Notes from Long-term Infrastructure Group

# Status:

## Decisions:

1. Combine two groups into one: Long-term Infrastructure
2. Combine all outputs into one document (for now); produce a DDI working paper out of the workshop (for internal discussion/presentation to the Executive Committee for input into the strategic plan) that can then be published as a paper for wider circulation

## Action items:

## Outputs (in same document):

1. Vision
2. Stakeholder analysis
3. Broad strategy
4. Coordination around funding
5. List of grant application components
6. Examples of publications needed

# Monday:

## Introduction to the issue

Funding proposals/infrastructure:

Funding:

* Bill proposed last year
* Description of purpose of DDI, Alliance, Community, etc.
* Infrastructure:
* George’s idea >Bergen; related to practices in astronomy (steal idea about physical infrastructure)

Also related: strategic plan and mission; next version in process; above two will be fed into strategic plan

George’s presentation:

* Pyramid of the following for infrastructure:
  + Vision of reusable metadata/software w/in an infrastructure; high level and more detailed lower level
  + Services: what does that really mean (e.g,. question/variable banks) in the long run, reusable code for data transformation steps
  + Broken down to a more technical level, what is required (e.g., ID, versioning)
* Then an institution could pick a (vertical) building block to put together; not an isolated thing of a local project but relate to an overall vision, have a longer life than just the project
* Library of building blocks for funding proposals could support proposals for the infrastructure
* Phases: data collection, processing, distribution, analysis
* Service example: Variable/element registry (from ontologies)
  + w/concepts, elements (representation), responses, and response mappings (elaboration of a question bank), similarity index (incl. translation)
  + everything has a PID (that’s what makes it a registry)
  + element list > CAI instance (reducing overhead of negotiation between designer and survey firm) > data/metadata/paradata
  + metadata as a byproduct of the data collection process
  + processing: data transformation script produces new forms of data/metadata
  + data lake: streams flowing in and out, doesn’t have to be based on RDF; new way to do analysis/create new datasets
  + discovery based on PIDs from variable registry
  + harmonization phase based on response mappings: a) simple between (equal) response schemas b) elements themselves (more complicated)
  + discovery across different entities (centralization and diversity/distribution); use concepts for discovery in addition to (known item) PIDs
  + central registry (virtual metadata pond)
  + response mappings for harmonization
* Include how does this apply to: administrative data (get into registries), other disciplines’ observational data; qualitative data

Discussion:

* Bill: strategy important, meat of funding proposals; role of influencing funding agencies and other groups (CAIs); make a case of how it would help them
* How does the Alliance related to DDI-related projects w/their own purpose
* How liaise w/funders (across borders); compelling vision of working together towards broader goals
* High level goals; major Archives retooling; build around DDI activities; working w/statistical agencies
* 2008 meeting among archives (Kevin, Myron); on how to coordinate in development of repository infrastructures (Matthew and George said they’re too busy); and how does CESSDA fit in
* Alliance: role tying together different organizations and entities (e.g., CESSDA & NA)
* Distinguish between: research proposal, PPP that produce software in combination w/companies
* Variety of funding contexts
* Entities: funding agencies, NSI, archives, survey organizations, [researchers], note taking and organizational
* Proposals: everyone’s doing their own thing; method to keep track of proposal options and what’s being put in
* Role of tools
* Role of the Alliance in implementing this vision vs. coordinating it
* Incorporating needs of small scale research
* add preservation step
* Why metadata useful
* Tools; how does the Alliance organizationally fit into the infrastructure
* Need more context for how it would be different in different contexts along the life cycle
* Data vs. metadata; when the latter becomes the former
* Real-time data?

Benefits to achieve in the infrastructure system: (& other goals)

* Automate capture of metadata (reduced costs and time)
* Capture better/more complete metadata
* Enable new data discovery and analysis tools
* New data harmonization, comparison and combination tools
* Systems that can be used/across organizations; Transparency across organizations across or w/in stages of the life cycle
* Encourage interoperability and comparability across studies, domains, and countries
* Infrastructure for small scale w/o benefits to overhead
* Lower cost of using/entry into DDI infrastructure (barriers to entry)
* Faster and more efficient research/data collection design design
* Reproducibility
* Credit for producing items in the life cycle
* Increased use of DDI; tool that’s used at (and enhances) all stages of research w/in infrastructure
* Purpose: support discovery, analysis, preservation, harmonization, reuse to enable future research more easily
* Inclusions and exclusions; can incorporate variety of observational data while intersecting with/relying up on other standards as appropriate
* Multi-lingual and multi-country environment

# Tuesday:

Feedback from plenary presentation:

* Research groups who only after a while found out that colleagues across the country were designing surveys w/the same questions for the same universe; registry is good
* Concern: building CAI instrument from a registry of questions; NCHS put together a grand idea question bank yet hasn’t been used that much b/wording of a question depends so much on a mode and the types of responses that are targeted; include something more sophisticated to account for the need to write test questions and not just pull automated, and metadata on mode of collection; \*\*talk to NCHS; maybe mode could be treated like language in having a similarity index
* Assumes that data linking can always happen; how will studies as a whole be represented? A: people will be able to have same kinds of files that we serve up now.
* Registry: centralized and maintenance? How this is done organizationally.
* What’s interesting: interaction on larger surveys between the study design and \_\_\_\_[?]; need broad assessment on the strengths and weaknesses of each parts as they stand so that we know what to work on?
* Concerning the vision, there’s a lot that we’ve been describing before; added some valid areas of expansion, interoperability across the life cycles; what make the standard applied is that it’s usable, different stages and need tools; we should ID tool gaps; if focus too much on the registry will leave out some other areas where the interoperability issue isn’t yet solved; alliance encourage people to work on these gaps
* DDI hole in field work, Ingo has worked on this
* What’s missing: tool to automate creation of similarity index (resource intensive)
* Liked how also included qualitative/digital humanities; big untapped area; explore how you could turn pieces of what you’d use to analyze a corpus of text into data elements; how reuse
* Do we have everything in DDI to support this vision? DDI doesn’t have to do everything, other standards can fill the gaps, vision should elaborate how it interrelates w/other standards
* Do we want to just try tackling a difficult problem, take a step (e.g., similarity index)
* There’s the similarity of concepts and also looking at existing descriptions of surveys where you want to define similarity among studies; distinction between data and metadata
* Arofan and example of 8-dimensional context measurement
* Should element registry contain elements from related/neighbour standards; and also an element type registry
* Be careful in looking at health ontologies b/their concepts are more sharply defined
* How do we get people enthused on developing it
* Problems (funding and data collection) are different among different countries and different types of actors (e.g., commercial companies)
* Criteria for a good standard: participation on people representing all the various stakeholder groups; will this project be a push or pull project to get people to be involved;\*\*this is a weakness for us; A: has to be an open process, take little pieces and do demonstration projects
* Looks like a pipeline for metadata ingestion/creation/sharing; ID points where can interoperate w/existing tools/systems/standards and can communicate about bringing all together; doesn’t seem to be an overarching infrastructure but rather as middleware; A: all of the junctures are independent from all of the others, redundancy across the system
* How to organize the effort writing document? Priority of projects?
* R.e. GSBPM process model to allow other people to talk to each other; what we could do w/this, where do we put it so that it’s visible/promote; NAS workshop on transparency of federal statistics
* Part of pipeline already done (ESS, SHARE); add to functionality of existing things, maybe leverage these more
* CESSDA question bank isn’t the same as a registry; \*\*work on this
* Low awareness of DDI among other standards
* Another community to benefit: researchers in institutions, infrastructure to support open science (e.g., OSF & other existing tools)
* If want new buy in, most people attracted to this aren’t using DDI already
* Copyright and licensing issues for questions
* Summary of holes:
  + Need to ID holes in DDI
  + List of issues to be addressed for each new tool
  + Inclusion of related standards

Registry models (Arofan):

* CESSDA: centralized catalog w/centralized metadata store, harvests via OAI-PMH from various metadata stores
* SDMX registry: can subscribe to notification events on updates in central system (or from distributed partners); concepts, code lists, metadata structures; don't support textual searchers; interfaces for interacting w/registry the same as for the distributed data stores (REST and SOAP)
* Australian Bureau of Statistics: XMl metadata model; XML metadata model, code generation turns into web services functions; when come across a new kind of metadata, describe as XML and then automatically generate search interface; Distributed storage and centralized catalog for now, but moving to centralized storage; event-driven; versioning is important
* IHSN survey catalog: open source example using DDI codebook,
* Issues: subscription notification vs. harvesting, standard services interfaces; link between data and publications (DOI in publications are ideal; link to those who are working on this); different levels of access and associated metadata
* Software available for registries: Eurostat, MTUK/MTUK (ask Pascal), ask ABS about MRR; worth looking into this more (even broader tools)
* Where discuss these issues more broadly: AAPOR, official statistics, health sciences (public health, epidemiology), Wellcome Trust data forum (public health funder council), IHSN/World Bank, RDA, learn from DWB project
* Kndefjgawi principles

# Wednesday:

# Thursday:

# Friday: