DDI Alliance Meeting
Monday, May 22, 2017, 08:30-17:00
University of Kansas Memorial Union
Big 12 Room -- Map

Agenda -- Meeting of Members				
Time	**Subject**	**Detail**	**Lead**	**Purpose**
08:30-09:00	Light Breakfast			
09:00-09:10	Welcome		Steve	Introductions
09:10-09:30	State of the Alliance 2017		Steve	Update group on last year's work
09:30-10:30	Panel Discussion	Updates from the following groups: Marketing, Training, Technical Committee, and Moving Forward	Amber Barry Wendy Achim	Review activities and get buy-in on future direction
10:30-10:45	Alliance Budget	- Current status and future projections - Member Forms	Jared	
10:45-11:00	Break			
11:00-12:15	DDI Vision and Strategic Plan	Detailed discussion of vision with the membership - Including the Infrastructure Manifesto	Steve	Get input and feedback
12:15-12:25	Executive Board Election	Discuss available positions and upcoming election (including Scientific Board vice-chair)	Steve	Inform about the upcoming election
12:25-12:30	Proposed Date for Next Meeting		Steve	Agree on best day to meet
12:30-13:30	Lunch			

Agenda -- Meeting of Scientific Board				
Time	**Subject**	**Detail**	**Lead**	**Purpose**
13:30-14:00	Scientific Board direction and goals for the year	-Reflecting on the DDI Vision -Specific activities for the Alliance (e.g. URN resolution, REST protocol, publications and best practices)	Chair	Set goals for what to accomplish
14:00-15:00	Work products and Moving Forward program	-Review the DDI Alliance work products -Overview of the DDI 4 timeline -Update on past reviews -Preparation for the codebook functional view	Steve Wendy Achim	In-depth discussion of DDI4 development
15:00-15:15	Administrative matters	Vice-chair election		
15:15 - 15:30	Coffee break			
15:30 - 16:00	Technical Committee report	Update of the Technical Committee on recent activity DDI Lifecycle and DDI Codebook updates	Wendy	Update group on progress
16:00 - 16:15	Related Initiatives	Report on related initiatives (SDMX and GSIM)	Steve	Update group on progress
Reports for Information (Discussion by Exception)				
16:15-17:00	EDDI Report NADDI Report Working group reports -Vocabularies -ADMP -DDI Dataverse	Brief (five-minutes each) reports Future activities and “where to next” for each group	Various	Update group on progress

18:30 - Informal DDI group dinner at Free State Brewing Company

20:00 - Informal IASSIST pub crawl

State of the Alliance 2017

Strategic plan 2014-17

http://www.ddialliance.org/system/files/DDIAllianceStrategicPlan2014-2017.pdf

Three core work areas:

- Standards maintenance and development
- Expanding the DDI Community – Marketing and partnerships
- Restructuring to achieve our priorities

Standards maintenance and development

- Manage and maintain the two existing product lines (Codebook and Lifecycle)
- Review and vote on RDF Vocabularies (2016)
- Develop a next generation model-based DDI specification (2016)
- Continue to publish new Controlled Vocabularies (2016)
- Gain ISO certification (2016)

Expanding the DDI Community – Marketing and partnerships

- Build partnerships and strategic alliances (2016)
- Assess the current state of DDI usage, community needs, and resources (2016)
- Improve the DDI website (2016)
- Create new materials explaining the value of DDI to people who are not DDI specialists (2016)
- Build a community around DDI training and increase access through innovative mechanisms (2016)

Restructuring to achieve our priorities

Review governance arrangements, including structure and Bylaws (2016)

Review revenue and funding request models (2016)

DDI Marketing and Partnerships Group

Report and Plan 2018

Team:
Barry Radler
Kelly Chatain
Jared Lyle
Steve McEachern
Ron Nakao
Dan Smith
Wendy Thomas

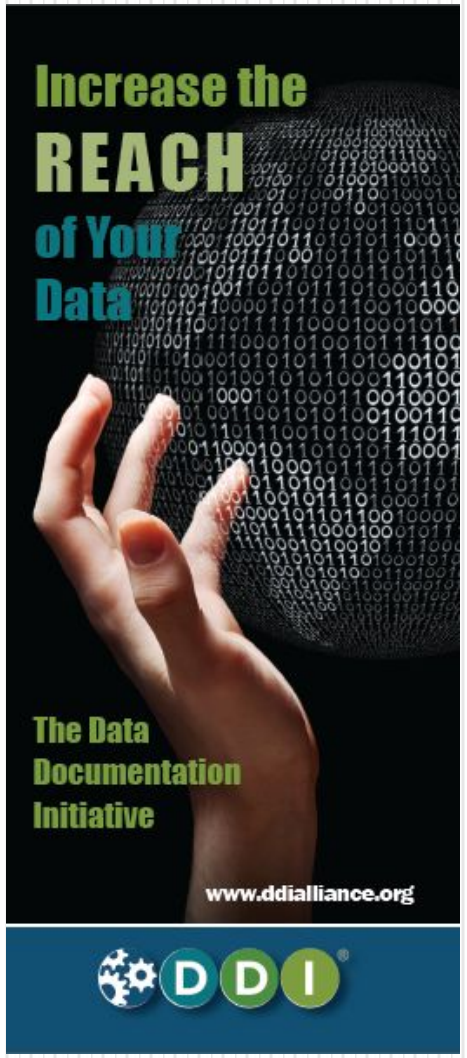
Mission Statement

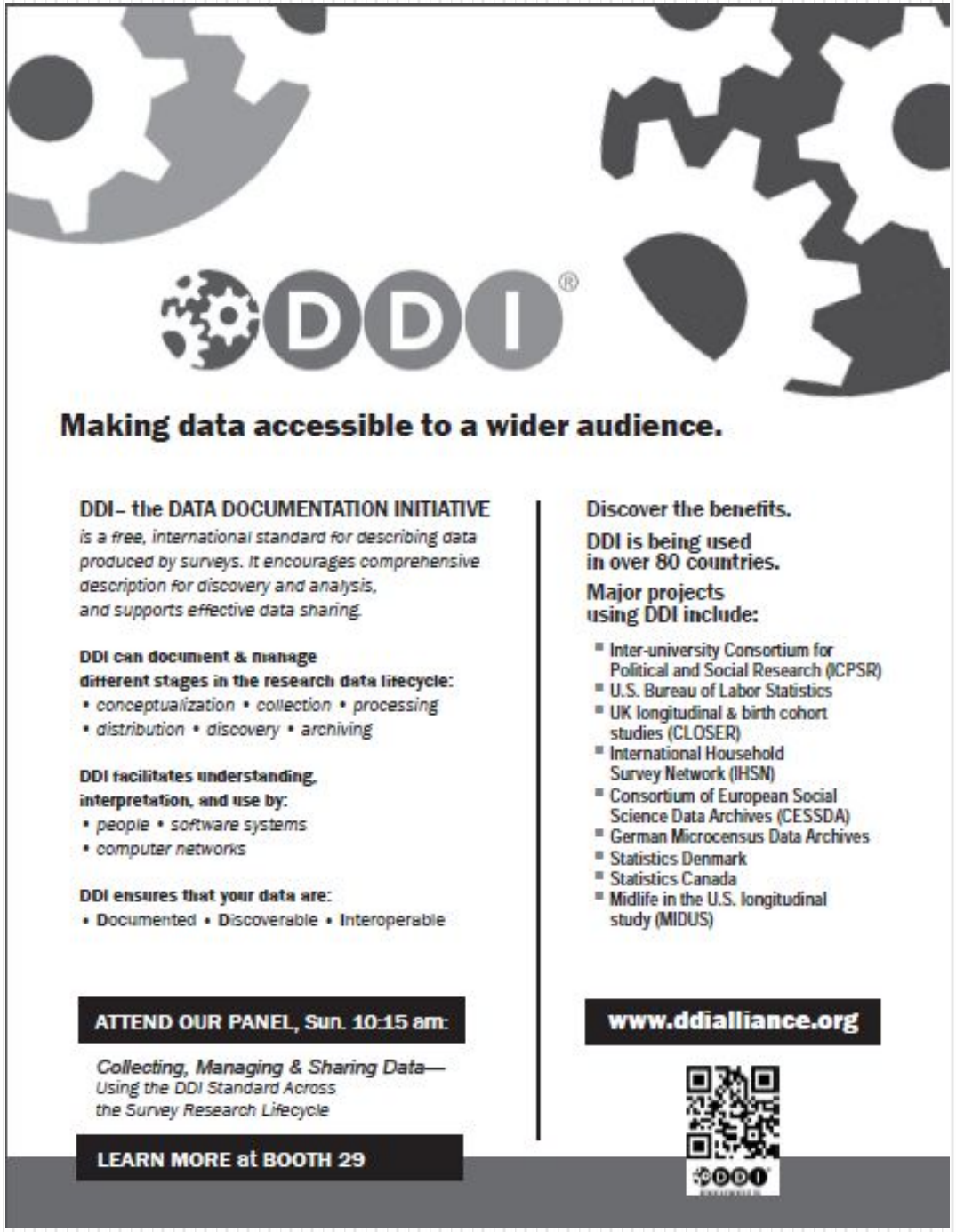
- Coordinate marketing activities, establish DDI brand, ensure consistent messaging
- Interface with other standards bodies
- Increase the DDI user community and DDI Alliance membership

What worked during the past year?

- Materials development
  - Updated tri-fold brochure
  - ICPSR Data Fair webinar (coordinated with Training)
  - Ongoing Website design and maintenance
    - Monitor with Google Analytics
- Conference attendance
  - New promotional materials, rolling presentation, conference schwag
  - Expanding promotion to new communities/conferences
    - Sponsorships and ads at AAPOR, IASSIST, ESRA
    - AAPOR Transparency Initiative outreach
    - IFDTC, ACSPRI, 3MC,

Brochures and Ads





What worked during the past year?

- Materials development
  - Updated tri-fold brochure
  - ICPSR Data Fair webinar (coordinated with Training)
  - Ongoing Website design and maintenance
    - Monitor with Google Analytics
- Conference attendance
  - New promotional materials, rolling presentation, conference schwag
  - Expanding promotion to new communities/conferences
    - Sponsorships and ads at AAPOR, IASSIST, ESRA
    - AAPOR Transparency Initiative outreach
    - IFDTC, ACSPRI, 3MC,
- Coordination with other DDI groups
  - Quarterly meetings
  - Establish online Help (Training)
  - Archiving material (Publications)

What didn't work during the past year?

- Formalize conference evaluation demo
  - Proposal to host evaluations at conferences attended
  - Performed at NADDI since 2015
- AAPOR invited reception
  - Expensive and sparsely attended
- In Development:
  - Establishing relationships with other standards bodies
    - AAPOR Transparency Initiative
    - ORCID Organization Identification Registry
  - Refining division of labor among working groups

Plans for next 12 months

- Continue to improve website, materials, message
  - Update time-sensitive content on website
- Continue/expand conference presence and attendance
  - Consider more sponsorships, “getting in the program”
  - Place ads in programs; distribute business cards at booths/exhibitions
- Formalize conference evaluation demo
  - Propose to host evaluations at conferences attended
  - How to proceduralize without incurring excessive cost?
  - Avoiding favoritism or conflicts of interest vis-à-vis tools
  - White paper, presentation, or brochure on NADDI evaluations?

Plans for next 12 months

- Promote improved (new) documentation for 2.5, 3.2
  - Coordinate with TC on releases
- Identify and target most relevant organizations
  - How to proceduralize outreach?
- Outreach to tools and software, not just other standards
  - R, SPSS, StatTransfer, Stata, etc.
- Evaluate proposal: Consider rescheduling DDI Members and/or Scientific Board Meetings
  - Promote and increase attendance at DDI user group conferences
    - Comingle members, current users, and potential audiences
  - Examine attendance numbers at NADDI, EDDI, and IASSIST

Resources required next 12 months

- 2018 budget ($15k)
  - Ongoing - marketing materials, producing ads, printing brochures, schwag
  - Ongoing - conference attendance, outreach, travel
- New: Outsource tasks not being accomplished by volunteer contributions
  - Website maintenance
    - Updating social media?
  - Conference evaluation tool
  - AAPOR Transparency Initiative tool
  - Educational videos (coordinate with Training)

Technical Committee

Members Meeting

May 2017

TO DO ITEM	6/16/2016	6/23/2016	6/30/2016	7/7/2016	7/14/2016	7/21/2016	7/28/2016	8/4/2016	8/11/2016	8/18/2016	8/25/2016	9/1/2016	9/8/2016	9/15/2016	9/22/2016	9/29/2016	10/6/2016
Finalize RDF for review																	
RDF Vocabulary page set up																	
RDF Vocabulary public review																	
Q2 set up pages/JIRA																	
Q2 request issues																	
3.3 question structure resolution																	
Q2 pre-announce																	
Q2 development release																	
Codebook request issues																	
3.3 schemas																	
3.2 documentation																	
3.3 documentation																	
3.3 review																	
Codebook set up pages/JIRA																	
Codebook pre-announce																	
Codebook development release																	

2016-2017

- RDF Vocabularies
  - Due to time constraints final modifications to DISCO have not been completed and are required for release.
  - XKOS underwent public review in January 2017
- DDI 4 Q2 Development Review
  - Completed Build was not received from MT until September 30
  - TC had all preparations for development review completed by October 8 and was presented for review on October 17
- DDI 3.3
  - Issues have been discussion with some decisions remaining
  - Four members of the TC will be meeting in Minneapolis in June to complete entry work
- Codebook Functional View
  - Has not been released by Modeling Team. Anticipated September 2017

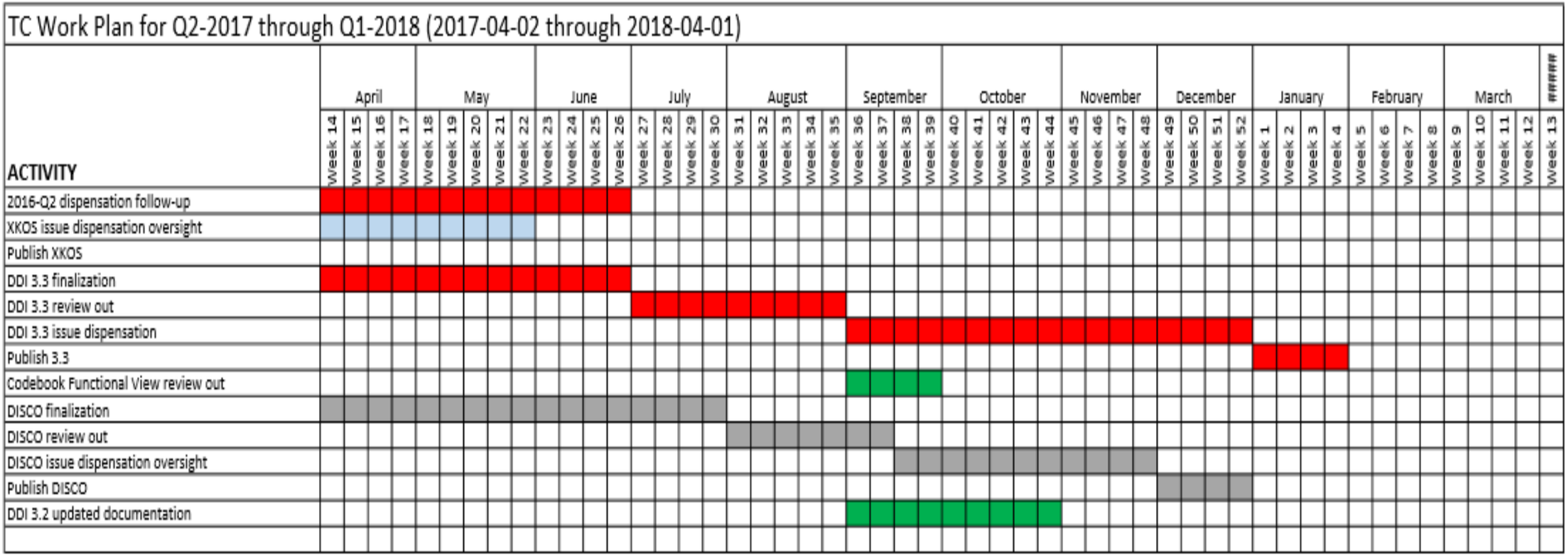
DDI 4 Q2 2016 Development Review

- Ran second developmental review of DDI 4 revising the approach for review to accomplish the following:
  - Faster response on bugs
  - Pushing broader issues back onto the developer groups
  - Tracking follow-up
  - Update on status
- Developer groups are still working on a number of issues
  - Modeling team will be addressing complex issues relating to collections, process model, and GSIM relationships during the sprint next week

XKOS RDF Vocabulary

- Issued XKOS for public review 15 January 2017
  - 51 issues filed by 6 reviewers
  - Franck Cotton is managing responses
  - Review approach: The decision can be to dismiss the issue (explain why and close), accept the issue for XKOS v1 (make corresponding modifications and close), or postpone the issue to XKOS v2.
- Pick up on this in June

TC Work Plan for Q2-2017 through Q1-2018 (2017-04-02 through 2018-04-01)



Activity Types

- Critical
- Primary
- External dependencies
- Oversight only

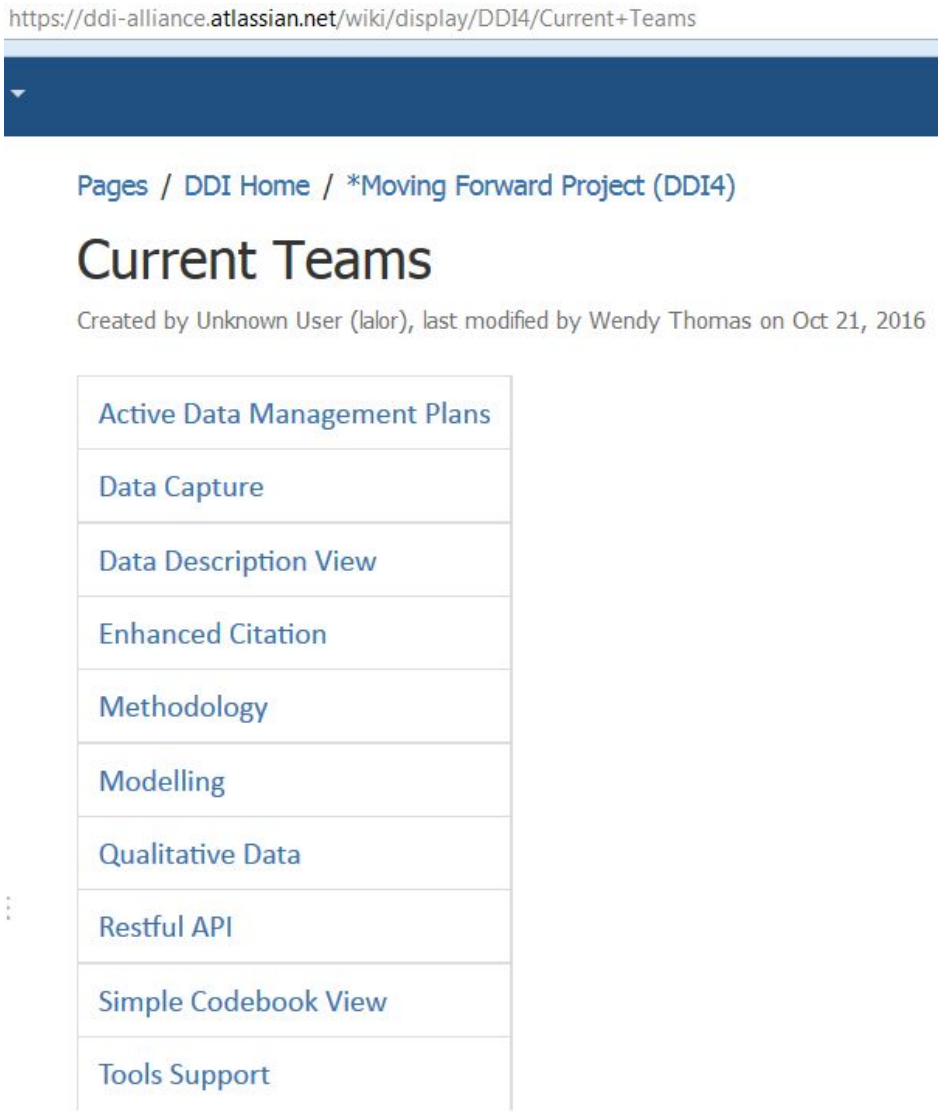
2017-2018

- TC has a new work plan with work identified as Primary, Critical, External Dependence, Oversight only
- New increased focus on supporting implementation of current and new users of published DDI standards
  - Best Practices
  - Updating and expanding documentation - reissued 3.2 following 3.3 review release
  - Long term managed shift in DDI Lifecycle from version 3.x series to version 4
  - Reinstating some form of the former TC on-site working meeting (3.3 focused meeting in June with 4 members to complete package for review)
- Managing Codebook Functional View development review

DDI Moving Forward Project

Status and Outlook May 2017

Group Work in Virtual Meetings and Sprints



Virtual Meetings

- Frequent meetings
  - Data Description View
  - Modelling
  - Simple Codebook View
- Meetings when needed
  - Active Data Management Plans
  - Tools Support
- Temporary inactive
  - Enhanced Citation (major work is already done)
  - Methodology (active thru Dec 2016)
  - Qualitative Data (open task: integration of model parts into DDI 4)
  - Restful API (not started yet)

Three Sprints / Workshops since June 2016

Venue: Schloss Dagstuhl – Leibniz Center for Informatics in Wadern, Germany

- DDI Moving Forward: Facilitating Interoperability and Collaboration with Other Metadata Standards
  - October 17 – 21, 2016
  - 21 participants, 9 from other metadata specifications and groups
- DDI Moving Forward: Improvement and Refinement of Selected Areas
  - October 24 – 28, 2016
  - 24 participants

Cologne after EDDI16

- December 12-16, 2016
- 6 participants

DDI Moving Forward: Facilitating Inter-operability and Collaboration with Other Metadata Standards

Specifications

- DDI – Data Documentation Initiative
- CDISC – Clinical Data Interchange Standards Consortium
- HL7/FHIR – Health Level Seven / Fast Healthcare Interoperability Resources
- SDMX – Statistical Data and Metadata eXchange
- GSIM – Generic Statistical Information Model
- W3C CSV on the Web (Comma-Separated Values)

DDI Moving Forward: Facilitating Inter-operability and Collaboration with Other Metadata Standards

Topics

- Across multiple metadata specifications
  - **Data Description Commonalities**
  - **Manifesto (Design Principles)**
  - Bindings
  - Protocols
  - Business Scenarios/Use Cases
- **Provenance**
- Design Patterns in DDI-Views (Version 4)

DDI Moving Forward: Improvement and Refinement of Selected Areas

Further development of DDI-Views (Version 4)

- Validation of Data Description
- Integration of Data Capture into full model
- Controlled vocabularies
- Re-usable structured documentation
- Long-term metadata infrastructure plan

Cologne Meeting

Work on

- Document on RDF task list (intended for external expert)
- Document on development tasks for model capturing environment (Lion server)
- Migration of integration server for the production framework
- First steps of migration of Lion server into the cloud, better separation of distinct tasks

Upcoming Sprint after IASSIST

- 5-days in Lawrence
- 8 participants
- Topics include
  - Codebook Functional View
  - Document for DDI 4 providing a common understanding of the goals of DDI 4 from current perspective
  - Review of package integration
  - coverage and gap review between DDI 4 and DDI-Lifecycle (DDI 3), and DDI 4 and GSIM

Current High-Level Status

- Past work focused on broad development on different levels
  - New content
  - New architecture
  - Structural modeling
  - Production system
  - Interoperability with other metadata specifications

Outlook

- Future focus should be on
  - Publication of core Functional Views
    - Codebook and related basic data description and data capture
  - Necessary tasks for the purpose above
    - Selection of mature elements
    - Filling in gaps and integration
  - New developments should have minor priority
  - Continuation and improvement of selected approaches of
    - Production framework
    - Structured documentation
  - Intensification of the creation of
    - Technical test cases
    - Business use cases

Possible Future Workshops

- Dagstuhl, October 2017
  - One or two workshops along the lines mentioned before, currently in planning state
- Chur, December 2017 (week before EDDI17)
  - Subject tbd

	Actual FY2012	Actual FY2013	Actual FY2014	Actual FY2015	Actual FY2016	Budget FY2017*	Actual FY2017	Budget FY2018
Total Revenue	$74,917.00	$84,807.00	$84,815.00	$87,419.00	$85,345.00	$108,000.00	$98,074.00	$101,000.00
Expenses								
Staff Salaries	$31,970.00	$22,549.00	$25,544.00	$29,633.00	$28,989.00	$27,584.00	$17,216.03	$28,412.00
Consultants		$4,970.00	$4,970.00	$27,426.00	$20,360.00	$20,000.00	$1,795.12	$20,000.00
Data Processing	$2,760.00	$2,217.00	$1,879.00	$3,003.00	$3,224.00	$3,224.00	$1,656.33	$3,500.00
General Expenses		$73.00	$15.00	$150.00	$113.00	$200.00	$136.83	
Marketing					$6,567.00	$15,000.00	$5,659.73	$15,000.00
Research Supplies & Services	$54,205.00	$2,900.00	$5,647.00	$5,876.00	$948.00	$8,000.00	$4,718.52	$4,000.00
Training					$1,073.00	$10,000.00	$0.00	$5,000.00
Travel and Hosting	$17,191.00	$28,814.00	$17,209.00	$22,218.00	$40,646.00	$31,000.00	$21,507.42	$31,000.00
Transfer	-$13,974.00							
Total Expenses	$92,152.00	$61,523.00	$55,264.00	$88,306.00	$101,920.00	$115,008.00	$52,689.98	$106,912.00
Revenue Over/(Under) Expenses	-$17,235.00	$23,284.00	$29,551.00	-$887.00	-$16,575.00	-$7,008.00	$45,384.02	-$5,912.00
Ending Fund Balance	$109,407.00	$132,691.00	$162,242.00	$161,355.00	$144,780.00	*$137,772.00*	*$190,164.02*	*$184,252.02*

Currency in USD.

*FY2017 Revenue estimate based on 36 full members at the OECD base contributor level.

Vision for DDI Long-Term Infrastructure and the DDI Alliance

Vision for DDI Long-Term Infrastructure

- DDI-based infrastructure for the support of empirical sciences in the social, behavioral, economic, and health domains
- Describing data in a structural and standardised way
- Based on a central element registry and distributed metadata/data repositories

Purpose

Providing the basis for a reliable framework in a global network in order to support …

- Exchange and long-term preservation of metadata
- Re-using metadata in a single data collection, across waves of longitudinal data, across data collections, and across institutions
- Metadata-driven data collection
- Transparent research
- Research reproducibility

Mutual Benefits

- An institution realizing a part of the infrastructure framework benefits from …
  - a larger plan with well-defined interfaces
  - existing components
  - referencing both in proposals for funding agencies
    - Such a proposal would be a part of a bigger picture and no isolated development
- The empirical SBE sciences benefit from a growing distributed infrastructure framework
- The DDI Alliance benefits from third-party contributions
  - The Alliance wouldn‘t have the resources (nor it is their objective) to realize all components of the infrastructure

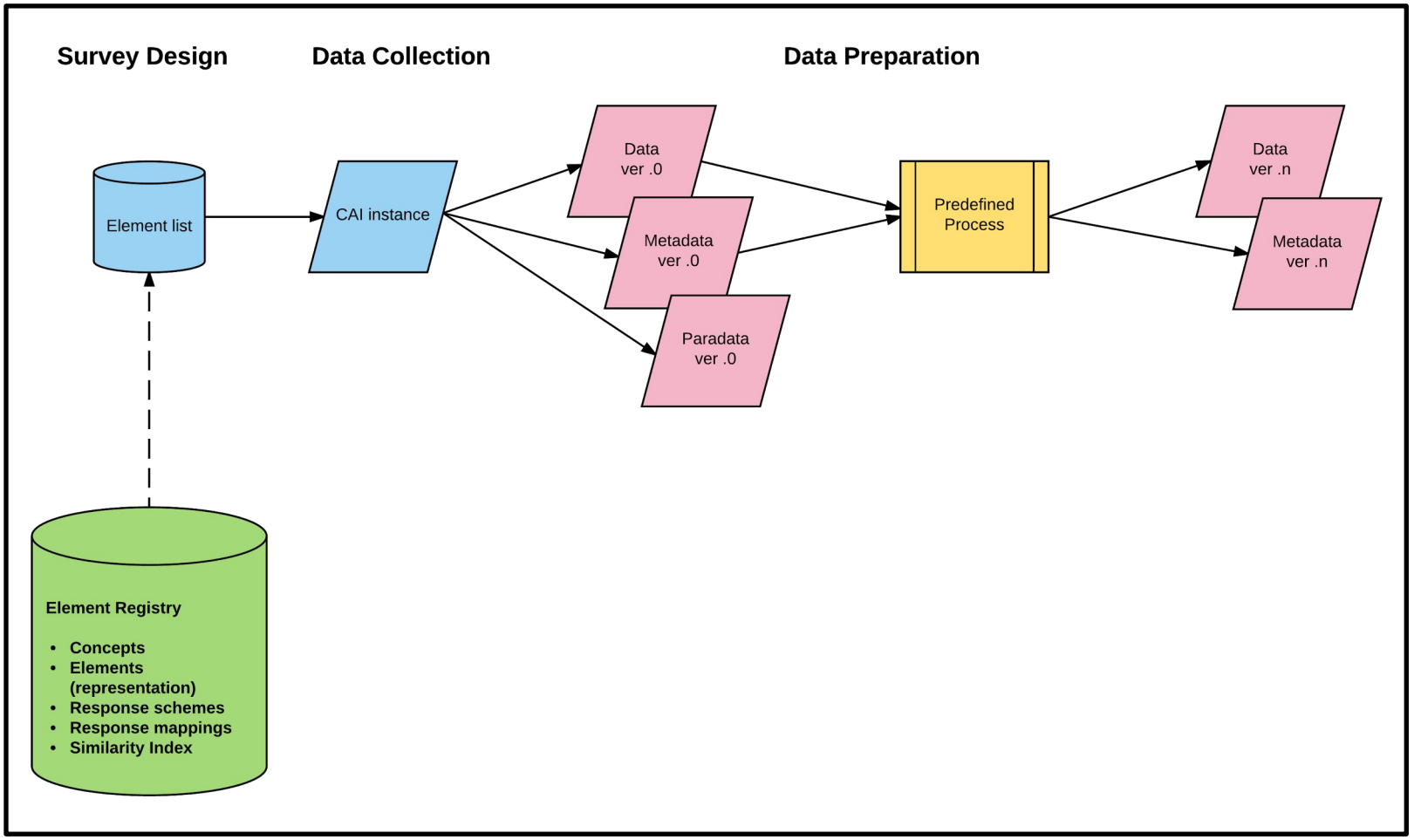
Limitations

- All components of the infrastructure framework …
  - would need a license which allows the public use of them
  - need to be compliant with the related rules
- Data and metadata elements could be provided with access restrictions if necessary

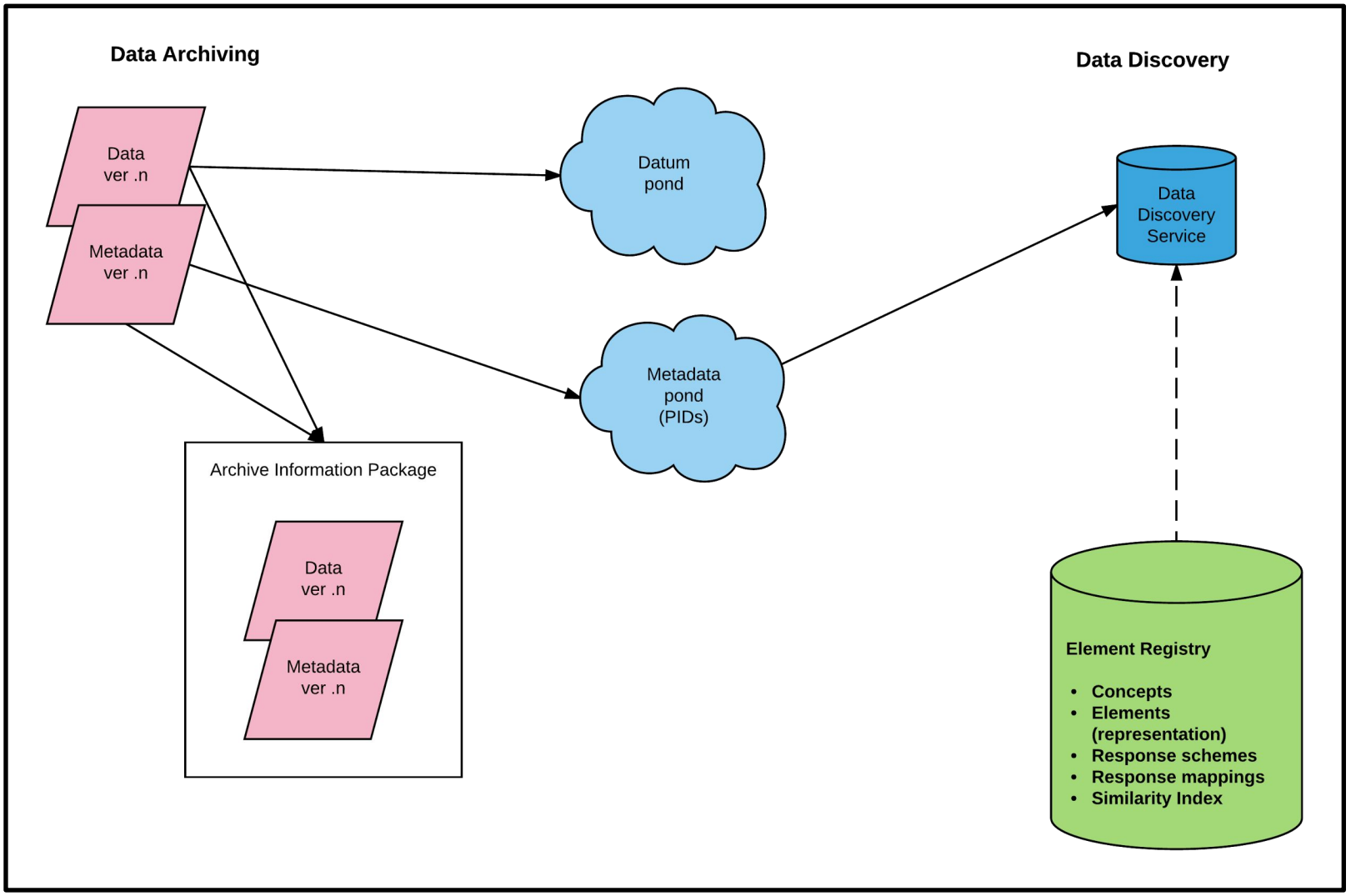
Background

- The idea of a long-term infrastructure plan is borrowed from areas in sciences, astronomy and particle physics
  - There research depends on expensive infrastructure and related work

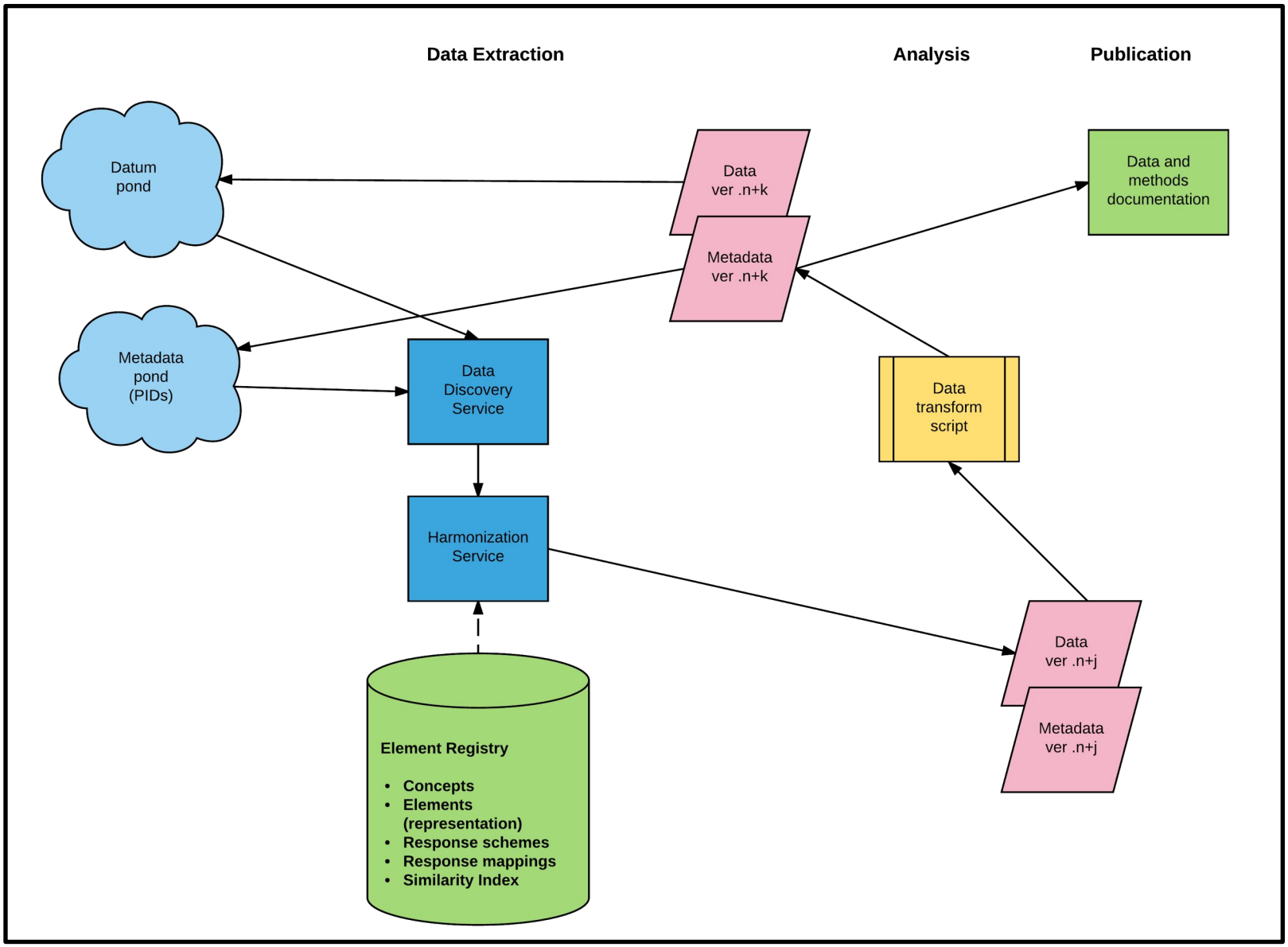
Element Registry and Survey Design



Element Registry and Data Discovery



Element Registry and Data Extraction



Pillars of DDI Long-Term Infrastructure

- DDI Specification
  - Definition by formal language and English documentation
  - Supporting material
    - Test cases – for the technical use
      - Technical instances to show the use of specific parts of the specification (i.e. in XML) for detailed purposes
    - Use cases – for the business use
      - Descriptions to show the application of the specification for business scenarios (not necessarily as technical instance)
    - Best Practices for achieving best results in using DDI
- Identification, query, and resolution of DDI objects
  - Definition of DDI URN
  - Definition of DDI query protocol (i.e. REST)
  - Prototype software components for query and resolution
- Rules and software for metadata registry
- Metadata repositories
  - Software for building and querying repositories
  - Content of repositories

Who is Doing What?

	DDI Alliance	DDI Community
Specifications including formal documentation	x	
Test cases	x	
Use cases		
Documentation structure	x	
Description	Prototype	x
Best Practices for achieving best results in using DDI		
Documentation structure	x	
Description	Prototype	x
Identification, query, and resolution of DDI objects		
Definition of DDI URN (persistent, location-independent identifier)	x	
Definition of DDI query protocol (i.e. REST)	x	
Software components for query and resolution	Prototype	x
Rules and software for metadata registry	x	
Software for building and querying metadata repositories		x
Content of metadata repositories		x

Complementary Core Documents

- Vision for DDI Long-Term Infrastructure
  - For the DDI community and the DDI Alliance
  - Purpose is to provide a reliable long-term planning framework
  - Parts are already realized or will be realized by the DDI Alliance
  - Other parts can be used by the DDI community
- Strategic plan of the DDI Alliance for 3-4 years
  - Translating the DDI Alliance parts of the vision into broadly defined goals and a sequence of steps to achieve them
- Mission and Principles of the DDI Alliance - unchanged over time
  - Mission: declaration of the core purpose and focus
    - Based on the Objectives of the DDI Alliance Charter
  - Guiding principles: Fundamental norms, rules, or values that represent what is desirable and positive in terms of developing DDI specifications for a global network
    - Broad audience: DDI specification developers, DDI users, other organizations in the field

Steps for Developing the Documents

- Discussion at DDI Alliance Annual Meeting 2017
- Panel session at IASSIST conference 2017
- Working group initiated by the Executive Board
- Development of mature versions of the documents at Dagstuhl workshop in October 2017
- Distribution to members and improvement of documents
- Approval of documents at DDI Alliance Annual Meeting 2018

Basis Documents

- Discussion paper „DDI Long-term Infrastructure Manifesto”
  - Started at Dagstuhl workshop in October 2016
- DDI Alliance Strategic Plan, 2014-2017
- Moving Forward Design Principles
- DDI Mission and Guiding Principles, draft from 2012
- Principles for developing metadata specifications
  - Started at Dagstuhl workshop in October 2016

DDI Long-term Infrastructure Manifesto[1]

George Alter
Ingo Barkow
William Block
Jared Lyle
Steven McEachern
Katherine McNeill
Katja Moilanen
Joachim Wackerow

1. Overview

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.

By outlining our overall vision, we hope to inspire the the DDI community to build DDI tools and services encompassing the entire research lifecycle. Development typically comes in stages; this document provides an overall vision and shows how DDI-based processes in different parts of the research workflow complement and reinforce each other. Connecting the dots from

[1] This working paper was written 24-28 October 2016 at Schloss Dagstuhl as part of a DDI sprint.

instrument development to data collection to data management to archiving to analysis to publication 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.

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.

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 of (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. 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.

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:

<u>Study design:</u>
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.

<u>CAI implementation:</u>

- 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:

<u>Data preparation:</u>

- 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.

<u>Data archiving:</u>

- 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.[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.

[2] 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).

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.[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"),

[3] 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.

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.

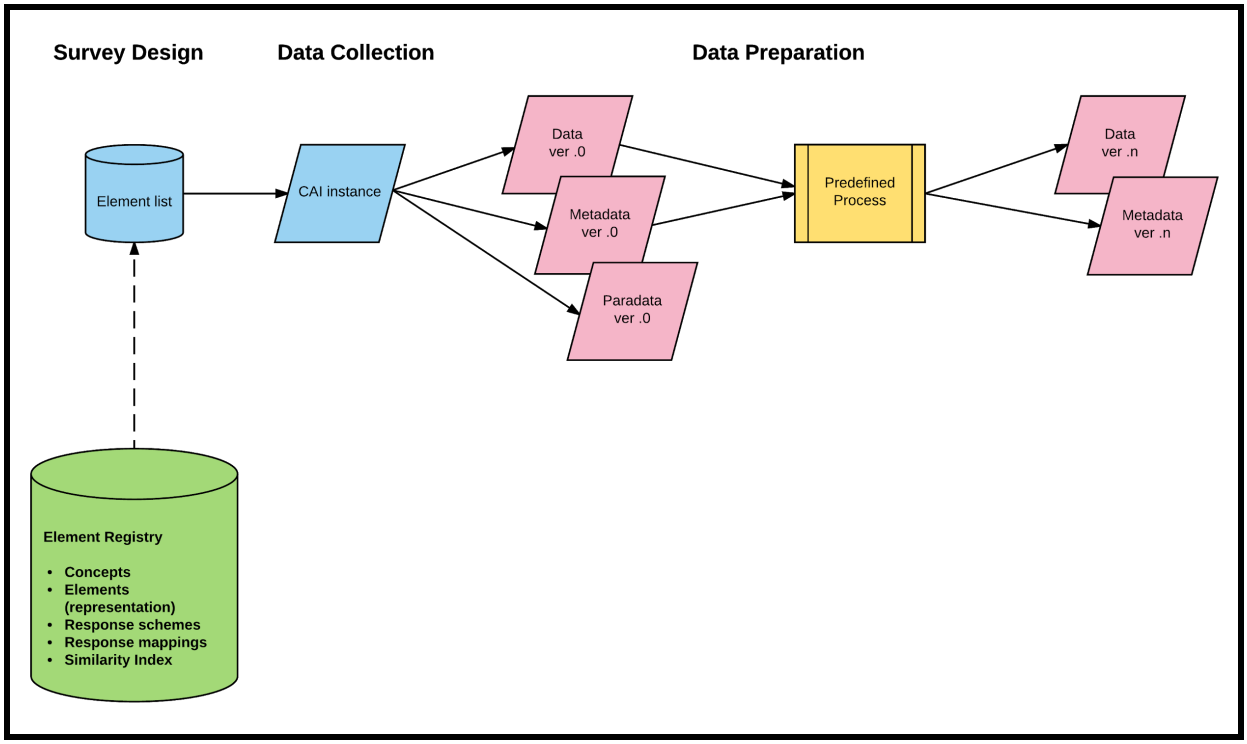


Figure 2.

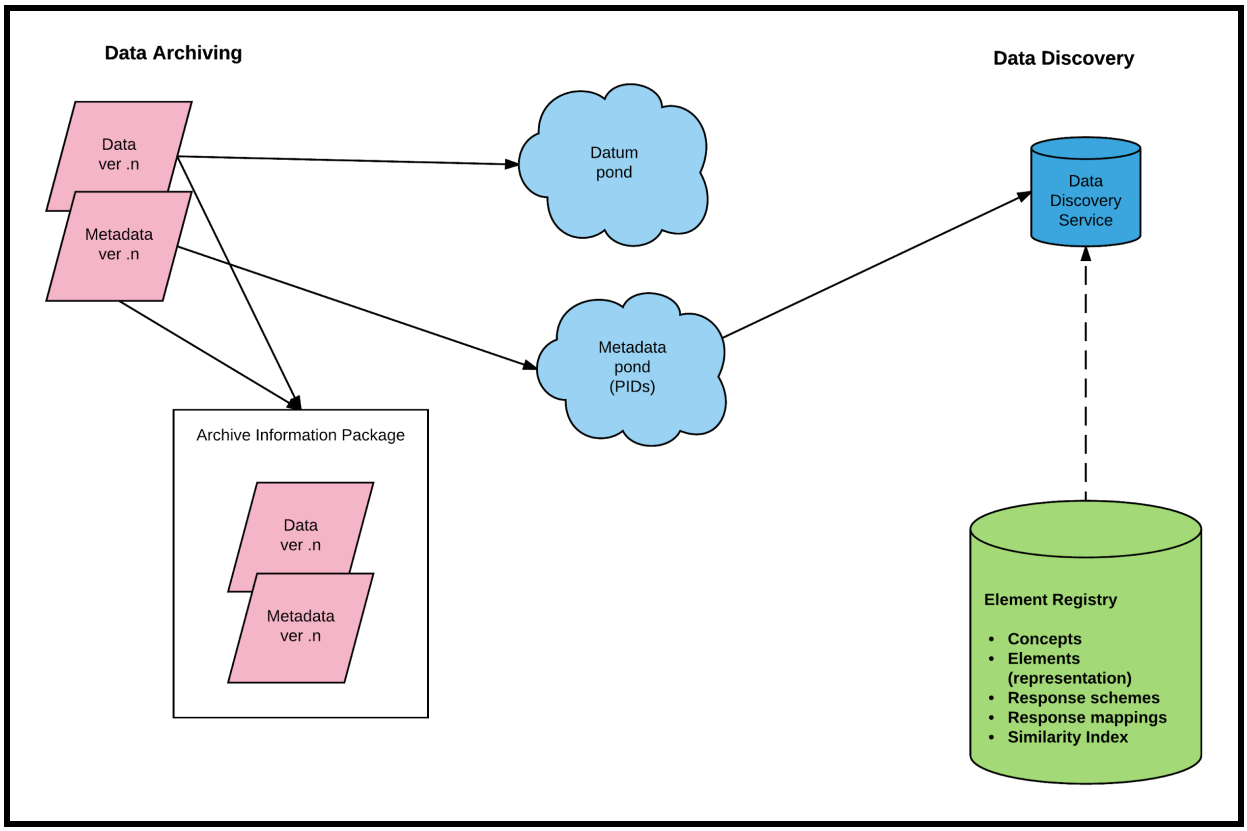
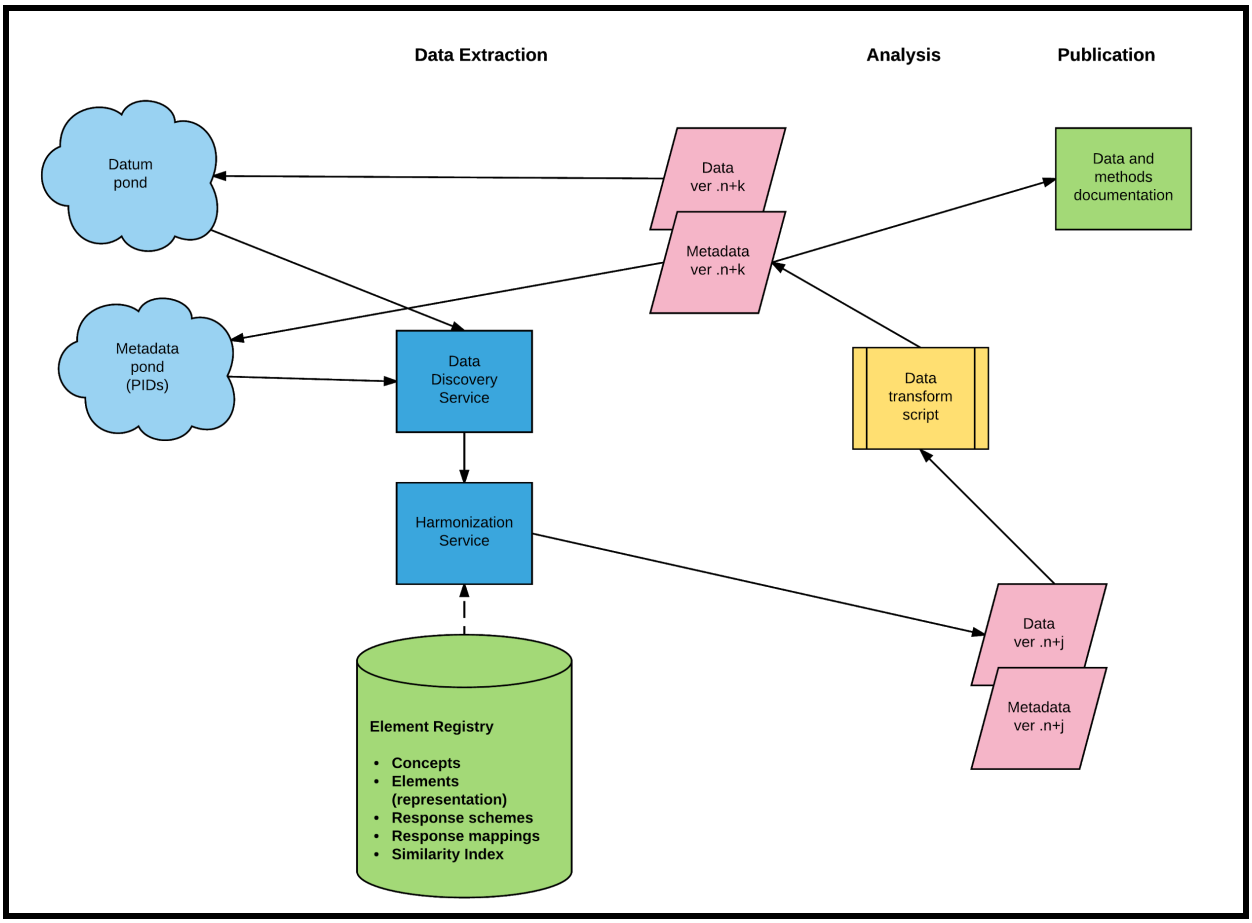

Data Archiving
Data Discovery
Data
ver .n
Metadata
ver .n
Datum
pond
Metadata
pond
(PIDs)
Data
Discovery
Service
Archive Information Package
Data
ver .n
Metadata
ver .n
Element Registry
Concepts
Elements
(representation)
Response schemes
Response mappings
Similarity Index

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

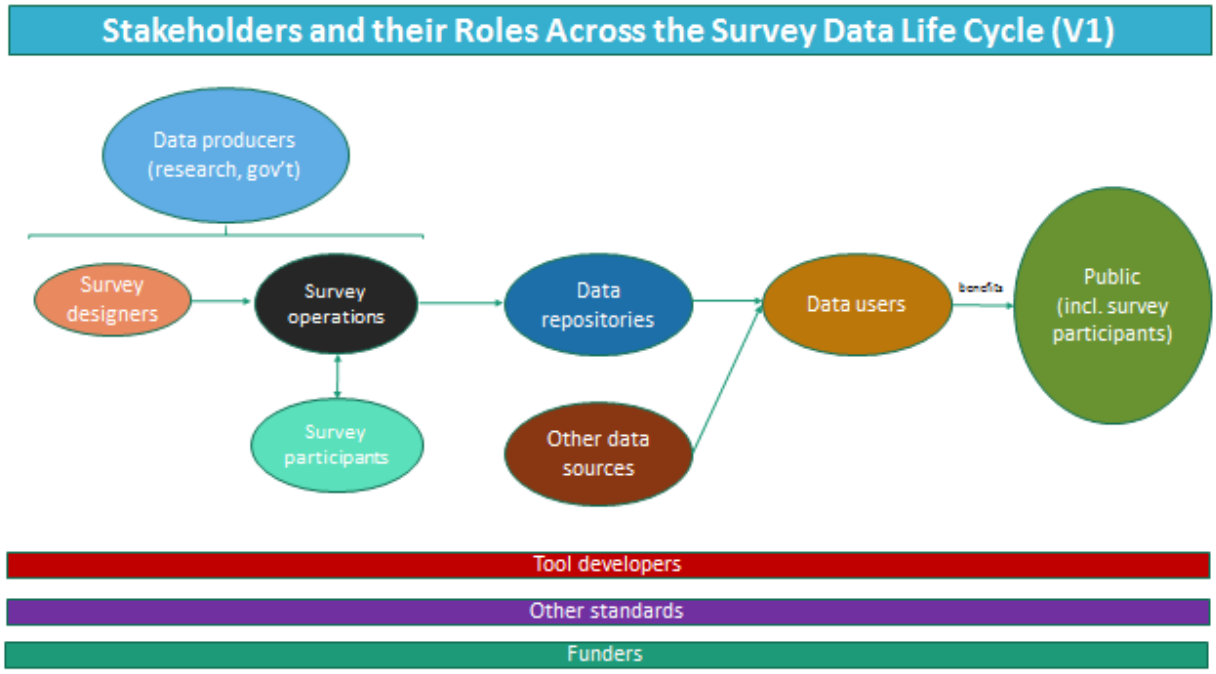
encouraging the use of Common Data Elements to standardize the collection of data, and initiatives like Logical Observation Identifiers Names and Codes (LOINC) are providing standards and tools. There has also been rapid progress in the development of standards for describing and exchanging biomedical 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.

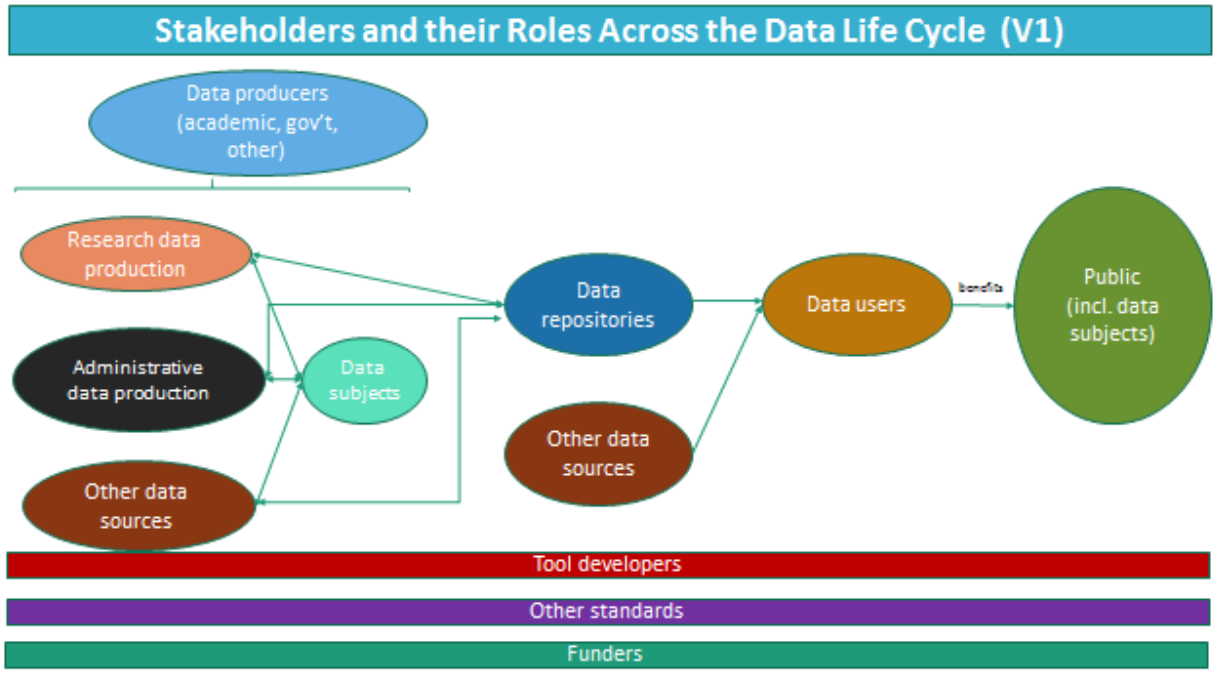
The growing demand in social science research for administrative and transactional data created by governments and businesses poses especially difficult metadata problems. The systems from which these data are derived were created for other purposes, and the documentation that social scientists need 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.[4] Such repositories take data from a range of producers.

[4] Further iterations of the infrastructure could expand to involve a greater variety of data repositories.

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.,
    a. Australia: Australian Research Council (ARC)
    b. Germany: Deutsche Forschungsgemeinschaft (DFG)
    c. Netherlands: Netherlands Organization for Scientific Research (NWO) and Royal Netherlands Academy of Arts and Sciences (KNAW)
    d. U.K.: Economic and Social Research Council (ESRC)
    e. 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.,
    a. European Commission (EC), including Horizon 2020, Eurostars, and European Strategy Forum on Research Infrastructures (ESFRI) programs
    b. Organisation for Economic Co-operation and Development (OECD)
4. Private Research Foundations, e.g.,
    a. Alfred P. Sloan Foundation
    b. Gates Foundation
    c. 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.[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)

---

[5] 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/.

- Data in related disciplines (various examples can be found on the DCC directory of Disciplinary Metadata[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 C^{2}Metadata ("Continuous Capture of Metadata") is an example of project that is addressing a specific gap in the metadata workflow.[7] Funded by the US National Science Foundation, C^{2}Metadata 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

[6] http://www.dcc.ac.uk/resources/metadata-standards
[7] http://c2metadata.org/

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.[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.[9]

[8] http://www.colectica.com/news/Open-Source-DDI-Converter-Project

[9] 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

Strategy development 2017-20

Timelines

Activity	Responsibility	Completed by
Environmental scan. Key issues, questions, and choices to be addressed	Community consultation: Member survey and online discussion through member's email list	October 2016
Review of Alliance vision and mission	Executive committee to draft for circulation (EDDI 2016) Feedback from members and community	December 2016 February 2017
Development of key strategic goals and action plan. Circulation of draft strategic plan	Executive committee to draft For circulation and discussion to members and community	 End of April 2017
Review of draft plan and final approval	All members	IASSIST 2017

Current status

- Strategic planning has not progressed
- Several reasons, but primarily shifting priorities in 2016-17
- Need for reconsideration of:
  - Strategic direction of the Alliance
  - Future funding demands
  - Organisational structure and growth
- Has lead to development of the long-term infrastructure model
  - For discussion later: to form the basis of of next Strategic plan