creation should be fully automated and integrated with study design, data creation, distribution, analysis, and publication. Every activity in the data life cycle should be documented as it occurs from conceptualization to publication. Today, data documentation is usually an afterthought, and DDI metadata is created by repositories at the end of the data life cycle. The result is costly, inefficient, and incomplete. The expansion of DDI infrastructure described here will improve the accessibility and usability of data across all research disciplines and reduce costs as well. This document lays out ideas and building blocks for knitting together a total documentation package.

The heart of our idea is called an “Element Registry” which is a curated repository of data elements (e.g., variables) with persistent identifiers (PIDs) stored in DDI. Data elements (e.g., “How old are you?”) are linked to the concepts that they represent and to response schemas. The Element Registry will also contain mappings to harmonize alternative response codings, and tools for finding related data elements, such as similarity indices. Other services and tools will rely on the Element Registry to embed DDI documentation across the data lifecycle. In this vision we describe tools for study design, computer-assisted interviewing (CAI) implementation, data preparation, data archiving, analysis, meta-analysis, and publication.

# 2. Envisioning a Long Term Infrastructure for Social Science Data: Survey Research

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:

Metadata should be recorded in DDI during the design phase of a project before any data are collected. Research is a team effort with specialized roles, and documenting the intended research process in DDI creates a record that can be shared with partners.

* A 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 by drawing 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.

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.

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.

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 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.

CAI implementation:

* Output in DDI format from the Survey Design and Instrument Design Tools will be passed electronically to the CAI application, which assures that the instrument in the field matches the design criteria.
* Modifications of the CAI application will be recorded in DDI metadata as they occur.
* The CAI application will export all data, metadata, and paradata in standard formats.
* Paradata (data describing the data collection process) is rendered into a data file with its own DDI metadata, so that it can be analyzed to inform future studies. As survey designers increasingly rely on paradata, the research community will adopt minimum paradata standards that all survey operations are expected to produce.

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 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:

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 combining data from multiple studies. For example, a time series of public 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 DDI-enabled data lifecycle introduces future software applications that rely on new public resources: the Element Registry, the Metadata Pond, and the Datum Pond.

# 3. Integrating DDI into the Data Lifecycle

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. Concepts can also be indexed and linked to data elements in the Registry. The Element Registry adds important features to the “question banks,” which already exist in several places.[[3]](#footnote-2)

1. 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.
2. 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.
3. 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”.
4. Fourth, the Element Registry includes a repository of “response mappings” that allow a machine to automatically recode two datasets to the same codes.
5. 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 pdescribe 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. We use “datum” to refer to the information generated by a single measurement, response, or event for a particular unit of analysis (person, country, year, etc.). By focusing on these elementary particles of information we draw attention to new ways of combining and reshaping data for analysis. Researchers have always concatenated, joined, and restructured data from different sources to create datasets for analysis. Our goal is to make discovery of data elements much easier and to create new services and tools for automating the process of creating new data structures. The new idea here is that elements within large data collections will inherit contextual and provenance metadata so that they can be accessed by automated services.

For example, suppose that one wants to study the relationship between attitudes toward the death penalty 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. Standardized services provided by data repositories will allow researchers to request data identified by element PIDs, and these data will arrive with metadata describing the process that produced them. We envision tools that will extract the relevant data from multiple locations and combine them into a dataset structured to the needs of the researcher. The Response Mapping repository will provide an automated way to harmonize responses when identical questions have been coded differently. This scenario can be extended to include other types of variable. If each respondent has been coded with a geographic location, we could add crime rates from other datasets to ask “Has the relationship between crime and attitudes about the death penalty remained constant over time and space?”

Our example was not chosen at random. The American National Election Study (ANES) and the General Social Survey (GSS) have asked exactly the same question about the death penalty since 1990 and 1972 respectively, but combining those data remains a tedious manual process. First, one must discover which waves of each survey contained questions about the death penalty. Second, questions must be compared to verify that they used the same or equivalent wording.[[4]](#footnote-3) Third, the data must be harmonized. In this case, GSS coded ‘favor’ as ‘1’ and ‘oppose’ as ‘2’ while ANES used ‘1’ for ‘favor,’ ‘5’ for ‘oppose,’ and sometimes added ‘3’ for ‘depends.’

Researchers discover, reshape, and combine data from multiple sources every day, but this process should be much easier. Countless studies have been abandoned because the data management phase was too difficult or time consuming. We don’t expect our students to write their own programs to create crosstabs or regressions, why should they be writing elaborate scripts to create new combinations from datasets that are fundamentally the same?

Actuality and Potential

DDI was originally developed to describe survey data, but we are just beginning to realize its potential. Communication from survey designers to survey operators, from survey operators to data processing, and from data processing to data repositories is typically done in text documents and spreadsheets. At the end of this chain the data repositories, who invented and adopted DDI, create structured metadata from whatever they receive. Data repositories have been using DDI for preservation, data discovery, and codebooks, and they have been implementing new tools, such as online codebooks and variable-level searching. However, the documentation and services provided by data repositories are limited by the incomplete metadata that they receive from data producers. Important information about a questionnaire, such as the order of questions, the internal logic of questions in the instrument (“skip patterns”), and data transformations are not available and too costly to reconstruct from static (e.g. pdf) representations of the questionnaire. Moreover, only a human can determine whether two surveys used the same question or coded the question in the same way. Harmonization across surveys is a laborious time-consuming process.

Figure 1.



Figure 2.



Figure 3. 

# 4. Lifecycles of Other Data Types

Our vision for a future in which metadata are continuously and automatically captured, updated, and transmitted across the data life cycle is not limited to the case of survey data described in the example above. Many of the insights that inform our vision have implications for other types of data, such as health records, administrative records, and data describing objects like images and texts. While the processes generating these data types differ, all of them can benefit from the new tools and services that we have described for data processing, discovery, dissemination, and analysis. Indeed, parts of this infrastructure are already in place in some fields. The DDI Alliance should work closely with other communities defining standards for describing and sharing data to assure that data can be transferred without information loss between standard representations.

The world of biomedical data offers important examples where data standards and ontologies have been integrated in the data creation process. The International Classification of Diseases has been used for more than a century, and biomedical researchers have created ontologies for many specific subjects and purposes. Recently, the U.S. National Institutes of Health has been encouraging the use of Common Data Elements to standardize the collection of data, and initiatives like Logical Observation Identifiers Names and Codes (LOINC) are providing standards and tools. There has also been rapid progress in the development of standards for describing and exchanging biomedical data, such as Clinical Data Interchange Standards Consortium (CDISC) and Fast Healthcare Interoperability Resources (FHIR). In some cases ontologies have been imposed for administrative convenience. Health insurance agencies, such as Medicare in the U.S., require hospitals and other health providers to use controlled vocabularies for diseases and treatments, which results in enormous quantities of data with standard terminology and codes. In effect, the biomedical world already has an infrastructure of “element registries” that are actively used in data creation. Unlike the social sciences, which have few agreed upon ontologies, the number and diversity of biomedical ontologies is itself a problem, which projects like biosharing.org and BioPortal (<http://bioportal.bioontology.org/>) are helping to coordinate. When biomedical data exist in standard formats and metadata, they can be translated directly into DDI for combination with other data and analysis.

The growing demand in social science research for administrative and transactional data created by governments and businesses poses especially difficult metadata problems. The systems from which these data are derived were created for other purposes, and the documentation that social scientists need may not be relevant to the objectives of the data owners. Since these data are often acquired directly by researchers, there is a need for better tools to help researchers create metadata from the documentation that they do receive. We can also hope that demonstration projects with the DDI community and research on provenance in computer science will generate new standards for documentation of administrative data systems.

DDI should also forge alliances with emerging standards for describing digital objects that are not within scope for DDI itself, such as images, videos, and text. Social scientists are rapidly incorporating other types of data into their research. A growing number of major data collections, like MIDUS and HRS/SHARE, include physical measures (e.g. blood pressure), blood and other specimens, genomes, and MRIs or other images. The research communities developing around these digital objects are increasingly aware of the need for their own standards and workflows. We do not advocate developing a way for DDI to describe brain scans, but we point out that measurements derived from brain scans, such as hippocampal volume, can be combined with other quantitative data in DDI datasets. Quantified attributes of any digital object can be described in DDI along with a persistent identifier pointing to the object itself.

# 5. Stakeholders

Data infrastructure involves a range of inter-related stakeholders:





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) and roles may have overlapping needs. A robust long-term data infrastructure, underpinned by DDI, would facilitate and improve the work of all of these stakeholders and help meet a broad set of needs. A main feature of a successful infrastructure: metadata is created as a byproduct of the work at all phases in the life cycle--created at the point in the process related to that activity--and thus can be more efficiently created and reused in the long-run.

It will be vital to engage stakeholders in two-way exchanges, discussing and contributing to this proposed infrastructure, so that the Alliance can have the most robust set of partners in advancing this, and takes forward an infrastructure that is optimal for the various stakeholders. A range of strategies would serve to engage a variety of stakeholders:

* Presentations and discussions at conferences which discuss the vision and how it could meet the needs of particular stakeholders
* Targeted outreach to leadership of professional associations
* Publishing the vision in journals and newsletters
* Designated mechanisms to get input and feedback from stakeholders (e.g., conference workshops, surveys, interviews) to refine and improve the vision
* Proposed pilot projects with various stakeholders, to develop and test requirements and infrastructure

In the sections below we outline additional particular strategies for engaging with specific stakeholder types.

#### 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 social media).

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. Moreover, large-scale producers are much more likely to have distributed actors playing different roles: different organizations may conduct survey design or 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. Many of these stakeholders are aware of DDI, but they have not yet seen enough benefits to build DDI into their work processes.

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 invest in DDI-related tools directly. These stakeholders would benefit if DDI were built into standard tools, like Survey Monkey and Qualtrics. In addition, they often engage 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, both for their own efficiency and for institutional, funder, or publisher requirements.
* 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 produced

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.[[5]](#footnote-4) 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.

Data discovery and data preservation, the core services of data repositories, are built on metadata, and data repositories have been the most supportive of DDI and other metadata standards. Data curation (i.e. creating metadata) is the most costly part of data archiving. Data usually arrive at the repository in a statistical package with minimal metadata. Important information, such as question text and questionnaire design, is provided in a separate document (often a pdf) or not at all. Richer metadata accompanying deposits will improve the speed and lower the cost of data curation and provide a higher quality product. If data repositories received complete and accurate DDI metadata from data producers, they could redirect their own resources from creating metadata to providing other services that benefit the research community.

#### Data users:

The raison d'etre of a research data infrastructure is to enable use by researchers, now and in the future. The primary audience that comes to mind is that of secondary data users, yet a robust infrastructure can support re-use of their own data by data collectors as well. An effective infrastructure will serve data users from a variety of settings (academia, public sector, nonprofits, for-profit companies, and the general public) and accommodate users with various levels of skill and expertise in working with data. The primary needs of data users are the following:

* Discovery of data: An optimal infrastructure would enable users to discover data across multiple data sources and repositories, at various levels of granularity (e.g., study or variable), and in different ways depending upon the research need (e.g., known-item searchers vs. discovery by characteristics (topic, time, geography, relationship to other concepts, etc.).
* Data documentation: Researchers need to understand how data were created (provenance metadata). Currently, important information about study designs and execution and the management of variables is not recorded. Data users are often referred to other documents, like questionnaires, to find out which subjects answered a particular question. More complete and accurate DDI metadata will allow data providers to present this information in new kinds of online services.
* Once the data is discovered, systems must provide access as appropriate (open, safeguarded, or controlled), based on the context of the user as well as the sensitivity of the data at various levels of granularity.
* Such systems will need to facilitate linking and combining datasets from varied sources in new ways that are either impossible or difficult to do at present; such linking will enable the creation of new kinds of knowledge. Data should be able to be linked based on the individual unit of analysis (e.g., through linking among administrative sources) or on common characteristics (e.g., geography).
* Once access is granted, researchers need to analyse the data to make meaning. Systems should allow such analysis to be done either locally for the user (by downloading one or more data files) or, increasingly, online in the environment of the data publisher. Increasing features for online analysis will enable analysis of data by a greater variety of users, including those who lack the local infrastructure or skills to do statistical analysis, and enable linked analysis of data from multiple sources. Data users will need to do various types of analysis, such as descriptive or inferential statistics, comparing change (over concepts such as time or geography).
* As researchers do their analysis, whether it be locally or on the side of the publisher, systems must support their ability to document their work and use of the data, both for their own efficiency and for institutional, funder, or publisher requirements. These requirements are increasingly proliferating in an effort to improve the management of data and reproducibility and replication of research results.
* An additional context for use is incorporating use and analysis of data as part of teaching research methods.

#### Funding agencies:

Funding agencies, both public and private, entities have a direct interest in the developments described here, because they will impact both the costs of data creation and the re-use of existing data. A core benefit of this improved infrastructure is greater efficiency, whereby metadata creation is automated and re-used across the life cycle. Such re-use holds great promise to minimize friction in the system and maximize the amount of research benefit that comes from data collection and preservation efforts. Furthermore, the infrastructure outlined here improves mechanisms for keeping track of the usage, outcomes, and impact of investments throughout the research infrastructure. In our vision funding agencies will see more science for each dollar invested in data creation, and they will know more about how their data are being used.

Current data creation workflows are inefficient and ineffective, because metadata created and re-created manually by data producers, data managers and data repositories. A continuous DDI-based workflow will eliminate redundant steps and produce more complete and accurate metadata. For example, DDI-based survey design will improve communication between survey designers and survey operators, reducing the costs of fielding new surveys. In other words, automating the creation of metadata across the data life cycle lowers costs and produces a better product.

#### Funding agencies will also be greatly interested in the improved comparability and harmonization that DDI-based data creation will enable. It will be easier to build new data collections that are directly comparable to earlier studies and to conduct meta-analyses across time and space. We also expect other agencies to follow the lead of the use National Institutes of Health in encouraging the use of designated “common data elements” that assure comparability across data collected at different times and places.

Funding entities with an interest in data infrastructure on the social sciences are many, and sit within several broad categories:

1. Large-scale national funding agencies, which have an interest in the efficiency of research and research infrastructure within their countries (some of which may have a dedicated research infrastructure funding agency or focus on funding information science as a discipline) e.g.,
	1. Australia: Australian Research Council (ARC)
	2. Germany: Deutsche Forschungsgemeinschaft (DFG)
	3. Netherlands: Netherlands Organization for Scientific Research (NWO) and Royal Netherlands Academy of Arts and Sciences (KNAW)
	4. U.K.: Economic and Social Research Council (ESRC)
	5. U.S.: Institute of Museum and Library Services (IMLS), National Science Foundation (NSF) and National Institutes of Health (NIH)
2. Government ministries, many of which not only fund research or research infrastructure, but also engage in data collection efforts
3. Large-scale international agencies, which have an interest in the efficiency of research and research infrastructure across country borders, e.g.,
	1. European Commission (EC), including Horizon 2020, Eurostars, and European Strategy Forum on Research Infrastructures (ESFRI) programs
	2. Organisation for Economic Co-operation and Development (OECD)
4. Private Research Foundations, e.g.,
	1. Alfred P. Sloan Foundation
	2. Gates Foundation
	3. Wellcome Trust
5. Universities: these provide baseline funding and hosting for activities throughout the life cycle, including: data collection and primary research; storage, preservation, and access; and secondary research

In addition to the general needs described earlier, funding entities have particular interests in an infrastructure which effectively can:

* Provide measurable impact of the aspects of the research infrastructure which they fund
* Gain efficiencies by facilitating exchange of information and data
* Enable new types of research (for their funded researchers) across borders, disciplines, and data types

Engaging with funding agencies will involve targeted outreach to funding agencies to discuss various points of relevance: potential avenues for funding the creation of this infrastructure; the way entities see the infrastructure fitting into their vision of their support for research; and how the infrastructure can be best designed to meet their needs in a landscape of changing research funding. This should be done by engagement both with individual entities and also with multi-agency organizations (e.g., Association of Charitable Foundations, Council on Foundations, European Research Council, Research Councils UK).

#### Members of the public (including data subjects):

In addition to being represented in the above categories, members of the public have additional benefits and requirements. In general, all of society stands to gain from research, based on data, which creates knowledge to improve society and solve its complex problems. This is especially important for those many members of the public who (as individuals or members of broader entities) are subjects of data (research or administrative), who merit a demonstration of public benefit resulting from their participation, which may be increased in an environment of easier data reuse. Moreover, key to issues that must be addressed (and discussed with the public) is that of confidentiality of their information, especially given that a new infrastructure has the potential to significantly increase the sharing and linking of personal data.

A robust public engagement program will be key to both the creation, implementation, and maintenance of such an infrastructure. The Alliance can learn from many such past efforts in planning such a program.[[6]](#footnote-5)

#### Other metadata standards:

DDI underpins the effectiveness of this envisioned infrastructure. Yet at the same time, the DDI stands to benefit from the existence of other distinct and related metadata standards, which can be deployed to complement it and document aspects of the infrastructure not well covered by the DDI. Such standards include those which describe:

* Information objects at a high level (e.g., Dublin Core, MARC)
* Statistical data (GSIM and SDMX)
* Preservation (METS and PREMIS)
* Geography (ISO 19118
* Metadata Registries (ISO/IEC 11179)
* Qualitative data (QuDEx)
* Data in related disciplines (various examples can be found on the DCC directory of Disciplinary Metadata[[7]](#footnote-6))

For the infrastructure to interoperate well, it needs to engage with other standards from the beginning, in order to ensure that systems are built to interoperate in the most appropriate ways. Experts in other standards should be involved in early dialogue and pilot projects throughout its creation. Moreover, as the infrastructure evolves into the future, the Alliance should monitor and engage with new and developing standards as they emerge.

#### DDI Tool/service developers:

Finally, many stakeholders exist within the DDI community whom are key participants in, and contributors to, this proposed infrastructure. Primary in this community are tool and service developers, many of whom sit within one of the aforementioned types of stakeholder organizations (e.g., data producers or archives). Developers have several areas in which they can benefit from this infrastructure, including having their tools widely known and adopted and the ability to build upon others’ tools.

Yet they also have particular needs which systems must meet, including:

* Awareness of other tools upon which they can build
* Finding collaborators to co-create tools when a need is shared by multiple organizations
* Long-term hosting and maintenance of tools
* Systems which facilitate development and exchange of information (rather than add a multitude of additional requirements)

In summary, an effective infrastructure for a complex research environment will depend upon effectively, understanding, engaging with, and incorporating the needs of this diverse set of stakeholders.

# 6. Strategies for realizing the vision

We anticipate a variety of strategies will be instrumental to realizing our vision.

First, this document can be viewed as a roadmap that addresses needs at every stage of the data life cycle. We do not see a single solution to every problem. Rather, we see many gaps that need to be filled and many areas in which development is needed. The C2Metadata (“Continuous Capture of Metadata”) is an example of project that is addressing a specific gap in the metadata workflow.[[8]](#footnote-7) Funded by the US National Science Foundation, C2Metadata is a partnership of two data repositories (ICPSR and Norwegian Centre for Research Data), two independent software producers (Colectica and Metadata Technologies North America), and two major surveys (American National Election Study and General Social Survey). The outcome of the project will be software applications that can read scripts from the four main statistical packages (SPSS, SAS, Stata, and R) and incorporate data transformation information into metadata in two widely used standards (DDI and Ecological Markup Language).

Second, 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.[[9]](#footnote-8) 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.

Third, 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.[[10]](#footnote-9)

1. This working paper was written 23-27 October 2017 at Schloss Dagstuhl as part of a [DDI sprint](https://ddi-alliance.atlassian.net/wiki/pages/viewpage.action?pageId=39911463). [↑](#footnote-ref-0)
2. Data access will be subject to conditions protecting the rights of subjects and data creators. [↑](#footnote-ref-1)
3. 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-2)
4. ICPSR’s “ANES/GSS Crosswalk” (<http://www.icpsr.umich.edu/icpsrweb/ICPSR/taxonomy/view/312/605102?actionType=view>) is an example of a DDI-based tool that provides facilitates data discovery and offers side-by-side comparisons of variables across different data collections. [↑](#footnote-ref-3)
5. Further iterations of the infrastructure could expand to involve a greater variety of data repositories. [↑](#footnote-ref-4)
6. See <https://adrn.ac.uk/research-impact/public-engagement/>, <http://www.esrc.ac.uk/public-engagement/public-dialogues/>, <https://wellcome.ac.uk/what-we-do/our-work/public-engagement-and-trust>, and <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5103425/>. [↑](#footnote-ref-5)
7. <http://www.dcc.ac.uk/resources/metadata-standards> [↑](#footnote-ref-6)
8. http://c2metadata.org/ [↑](#footnote-ref-7)
9. http://www.colectica.com/news/Open-Source-DDI-Converter-Project [↑](#footnote-ref-8)
10. 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](http://www1.unece.org/stat/platform/display/metis/Existing%2Bresources%2Brelated%2Bto%2Bthe%2Brelationship%2Bbetween%2BSDMX%2Band%2BDDI)

 [↑](#footnote-ref-9)