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 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 DDI. Our vision is for DDI to be the basis of building a large-scale distributed infrastructure for all of empirical research. Embedding DDI documentation into the research lifecycle will lower costs of re-use while increasing DDI users.

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.

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 through tools and services, metadata are improved and more complete.

Our vision supports the research process from the beginning of data collection to re-use. Designing and implementing of the new data collections from conceptualization to questionnaire production will be aided making it faster and more efficient. Data prosessing phase will be supported by metadata capture. New data harmonization, comparison and combination tools will encourage interoperability and comparability across studies and even between different domains/disciplines. Findability of the data will be improved by new data discovery tools. DDI community is international and covers variety of domains. By utilizing also related standards with DDI and building on existing tools we will be able to implement this multi-lingual "environment"/system.

- and analysis tools (us and other ones to build)
- Credit for producing items in the life cycle
- Reproducibility
-

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.

2. Vision for long term infrastructure for the social sciences (George)
    a. Example use case: Survey design

A continuous, 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 production 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 exists *in silica*. 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 currently stillborn. 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.

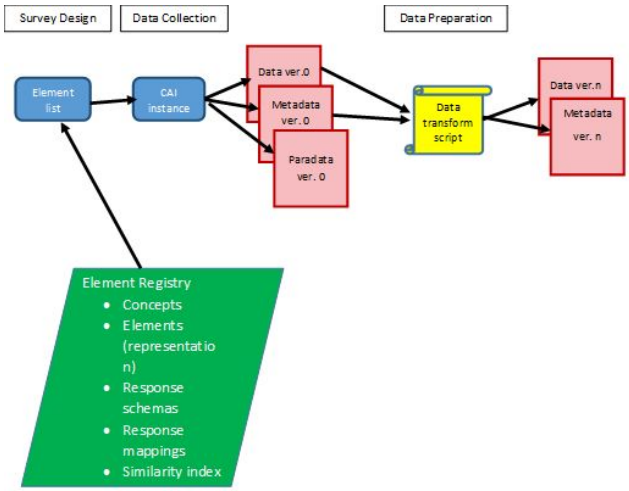
Our tour of the data lifecycle begins with new resources for survey design. At this stage a researcher wants to turn concepts and hypotheses into questions and instruments. The survey designer wants to know: "How has this concept been implemented in the past?" "What were the results of previous surveys?" "How long does it take for the average subject to answer this question?" We offer a new resource for answering these questions -- the Element Registry.

The Element Registry is 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] 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. 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.

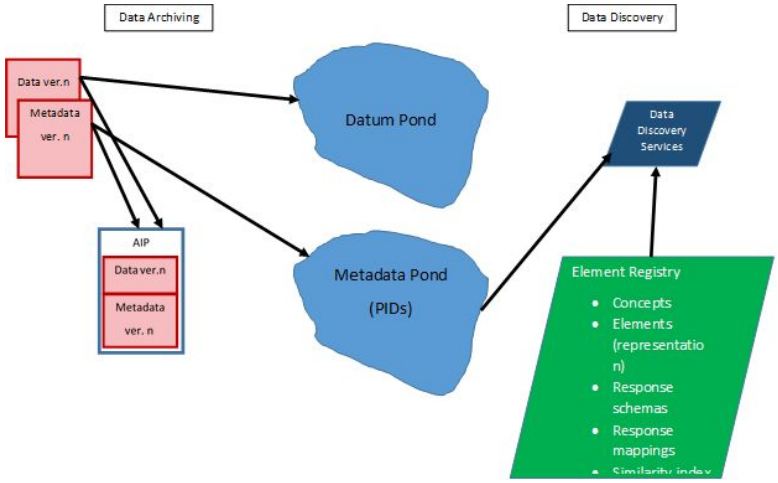
A survey designer may begin by browsing the Concept list in the Element Registry to discover how previous researchers have measured the concepts in her study. She may also use a

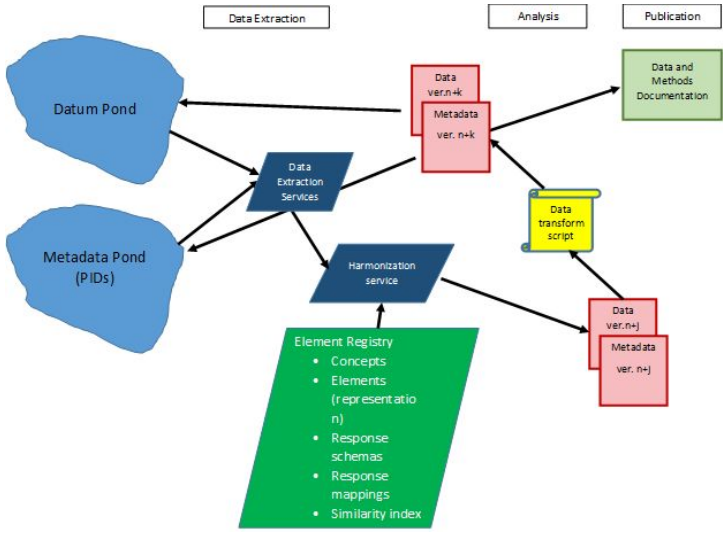
[1] See ICPSR's Social Science Variables Database (http://www.icpsr.umich.edu/icpsrweb/ICPSR/ssvd/index.jsp) and CESSDA's Euro Question Bank (http://cessda.net/About-us/2016-Work-Plan/Euro-Question-Bank).

similarity index to create a list of closely related elements that deserve further consideration. PIDs make it possible to find datasets that have already used a specific question as well as publications based on those data. Since data repositories expose their metadata in the "Metadata Pond" (see below), a "PID Discovery Tool" will return a list of datasets that include each question, as well as publications that have used each dataset. The PID will also link to paradata from previous studies. Paradata provides information about the administration of a survey, such as the average time for a response and the number of non-responses, which allow later surveys to build on previous experience. Thus, the Element Registry provides valuable information for designing new surveys and encourages researchers to reuse successful questions.



Having decided on a list of questions for the survey, the researcher will use the "Survey Design Tool" on the Element Registry to view her questionnaire.


Data Archiving
Data Discovery
Data ver.n
Metadata ver. n
Datum Pond
Data Discovery Services
AIP
Data ver.n
Metadata ver. n
Metadata Pond (PIDs)
Element Registry
Concepts
Elements (representation)
Response schemas
Response mappings


Data Extraction
Analysis
Publication
Datum Pond
Data ver.n+k
Metadata ver. n+k
Data and Methods Documentation
Data Extraction Services
Data transform script
Metadata Pond (PIDs)
Harmonization service
Data ver.n+j
Metadata ver. n+j
Element Registry
Concepts
Elements (representation)
Response schemas
Response mappings
Similarity index

3. Integrating DDI into the Data Lifecycle
    a. Element registry
    b. Instrument Design
    c. Data transformation
    d. Archiving
    e. Data Discovery
    f. Data extraction
    g. Data merge
    h. Data analysis
4. Lifecycles of other data types
    a. Administrative data (Katy?)
    b. Other data with existing metadata standards (e.g. EHR)
    c. Other object types: images, text, video, etc.
5. Stakeholders (Katy) (includes types (T), needs (N), how to engage them (E)
    a. Researchers
        - T: Producers: large-scale (may overlap w/survey designers and/or survey operations below)
            1. N: Reproducibility (incl. ability to track their work)
            2. N: Greater efficiency in doing their research
            3. N: demonstrating use and impact of the data they produce
            4. N: Discoverability of their data
        - T: Producers: small-scale
            1. N: Ability to document their research/create metadata as they do their work w/minimal additional cost of using DDI (i.e., integrating into existing tools they use)
            2. N: Reproducibility (incl. ability to track their work)
            3. N: Greater efficiency in doing their research
            4. N: demonstrating use and impact of the data they produce
            5. N: Discoverability of their data
        - T: Secondary data users (note: includes academia and beyond (public sector, non- and for-profits, etc.) and people with various levels of skill and expertise
            1. N: Discovery (known-item and by features (topic, time, geography, etc.)) across multiple data sources and repositories; includes discovery at the variable level
            2. N: Analysis (remote and on desktop) (given varied skill levels)
            3. N: Linking and combining datasets from different sources (esp. in new ways)
            4. N: Comparing change (over time, geography, etc.)
            5. N: Reproducibility (incl. ability to track their work)
            6. N: Using data as part of teaching research methods
    b. Survey designers

- T: Academic and government
- N: Re-use of existing survey components
- N: Design new survey components
- N: enable changes in measurement while maintaining comparability over time
- N: ability to integrate various types of measures (biometric, open-ended responses, etc.)

c. Survey operations
   - T: may/may not be one and the same as survey designers
   - T: Academic, government, and commercial
   - N: ease of receipt of requirements from designers

d. Data repositories
   - T: broad, specialized, archives, self-publishing platforms
   - N: provide access to data at various levels (given issues of licensing and confidentiality)
   - N: Enable discovery and analysis by end-users; incl. enable discovery in tools beyond their own catalogs (e.g., virtual metadata pond, but also linking data and publications)
   - N: Preservation (for some)
   - N: 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)
   - N: demonstrating use and impact of the data they make available
   - N: ability to build upon/learn from the work/systems/tools that others have done to better leverage resources

e. 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)
   - 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
  3.
- Universities

f. Research subjects/survey participants
   - N: confidentiality
   - N: demonstration of public benefit resulting from their participation

g. Other standards:
   - N: having DDI complement their standard by documenting aspects they don't cover

h. (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

i. Members of the public:
   - N: to gain a benefit for society from data gathered and used

j. Other communities

6. Services and Tools (Ingo)
   a. Survey Design Tool
      - Management tool to describe the concept of a survey, the sample sizes and processes and sets quality requirements for the survey (see Evaluate Process - GSBPM or GLBPM)

   b. Instrument design (part of the Survey Design Tool?)
      - Tool to enable researchers to create instruments ideally via a graphical user interface by re-using elements from the element registry
      - Looks like a survey instrument
      - No need to learn CAI-specific questionnaire description languages (e.g. Blaise, MMIC)
      - Searchable
      - Exports DDI in a format that a CAI instrument can read
        1. Need a tool that can design surveys
        2. Need an transport function
      - Alternatively might be embedded as a module into a CAI system which uses DDI natively (e.g. Rogatus)
      - Types and examples of CAI systems

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

- Element registry should contain elements for all CAI modes including paper&pencil interviews
- Typical elements of CAI systems
    1. Graphical Questionnaire Designer or Questionnaire Design Language (overlap with instrument design tool?)
    2. Survey Management System / Case Management System (e.g. Disposition codes, case assignment, sample management, interviewer assignment and tracking, synchronization mechanisms)
    3. Logging mechanism / Audit trail
    4. Reporting / Field Monitoring
    5. Export of the results (data and paradata) into formats of statistical packages (e.g. SPSS, Stata, R, MPlus)
- Ideally seamless integration between instrument design software using metadata standards plus element registry and CAI system

c. Data creation
d. Documentation tools
e. Need to define format for Datum Pond
f. Need to define format for Metadata Pond (PIDs)

7. 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. and [Should we go into specific tools as examples of things we can build upon?]

a. Building on existing tools and projects

    b. Utilizing related standards
    c. Research and demonstration projects
8. Next steps
    a. Grant application components
    b. Publications
    c. Coordination

- Tools needed for Element Register
    - Submission
    - Curation
    - Discovery
    - Survey Design