**DDI-CVG Meeting Minutes, September 21, 2016**

Present: Alexander, Anne, Hilde, Sanda, Stefan, Taina.

1. Proposed amendment to the Sampling CV:
   1. Alex has prepared a draft that we discuss, and approve with some changes. Sanda will have to publish a new version.
2. Definitions for the Data Collection Instrument CV:
   1. All draft definitions are reviewed, edited and then approved. The CV appears ready for publication! Another task for Sanda.
3. We did not discuss what comes next – perhaps the Data Type by Provenance?
   1. Group, please provide input.
   2. No specific tasks were assigned, we might do this by email when we know for sure what comes next.
4. Anne, Taina and Sanda will be in Dagstuhl the last week of October.

Next meeting: November 9, at 13:00 GMT

**DDI-CVG Meeting Minutes, August 10, 2016**

Present: Alexander, Anne, Hilde, Sanda, Stefan, Taina.

Online: Sebastian

1. Dagstuhl: Right now it looks like Anne, Sanda, and Taina are going for sure.
2. Nordic Health Data Project requests:
   1. Sanda’s new proposal for an amended definition for Mode of Collection term Physical Measurements and Tests I reviewed and approved. We also have an approved amendment for the definition of Observation. So at this point Sanda can go ahead and issue a new version of Mode of Collection.
   2. Randomized Control Trial, Ecological Study, and Case-control Study will be considered as sub-terms for Clinical Data in the Kind of Data Provenance list, when we get to work on it. This suggestion is entered in an updated document that includes the Provenance list.
   3. The rest of the requests seem to have been dealt with, but we should return to the Nordic Health Data spreadsheet when we consider the Provenance terms. There is still an issue with a desired term “Register data” although at first sight it seems too vague and difficult to define (and we have registers in Data Source).
3. Instrument Type list
   1. We have talked about this as Research Instrument, but in the DDI it appears defined as “Data Collection Instrument” so we will probably need to include “Data Collection…” in the list title. It helps clarify the content of the list – “research” might be too broad or vague (this is just Sanda’s comment, perhaps we can confirm it with the group at our next meeting).
   2. Sebastian has provided very useful comments, mostly regarding the definitions. We will look at them as we edit the definitions.
   3. We decide to have an upper level term for questionnaires (i.e., Questionnaire) that will include all different types of questionnaires as sub-terms.
   4. For the guides/instructions terms, we decide that we prefer Data Collection Guidelines as an upper-level term. It is broader than Guides, and will also allow us to avoid repeating the same word at both levels. The proposed term “participant tasks” is changed to Task Settting and included as a sub-term under Data Collection Guidelines.
   5. The proposed term Stimulus Material is deleted, as it can be considered supporting material for other types of instruments, already included in the list.
   6. A new term “Technical Instruments” is added to account for measurements taken by various measuring instruments. Lab tests is deleted, as the new term “Technical Instruments” is broader and can include medical instrumentation (this will also be explicit in the definition, with examples from various domains.)
4. Instrument Type Definitions: We will start discussing them at our next meeting. Until then,
   1. Taina and Anne will review/revise/comment on the definitions for all of the Questionnaire terms, as well as Interview scheme or themes and the proposed Programming Script entry.
   2. Alex and Hilde will look at all of the Guides terms
   3. Sanda will revise/complete the definition for Technical Instruments.
   4. Stefan will make sure that his temporary move to the Asian continent is successful, and that he does not miss any of our future meetings.
5. Sampling Procedure definitions:
   1. Taina signals user confusion in classifying sampling method as Mixed Probability and Non-probability. We agree that this term should be used when the method is used for the same sample, not two different samples in a study. Alex will try to suggest a way of improving the definition to make that distinction clear. It might be enough to replace “the same target population” with “the same sample”. The last but one sentence in the definition may also be the source of some of the confusion.
6. Next meeting: September 21, at 13:00 London Time.

**DDI-CVG Meeting Minutes, July 6, 2016**

Present: Alex, Anne, Hilde, Sanda, Stefan, Taina

1. IASSIST, DDI Alliance Annual Meeting, and Dagstuhl 2016 sprint.
   1. Sanda has presented a poster about the CVs at IASSIST. She will email a PDF copy to the group, and all members are encouraged to take it, and use it as fit at any of the meetings they are attending. This would increase our visibility, and people’s awareness of our work.
   2. The DDI Alliance Technical Committee would like us to use a common workspace for our activities. The TC is now in the process of creating a space dedicated to our group, but visible to TC and other working groups. We will look at it together when it is finalized. For the time being it is not yet clear how it will work, or what exactly we are required to post in this space.
   3. The TC will soon send an official invitation to the Dagstuhl sprint (last week of October 2016) which will focus on Controlled Vocabularies. Several members of our group will be invited, and they are encouraged to attend. The invitation should also clarify the scope of the planned discussions.
2. CV amendments recommended by the Nordic Health Data Project.
   1. Mode of Collection: Measurements and Tests.Physical. We find that “examination” cannot be added to the term itself given the way we built the hierarchy, so we decide to edit the definition to include “medical examination (e.g., palpation or auscultation)”.
   2. The definition for Observation is also amended to include “development of condition or disease” in the examples.
   3. A more difficult task is including in our CVs terms to cover for “Randomized control trial (RCT)”, “Ecological study”, and “Case-control study”. These are really types of study designs, and we don’t have an element for study design in DDI. Data Kind by Provenance lists types of data, not of studies. A possibility might be to try to include them in Mode of Collection (?) Alex and Hilde will look into this for next time, and present some options.
3. Type of Research Instrument list.
   1. The “Questionnaire” terms and hierarchy seem ok.
   2. Should we create an upper-level for the rest of the terms that apply primarily to qualitative data collection? Perhaps “Guides and Directions”?
   3. The terms mentioned in b. above include the word “guide”, with the exception of “Interview schemes and themes”, which cannot really be changed to “interview guide(s)” because it would have a different meaning.
   4. We do agree to change “Data collection scheme” to “Secondary data collection guide”.
   5. Obviously, there are other types of “technical” instruments for collecting data, like those used for physical measurements and test. How do we deal with these, as there are many different types? Should we try to find a more “generic” term to include all? Or simply ignore them, as they can be specified using “Other”?
   6. Anne and Taina will review the list trying to figure out which terms make sense, and how they should be organized, keeping in mind the issues mentioned above. They will look at the definitions, but mainly for reference. We will review the latter in-depth when we have an agreement on the terms in the list.

Next meeting: August 10, 2016, 13:00 BST

**DDI-CVG Meeting Minutes, May 10, 2016**

Present: Alexander, Anne, Hilde, Sanda, Stefan, Taina.

Data Source Type:

We review some of the definitions that were still questionable and finalize this CV. The differences and final decisions may be identified by comparing the final version with the previous ones.

The goal is to publish Data Source, as well as Sampling Procedure and Kind of Data Format this month.

A “data entry” error found by Taina in ModeOfCollection V 2.0 will also be corrected, however, there won’t be a new version of this CV, Sanda will just have the files for V 2.0 replaced.

Restructuring the CV view/download page.

The current setup is found to be confusing, with people tending to download the files in the column closest to the definition, although these are the oldest versions. It is suggested that we present a page with just the most recent version, and create a link to a different page that will include a table similar to the current one, containing all versions. From the latter, the definitions may be deleted to make more room for the listing of multiple versions.

The package download should sit on the first page, but needs a statement that contains the most recent version of all CVs.

Sanda will look into this in June/July and will talk to our Web designer to see what can be done.

DDI-CVG work sprint at Dagstuhl.

Sanda will send the group the preliminary information she got from Wendy. She will also make further inquiries at IASSIST, as she meets with some of the TC members in person. She will provide more information as it becomes available. In the meantime, members from our group should look into possibilities to attend this week-long workshop at Dagstuhl, just to get a feel of what is planned, and hopefully also put forward our own views/needs/requests. Any practical decisions taken at Dagstuhl ,or as a result of the discussions held there will impact our work, even if this is not likely to happen in the very near future.

Membership:

Bodil Stenvig from DDA has just notified us that she can no longer be an active member of our group.

Next meeting:

Under Taina’s guidance, we will need to revisit the issue of the Nordic Health Data Project, specifically how it relates to our own work. There may not be much to do or discuss, but we need to make sure we have covered everything that needed attention before we move on to other topics.

It looks like the next product we should be working on is the Instrument Type list. Unless there is a good reason not to.

A Doodle poll has been set up for our next meeting date. Taina returns from vacation at the beginning of June, however we may hold a meeting without her at the end of May (hosted by Anne) depending on our other members’ availability.

**DDI-CVG Meeting Minutes, April 20 2016**

Present: Alex, Anne, Hilde, Sanda, Taina

Today we focus on finalizing the lists we’ve worked on lately.

1. KindOfDataFormat:

Anne suggests adding a 3D entry, and we include one.

We consider whether it would be necessary to add a sentence in the Geospatial definition that would address a situation where geospatial data would come in the form of images or 3D representations, but decide against it because the current definition for Geospatial appears to preclude a possible confusion.

Taina suggests that we review all the examples in this list and try to put as many as them in the plural, for consistency. Sanda will do that.

This list is approved, and closed. There will be a final review mostly for typos and other minor errors, but we do not plan to re-open discussion on more substantial aspects of the CV (unless, of course there appears to be a serious problem that we overlooked).

1. SamplingProcedure:

Probability.Stratified.Disproportional.

We amend the definition to include optimal allocation, without an explanation of what it is, which would be too technical/specialized for metadata people. But the term is there and can be found in searches. The examples are slightly edited for clarity/brevity but we think it is necessary to maintain both, because they explain how one can get different combinations of percentages in disproportional samples.

Probability.Cluster.StratifiedRandom.

We edit the definition to clarify the stages in which we have clusters, and strata in this procedure. We think it is clear now.

Probability.Multistage.

Alex clarifies that in this type of sampling there is a type of probability selection at each stage, including the last one (as opposed to cluster sampling in which all units are selected in the last stage). The definition is edited to indicate this, and considerably abbreviated, as some of the details seem unnecessary. The example is good, no changes.

Nonprobability.Purposive.

We discuss Sebastian’s suggestion to include additional explanations for the typology that is simply listed at the end. In the end it is agreed that they do not seem necessary for the purpose of this CV, which is basically to serve in metadata creation and search and retrieval operations.

(For the same reason, we had previously agreed not to provide subterms for this entry).

Mixed…

We had somehow lost this one but now it is happily retrieved. We review the definition, and find it okay. No changes needed.

This list is also approved, and closed. Same status as described above for the KindOfDataFormat list.

Sanda will send out both the .xls and the .xml documents for review when she has finished writing them. The review cannot take too long, so we will communicate by email regarding this.

Still pending is the Data Source list, and for this reason we decide to meet again in May, although it is a busy month for many of us.

Another important topic we have not touched upon is the Controlled Vocabularies working sprint at Dagstuhl. But there is still time for that. This is a week-long sprint, scheduled for the last week of October. Cessda has a stake in this, so relevant people should try to participate.

Our next meeting will be in the second week of May, date TBD.

**DDI-CVG Meeting Minutes, March 9, 2016**

Present: Alex, Anne, Sanda, Stefan, Taina

We have had again problems receiving emails sent to the list. Sanda will look into this, but since the list address does not seem to be 100 percent reliable, she recommends that people create a group contact in their own email software.

Here are the email addresses of our members:

[aether@essex.ac.uk](mailto:aether@essex.ac.uk)

[Taina.Jaaskelainen@staff.uta.fi](mailto:Taina.Jaaskelainen@staff.uta.fi) or [Taina.Jaaskelainen@uta.fi](mailto:Taina.Jaaskelainen@uta.fi) *(recommended to add both)*

[stefan.ekman@snd.gu.se](mailto:stefan.ekman@snd.gu.se)

H[ilde.Orten@nsd.no](mailto:ilde.Orten@nsd.no)

[sandai@umich.edu](mailto:sandai@umich.edu)

[Alexander.Jedinger@gesis.org](mailto:Alexander.Jedinger@gesis.org)

[bs@sa.dk](mailto:bs@sa.dk)

[sebastian.kocar@anu.edu.au](mailto:sebastian.kocar@anu.edu.au)

[Atle.Alvheim@nsd.no](mailto:Atle.Alvheim@nsd.no)

*(Atle has asked to stay on the list so that he can be informed of our progress. I have not heard from him in a long time, but I think he is still at NSD (Hilde?) so please add him too unless we hear otherwise from Hilde.)*

Please let me as well as the group know if I made any mistakes/typos in the list above. If there is no response in the next few days, I think it is safe to create the list.

At this meeting, we deal almost exclusively with the definitions for Sampling Procedure, more specifically the definitions for the Probability terms.

Definitions for Total Universe, Probability, SimpleRandom, SystematicRandom, Stratified, and StratifiedProportional, Cluster and Cluster.SimpleRandom are approved.

Definition for StratifiedDisproportional is tentatively approved. Group members are invited to show it to colleagues to find out whether it is clear enough.

Definition for Cluster.StratifiedRandom still needs to be reviewed for clarity.

Definitions for Multistage and Mixed (probability and non-) need to be discussed next time, but the discussion will be more productive if they were reviewed in the meantime.

Also for next time, we need to revisit and approve the definition for NonProbability.Purposive.

Plans to publish:

It looks like DataSource and DataKind by Format are ready to be published. Sanda will try to put together the spreadsheets and then send them over for review before she writes the xml.

Taina has pointed out another correction needing to be made to ModeOfCollection. Plus, there needs to be an edit to this list requested by the Nordic Health Project. Taina will get back to Sanda on this, and she will also send the spreadsheet indicating the changes we agreed to make in order to better align ourselves with the Nordic Health Project.

It would be ideal if we could publish the lists mentioned above as well as Sampling by May. If not, at least the two that are finalized. We will stay in touch on this.

Next meeting: April 20th at 13:00 GMT.

**DDI-CVG Meeting Minutes February 17, 2016**

Present: Alex, Anne, Hilde, Sanda, Stefan, Taina.

This meeting deals with the Sampling Procedure list, focusing mostly on the choice of terms and definitions for non-probability sampling.

1. Non-probability terms and definitions.
   1. We agree to globally change the word “elements” in the definitions to “units” or “sampling units”; group members please have an eye open for instances that might have been missed.
   2. Non-probability, upper level entry. We note Sebastian’s proposal for expanding the definition, but end up agreeing that it is not really necessary, as the shorter version is clear enough and seems complete. We do add a sentence about recommended usage of this term. Approved
   3. Convenience or availability. We continue the discussion regarding the possibility of creating sub-terms for this entry. It appears that very often “convenience”, “opportunity”, and even “accidental” may be used in practice for similar types of sampling, without a clear-cut distinction among them, and we decide it is better to not separate them. We do add a sentence in the definition to mention that this type may also be called “convenience” or “opportunity”. Self-selection is also mentioned in the definition, so this should facilitate searches on those particular terms. Approved.
   4. Purposive. We decide to forgo subdividing this term too. The various subtypes can be described in a text field if the information is available. We also find that providing one or two examples at this level might be misleading, because they would only reflect some of the methods that can be used and could thus unintentionally limit the scope of the definition. We tentatively mention some of the possible sub-types in the definition, and *Hilde* will try to show this part to some of her colleagues, to see if they clarify things, or perhaps create more confusion.
   5. Quota. We decide on no sub-terms for this one either, and attempt to explain the difference between proportional and non-proportional in the definition and example.
   6. Respondent-assisted. We are happy with the term, as it is more generic. No sub-terms, and we include “snowball” in the definition/example.
2. Main Sampling list.
   1. We decide to add a term for “Mixed probability and non-probability sampling”. *Alex* will try to work out a definition for next time. He will also go back to the Probability terms, where some examples are still needed. We may have to briefly review that half of the list, because some of the definitions may not be final.
3. If all goes well, at our next meeting we might be able to resume discussion on the tasks required from us from the Nordic Health Data project. *Taina and Stefan* will bring us up-to-date.
4. Next meeting: March 9, 13:00 GMT

**DDI-CVG Meeting Minutes January 13, 2016**

Present: Alex, Anne, Hilde, Meinhard, Sanda, Stefan, Taina

1. New members:
   1. Alex Jedinger from GESIS has graciously agreed to join our group. Thanks, Alex!
   2. Sebastian Kocar, from the Slovenian Data Archive, has also expressed interest in participating in our work. He is currently on a fellowship in Australia, so he won’t be able to attend our videoconferences, but he can participate by email, like Michaela Olde used to. We will give this a try. Sanda will add him to the mailing list, and will give him a brief introduction.
2. Sampling Procedure CV
   1. Probability sampling.

We look at Alex’s definitions and examples for the probability terms. Most are approved with minor edits. Upper-level “Probability” and “Simple Random” might not need examples.

Some work still needs done on:

* Stratified.Disproportional. We need to re-examine the example to make it as clear as possible
* Cluster.Simple, Cluster.Stratified, and Multistage. Alex will attempt to provide examples for these terms that will be all drawn from the classroom example.
  1. Non-probability sampling.

In a brief discussion, we agree for the following higher-level classes: Availability, Purposive, Quota and Respondent-driven. Taina will provide definitions for these, while Sanda will continue to look at the subclasses and do some research for the definitions.

We already have GESIS definitions, so these don’t need to be attempted from scratch.

1. Nordic Health Data Project.

We don’t really have time to talk about this topic, despite recent contributions to the discussion by Stefan and Anne. It will be important to go back to this topic before we finalize the Kind of Data lists.

1. Next meeting:

February 17, at 13:00 GMT.

February 10 is also a possibility, although not preferred because Hilde cannot attend.

People should let Sanda know if they find out they have schedule conflicts for February

17, because we want as many live participants as possible.

**DDI-CVG Meeting Minutes November 24, 2015**

Present: Anne, Hilde, Meinhard, Sanda, Stefan, Taina

Alex Jedinger from GESIS graciously joins us for the Sampling list discussion (he worked on the GESIS list).

1. Sampling Procedure CV:

FSD and GESIS have already done extensive work on this CV and there is a draft available from the DDI-SDI (Survey Design and Implementation) working group as well. Our discussion starts from examining and comparing these lists.

At this meeting we decide to include a term for “no sampling” studies and tentatively name it Total universe/Complete enumeration

We also come to a general agreement regarding the Probability terms:

-We decide to include the word “sampling” only in the uppermost level, to assist in searches, but will not repeat it in the subterms.

-We also decide to drop the subterms from “multistage” sampling that appear in the GESIS list. We will try to come up with a definition for “multistage” that will mention the possible variations of this sampling technique, so that we do not entirely lose the detail provided by the GESIS subterms.

Nonprobability terms:

-We will deal with these at our next meeting, however there is some discussion aabout the Availability subterm. Most people think that “availability” is a better term than accidental, or convenience (?) but we cannot come to a conclusion on whether this includes “self-selection” or we need to make “self-selection” a separate entry. We will resume this discussion at out next meeting.

Alex agrees to join us for another two meetings or so (as needed), and share his expertise in this field. He also offers to work on the definitions for the Probability sampling terms until our next meeting.

1. Nordic Health Data Project search terms and the DDI CVs.

With the help of Elizabeth from SDA we look at the search categories and terms proposed by the Nordic Health Project to see if and how we can match them up with DDI CV terms. The preliminary results are captured in the Nordic Data Health Project wishes vs DDI CVs V3 – mostly Taina’s work before the meeting, but also includes some of the group’s decisions/comments. Taina with Stefan and Elizabeth will do some further work on this and let us know before our next meeting. It looks like some of our published lists will have to be amended.

Next meeting: Wednesday, January 13, at 13:00 GMT.

**DDI-CVG Meeting Minutes October 21, 2015**

Present: Hilde, Meinhard, Sanda, Stefan, Taina

1. Data Kind by Format list and definitions.

We like the new definition for “geospatial” contribute by Stefan (and co-worker) and we move it to the final draft. We review some of the terms – can there be overlap between “interactive resource” and “software”? We think not, the “resource” would be the ‘product’ of a piece of software, not the program itself. Is “still image” translatable in most languages? There might be a problem with Finnish, but the other languages seem to support an accurate translation, so no changes are suggested, or made. This list is ready for publication, and we will not be discussing it again, unless somebody brings it up for some reason.

1. OtherMaterials CV.

We review the history and context for this list so that we can better decide on a strategy. The list was created to assist in creating metadata for three DDI-C fields, and includes many of the terms suggested in the DDI-C field documentation, and others that were considered necessary, but not found. DDI-L uses the OtherMaterialType format for six elements that are conceptually different (and would likely need different CVs for the Type). It is also suggested, and agreed, that it might not be such a good idea to publish an unpolished list because then people at different institutions would start creating local lists to answer their needs, and the purpose of a CV is precisely to prevent that (in order to ensure interoperability).

So it is decided that we will work on this list to make it acceptable to us, and publish it only when it meets the same quality standards we have been using for the other lists. We might still make the Slovenian co-authors of the list, since they initiated it. But Sanda will have to let them know what we’ve decided, because they expect their list to be published as such.

1. Order of priorities at this time:
   1. Instrument Type
   2. Sampling
   3. Other Materials
   4. Data Kind by Provenance
2. Instrument Type.

We look at the DDI 3.2 field definition. This is the Data Collection Instrument, so we need to keep this in mind when editing the list. And the definition for the field itself needs to be changed to reflect this understanding. We look at the list and are able to decide that some of the terms need to be eliminated. These are greyed out in the draft document, and two about which we are not sure are greyed out but also moved to column B, which tentatively reflects the edited list. Next week Taina will participate in a meeting with people from the medical sciences and she will bring us their input on specific instruments they might use to collect their data. We need to cover this area of collecting scientific data through measurements, hopefully by finding just a few terms that might cover classes of instruments (we do not want excessive detail).

Tasks for next time:

Sanda – definitions for Questionnaire and sub-terms

Hilde – definition for Interview Scheme or themes

Anne – definition for Observation guide

Stefan – please spend some time with the most recent Other Materials draft and provide feedback. Settling on a list of terms is a priority, we can work out definitions later.

Meinhard – please look at the existing Sampling CV draft and comment – here I think definitions also need attention, because some may be reworked to include things that appear to have been left out?

Taina – in addition to the task mentioned at point 4. above, please provide input on the Sampling CV as soon as you have any results from the on-going discussion at FSD.

**DDI-CVG Meeting Minutes September 23, 2015**

Present: Hilde, Sanda, Stefan, Taina

1. Creative Commons license.
   1. Taina will be looking into any issues that may be raised by using this license (or some of its versions) and will let us know when she has reliable information. This may be relevant to archives intending to publish translations of our CVs on their own websites.
2. Data Kind by Format list.
   1. We review the list, currently the agreed upon version includes the following terms: numeric, text, still image, geospatial, audio, video, software, interactive resource.
3. Data Kind by Format definitions.
   1. We come up with what we think are good definitions for “numeric” and “text”; the definitions, and the examples offered under “text” are clear enough and suggest the differences between “numeric” datasets and text files without mentioning the quantitative vs. qualitative distinction (is already covered in DDI) or the ways in which these data are supposed to be analyzed, which is what we wanted.
   2. Geospatial. There is more discussion around this definition, including on whether ‘maps’ are images or geospatial data. Depending on their features they can be both but if we use the definition from EPA's National Geospatial Data Policy they can be included in geospatial data because the definition provides for a variety of forms, both digitized and non-digitized. This is still under review, Anne, Bodil, Meinhard please provide feedback.
   3. We proceed to modify the DataCite definition for “Interactive Resource” in an attempt to make it clearer. We think it works in this edited version. It’s probably ready for approval.
4. Instrument Type.

There is some general discussion about the list put together at FSD.

* 1. The term “instructions” is probably not the best because it may be confused with the “interviewer instructions” that are part of questionnaires. For instructions on conducting a discussion (only general outlines provided) or on writing an essay or narrative (only theme and general guidelines provided to respondent) we will probably want to use another term like “guide” or “protocol”.
  2. Another important point that we need to consider is the difference between an instrument as a whole and components of an instrument. Visual stimuli for instance or other stimulus material are most likely only part of an instrument because they would be accompanied by instructions on how to use them or react to them…
  3. All group members need to take a close look at this list, and be able to comment on both the list and the definitions at our next meeting. Email discussions are welcome, but not the day before the meeting.

1. Next meeting: October 21, at 13:00 GMT (same time as usual)

**DDI-CVG meeting, August 19, 2015**

Present: Anne, Bodil, Sanda, Stefan, Taina

Source Type.

The Source Type list is closed. Sanda will try to publish as soon as she finds some time. If anybody has any changes to suggest, please hurry. (Sanda also has to produce corrected versions of TimeMethod, just to edit a “typo” found by Taina).

1. Creative Commons licence V. 4.0

We will be using V 4.0 moving forward but will not go back and fix the CVs that are already published. It is okay to publish a translation using V 4.0 even if the “original” US-English version states a different version.

1. Data Kind by Format: definitions, list.
   1. Numeric and text: We still do not have a clear idea/decision on how to treat these two entries, especially with regard to those datasets that are numeric but also contain “text” data (“alphanumeric” variables). We could define the latter as simply “numeric” because they can be analyzed with statistical software. Do we need to add something to the definition to specify this? Do we want to mention in the definition for “text” that this entry should be used only for files that contain mostly text (i.e. qualitative data, see examples in table)? Or perhaps leave the definitions more general (and let users/archives decide on how they use the two terms to describe their data? Taina, Stefan, Anne, Bodil will talk to people at their institutions who are more familiar with processing/analyzing data and will report on the discussion results. Hilde, please let us know what you think as well.
   2. Alpha-numeric: it is generally agreed that we should leave this term out, as it seems to apply to individual variables rather than datasets, but we still need to keep the concept in mind when we talk about numeric vs. text (point a. above).
   3. GIS: this is a difficult issue and needs more investigation. Basically, “GIS” data comes in multiple “formats” but they are delivered and processed together, they “make sense” only taken together. It appears however that this type of data can also be processed by other means or systems than GIS. After consultation with a colleague Stefan suggests we enter “geospatial” in formats and GIS data in the Provenance list. But we are still lacking a clear and convincing definition. Stefan will look further into this and make concrete suggestions. We all need to think about it though, and try to better understand these data as they are becoming increasingly popular and widespread.
   4. Still image, Audio, Video, Software: we think we’ve nailed the definitions for these terms. Hilde: please review for approval.
2. Instrument Type: as agreed, we will turn to this list after finalizing the Data Type by Format, because FSD has already put together a list, with definitions. So we will keep in mind our original proposal (page 6 of “Rationale for new approach…” document) but will concentrate on reviewing FSD’s list, provided by Taina. Those of us who work in archives holding a lot of qualitative data are invited to take a close look (Bodil, Anne, but also others as appropriate).

**DDI-CVG Meeting Minutes July 8, 2015**

Present: Sanda, Stefan, Taina.

1. Source Type definitions.
   1. The latest drafts of the definitions for Communications, Communications:Public, and Biological Samples are approved, with just one minor edit on the latter. The rest of the group is invited to take a look at the most recent version of the working document, and if there are no voices against them by our next meeting, we will consider them final.
2. Using the Creative Commons license V 4.0. We forgot to talk about this, but basically there is agreement that we would use V 4.0 moving forward, but will not go back to edit the published CVs for this reason only. Hilde raises the question of whether it might be possible to just publish the license once, on the same web page as the vocabularies (see for example <https://github.com/DASISH/survey-tools-metadata>). We did not talk about this, but I am not sure. I would have to ask TIC, I suppose, but even so the current vocabularies have the text in each xml instance, so we would not be consistent. We can go back to this issue at some later date as it does not seem urgent.
3. Regarding the order in which we should approach the lists, we would prefer to look at “(Research) Instrument Type” and then “Sampling” after the Data Kind by Format list. Data Kind by Provenance appears to be less urgent, and there is existing work that we can capitalize on for both the Instrument and the Sampling lists.
4. Data Kind by format

Terms in list:

* Before we try to decide which terms make sense, we also look at ICPSR’s solution and a list of quantitative data “formats” or types put together in a paper by Achim and Larry Hoyle. We also look at some of the DataCite CVs, although their approach is a little different.
* In our own list, we start by highlighting the terms that we think \*must\* appear: numeric, text, (still) image, GIS, sound/audio, video/(moving image), program source code and interactive resource.
  + Regarding “web pages” Stefan brings the argument that in fact they are stored as text+(image)+(sound)+(video), so they would be placed at a different level than simple formats (and would be composites of simple formats). So we decide to tentatively move web pages to the “provenance” list.
  + Maps per se might also not have a place in this list because if stored as “maps” they would in fact be images, whereas GIS is a specific format for files that are processed by software before they can be viewed as renditions of geographic places.
  + Program source code and software: as Stefan points out, there is a difference in that the software would be compiled, and the source code would be distributed as such, amenable to modifications. Do we want to include both terms as a single entry with a slash and mention the difference in the definition? Or is there another option?
  + Interactive resource. This can be included among “formats” because it requires interaction in order to be brought to life (whereas software is “static”, and using it does not change it).
  + There is also discussion around a potential “alpha-numeric” entry. We think that this term might be more appropriate for a variable-level description, but at study or file level we would tend to use either “numeric” or “text”. This is a fuzzy area and we can, and should continue this discussion with the other members who are not present today. Our tendency would be to delete it.

When we consider these terms it is also important to always remember the attribute @type that qualifies this DDI field, with 3 possible entries: qualitative, quantitative, mixed. Any term can be used in conjunction with this attribute.

Because there are about 6 weeks until our next meeting, I am assigning tasks to people, and am reassuring everyone that it will be “okay” if you are on vacation or otherwise unable to do much. Any contribution, however small, will help us move forward faster.

Definitions for:

Numeric: Taina

Text: Anne

Still image: Sanda

GIS: Hilde

Sound/Audio: Anne

Video: Sanda

Program source code/(Software?): Stefan

Interactive resource: Taina and Hilde (there is already a tentative definition for this one)

Next meeting: August 19, 2015 at 13:00 GMT

**DDI-CVG Meeting Minutes May 20, 2015**

Present: Hilde, Sanda, Stefan, Taina.

1. It turns out that FSD has produced a CV for Instrument Type and is now working on a CV for Sampling Procedure.
   1. Instrument: Since creating a list for Instrument is also on our to-do list, Taina agrees to send FSD’s CV to our group for evaluation. We can either edit it and publish it as our own, or publish it as FSD’s, on our site, if both sides agree to a unique format. The latter would be ideal. We should avoid having multiple variants of a CV published, or in use, if we can come up with a unique version.
   2. Sampling Procedure. The DDI-SDI (Survey Design and Implementation) Working Group has already put together a draft CV for sampling procedure. Sanda has it and will send it to the entire group. Taina will look at it, see how it differs from FSD’s draft, and we will discuss a possible outcome like the one described in a. above – i.e., end up publishing and using a unique version.

Taina and Stefan also need to remember to send over any finalized translations so that we can post them on the DDI-CVG (sub)site.

1. Finalizing the “Source Type” list.
   1. There are no strong suggestions for modification in the actual list (although some people might be uncomfortable with some of the terms, mostly for “philosophical” reasons; the list as it stands does reflect a change in perspective, or in understanding of what a data source can be).
   2. All of the definitions are reviewed, paying special attention to the ones that are still unfinalized. Hilde and Anne have contributed updated comments and these are also carefully considered. A “history” of our discussions, with past and more recent comments, as well as various suggestions for definitions are preserved in the file named “SourceTypeAndDataKindProposal 2015-05-15”. The new working document is dated 2015-05-20 (in the file name).
   3. Most pending definitions are approved with only minor adjustments. Two are left for further consideration:
      1. Communications:Public. We all agree that we need to reflect the fact that mass media can take different forms – print, air wave (conventional radio and TV) or internet (Web?) – but cannot come up with a satisfying definition that conveys this idea. Hilde finds a Wikipedia entry that might help us, so we decide that Hilde and Sanda are going to look further into this. The current draft has the wording “for example, in person… etc.” in response to Taina’s observation that some types of public communications, like leaflets/flyers or public reports (?) were not covered in the previous draft. We feel that the current wording implicitly includes them, without naming them, which is what we want.
      2. Biological samples. Taina presents a more detailed examples section that we like. However, the current definition only seems to cover for samples from humans and animals. Stefan points out that samples may be taken from plants as well, so how do we extend the definition to cover all living beings? Taina and Stefan will come up with some proposals at, or before, our next meeting.
      3. As a note, in the definition for “Population group” we have replaced “people” with “individuals” to allow for applying this term to groups of animals, if necessary. This definition has been approved. Should we also think in terms of plant populations? We are walking a fine line here.

Next meeting: Sanda will create a Doodle poll, because many people were not present today, and some may already start taking vacation time. Please respond as soon as you can.

**DDI-CVG Meeting Minutes, April 22, 2015**

Present: Anne, Bodil, Hilde, Meinhard, Sanda, Stefan, Taina

1. Revised Time Method: approved without any new changes; Taina will send the final spreadsheet to Sanda, just to ensure it is the correct version. This file will be used for publication.
2. Revised Mode of Collection: some details, mostly in the definitions, are being discussed:
   1. Interview. Face-to-face: CAPI/CAMI: can CAPI and CAMI be combine in the same entry? Is CAMI always “face-to-face”? A web page (<http://www.soft-concept.com/surveymag-en/collection-data-methods-4.htm>) found by Meinhard seems to suggest “yes”: “*The offer of mobile devices incredibly increased on these last few years, and phones and digital tablets went over the CAPI tools in each technical and ergonomic sector. These defaults and the popularity of these new digital devices lead to a disappearance of the CAPI and the appearance of the CAMI, the mobile survey technology that we are going to detail in the next slide.”* We are satisfied to leave them together.
   2. Other minor edits meant to increase definition clarity are made and accepted for Self-administered questionnaire: Email, SMS/MMS,Web-based, and Computer-assisted
   3. Edited definition accepted also for Observation.Computer-based. We also review the new definition for Observation.Computer-based in conjunction with the definition for Recording to see if there is a conceptual overlap that can lead to confusion regarding their specific usage. We decide that misunderstanding or misuse is unlikely.

Revised Mode of collection approved. Taina will send final spreadsheet to Sanda for publication. FSD (and others, as needed) should feel free to use these two new list versions as if they were published.

1. Type of video clip and audio clip: we do need to publish these even if we are not experts and therefore not able to produce “perfect” lists. We look at Meinhard’s draft, and there is a question whether Smil format is really appropriate for this list, as it appears to be XML metadata for audio and video (rather than audio and/or video per se). At this time, we are still unsure whether to delete the entry or leave it in. Stefan notes that some of the definitions need better/more complete phrasing and agrees to edit them. Sanda will review his edits and then we will present them to the group.
2. Unfinalized Source Type definitions. These are listed in the agenda for April 2015. Anne and Hilde are charged with reviewing them, and next time (in May) we will try to approve all or most.
3. Taina expresses an interest in the Slovenian contribution. Sanda will try to find the original email and send it to her.

Next meeting: May 20th, 2015, at 13:00 GMT

**DDI\_CVG Meeting Minutes, March 25, 2015**

Present: Sanda, Stefan, Taina.

Today’s meeting is focused on finalizing the new version of Mode of Collection.

Taina has put together a draft for new entries as well as some new or edited definitions, based on the discussion from the February 2015 meeting.

Based on the observation that many computer-assisted personal interviews are now conducted using mobile devices like tablets or smartphones, we consider whether we should have a separate entry for these (CAMI) but in the end we decide to combine CAPI and CAMI within the same term and adjust the definition (the difference between laptops and tablets is becoming blurred as new devices are produced that combine their features for enhanced usability).

New definition for SelfAdministeredQuestionnaire was amended as a consequence of eliminating “fixed form vs. interactive” as the next level. It looks good.

Regarding Self Administered Web Based, we decide to eliminate the “fixed form vs. interactive” distinction, mostly because a) most web based questionnaires are interactive, and b) sometimes it is not known whether the questionnaire was interactive or not. Taina will edit the definition for Web based to include more text from the Wikipedia definition which allows for the possibility of the questionnaire being interactive, or not.

A similar decision was taken regarding Self-Administered Computer-Assisted: we eliminate the “fixed form vs. interactive” distinction and Taina will rephrase the definition to allow for either form. She will also mention in the definition that CASI covers all forms of CASI – meaning VCASI, ACASI, TACASI: there will be no additional entries for these.

We also discuss the possibility of removing some, or all examples of aids or measures used in projective methods (for MeasurementsAndTests.Psychological) but in the end decide to leave them in. We think that more detail may be useful if users are not quite familiar with projective methods and how they are conducted.

Taina will send a final draft to the group as soon as she is able to, and we will not wait until next meeting to consider it approved. When it is sent, people will have a couple of weeks to comment and/or disagree, and if there are no comments, we will consider this new version approved.

Next meeting: April 22, 2015, at 13:00 GMT.

**DDI-CVG Meeting Minutes, February 25, 2015**

Present: Anne, Hilde, Sanda, Stefan, Taina

1. Revisiting ModeOfCollection. We examine Taina’s proposals. The suggestion to add “Paper.Mail” as sub-terms for SelfAdministeredQuestionnaire is rejected; the various ways in which paper questionnaires might be distributed and collected are mentioned in the definition for “Paper” and we feel that if we create a sub-term for “mail” we would need to create sub-terms for the other modes (fax, in person, etc.) and this would make the list even longer.  
   We also consider whether we might combine the entries for “web-based” and “computer-assisted” into just one category that would say something like “computer-based”. But, after some discussion we agree that perhaps the distinction web-based vs. computer-assisted should be maintained, so we end up approving Taina’s proposal 6), minus the “Paper.Mail” entry.

Taina will now examine the definitions for all the approved terms in proposal 6) to see if and how they need changing, to accommodate for the edited list. One important thing to capture would be the difference between Web-Based and Computer-Assisted, which is mainly that in the second case the interview software would be saved onto a computer or on a local server, and not accessible through the Internet. The respondents would need to access a computer or memory device handed over to them by the survey administrators, or sit in a lab in order to answer the questions.

1. Revisiting Time Method. We discuss Taina’s proposal and revisit the existing list to see how the various methods are represented. All possibilities seem to be covered even though perhaps not ideally. It does not seem really necessary to add new terms, but perhaps we can edit the definitions to make the terms clearer and easier to use. We agree to change the first sentence of definition of Cross-Section to say: “Cross-sectional data is collected by observing subjects within the study period, without regard to differences in time….” Taina has also suggested some changes in other definitions in her new proposal, but we haven’t really had time to look at them.

So in the next few weeks she will compile a new spreadsheet that will include all of the proposed changes, so that we can take another look at it. Some examples will also be added, to further clarify usage. Intervention/clinical studies would be examples of panel studies, and Hilde offers to send Taina one or two more examples, to be included in the edited spreadsheet.

Sanda has found a new type of time method: cross-sequential. We did not discuss this as we were already overwhelmed by the existing list. Perhaps we can bring this up as an option next time?

1. Type of Address. Definitions are approved, list is closed.
2. Type of note. For “problem”-type, there are a number of options, though none is completely satisfactory. We end up agreeing on “Observation” with a clearer definition suggested by Taina, and edited and approved in the meeting. We will consider this list closed as well.

**DDI-CVG Meeting Minutes** January 21, 2015

Present: Anne, Hilde, Sanda, Stefan, Taina

Externalized lists:

1. Type of address. More discussion on the list itself, including proposal from Hilde. Since Individual and Organization are already defined in DDI at the main entry level, it would be redundant to keep these words in the CV terms. It is however noted that the “residential” piece of the list will never apply to organizations, therefore it makes more sense to keep “mailing vs. visiting” at the highest level. We will make sure that the definition for “residential” makes it clear that this subterm does not apply (will not be used?) when defining contact information for organizations. Sanda will draft some definitions for next time.
2. Type of telephone. The discussion focuses mainly on making a distinction between a telephone number used to access the organization as an entity (a sort of a “central” number, or switchboard if there is one) as opposed to personal numbers that are not private, but used to access people “at work”. Definitions are adjusted to reflect this. All definitions are approved, so we can consider the list closed.
3. Videoclip and audioclip. Meinhard has sent an .xls file that will allow us to examine the definitions, but we are not ready to discuss them at this point. Will try next time, hopefully Meinhard can participate in the discussion. Stefan will show the lists to any “technical” people he can find at SND just to get a sense of whether they seem “good enough” (“perfect” is not necessary).
4. Type of note. Terms and edited definitions are approved, with the exception of “Problem”. We generally agree with Hilde’s argument that this note could be about an “issue” that is not necessarily negative, or a problem or error that needs correcting. It could also be about an irregularity, or inconsistency in the data (or documentation) that is not expected, but cannot, or should not be fixed. So we need to find a better term. “Issue” is an option, but not ideal.

We will be exploring other possible terms for next time. Everybody needs to contribute.

1. Concept group. “Administrative” is acceptable, for situations in which the data are re-organized according to certain groups of concepts. “Thematic” is the preferred alternative to Conceptual (for the 2nd term). Definitions approved.

Reviewing Time Method and Mode of Collection.

1. Time Method: Taina is preparing a proposal, she might be ready by February
2. Mode of Collection: Taina has sent a few suggestions to cover for the absence of a separate entry for “Online questionnaire” when no information is available on whether it is interactive or not. We have not had time to look at these options in detail, entire group please do so for the February meeting.

Source Type definitions; Assign Data Kind definitions: We did not get to these, hopefully next time?

Next meeting: February 25, 2015, at 13:00 GMT

**DDI-CVG Revised Meeting Minutes** December 10, 2014

Present: Anne, Bodil, Meinhard, Sanda, Stefan, Taina.

Today’s discussion concentrates on the externalized vocabularies.

1. Type of address. While we agree that the distinction between residential and nonresidential is important, there is discussion whether the second distinction we are making should be between mail going to a physical address vs. being held at the post office (PO Box or poste-restante) or rather between a visiting vs. mailing type of address. People generally agree that the latter might be more important so we redraw the list to reflect these two types of distinctions – residential vs. nonresidential on the first level, and mailing vs. visiting on the second level.

Sanda will try to edit the definitions so that they conform to the new list structure.

1. Type of telephone. After some discussion on the need to cover online/Internet connections or other more modern types, we decide to leave the list as is -- but include satellite-type connections in the definition for mobile/cell phones. Anne will try to edit Hilde’s definitions, just to make them more concise. The first three are tentatively edited during the meeting.
2. Type of Instant Messaging. Most members express hesitation to publish a list taken directly from Wikipedia. This would not be our product, and we could not assume responsibility for it in the same way that we do for the lists we normally produce. Plus, a vocabulary for instant messaging might not be that useful in the context of documenting contact information, given the multitude of applications that appear and disappear, and the potentially transitory nature of accounts. So we decide to temporarily drop this CV from our to-do list.
3. Type of email. We cannot think of use-cases in which knowing the type of email might be necessary for contacting someone by email. Just knowing the address seems to work very well in all circumstances. Therefore we decide to drop this CV from our to-do list as well.
4. Type of videoclip and audioclip. Despite the fact that we are not experts in this field we agree that these lists might be useful, as they would be used to help identify specific segments in video or audio data (likely qualitative data). So we all need to try to find and consult experts to see what would be a good solution for these CVs. Organizations that actually store audio and video data would be good places to look for advice. Meinhard has sent a message with some ideas on November 13, resent on December 10. We have neglected to look at it so far, but will keep in mind to do so, in conjunction with any new information we might find.
5. Type of note. The old (externalize) vocabulary seems to cover the type of note “content”, which is fine, except that “footnote” does not really fall into the same category as the other terms, because it indicates the placement of the note in a printed document rather than its content. We therefore agree to delete “footnote” from the list, and publish the rest as is. Sanda will try to have a look at the definitions to see if they seem ok.
6. Type of concept group. We are not comfortable with the current list. Sanda suggests some changes in wording (“conceptual” type only repeats the field defined, so it is not suggestive enough) and proposes a modified definition. However, the main problem for us is the dichotomy conceptual vs. administrative. We cannot think of a good example of an “administrative” concept group. Sanda thinks this distinction might have been inherited from DDI\_C that allows variables to be grouped together for administrative purposes (subsetting ? or lack of conceptual grouping, in which case variables are just grouped sequentially for better manageability). While this may work for variables (describing data), it is unclear when and how concepts may be grouped for administrative purposes, when concepts are purely theoretical constructs. We all need to think more about this.

Other issues:

1. Taina suggests that we need to revise TimeMethod to better account for studies that are not typical longitudinal projects, but do have follow-ups. She will try to come up with a tentative proposal by next month (January 2014)
2. Taina has created, and has offered to maintain, a document that brings together all of the minutes from our meetings in the past few years. This will be very useful for searching, and finding discussions and decisions of various topics. The Next meeting: January 21, 2015, at 13:00 GMT

**DDI-CVG Meeting Minutes**

November 12, 2014

Present: Anne, Hilde, Meinhard, Sanda, Taina.

(Bodil has trouble with setting up the audio system. Taina has kindly offered to try to help with this before next meeting, if possible)

Externalized vocabularies:

Language proficiency.

Current version unanimously approved, final.

Type of address.

A shorter list makes sense, but people feel it could be organized better. We agree that the distinction between residential and non-residential is more important than physical vs. postal, so we will keep the former at the top level in our list. Alternatives like private/personal for residential and business/organization for non-residential were considered, but in the end we agreed to introduce them in the definitions for the terms. The distinction physical vs. postal will constitute subcategories for both residential and non-residential. We will not have a third level for P.O. Box and poste-restante, but rather mention these in the definition for postal.

Sanda will prepare a new list and will also attempt to draft definitions for these terms.

Type of telephone.

The general layout in the new proposal makes sense. However, a term for videophone might not be necessary. It does not seem to be commonly used and could be mentioned under Other, if need be. So we delete it from the list. Pager does not make sense in this list because it is used for specialized purposes, not as a general means of establishing contact. It is also deleted. Meinhard mentions the possibility of telephoning through the Internet (rather than dedicated phone lines). He will research this for next time and will come up with a proposed term and draft definition if he finds this merits an entry in the list. Hilde offers to draft some definitions for the other (approved) terms on the list.

Instant Messaging.

There is general agreement that we can go ahead and publish the list from Wikipedia as is, especially as Wikipedia also provides definitions. Sanda will complete the spreadsheet. Since this is a list for which we cannot really claim authorship, it is suggested, and approved, that we make a note in the description section for the entire list showing that it was taken as such from Wikipedia, give a URL, and mention the date. Hilde offers to send a standard way of citing Wikipedia as a reference. Sanda will try to insert this type of citation in the spreadsheet.

Concept group type.

The existing list looks fine at first glance, but less convincing as we try to figure out how the two terms would be actually used. We might need to clarify the treatment of concepts and other related elements, like “subject” and “keyword” in the DDI 3.2 specification. Sanda offers to look into this, and send out a brief ‘report’, hopefully before the next meeting.

Type of video clip and type of audio clip.

We really have no expertise in this, but there is a feeling that these lists might be useful in locating segments of video/audio clips stored in archives as qualitative data. Meinhard will contact Stefan and they will try to work together toward a more final proposal (also, please point out, if possible, what is wrong with the current lists?). It would be good if Stefan could find and talk to some colleagues who would know more about this.

Type of note.

Following Stefan’s comment about “endnote” Sanda has looked up “footnote” and “endnote” and found that they include the same type of information, and differ only by their position in the document. This position relates to printed materials (end of page, end of document) so for DDI-XML documents it does not seem to be relevant. Taina suggests that we might want to pick another term that would be more suggestive of the content of the note, like the other terms in the list. She will look into this further, together with Hilde. [Sanda’s note post-meeting: we can also consider deleting the term if the others seem to cover most possibilities].

Next meeting: December 10 at 13:00 GMT.

Note: Please email Sanda or group if unable to attend. Today’s attendees have already approved this date.

**DDI-CVG Meeting**

September 24, 2014

Minutes

Present: Anne, Bodil, Hilde, Lucy, Sanda.

Current status of externalized CVs.

Language proficiency: Sanda feels this is ready for final review and approval; however, not all members have had a chance to examine it. We won’t be meeting in October, so everyone should use this time to have a good look at this list, consult with others as needed, and have a clear point of view by the time we meet in November.

Remaining lists: regarding the IM CV, there is some agreement that it makes more sense to list clients rather than protocols. If we decide to publish this, Sanda might just copy the list available on Wikipedia.

The other remaining lists are still “work in progress”; we will have to re-evaluate them in November. Each of the assignees (entered in the spreadsheet) will have to be able to present us with an update on the progress made and the current status in November; otherwise we could be losing track of these.

It is also unclear how useful some of these lists are. Sanda will try to talk to available TIC members at Dagstuhl, if she gets a chance. Group should not wait for Sanda to come back, assigned work needs to proceed as normal. It is almost certain that TIC will continue to require that all of the externalized CVs be published by us.

Data Source list:

Personal records: Sanda’s suggestion is preferred, still needs editing (Sanda)

Economic/Financial records: reviewed and approved (final)

Geographic area: reviewed and approved (final)

Physical objects: Stefan’s suggestion is incorporated, the definition edited, approved (final)

Legal records: reviewed and approved (final)

The remainder still need attention, as previously indicated.

Next meeting: November 12, 2014, at 13:00 GMT.

Please let Sanda know on an individual basis if you think you will be unable to attend. A Doodle poll will only be set up if it turns out that the date above is inconvenient for too many.

DDI-CVG Meeting

August 20, 2014

Minutes

Present: Bodil, Hilde, Lucy, Sanda, Stefan, Taina

Membership: We welcome Bodil Stenvig, from the Danish Data Archive (DDA) as a new member of our working group.

Current status of “externalized CVs” (Wendy’s list):

Time zone – this list appears ready to publish

Aggregation method – ready to publish as well.

Language proficiency. There is extensive discussion around this list, mainly trying to reconcile the presence of terms indicating proficiency with the term “native” which is conceptually different, and is not really an indicator of proficiency. It is recognized that “native” needs to stay on the list, because it is often used in this context, and analysts of translations, or other types of use of multiple languages by the same person, almost always differentiate between native and non-native users. After discussing several options, the group agrees with a solution that places “native”, “non-native” and “notknown” at the first level of a hierarchy, with proficiency specifications as subcategories for non-native and notknown. [perhaps also for native?]

Definitions for the terms do not seem satisfactory at this point because they have to be phrased in such a way that they can be applied to all three aspects of using a language – read, write, and speak (comprehension is not included in the DDI schema). Taina and Sanda will work to improve the definitions this month.

Date type. Still needs work, Sanda will work with Stefan to get this right and in line with the DataType list. A question is raised about including the week as “nth week in the year”.

It is not clear that this is an appropriate use for “date” – perhaps more so for “frequency”? Sanda will examine the W3C schema for data types and if week is not listed, it will not be included.

Numeric type also needs similar work, to make it comparable to DataType. Assigned to Sanda and Stefan as well.

The remaining lists on our agenda (type of address, type of telephone, type of IM, type of video clip and type of audio-clip) have already been assigned. But they all need to be revisited by the respective assignees !

Data Source type:

The presence of the terms “population group”, “geographic area”, and “Physical objects” is still debatable. Taina will give additional consideration to “physical objects” both as a term and regarding the definition. Lucy will produce a definition for “geographic area”. We all like the definition for population group(s?) as contributed by Sanda and Hilde but there is still uncertainty on whether this term sits well in the list.

The definitions for research data: published/unpublished are reviewed and tentatively approved

Sanda gives some clarifications on her most recent edits: her recent additions will generally appear in blue font, and she has highlighted in yellow the areas that still need attention. The old comments on the communication terms can safely be ignored, or read for reference, or background (?). Sanda has entered new suggestions, but they are only drafts.

Task for entire group: it should not take more than ½ hr to 1 hr for individual members to review these definitions. It’s important that everybody does that before we meet next month, so that we are all familiar with the content. This will help speed up discussions.

Next meeting: September 24, 13:00 GMT. Lucy will host the meeting, as Taina will be on vacation.

**DDI-CVG Meeting,**

July 16, 2014

Present: Meinhard, Sanda, Taina.

Ways to make more progress.

We now have two “parallel” tasks – our on-going efforts to create lists for Source and Data Kind, as well as the need to expedite publishing the finalized vocabularies and the externalized lists from DDI-L. As a way to save ourselves time, since we only get to talk once a month, we will try to assign each of the externalized lists to a second person to review the first assignee’s draft, and then possibly to a third person, until we feel we have come to an acceptable result.

We will still briefly review the progress made on all lists, but as a group we will not spend too much time trying to improve/finalize them. This effort should take place between meetings, hopefully allowing us to spend most of our time together on the Source and Data Kind lists.

Externalized CVs:

Read, Write, Speak.

Three lists were brought up – Sanda’s findings from the American Council on Teaching of Foreign Languages, as well as two pointed out by Taina: <http://www.international.gc.ca/ifait-iaeci/test_levels-niveaux.aspx?lang=eng>  
  
<http://www.jobline.uni-muenchen.de/app_language/describing_language_skills/index.html>

We find these may be too detailed/specialized for our purpose. In DDI the intention is rather to give a general idea of a person’s mastery of language skills, not to assess them thoroughly. Therefore, we end up deciding to use the existing list, and publish the terms without definitions if these are not easy to come up with. The existing terms are widely used and self-explanatory.

Type of Address list is reassigned to Taina for review. There are also some comments in the spreadsheet made before the meeting by Sanda.

TypeOfTelephone: list is reassigned to Sanda for review.

TypeOfInstantMessaging: Hilde will look at this list for our next meeting.

AggregationMethod: Sanda will review this list against our own finalized list. She will come up with a final version in the next month.

TypeOfVideoClip and TypeOfAudioClip will be reconsidered by Meinhard. We think that the long lists he has found are kind of cumbersome and Sanda suggests that we keep the existing terms, and Meinhard try to find definitions for them. Not sure how this will work in the end though, so Group: please comment, forward ideas.

TypeOfNote: assigned to Lucy for review. As a heads-up, this is probably a good enough list. But please review the wording/phrasing of the definitions for clarity and accuracy, and consider whether any additional term(s) is/are needed.

SourceType list and definitions.

Physical objects and Population Group. There is still doubt as to whether these terms are really needed, or make sense in the list. Taina will continue to consider this issue, particularly in relation to “Physical objects” – she will try to come up with an improved definition that will suggest that these are sources for descriptive data, or she will present a strong argument in favor of deleting the term from the list. This issue will be discussed again during our next meeting.

Meinhard’s definition for “biological samples” is approved.

Research Data: published/unpublished: Sanda’s brief alternatives seem to be preferred. To be considered for final approval.

Processes. Definition okay conceptually but we need some good examples, preferably from the social/political/economic arenas. Hilde and Lucy will try to come up with some examples.

DataKind lists:

Taina suggests that the term “Model” as defined would probably fit better into the “provenance” list since it appears to describe content rather than physical format. That is generally agreed and the term is moved into the Data Kind by Provenance list. It is just greyed out in the “format” list as a reminder of this decision.

Translations. We neglected to talk about translations, but my understanding is that we are all agreed on the specs from last year’s minutes, so when I have time I will just proceed to publish as planned. If anybody has any problems with this, they should speak ASAP.

Next meeting:

Preferred date for our next meeting: August 20th, 13:00 GMT, but will also do a Doodle poll

As a reminder, participation in the actual meetings remains a priority.

DDI\_CVG Meeting

June 18, 2014

Minutes

Present: Lucy, Stefan, Sanda, Meinhard, Hilde.

“To do” list and future tasks.

Sanda has put together a “To do” list, as agreed in our April meeting. Everybody is welcome to suggest additions to the list, and, if Taina saves it on “Thesauri”, to make changes, on the condition that they announce any edits to the group.

Achim (and TIC) have requested that we speed up our work, especially as a number of CVs have been externalized from DDI-L, and we need to publish them so that they are available for use.

Also, DDI 3.2 has over 100 codeValueTypes, and we will need to review them and try to generate CVs for some, if not all. In addition to this, there are some revisions pending to the existing lists.

We will need to prioritize tasks, and organize ourselves so that we can satisfy some of these multiple requirements.

In the next 2-3 months Sanda will try to:

Publish revised versions of ModeOfCollection and SummaryStatistics.

Publish the finalized CVs for AggregationMethod, DataType, CategoryStatistics and Calendar.

These will have references to 3.2 and 2.5 but retrospectively fixing the others will be a different task (perhaps less urgent)

Note that Contributor:Role will need a reevaluation, as a result of the current efforts to improve coverage of discovery elements in DDI 4. Sanda has been specifically asked to try to come up with a more comprehensive list of roles, applicable to most phases of the data lifecycle; while this is pending, we will **put Publisher:role on hold as well ?** [this has not been discussed, but needs to be considered]. Hilde will provide links to some documents that might help us in developing the Contributor:Role list. The DataCite documentation will need to be consulted.

As a group, we will try to produce publishable CVs for the lists that were externalized from DDI-L. We decided to assign individual lists to group members who would work on them between meetings, and present acceptable lists and definitions for group review. These should be comparable to our original productions, but the standards need not be too rigorously applied. The idea is to produce lists that are up-to-date, useful and that make sense, but not spend too much time on refining them.

Sanda will work on the “Fluency” list.

Lucy on “Type of Address”

Stefan on “Type of Telephone”

Hilde on “Type of Instant Messaging”

Meinhard on “Type of audio clip” and “Type of video clip”.

Source type (and DataKind by Provenance) list:

Processes, workflows, paradata, process data (and metadata). There is agreement that “workflows” should be on the list, because we do want to include “paradata” in the DataKind list (paradata is a “hot topic” right now). We also agree that “processes” (understood as a series of related events that lead to a change) \*can\* be a source of \*primary\* data, so we include them on the list and make workflows a subcategory.

Stefan brings up an argument from a professional meeting/paper (?) that basically suggests that “paradata” may be also described as “process metadata”. This kind of approach can lead to confusion so we spend some time looking up definitions and trying to disambiguate terms. As a conclusion, for our purposes, and for our main area of coverage (social, behavioral, political, and economic science data) we will define “paradata” as “data about the data collection process”. “Metadata” is information about data, and we will not include any type of metadata in our lists, because our lists are in fact (metadata) about data… To include a term like “process data” in DataKind would be redundant, since all of the entries are “data” and we have processes in the Source list. Therefore we will just stick with “paradata”.

As a side note, Wikipedia carries a second entry for “paradata” that refers to data collected (by software) about software usage (online learning programs). According to our own views, this would not be “paradata” because it is not about data collection, but about usage (of some tool or application, by extension).

Other terms added to the SourceType list:

Biological samples – this is necessary if we include biomarker data in the DataType

Research data published/unpublished. This is a recommended entry in the current 3.2 schema. **Should we add “primary” to make this clearer??**

Population group. This is a recommended entry in the current 3.2 schema. We think it makes sense, for certain types of data, like demographic data, or data on specific groups, like retired people, or high school students, etc.

“Geographic area” and “Physical objects” also included because they seem valid sources of descriptive data (in the DDI 3.2 schema the term “environmental object” is entered as an example, but we find “physical object” to be better/clearer).

For next month, the group will review all of the definitions that have already been created for the terms on our list.

The following members will provide draft definitions for the new terms added:

[Primary] research data: Published/Unpublished – Stefan

Population Group – Hilde

Geographic area – Lucy

Physical objects – Taina

Biological samples – Meinhard

Processes and workflows – Sanda.

Next meeting: Wednesday, July 16, 2014 at 13:00 GMT

**DDI-CVG meeting, April 9, 2014**

Minutes.

Present: Hilde, Lucy, Sanda, Stefan, Taina.

Mode of Collection published. Sanda has published Mode of Collection according to the new procedure that entails creating the master list directly in Genericode rather than Excel. An Excel spreadsheet has also been published, but only for maintaining consistency in presentation.

We had previously discussed making a note of changes in the spreadsheet format (but not in the CV content) like adding a new row or column. This may no longer be necessary since Excel is no longer our input format. The group agrees.

Usage tabs – previous CVs document usage in DDI 3.1 and 2.1. For Mode of Collection, Sanda has updated the DDI-L usage to DDI 3.2 but has chosen to leave in the DDI 2.1 usage rather than change it to DDI 2.5. The edits would have been more significant and it is not clear that most DDI-C users have switched to DDI 2.5. Also, the elements involved are the same. Taina notes that 2.5 may include changes that allow CV use in certain fields that previously did not support CVs. Sanda will look into this and will try to document the changes for the group to allow an informed decision.

Related to this, we also need to consider the need to change the references to the two DDI specifications in the existing CVs. This was not discussed, but is an issue. If we did that, we would have new minor versions for all vocabularies. Is it really worth it? Especially if there is no change in the elements that support the CVs in question?

Taina reminds us that there are four finalized lists waiting to get published: Aggregation Method, Data Type, Category Statistic Type, and Summary Statistic (Type?). It is not clear if we had been waiting for 3.2 to get published, or they were not supported in DDI 2.1? Sanda will have to look into this as well, and proceed to publish as existing questions are being clarified.

Mode Of Collection content:

Stefan raises the issue of diaries, food diaries in particular, that might be created in other ways than writing text (selecting from a list, for example, including an online list in which the selection would be made by clicking on existing entries). The current definitions for Self-Administered Writings and Diaries seem to suggest that these are created exclusively in text form, which would make it difficult to describe some types of diaries using these entries. Most participants agree that this seems to be an issue. The definition could be edited; however users of the CVs need to recognize that the act of selecting from an already prepared list can also be described using “Self-Administered Questionnaire” entries, or if photographs of food are submitted, this method of data collection could be described as “recording”. When we interact with users, we need to recommend a comprehensive approach to the list, where all possibilities are examined as opposed to a narrow selection of terms. In the end, we decide to edit the definition for “Self-Administered Writings and Diaries” to read: “Narratives, stories, diaries and written texts created by the research subject.” This will be applied creatively (i.e., with the necessary adjustments, to the sub-categories as well).

Taina suggests that we could offer clarifications regarding the CVs usage in a FAQ section. We all agree that’s a great idea, but Sanda points out that we currently do not have the resources to create, or maintain, such a “users’ tool” on the Web. As a temporary solution, we agree to keep “private” notes about such issues, so that they can be posted on the web when we find the resources to organize and publish them. Stefan suggests that some people at SND might be able to help in the not too distant future.

It is also suggested, and generally approved, that we create a “log” for tasks that are pending – like editing an already published list, or publishing new ones – so that they don’t get overlooked. Perhaps we should try to assign this as a task to one of the group members?

Additions to the Data Kind by Provenance list:

Paradata – generally approved

Biomarker data – mostly approved

Microdata – not sure yet, but noted

Multilevel (i.e. hierarchical) data – not sure yet, but noted.

We need everybody’s opinion on this.

Possible additions to the Source Type list:

Processes? Hilde mentions “workflow” data (the data kind would be “paradata”).

But having an entry like “workflows” does not seem to work well with the other entries. Perhaps processes would work better, because they would include workflows, but could cover other types of processes as well.

Biological samples. This entry seems needed, if we have “biomarker data” in Data Kind.

“Geographic areas” does not seem to appeal to the group.

But “Physical objects” might be a possible entry, although very general…

To consider: new entries like processes and/or physical objects will considerably widen the scope of the vocabulary, when we just had decided to eliminate terms that seemed to general. This particular list needs more thinking and all group members are invited to consider it more carefully and try to talk to other qualified people from their respective archives.

As an additional comment from Sanda – one easy way to keep it clear and clean would still be to limit this to “secondary” research. Perhaps add “Publications” to the first main entry and have “primary research data” as the second entry and close the list. If we make a clear decision in that sense, we should specify that this is about “secondary” collection in the list description. But this would be the easy way out – not necessarily the best solution

Next meeting: June 18. Lucy has offered to set up the meeting through “GoTo Meeting”. As a backup, we can still use FSD’s videoconference system, but will have to ask Jani to help facilitate that.

DDI-CVG meeting

February 12, 2014

Present: Hilde, Lucy, Meinhard, Sanda, Stefan, Taina.

Publishing the CVs. Jani is no longer able to help with publishing. Sanda has talked to a few people involved in Web design and programming at ICPSR. They recommended that the input format be changed to Genericode (from Excel). This would be best practice, as the CVs are actually expressed in Genericode. ICPSR people can then write a simple stylesheet converting Genericode to html. That makes sense; however, we have .xls files on our web page right now and we would not like to get rid of them (???). For now, the solution would be for Sanda to manually create a Genericode file in addition to the .xls so that we can continue to post all three formats. Sanda will try this, and will report on the results. If she is not successful, or encounters problems, we will re-address the issue.

Source Type list. We have a new proposal from Taina and Stefan that basically restricts the list to sources for “secondary” data collection. That seems a viable option that we proceed to discuss.

It is agreed that even if we decide to focus on sources for “secondary” data, we will not make an official statement to this effect. It is very hard to draw a clear line between what is already collected data, and a source of “raw” data, and in fact it is not easy to define “data” because this concept has been expanded in the last few years, and continues to be so. Basically anything that carries information can become “data’.

Data Kind list. It will be necessary to look at Source Type in conjunction with Data Type (Data Kind) because in our current approach we have simplified the DataKind list to include only formats, and have “moved” some well-established terms to Source Type. The main problem here is that some of the terms are commonly used by researchers both in literature and in searches. We agree that we have to continue to make them available (examples are events/transactions, or survey data, voting results, or census/enumeration data (returns for the UK?)). Taina suggests that one possibility for maintaining these would be for each organization to internally create their own list and make it available on a search interface, even if it’s a combination of terms from multiple CVs. That’s certainly feasible, but we feel that it would be better if we attempted to at least suggest such a list. In DDI 3.2 KindOfData is repeatable, so two lists may be used, but there is no attribute available to distinguish between them. Lucy indicates that the UKDA actually uses two lists for data type, one for identifying formats and an “unofficial” one for content-based searches. She will send us this list. Based on this discussion Sanda, Hilde, and Lucy will try to come up with suggestions for two lists for DataKind before our next meeting.

Source Type list. We examine the new draft prepared by Taina and Stefan. Processes have been taken out and we are all okay with that. Interactions have been taken out as well. Sanda feels however that events and/or interactions can constitute a source for data, if they are recorded and stored and distributed as such, without any changes. There seems to be some agreement on this so we might just keep events/interactions as a source (it seems better than events/transactions?) without any subcategories. Communications are also accepted as valid “sources”. Sanda proposes to “re-arrange” the list by making Registers/records a main entry with a lot of subcategories by type. There does not appear to be opposition to this, so a new draft will be available for discussion before our next meeting.

All group members: it’s important to give some thought to the issues raised - any ideas, comments, references, suggestions, even unpolished, are welcome.

Next meeting: March 12, 13:00 GMT.

**DDI\_CVG meeting**

January 15, 2014

Minutes

Present: Hilde, Sanda, Stefan, Taina

The discussion today centers on the actual Source Type list.

The starting point is Taina’s message (captured in the Source Type working document) in which she describes some examples of studies held at FSD and tries to classify them according to the current draft of our list.

We come to consider to what extent processes – which develop in time and are defined as sequences of events, activities, etc. – could actually constitute a data source. There is agreement that this is unlikely – rather, the events themselves would be somehow captured in the stored/distributed data and the process would be evaluated in the data analysis. Looking at the tentative examples we found also confirms that these would hardly constitute a data source. This leads to an agreement to delete the processes altogether from the list.

Taina also suggests that it would be easier if we limited the Source list to those type of sources that are involved in “secondary” data collection activities (i.e. the data already have been “collected” once, and the “secondary” collection involves a restructuring of the existing sources, or a change in the medium, or the presentation). Assuming we would agree to this, what other terms would we delete from our list?

Sanda would be reluctant to delete terms like court proceedings, or events/transactions because these have been incorporated in metadata records for a very long time, and people actually search for them.

In this new approach, we have moved these terms to “Source” from their current positions in “Kind of Data” (or data type) but we would not want to get rid of them altogether.

This train of thought leads us to briefly re-examine our KindOfData list. Currently we have a draft that limits the list to describing the data format. Some archives use this type of content (UKDA) while others currently use a mix of terms that also indicate the “content” of the data (ICPSR, NSD) (see KindOfDataType, v27 document). The current thinking at SND is that a formats list is preferred. Taina agrees, and so does Sanda.

Hilde is more hesitant, and points out that right now there is a tendency to reconsider this “element” (or concept?) and its importance in describing data. She suggests examining the efforts behind modelling DDI4, and/or submitting our lists to TIC. Our lists may be submitted to TIC, however Sanda points out that DDI 4 might not come out for several years and we should probably not get tied down by an event that seems too distant and yet uncertain.

To aid us, or at least keep us informed on current trends regarding Kind of data, Hilde agrees to send to the group some of the readings she may have available on this topic.

Lastly, and this is Sanda’s post-meeting request – when we have Hilde’s comments/findings, can we also have a (brief) discussion regarding the term “survey”? we have not included this word at all in any of the lists, but people might search for “survey data” in the same way that they search for “court proceedings” or other things that we took out from the KindOfData list.

At any rate, just by examining the 4 lists side-by-side in the table (included in the KindOfDataType document) we seem to agree that if we include biological samples, it should be listed under KindOfData, and not as a Source. So Sanda greys it out in the SourceType list.

We also talk about how “communication’ can be a source of data. This seems to make sense for public speeches or mass-media products, but there is hesitation in accepting that survey data may have ‘personal communication’ as a source (especially in Taina’s view, that favors including only sources for secondary collection in this list).

Stefan also mentions that he would restructure the “records” piece under Sources, which makes sense.

Perhaps Records with sub-categories: administrative, historical, clinical, archival (?) but other types would make sense as well, especially economic/financial (???) Stefan agrees to present a suggestion on this.

Before our next meeting, Taina and Stefan will also try to work together on an improved draft of the SourceType list. Hilde will also give this some thought, in conjunction with everything else that we touched upon. It is recommended that they come up with suggestions and argumentations sooner rather than later, so that the discussion is still fresh in our minds and we have some time to think these things through and react.

Next meeting: Wednesday, February 12, 2014 at 13:00 GMT

**DDI-CVG Meeting**

November 27, 2013

Present: Hilde, Meinhard, Sanda, Stefan, Taina

Mode of Collection spreadsheet.

Sanda has produced a first draft. She has entered codes, terms, and definitions. These will need a careful review by the group, especially as we plan to publish without additional technical/public reviews. Even a missed period matters!

In the definitions, she has made minor stylistic edits, so these will not always be identical to our last approved document (in case people double check for consistency). But they will have to make sense!

In the terms section, she has changed the word order in focus group subcategories and physical measurements subcategories, to reflect our previous agreement that these can be in natural language order, and make them consistent with other sub-terms.

Regarding the “Usage” sections (DDI 3 and DDI 2) it was agreed that we should wait until 3.2 and 2.5.1 are publish and make reference to those versions, especially as these sheets include links to the field level documentation. Regarding the element number in the DDI 2 section, this will no longer make sense as in the newer versions the elements are no longer numbered. (or we could keep it and just change the title of that section to “Element number in DDI 2.1”).

These adjustments to match the newer DDI versions will be changes in the spreadsheet, but not in the actual CVs. How do we account for these?

It is agreed that such changes in the spreadsheet content (but not the CV) should not affect the CV version. It is suggested, and approved, that we should instead insert a new field in the “General” section with a heading saying “Last Updated” and a date in the content. We prefer to use the “General” worksheet and not the “Identity” sheet because the latter really refers to the version of the actual CV.

The changes in the spreadsheet will trigger some necessary edits in the script that converts it to various formats. Jani is really busy but perhaps he will find some time to take care of this in January or February (when the spreadsheet is final). Taina will kindly talk to him about this.

Mode of Collection terms:

-Clinical data. Lucy sends a comment regarding the perceived need (at UKDA) to include a category that would describe the collection of “clinical” data. (Physical measurements would include clinical measurements, but would be a wider category). Stefan confirms that at SND they felt a need to include the term clinical in the vocabularies for their metadata, knowing that many people would search on this term. They ended up including it in data type (clinical data) but are not 100 percent satisfied with this solution.

A long discussion ensues around this perceived problem. To make “clinical” a subcategory of Physical measurements would not be quite right since 1. Clinical measurements can also be psychological, and 2. Collecting a sample (blood, saliva, etc. ) is not technically a measurement (if no data are extracted, but the sample is just stored) but rather a “procedure”.

If we create another term like “Clinical Procedures” at the same level as Measurements (uppermost level) it will by default include physical measurements so there will be overlap between terms, creating confusion for both metadata creators and users (in searches).

It is kind of agreed that we will need to include “biological samples” in data source. Would the combination source= biological sample, collection=physical measurement suffice? Do we need the word “clinical” somewhere to make sure that people searching by keyword will find it?

There’s also the question of what constitutes the data we are distributing?

(data Kind describes the distributed data) are we distributing the blood samples, or some numeric data extracted from them (then they are the source). “Clinical data” could be inserted in a different classification of data type, similar to what is used currently at some archives, including ICPSR and NSD (survey data, aggregate data, etc.) – but we decided to depart from that, and only focus on the medium at the distribution stage (we attempt to capture more of the data “content” in Data Source). At this point we decide we do not have a good solution and it is up to Lucy and/or Stefan to come up with a better one, i.e., that will take care of (or minimize) the concerns summarized above, or we will leave the list unchanged.

-Experiment as a mode of collection. Meinhard sends two comments from Gesis colleagues recommending the removal of ‘experiment’ from the list.

Argument 1 ([pascal.siegers@gesis.org](mailto:pascal.siegers@gesis.org)): “This category is problematic as a mode of data collection because it is in fact a research design used to establish causality between a treatment variable and an outcome. Data collection within experimental settings is done by questionnaires (in psychology), observation or physical measurements. This category should be deleted from the controlled vocabulary for Modes of Collection”.

 Argument 2 ([markus.quandt@gesis.org](mailto:markus.quandt@gesis.org)): “… The distinction between experiments and other methods is indeed about how a stimulus is set - in a controlled and logically exhaustive fashion or not. Therefore, questionnaire-based surveys have been called quasi-experiments. Data collection on the responses can then be performed in different ways: questionnaires, observation or direct recording of response behaviours, etc. STILL, many researchers would see e.g. laboratory experiments in game theory as a self-contained data collection approach. How do we talk to these people if we drop the term 'experiment'?”

There is discussion around this too, but mostly Sanda’s opposition to removing „experiment“ from the list. Even if it is a research design (true, but there is no field to document research design in DDI) it is also a way of collecting data. It cannot be accurately replaced by using „interview“ and „observation“, for example, because they are different modes. In observation, the independent variables are not changed, whereas in experiment they are. And interview is also different (there are contextual factors for instance, that vary within the same sample, etc.).

Taina also makes the point that „mode of collection“ is both repeatable (multiple terms may be used) and has a free text field in which any additional comments or information may be inserted.

In the end, we decide to leave things as they are.

Another issue raised by Meinhard that was not discussed in the meeting:

THIRD point: The same colleague (argument 1) adverts to a (probably minor) contradiction in our definitions: On one side we define an interview in general as a “pre-planned communication between two (or more) people - the interviewer(s) and the interviewee(s) - where information is obtained by the interviewer(s) from the interviewee(s).” On the other side we have sub-terms referring to different kinds of “SELF-interviewing”…  (mm: Would “questioning” be a better high level term? I found the latter on the once already mentioned SRM thesaurus of social research methodology.)

Sanda’s first reaction to this: in self-administered interviewing the information is still obtained by the interviewer so the more general definition of interviewing is correct. The difference is that the interviewer is not “physically present”. Regarding the term itself, we would need to keep it for the sake of keyword searches. As a reminder, we decided to remove “survey” as a mode of collection and that has been criticized on the same grounds (that people would use “survey” as a keyword in their searches.

In conclusion, changes are still possible to our list – even after publication. But we need very solid arguments, and a workable suggestion for a change to be implemented.

Next meeting: January 15th, at 13:00 GMT.

**DDI-CVG Meeting**

October 30, 2013

Present: Hilde, Meinhard, Taina, Sanda, Stefan

ModeOfCollection codes:

In general, they seem ok. Meinhard suggests that SAQ is not needed for Paper Self- Administered Questionnaire. It is agreed and SAQ is removed.

Sanda asks the group to review Computer based observation because it does not appear consistent with all of the Web-based subterms. We do review it, and decide that having computer instead of web-based is appropriate, because the observing software may also be installed on the computer hard drive or local server (not accessed through network). We add “local, or online” to computer programs in the definition, and remove the last sentence which mentions e-observation as an alternative term. This seems to make things clearer.

ModeOfCollection terms:

It is noted that in many instances the terms do not accurately reflect the codes’ hierarchy (eg., Participant Field Observation instead of Observation:Field:Participant) Some of us seem to prefer the current approach as it seems to be closer to natural language. While the codes need to follow a fixed pattern (structure) there are no similar constraints regarding the terms – or, we do not seem to be aware of any such constraints. Plus, an organization may choose to work with the codes if using English or choose a different approach when translating the terms, if using another language. So, technically this does not seem to be a big concern. Still, if anybody can think of a very strong argument against the current usage, they can bring it up for discussion in the next few weeks.

ModeOfCollection list:

Taina suggests removing at least part of the hierarchy under Observation, which seems to go into too much detail. We agree to take out the two Overt/Covert pairs, since they seem most unlikely to appear in the original PI documentation. Some archives like FSD and SND would not use them anyway.

ModeOfCollection definitions:  
Stefan has reviewed the definitions and suggests some improvements: adding “geospatial data” to cover for GIS data in the examples for Physical Measurements and Tests; and adding “using RSS to collect blog posts or tweets” in the examples for Compilation/Synthesis. Both suggestions are approved and included.

Hilde suggests adding “aptitude” in the definition for Educational Measurements and Test. This is also approved and added.

Publishing ModeOfCollection:

Sanda will attempt to fill in a spreadsheet in the next few weeks. As soon as she is done she will send it to the group for review. According to the latest document issued by the DDI Alliance regarding the review and approval process, we might be able to publish without further reviews (the Controlled Vocabularies are not subject to the same regulations as the DDI specification).

SourceType list:

This comes next on our to-do list. We will start by re-examining the definitions that are not finalized.

Assignments:

For next time, Stefan will look at Social and Economic Processes, Taina will examine the Political and Biological Processes, and Meinhard will review the Behavioral and Natural Processes. Lucy will review the entire Interactions piece, and Hilde will look at Communications.

Please note that Sanda has added some new found definitions in Processes. If these are acceptable, there is no need to look further, assignees please move on to examining Interactions and/or Communications, which are the least developed.

Obviously, everybody will have to offer feedback. The table format of the existing document may be used to enter comments or proposals, unless they run too long. It would be best if people sent out the results of their research at least a few days before the meeting, so that we have some time to consider them, perhaps react to them.

Next meeting: November 27, at 13:00 GMT

**DDI\_CVG meeting**

September 25, 2013

Minutes.

Present: Meinhard, Sanda, Taina.

Topics discussed:

Membership: Joan has just announced her withdrawal from the group. We are really sorry to see her leave. Michaela has been out of the picture as well. Hopefully she will be able to participate again in the near future. However, this brings up new concerns about membership. We really need a stronger group, with more members, and most importantly with wider representation.

Since we just heard that a group of people at SDA are also interested in developing controlled vocabularies for their metadata, Sanda will contact the person who wrote to her, inviting him, or another representative from their archive to our group.

At the same time, Taina will prepare and send a message to the CESSDA expert list with an appeal to people from the various participating archives to join our group.

Attribute type on KindOfData (DDI 3.2) and dataKind (DDI 2.5.1). In DDI 3.2 the attribute “type” has been restored, with a fixed choice of “quantitative, qualitative, mixed”. “Mixed” will not be too useful if entering CV terms (the field is repeatable, one iteration per term) but it might be used by those choosing to enter free-text information, listing multiple kinds of data within a single instance of the element. In DDI 2.5.1 there will also be a “type” attribute for this element, text type, but able to support a CV (CV usage in DDI-C is documented differently than in DDI-L).

Website: Linked bullets were added on top of our main page for easier navigation. Sanda also added a paragraph about current CV users. Edits may still be suggested and implemented. We will attempt to post translations when Taina produces spreadsheets with the Finnish translations (probably next month or so).

Wendy has sent a list of all of the CodeValueType elements in DDI 3.2, along with a list of all of the CVs that were externalized. We will have to publish all of these externalized CVs. Most of them do not have definitions, so that might slow us down somewhat. We will look more closely at these when we have finalized at least Mode of Collection and Kind Of Data. Finalizing these is our top priority right now.

Mode Of Collection definitions:

Simulation: Lucy’s and Hilde’s latest suggestion is approved with minor edits. It can be viewed in the new version of Mode of Collection, dated September 25. Hilde and Lucy, please have another look.

Recording: Taina emphasizes that we need to make a clearer distinction between “measurement” and “recording”: any measurements made would be presented as recordings so when would we use one term vs. the other? We agree that recording should be restricted to phenomena or processes or events as they occur naturally – a video of an earthquake for example, as opposed to the shocks registered by a seismograph. So we agree to suggest that in the definition, by taking out all references to recording measures. Approved, but Hilde and Lucy please review.

Measurements, and projective methods: Sanda’s basic proposal is accepted, with some changes:

We decide to call the main term “measurements and tests” .

We decide to have three sub-terms – educational, physical and psychological. Clinical can include both physical and psychological so there would be overlap, so we decide it should not be included.

We also decide to include “projective methods” in the examples for psychological, rather than make it a sub-category.

We also discuss possible definitions, and come up with some drafts that are also posted in the most recent version of Mode of Collection. This whole new section is highlighted in yellow. Hilde and Lucy, please review. We would like to make these final at our next meeting.

Task for all – please try to have a final look at the entire list. At the definitions too, but the list itself is more important because changes in terms imply a new major version. We would like to expedite publishing this list, as there are people already using it…

Possible dates for our next videoconference: October 23 or 30. Sanda will create a Doodle poll. (Might end up choosing October 30 because 23rd is uncertain for Sanda.)

**DDI-CVG meeting**

August 14, 2013

Minutes.

Present: Hilde, Lucy, Meinhard, Sanda, Taina

DDI-CV web page - CV usage announcement and hosting translations.

We decided a while ago that it would be a good idea to post an announcement about the organizations using our CVs. Where on the web page should we place this?

Lucy suggests first paragraph (subsection) for maximum visibility. It is agreed. So far we have identified FSD, GESIS, ICPSR, UKDA, The University at Bielefeld (Germany), and Mathematica (U.S.) as users. Doubtless there are others. Group members, please notify group if other users are identified. Or ask around if you go to European meetings, etc.

Sanda will try to insert a sentence about this in the first subsection on the main page.

Our group currently lacks the capability/resources to “officially” manage and maintain translations. But we could post existing translations on our web page just to make them available to others for consultation. The ensuing discussion concerns mainly their presentation on our web page(s).

We agree to present each translation separately, with its own spreadsheet.

The translations will be made available only in .xls format.

We (Sanda) will add a sentence about translations being posted in the “Publication, etc…” paragraph, and create a new tab for translations (download?) beneath the current tab for CVs download. From both the tab and the paragraph there will be links to a new page presenting a table in the format suggested by Taina:

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Title** | **Languages available** | **File Type** |
| AnalysisUnit | Analysis Unit | DE, FI | xls |
| TimeMethod | Time Method | DE, FI, NO | xls |
|  |  |  |  |

The language codes would be ISO 2-letter codes and each code would provide a link to that language version of the CV.

Each language version would include a note field in the language worksheet allowing authors to specify whether the translation is a draft or a final version, the name and contact information of the person to whom any comments or queries can be sent, as well as the date of the last update.

Taina has to produce a spreadsheet for the Finnish versions of the CVs translated, and used, at FSD.

Meinhard should try to find out the appropriate information for the note as mentioned above. Sanda can enter that info in the German translations spreadsheets.

Structure of main page : is there a way to increase the visibility/accessibility of each sub-section when this page is already too long? The simplest solution would be to create a list of all of the subtitles on top of the page with links for direct navigation to the respective paragraph, without having to scroll down. We will try this solution first.

Mode of collection definitions.

Synthesis, harmonization, and compilation. We have been going back and forth with these terms for quite a while now. Harmonization does not sound like a good choice after all because it is usually associated with data processing. Lucy has produced two examples from the UKDA records that use “compilation or synthesis of existing material”. This sounds like a good option when it is hard to establish whether pre-existing sources were compiled or synthesized, or both. So we replace compilation with Compilation/Synthesis and take out the sentence that made reference to the identity of the original sources. We are all quite happy with this.

Joan, Michaela, please review and approve.

Simulation. There is some discussion around this. From the links sent mainly by Hilde, it appears that simulation is mostly achieved by computer programs. Also according to literature, it is a type of modeling. (modeling may be more static? Simulation may use models to imitate a process?).

Hilde and Lucy will do further research on simulation and will try to suggest definitions and examples, preferably before the September meeting, so that we have a chance to ponder over them.

Projective methods/techniques. These need to be included but we still have to figure out a good place for them in our list. Sanda suggests that we may want to include “test administration” as a term and mention these techniques within that entry, because they are, in essence, personality (psychological ) tests. This also needs further investigation. Sanda will continue to look into the matter and will try to come up with some suggestions.

Next meeting: Wednesday, September 25, at 13:00 GMT.

**DDI-CVG Meeting**

July 10, 2013

Present: Joan, Meinhard, Sanda, Taina

“To do” list:

In addition to the CV lists and definitions, we have a number of other tasks to accomplish. Sanda has compiled a list and sent it to the group so that we all remain aware of these. There is not much time to talk about any of them in detail, but one that seems more important is the issue of translations.

GESIS is preparing to use at least two lists (TimeMethod and AnalysisUnit) and Meinhard has sent us a German translation of the terms added to the English-US spreadsheet. FSD has also been using TimeMethod and there is a Finnish translation of the terms and definitions.

When translating, Taina has also added some more explanation to the definitions to aid in correct usage. We would like to provide these on the Web, preferably not as officially published versions, but just for people to be able to look at them. This would be especially important for the German ones, as other German-speaking countries/agencies might want to know about them and consult them.

Before we provide them we need to figure out the correct format. Probably a separate sheet for each language in the main spreadsheet. Taina will look to see if she still has a spreadsheet with an additional sheet for German. Should we contact TIC to ask about input format? This is unclear at this point.

In the meantime, we need to think about how we present the projects using our CVs, and how we present the existing translations and link to them. The translations might be more unstable in the first months of usage, but if we do not publish them officially, we can replace the actual files without much trouble.

It would be best if we talked about this issue again at the next meeting. In the meantime, group - please come up with ideas on web presentation, as described above.

A proposal by Jon Johnson (Centre for Longitudinal Studies, UK), to add the term SIR to the software package list has been approved. Next step is for Sanda to update the spreadsheet and produce a new version.

Mode of Collection definitions:

Compilation: agreed

Summary – seems clear and example is satisfactory: agreed

Aggregation – Taina suggests a different example, that seems clearer/easier to understand. It is a good example, so the draft is edited to include it, and the definition is agreed upon.

Group: people who did not participate in today's meeting, please read these three definitions - your last chance to disagree.

Synthesis – there was quite a bit of disagreement on this definition in the discussions

that preceded the meeting. The concept itself seems fuzzy because it is used in different ways. It is also frequently associated with data analysis, rather than data collection. The examples we found are not very convincing, even to us. It’s very doubtful that the authors of a harmonized dataset would actually describe this process as a synthesis. So we agree to drop synthesis and add harmonization to the list. Joan and Sanda are charged with producing a definition for harmonization.

Next meeting: we will set up a Doodle for those who did not participate today, but our favorite date is August 14.

**DDI-CVG meeting, June 6, 2013**

Minutes

Present: Hilde, Lucy, Sanda, Taina.

Mode of Collection Definitions:

Self-administered questionnaire: Interactive:VCASI. This one is quickly reviewed, we had agreed on it in the previous meeting. It remains final, including the part that specifies that it is the earliest method of CA self-interviewing, so that it does not get confused with more modern types of “videos” that may include other types of presentations on the screen. This one only shows the questions in written form. Finalized.

Lucy recommends that we review the Computer-assisted entries for consistency in the subtitle.

We decide to keep them lower case and review them. If any inconsistencies are still found in future working documents, people should point them out.

Experiment: Web-based. We agree to go for Meinhard’s second (more recent) suggestion. We also agree that it is okay to have this entry, because in a case like this the “tools” or “means” used to collect data do make a difference, in that the participant is placed in a different environment, which can affect the way the data are collected. To better distinguish between web-based and laboratory experiments we specify the controlled setting as being “physical” in the definition for lab experiments.

Recording – discussion is postponed for next time. Taina will try to find out why we seemed to have a problem with the term ‘electronic’ means in the definition.

Compilation – also postponed for next time.

Synthesis. We have two different takes on this definition and neither seems fully satisfactory.

Sanda’s suggestion focuses on the fact that there are multiple sources and their identity can no longer be traced back to the original sources. Her example is a harmonized dataset (but we would prefer “variable” if we keep this example) and relates to quantitative data. Taina’s suggestion applies the term to qualitative data, and introduces the concept of summarizing. Lucy questions whether summarizing and synthetizing can be used together, or perhaps should have separate entries.

A discussion follows that comes to the following conclusions:

A summary can be produced out of one or multiple sources

A synthesis is produced out of multiple sources

We should have separate entries for summary and synthesis (i.e., add summary to the list)

A summary is produced from qualitative data

A synthesis may be produced from either quantitative or qualitative data (if we agree with Sanda’s example of a harmonized variable). We will \*not\* specify this in the definition but will try to give at least one example from each type of data. Hilde will try to find a good example for synthesis of qualitative data, either from real life or imagined (but has to be convincing)

How is harmonization of quantitative data different than aggregation? This question was raised, what follows is my own lame explanation:

In harmonization, a number of “source” variables that contain the same information in different formats are recoded, or otherwise edited, to produce a single variable that will contain the information from all of the source variables in a single format. *A variety of measurements are mapped to a unified measurement scheme*. For instance, to indicate “age” some studies may use a date of birth variable instead of respondent’s actual number of years. The “date of birth” variable is recoded to indicate number of years, and the resulting data are combined with the second source variable that already provides number of years.

Aggregation is a process by which micro data (collected at the individual level) are presented at higher levels (population group levels) after being subjected to mathematical calculations (sums, totals, percentages). In harmonization the data are changed so that they can be compared; in aggregation they are not changed, but only presented at a higher (group) level, usually for confidentiality reasons, but also (in some cases) because the analyst is interested in totals (ex: number of workers looking for employment in a certain month/year compared to another month/year).

Meinhard, Joan, Michaela, please give feedback on this discussion around synthesis and the proposed entry for summary.

Sanda, as well as Lucy and Taina will continue to look into the matter, and hopefully come up with some workable suggestions.

Topics we were not able to discuss:

Using DOIs/URNs/URLs for referencing (update from Sanda)

Jon Johnson’s proposal to add the term SIR to the software package list

Looking at the sampling and weighting draft lists from the Survey Design and Implementation Working Group

Mentioning on our website the projects that use any of our CVs. (Should we only mention them when they are in place, or is it okay to mention them even in the preliminary/beta phases?)

Mary’s proposal that we include “biological samples” in our Data Source list.

XML catalog on the website. (Sanda will try to deal with this)

Next meeting – there will be a Doodle poll for scheduling it, since we don’t know about the availability of Meinhard, Lucy, and Joan.

**DDI-CVG meeting, April 17, 2013**

Minutes

Present: Hilde, Meinhard, Sanda, Taina.

Recommendations for citing the CVs outside DDI.

It has come to our attention that there are organizations already planning to incorporate our CVs (including perhaps some of the unpublished ones) in their metadata records. When these records are not created in DDI-L or DDI-C V2.5 the CVs need to be referenced with a regular citation. Following some brief email exchanges, we have come up with an agreed model for citations, both for published, and for unpublished CVs, and Sanda has posted the information on the Controlled Vocabularies Usage page.

The content is reviewed and approved by the group.

XML catalog – Taina has agreed to ask Jani if he would write such a catalog. If he does, we will put it up on the Web, downloadable from the Usage page. (He should also think about the accompanying explanation – maybe we can make it clearer for those of us who are not so technical).

Mode of Collection definitions:

Definition for Self-administered, interactive; we include a reference to the fact that the question sequence is managed by computer software. Definition is approved.

Michaela suggests that the terms VCASI, ACASI, TACASI and VACASI should be written as initials, no dashes. Approved, and we make the changes where appropriate in the list and definitions.

Taina finds usage of the term VACASI (i.e., video-audio) and suggests that we add that to the VCASI definition. This is discussed, and while we agree with adding a note about VACASI, we end up agreeing that VACASI is rather an alternative name for ACASI (historically, the video (i.e. questions displayed on screen) appeared before the audio (questions being read out loud by a voice in addition to being displayed)). So the note is added to the definition for ACASI

E-observation or computer-based observation. There is a long discussion about this. Using e-observation might not be clear enough, and using electronic observation might suggest too wide an application (computers are just one type of electronic device, but there are many others, that might not be used in observation). We end up agreeing with computer-based.

Web-based experiment: there is a lot of hesitation about this. This type of experiment has the same features as a lab experiment, except that the “lab” (i.e., setting) is virtual.

We agree however that we cannot include it as an example, or an alternative, in the definition for lab experiment. The latter type is a classic, it is widely used and we would be messing up the concept. Sanda also mentions that while the virtual setting is created and controlled by the researchers, there is also a “physical” environment, and in fact the researcher is unable to control the participation feature. The respondent might not be alone, and his/her answers or reactions might be influenced by other people or other factors in the real, physical environment that are unknown to the researcher. Meinhard also emphasizes, as a special feature to this type of experiment, that the respondents are spread out geographically (in the physical space) while in a lab experiment they are all in the same room. No decision is made about this definition. Meinhard, Taina and Hilde are going to further examine the issue until our next meeting but we need everybody’s input on this (this is a task!).

Transcription – we add “different source” because it looks like things can be transcribed even if they are originally in written language. We also add, at Taina’s suggestion, a note seeking to clarify that for transcribed interviews or observations, transcription is not a mode of collection. These should be classified as interviews and/or observations (in these cases, transcription is data processing). We also decide that examples are not needed for this definition, and might in fact create more confusion because transcription is oftentimes used in combination with other modes of collection (like compilation, or others)

Physical measurement. We add measurements specific for the human body. Examples are part of the definition.

Tasks for next time: *Everyone* in the group will have to send in a comment of sorts about *each* of the remaining definitions \*unless\* the current document (updated 4/19) already includes their comment. The comment may be short. But we need to know your opinion.

Also for next time – we need to discuss at some point about the possibility of including biological samples in sources.

Next meeting: Thursday, June 6, at 13:00 GMT

**DDI-CVG Meeting**

**March 20, 2013**

**Present: Hilde, Lucy, Meinhard, Sanda, Taina.**

CASI subcategories and how to deal with the proliferation of new computer based devices/applications.

We first look at the possibility to include in the list some CASI subcategories, as described in a paper found by Sanda (Thomas Harmon, “Impact of T-ACASI on Survey Measurements of Subjective Phenomena”, Public Opinion Quarterly, vol. 73, issue 2, pp. 255-280). Sanda was looking into this as a result of ICPSR having introduced T-ACASI as a term in their own Mode of Collection list. Since T-ACASI seems to be used and ACASI is already in our list, we agree to make them both children of CASI, and add video-CASI at the same level.

Sanda will attempt some definitions for these for our next meeting.

Note however that we forgot to discuss whether the terms should appear as initials (VCASI, TACASI, ACASI) or prefixed with words (i.e., video-CASI, audio-CASI, telephone audio-CASI) – so everybody will need to form, and express, an opinion about this before our next meeting.

A few Google searches made by Taina reveal that data may now be collected using practically any of the new computer-based devices that have been launched in the past few years – tablets, smartphones, cell phones, etc. etc. In one of the search results these are referred to as “mobile computers” (but we do not necessarily want to use that term). We decide that we cannot include all of these in separate entries, so Taina suggests – and it is approved – that we modify the definition to “Self-administered questionnaire: Interactive” to reflect the fact that they use computer technology that can be built into a variety of tools, or devices. There is no need to list them, but we would want to include a couple of examples in parentheses, followed by etc. Lucy kindly accepts to produce a new draft definition for the next meeting.

Continuing with ModeOfCollection definitions:

Definitions for Observation: Field:Non-participant, as well as Observation: Laboratory, Laboratory:Participant, Participant:Overt, Participant:Covert, and Non-participant were finalized and approved.

Observation: Web-based. There is a lot of discussion around this. We are trying to make sense of what we are really describing here. We find that the use case in which a researcher signs up to social media and observes exchanges among members, collecting qualitative data about the nature and/or content of interactions is really very similar to “field observation” except that it is conducted in a virtual environment. Despite being virtual, this kind of online environment may be described as a “natural” environment in that it is not controlled, it is created willingly and spontaneously by real people, etc. *[should we include in our definition something to this effect? – just food for thought for now]*.

The example sent out by Lucy describes a completely different use case – in-built software collects data about the way students use online teaching materials developed for online courses; teachers then use these data to modify /improve these materials for the next courses offered. Presumably these data are about time spent on a page, how they navigate between pages, how they use the exercises in conjunction with the manuals, etc. etc. (Here students interact with online content that has been prepared and posted for some purpose. In social media, people interact among themselves; if they make comments about some external content, the comments are addressed to other people.) Sanda finds this second use case similar to what Google analytics does.

Those present agree that if we have an entry for web-based observation, we should be referring to this second use case, and not the first one. A good definition needs to be produced and we will need everybody to think about it and offer suggestions. We should try to work something out by email before next meeting. One concern raised by Taina is how to clearly differentiate between this type of observation and recording as a data collection method. Intuitively, they seem different. But we need to be able to make a clear distinction. Finally, it is suggested and agreed that we need to think about possible alternatives to the term “web-based” in an effort to better represent our concept. Some possibilities are online observation (we don’t like this one because it is more suggestive of a community, people being connected to, or present in, and online environment), e-observation, or computer-based observation. The latter appears as preferred, but we need everybody’s thoughts on this.

As they think about, and comment on, web-based observation, group members should also pay attention to possible definitions for web-based experiments, and how both concepts fit within our list.

Next meeting: April 17, at 13:00 GMT

**DDI\_CVG Meeting, February 6, 2013.**

Minutes

Present: Hilde, Jani, Meinhard, Sanda, Taina, Joan (?)

Publishing new versioning policy, guidelines for collection migration, and new versions of old CVs on the web

This has been done, with the exception of the XML catalog, on which Sanda is still making research and inquiries.

The content of the new pieces is okay, has already been discussed multiple times and approved at all levels. However, the presentation is perhaps not perfect. Group members please review the new pieces. Ideas for improvements are welcome (by email to group) as long as they do not require too much staff effort. The Web team at ICPSR can certainly help make some changes, but we can’t ask for complicated / complex changes because time is very tight.

Definition of Mode of Collection. Opinions are generally split between keeping it simple and very general (as it is right now) and perhaps adding a few words about both primary and secondary methods of collection being covered in the list. We would be reluctant to expand the definition if it becomes more open to interpretation. There may be some ambiguity as to whether some sources (for instance, photos or various print materials) yield primary or secondary data. For the time being , we decide to leave the definition as is, and Sanda adds the proposed expansion in greyed out font, so that we do not lose it – in case we decide to go back to it.

Content analysis – it is agreed to change this term to “content coding”. The latest proposed definition (Sanda’s, based on, and with amendments from group) is also approved.

We finalize definitions for: observation, observation:field, Observation:field: participant, and observation:field:participant:overt/covert. We decide that we prefer the terms overt/covert to obtrusive/unobtrusive. Meinhard rightfully points out that the former are more neutral whereas the latter may take a negative connotation. Overt and covert are also more frequently used (Taina counts Google hits)

For next time

The definitions for Web-based experiments and observation need some attention from everybody. There are some suggestions from Sanda, Taina and Meinhard in this meeting’s final agenda. But everybody needs to think about them.

We need suggestions for Observation:Laboratory – all subcategories. We currently have no definitions for these. Hilde and Sanda will come up with ideas for next time.

If we make sufficient progress, we might be able to go even further down the list.

It would be helpful if people gave some thought to any or all of the definitions that have not been finalized.

Next meeting: February 27, 2013.

(There won’t be a detailed agenda, since we already know what we are doing. And the outstanding tasks are already outlined in the minutes above. Interim communication is encouraged ☺)

**DDI\_CVG Meeting Minutes, February 27, 2013.**

Present: Meinhard, Taina, Joan

Since there were so few of us, we did not make any decisions, only some suggestions for discussion. We started by discussing *Observation: Web-based* but became quickly stuck, not knowing whether it should refer to observing online textual exchanges or data collected by software. In either case there may be overlapping with some other terms in the CV (see comments in the ModeOfCollection CV document). Difficult distinction between the mode and the media of collection. And if we assume it to be the latter option (data collected by software) we need to check that the definitions for other observation terms allow web-based observation.

Observation: Field: Non-participant

We suggest rephrasing “in a natural setting” to “in a non-controlled setting” if we decided this includes for example, observing chat, blog, forum texts etc. (But according to its definition, these could also easily be Self-administered writings: Web-based??)

“Observation that is conducted in a non-controlled setting without any interaction between the researcher and his/her subjects”.

Observation: Laboratory observation

Observation: Laboratory: Participant

Hilde’s definitions were approved.

Observation: Laboratory: Participant: Overt

We suggest shortening the example a bit to avoid tautology

“Type of laboratory observation in which the researcher participates in the social situation and acknowledges and explains the reason for his/her presence in the laboratory setting. Example: Observation of children’s play in a laboratory playroom where the researcher takes part in the play and explains the reason for his/her presence”.

Observation: Laboratory: Participant: Covert

We suggest changing the example by replacing “in a disguised manner” at the end to:

“Type of laboratory observation in which the researcher participates in the social situation without informing the subjects of the reasons of his/her presence. Example: Observation of children’s play in a laboratory playroom where the researcher takes part in the play assuming the role of a normal participant.”

Observation: Laboratory: Non-participant observation

We support including this term to the CV and agree with Hilde’s definition.

Taina had noticed a few presentations and texts on qualitative data collection methods which included Projective Techniques. We decided to leave to discussion for next time whether this term should be included or could be covered by ‘Other’. FSD now has data collected with this method.

A projective technique is defined as an instrument that requires a subject to look at an ambiguous stimulus and to give it structure (SAGE Encyclopedia of Social Science Research Methods). May include word association, sentence completion, vignettes, cartoon test, thematic apperception tests, role play, bubble drawings.

Sanda will set up a Doodle for next meeting. Taina is away from office 25 March – 5 April, so maybe on 20 March? Jani is available on 27 March to let people into the webinar facility, so that is also another possibility.

**DDI-CVG Minutes: Jan 9, 2013**

Present: Hilde, Taina, Lucy, (Jani)

Consensus on the ModeofCollection list order:

Interview, Self-administered questionnaire, Focus Group, Self-administered writings etc. We discussed the alternative of having the terms in alphabetical order but concluded that because of the hierarchical structure it is not possible.

‘Content analysis’:

We agreed that the currently suggested ModeOfCollection definition is good (though did not discuss the alternatives ‘use’/’apply’ in the wording) and that we need this mode. However, the descriptive term name should preferably be something else. Please discuss in your respective organisations Hilde’s suggestion for a new term name ‘Content coding’, which we thought sounded a good option. Decision next meeting.

Web-based observation and web-based experiment:

We were not sure whether definitions would be possible, as information on the net seems to indicate that people include a great variety of things under these terms. Meinhard, could you try to produce draft definitions and examples for the next time?

Definition of the ModeOfCollection CV:

To clarify that the scope of the CV, we decided to submit a new draft definition to be discussed the next time:

"The procedure, technique or mode of inquiry used to attain original data (primary data collection), or the method of producing data from already existing sources (secondary data collection)”.

-The advantage would be that we would not have to explain in the definitions of the specific terms what we mean by primary and secondary data collection. Decision next meeting, amendments welcome.

Next meeting: perhaps 23 or 30 January? Sanda, please set up a Doodle for the dates suitable to you.

DDI-CVG Meeting, November 21, 2012

Minutes

Data Collection Methods List: Meinhard has sent an edited list where he has further categorized data collection methods, while adding possible new terms. We start by discussing this document.

We agree that the most useful “new” distinction in this list is between primary and secondary data collection methods. This is really helpful in placing and defining methods, however we agree that we should not create an additional level in the hierarchy with these concepts. Rather, we will add in the definition for the secondary types the fact that they are secondary data collection modes. In this category we currently have Content Analysis, Transcription, Compilation, Synthesis, Aggregation, Simulation [do we have a good definition for Simulation? Need to talk about this in a future meeting]

“Test and assessments” – it is noted, and generally agreed that these would qualify as instruments, rather than modes of collection. We could have a mode of “interview” or “self-administered questionnaire” or “experiment” and tests as instruments. We will tentatively add these to the “instrument type” list. [Sanda – my preference would be for “tests and evaluations” – group please comment whether “assessments” or “evaluations”]

Distinguishing between “quantitative” and “qualitative” interview. We have already decided not to make this distinction, and for that reason “interview” only is now defined as a mode of collection. It can be either qualitative, or quantitative, or both and most subcategories do apply to both types. If needed, the qualitative vs. quantitative distinction can be done by using the KindOfData@type attribute.

Taina reminds us that we have asked TIC to introduce a text field as a child of Mode of Collection that would parallel the coded field and would allow users to enter additional info about this operation. Sanda will check if this has indeed been done in V3.2 and will report to the group.

“Projective methods” – these may fall under experiments and therefore may not need a separate entry.

Web-based observation – at first glance, this may be a useful addition to observation types. If we add this, it would be at the same level as “field” and “laboratory” [group? – need more discussion on this].

Web-based experiment. Not sure what these would look like. We need to further clarify. Meinhard will provide examples. Need to make sure that such experiments are not really “field” experiments. [group?]

Types of recording (audio, video, mechanical, etc.) – these are covered in Kind of Data. But we could, and probably should, mention them in the actual definition for recording.

Diaries, journals, etc. These are covered in self-administered writings. Perhaps we should add “self-reported checklists” in that definition? [group? – if agreeable, Sanda will add]

Meta-analysis / synthesis. Sanda disagrees with including meta-analysis as a term, because that is a dedicated mode of analysis, not collection. However, synthesis may well qualify as a mode of collection for the data used in meta-analysis. All: we need clear info on how the original data (from multiple studies) are re-processed in order to apply meta-analysis. We could then figure out whether they are in fact synthesized, or just compiled (i.e., new identity, or preservation of former identity?)

Content Analysis. We note Sharon’s disagreement, however the group’s preference is to keep this term, and avoid Coding, which is found too broad, can be applied in several different ways to different ends, and therefore can create confusion. We note that there is useful clarity in the excerpt sent by Michaela, so Meinhard will try to take some pieces of that quotation and include them in the definition for Content Analysis, to make it more clear and complete.

All: please review definitions for transcription, compilation, synthesis, aggregation, recording, simulation in view of our next discussion. Taina will try to find relevant pieces in the recent emails and add them to the draft document. Sharon’s proposal on aggregation has already been added.

Observation types. We agree on Field and Laboratory as second tier terms and Participant and Non-participant as the third level. After some discussion we also agree that the distinction between overt and covert participant observation is quite important in data analysis. We end up by adding a fourth level with these terms; in general we do try to keep the hierarchies at a maximum of three levels, but this seemed too important to miss.

Experiment – we will talk about this next time.

Next videoconference: Wednesday, January 9, 2013.

Best wishes for a happy holiday season and an excellent New Year!

**DDI-CVG Meeting, November 7, 2012**

Minutes

Present: Hilde, Jani, Meinhard, Sanda, Taina.

Email contributions: Michaela, Sharon.

We discuss the possibility of replacing “Coding” and “Synthesis” with Content Analysis in the Mode of Collection list. Pros and cons are being expressed, and the removal and replacement of the current terms is discussed in conjunction with possible definitions for all of these terms. Coding seems too broad a term, and can be misleading. It is frequently used as a data processing method (surveys, open-ended questions), not as a data collection method. If these terms are used, we would have to restrict their meaning in the definitions. Jani also points out to a useful distinction between “coding” and “encoding” (coding implies categorization and some loss of content, while encoding normally does not lose content).

The final consensus tends toward replacing “coding” and “synthesis” with Content Analysis, and we come up with the following draft definition for this term: **“As a mode of data collection, content analysis is a set of methods for systematically coding and/or synthesizing qualitative data with the purpose of producing new data”.**

Task: Michaela and Sharon, please comment. Taina will contact people from the CLARIN group to get their feedback.

Website and publication. There are some restrictions from TIC and the Alliance regarding publication, presentation, downloads and file URLs. For now we will try the following compromise solution: Jani will generate the new files, and Sanda will try to get approval for the website work to be done at ICPSR.

Task: Sanda will approach the assistant director at ICPSR as well as the Web team to see if they can help with publication

Next meeting: November 21, at 13:00 GMT.

Agenda for next meeting: we will continue to discuss Mode of Collection definitions, as outlined in the Agenda for October.

We anticipate no meeting during December.

**DDI-CVG Meeting, October 17, 2012**

Minutes

Present: Hilde, Sanda, Taina (with thanks to Michaela for her email contributions)

Today’s discussion was centered primarily on how we treat translations.

While nobody disputes the need to version translations, we are not able to come to an agreement regarding an optimal, or recommended, format for publishing translations (each language separately or all together in a single document?). The publication format impacts the versioning so, at this point in time, we find ourselves unable to make recommendations about that either.

That should not affect our work, however, because it has already been decided that our job is limited to creating and publishing the master CVs in American English. We will neither produce, nor maintain translations because we do not have the necessary resources, and are not qualified to do such a job.

Since currently there are no translations of the CVs, and we are not aware of any concrete plans to create them in the near future, we decide to leave the matters of publication format, versioning, and perhaps even selection of the hosting site(s), to be decided when such plans actually come into being.

Taina notes that it is very likely that CESSDA-Eric will initiate these translations. At that time, we can talk with whoever is in charge of this project and see how, and what, we could contribute, even if just opinions or advice.

In the meantime, we will refrain from making detailed comments or recommendations regarding translations on our current web site. Under “Publication, Maintenance and Management” we will add a sentence stating that our group will not produce translations, but we would like to be informed if other agencies are interested and /or plan to do so.

We will publish the new versioning policy without any reference to translations.

It is still unclear if and how we need to publish Achim’s catalog, and what it does. Also, how do we proceed to publish an update to TimeMethod? Probably regenerate the entire package with new version numbers (2-digits), while the old package will remain accessible? Sanda will ask Mary to approach Achim at Dagstuhl and try to get some clarification directly from him.

We want to have another meeting soon, since we feel we did not make enough progress today. Sanda will set up a Doodle for our next meeting, suggesting either October 31st or November 7.

**DDI-CVG meeting, August 15, 2012**

Minutes

New versioning policy: we have agreement from everybody to proceed with this.

Guidelines for migration will be published separately from the actual policy text.

We will try to avoid mentioning issues like compatibility and flat lists in the policy text, as well as in the migration guidelines.

Sanda will put together a policy text (to go on the web) as well as a request to TIC for policy approval. She will send the text to the group and then to TIC before the next meeting.

“Mode of collection” definitions:

Focus group, upper level – approved, final.

ACASI – reviewed to incorporate part of Michaela’s suggestion. Final.

Coding – we lack good/convincing examples. Discussion follows – is coding really a data collection method or is it data analysis? Sometimes it is hard to draw a clear line, especially when qualitative data/methods are involved. What is data collection? How do we define it? Sanda suggests that perhaps we can use data collection in conjunction with secondary sources when new data are being produced. Could we use content analysis as a data collection method, perhaps to replace coding and synthesis?  
Meinhard finds excellent support for including content analysis as a mode of collection.

Task: Sharon, Hilde, Taina, Chryssa, Michaela – please do some more thinking/research on this and send out your findings (including examples) before our next meeting.

We resume next time with Transcription, Compilation, Synthesis, Aggregation, Recording, Simulation, Observation (+list), Experiment (+list), Physical measurement.

Task: all – please review comments included in May 23, 2012 “Tentative Agenda”.

Next meeting: October 17, 2012, 13:00 GMT

DDI-CVG Meeting, May 23rd 2012

Minutes.

Present: Chryssa, Hilde, Sanda, Taina.

This meeting was dedicated to the CV versioning issue.

The opinions expressed by all members, as well as Achim’s and Jani’s contributions, may be found in detail in the document “CV Versioning Discussion.docx” posted on the Thesauri site.

After reviewing the various positions, we came to the following agreement:

We will issue a new versioning policy

The new policy will be based on an intellectual, or logical, assessment of the nature of change, and will only distinguish between substantive and non-substantive changes in the CVs, as described further below.

Since we only distinguish between two types of changes, we will change the version numbering system from three digits to two digits (separated by a period, both digits will always need to be expressed). The DDI 3.1 schema includes a two-digit default for CV version number, so moving forward it will not be a problem to reference the CV in the DDI instance. Problems related to previous usage: We will post an announcement on the DDI users list as well as on the Controlled Vocabularies page alerting users to retroactively change any reference to V1.0.0 into V1.0. From that point on, new versions can be used and referenced normally.

To the CVs documentation, we will add more detailed information on the nature of the change as suggested by Jani (see details below), to assist in software creation/development.

Some of the CVs use hierarchy to place terms in a conceptual context. In applying the versioning policy, however, we will treat the CVs as “flat” lists, meaning that changes to any code, term, or definition will have the same status, regardless of the hierarchy level at which the codes are placed.

Substantive changes: any change in (list) content or (code) meaning

Addition of new code(s) (change in list content)

Deletion of existing code(s) (change in list content)

Widening the definition of a code (change in meaning)

Narrowing the definition of a code (change in meaning)

Change in “name” of a code, including change in spelling; since the codes are the “official” or “legal” entries (“terms” and “definitions” are documentation for the codes) a change in name really amounts to a change in code, i.e. change in list content, therefore this is a substantial change)

Merging codes (amounts to deleting codes and adding new one(s))

Splitting codes (amounts to deleting codes and adding new ones)

Non-substantive changes: Changes in wording, spelling, etc. (i.e. “form”) that do not involve changes in content or meaning:

Rephrasing a definition to make it clearer, or adding examples without changing the meaning of the code

Rephrasing a “term” (the natural language “label” for the code) for clarity without changing the meaning of the code

Correcting spelling errors in both “term” and “definition”.

In addition to a change in the version number, each new version of a CV will contain detailed documentation about how the new CV compares with the previous version. The documentation will cover the following possible changes, as proposed by Jani, but we will take out the parts that refer to migration measures, because we don’t want to be telling people what to do with the information provided (they might use it for other purposes than migration, depending on their own specific needs and goals):

UNCHANGED (X): The code and its definition have remained unchanged. No migration measures needed.   
  
RENAMED (X, Y): The definition has remained the same but the code itself has changed and is now Y. X can be changed to Y with automatic migration.   
  
REPHRASED (X): The definition has been rephrased without any real change in the meaning. No migration measures needed.   
  
WIDENED(X): The definition of the code has been changed to enlarge the meaning of the code. All old codes are included in the new definition. From the point of view of migration, old X can be transferred to new X with automatic migration.   
  
NARROWED(X): The definition of the code has been narrowed from which follows that a human being has to check all the data where this code has been used and decide whether the code is still valid for the dataset or whether it has to be changed.   
  
MERGED (X, Y):  The old code X has been removed and all the data classified with it now are included in code Y. The difference between MERGED and RENAMED: RENAMED denotes that the code has changed but the definition has remained the same (e.g. a mistake in the code spelling). MERGED does not guarantee that the definition has stayed the same but only says that all the old X codes can be automatically migrated to Y.   
  
REMOVED (X): The code X has been deleted. A human being has to go through all data documented with it and decide what code to use instead.   
  
ADDED (Y): A code has been added to the CV. No migration measures necessary. (If the new code causes migration measures, these are included in the specifications for the old codes).

The spreadsheet itself (therefore the downloadable package) will only contain brief “codified” information about the changes. For example, RENAMED (Individual, Person). Additional information about the meaning of the entries will be posted online, on the Controlled Vocabularies page, in connection with a description of the versioning policy.

Members of DDI-CVG have until June 25, 2012 to agree or disagree with the general course of action described above. Minor modifications that do not change the essence of our decision may be requested, but will not be accepted after June 27, 2012, 14:00 GMT.

Members who substantively disagree with the proposed solution are urged to present a fully developed alternative that takes into account all of the issues raised in our discussion (types of changes, version numbering, etc.), also no later than June 25, 2012.

If we do not hear from any member until the deadlines stated above, we will assume that he or she fully agrees with what has been proposed, including any subsequent changes that might occur.

When we have agreement from a majority (4 members, I suppose), but not sooner than June 27, 2012, 14:00 GMT, we will proceed to implement the changes, re-write and re-publish the versioning policy, and re-generate and re-post all CVs (with Jani’s assistance)

Next videoconference: June 27 at 13:00 GMT

**DDI-CVG Meeting, April 25, 2012**

Minutes.

**Time and timing issues.**

Due to the time zone difference, Sanda is unable to read suggestions or comments sent the day of the meeting before the meeting actually starts.

Therefore it was agreed that the deadline for emailing comments would always be the day before the meeting. Those who cannot meet this deadline should bring (voice) their comments directly to the meeting, or wait until minutes are sent, to see if their comments are still relevant or not.

**Updating TimeMethod.**

Sanda can send the updated spreadsheet to Jani; he will take care of everything.

**CV Versioning rules.**

Discussion postponed for next time, since the issue has increased in complexity as a result of Achim’s recent comments, followed by Jani’s extensive response. There is also a message from Larry Hoyle. Sanda will put them together and will send them over to the group. Taina will post them under Web pages and publications on the “thesauri” page.

Task for group: these messages need to be read quite carefully before the next meeting. They are not easy to understand, and we must come up with some kind of position. I would not publish any new version until we have a policy.

**New definition of ModeOfCollection.**

Approved. Final. Sanda will enter in working documents.

**Focus Group definitions.**

Regarding the first one (upper level) Chryssa thinks we may need to mention something about how the group was created / the participants selected. The general feeling is that this may not hurt, but it may not be mandatory either, because there is a Sampling element where this kind of information may be entered. (This is a free text / descriptive element and will probably remain so in DDI3.2, while an extensive sampling section is projected for a future DDI 4.0)

With Chryssa’s approval Sanda has also entered a proposal in the document called ModeOfCollection\_CV\_2012\_04\_25 . If the alternatives don’t look good, the old one will stand.

Definitions for subcategories are good, approved. Final.

**ACASI**

Definition slightly amended but needs a last review. Sanda will give it another look.

**Coding**

New definition approved, however the example is not good enough. Sharon will try to come up with 1-2 better examples for next time.

**Observation and Experiment plus subcategories. Compilation and Synthesis.**

Postponed for next meeting. Group please review existing comments and suggestions.

**Next videoconference: Wednesday May 23rd, 13:00 GMT**

**DDI-CVG meeting, February 1, 2012**

**Minutes**

Present: Hilde, Sanda, Taina

CV versioning, website, tools, Michaela’s suggestions.

Trying to incorporate Michaela’s suggestions for the introductory paragraph on versioning, we come up with the following suggestion: “CV versioning follows a three-level structure (example: 1.0.0). A CV will include the version number, as well as a note about the changes it contains compared to the previous version. A version will always be referenced with a three digit number, expressing all its levels (i.e. Version 1.0.0, not Version 1). Version numbering will be made according to the following rules:” – Group: please read and approve, or suggest edits if found necessary.

We need to incorporate information about version changes in the actual spreadsheet, which is the base document for data input and subsequent conversion. We agree that the best way to do this would be to create an additional row on the Identity tab. The row would be labeled “Version Notes”. (Version history, or Version Rationale, or Version Changes, were also suggested, but it was decided that Notes would be best).

Jani will need to modify the spreadsheet to incorporate this new row. Taina can talk to him, when he is ready to look at the CV tools. Sanda should also email him to this effect.

Definitions for types of focus groups.

(Michaela) definitions may sometimes seem redundant, however it is a rule that we have a definition for each term. (Redundancy – probably okay in the age of electronics with reduced attention span.)

Sharon’s suggested definitions: we try to make them more concise. We replace interviewers and participants with “participants” given that the latter may change roles in some focus groups. We remove the synchronous/asynchronous dichotomy from all definitions, because it would make “telephone” similar to “online” in that respect. The revised definitions are in the document ModeOfCollection CV dated 2012\_02\_01.

Group: review and approve, or suggest edits if found necessary.

Subcategories in Observation. Yes these are conceptually different, so we need a new sub-categorization. Here are two possibilities:

Observation

Field

Participant

Non-participant

Laboratory

Participant

Non-participant

OR: (see next page)

Observation

Participant

Field

Laboratory

Non\_participant

Field

Laboratory

Group: please advise. Consult colleagues if possible/appropriate. Need opinions on this.

Taina’s remarks regarding the conceptualization of methodology elements.

We all agree there was some hard thinking involved, and this is not an easy matter to resolve.

Practical implications:   
 a. The definition for ModeOfCollection seems very good (“The procedure, technique, or mode of inquiry used to attain the data”) We could ask TIC to include it in the inline documentation for DDI 3.2 (?) but we will also use it as the definition for the main CV entry.

b. We probably do not need an additional DDI field for “Capture method”. An important reason is the presence of some terms in our ModeOfCollection list that may also describe capture methods. We want to avoid creating further confusion. But, a field where the capture method could be described in free text if needed or desired, would be welcome.

On December 12, 2011, Sanda has sent the following request to TIC:

*“On behalf of the Controlled Vocabularies Working Group:*

*For DDI 3.2 we have requested that ModeOfCollection be changed to a “codevalueType” because we are in the process of building a controlled vocabulary for this element.*

*However, we would also like to have a “descriptive element” accompany this coded content, for those who would like to add comments/specifics on the CV term used.*

*So we would like to suggest that ModeOfCollection have two subelements in the next iteration of DDI 3:*

*1. ModeOfCollection:Type (would be a codevalueType)*

*2. ModeOfCollection:Description (would be a string type for entering free text)*

*Thank you so much for your attention.*

So, when it is implemented, the “Description” field may be used to include information about data capture. However, we do need to remember to recommend this kind of usage somewhere in the documentation (for CVs, for DDI 3.2) Taina will be in charge, because she is the most organized.

Sanda also checks DDI 2.5 (freshly published) and confirms that collMode supports both a CV and a free text entry (see explanation of conceptualTextType in the DDI 2.5 schema)

Definitions for compilation and synthesis – discussion postponed for next meeting, we prefer to hear from Chryssa and Meinhard before we produce anything close to final.

Jani’s comment on quantitative and qualitative being research methods rather than types of data.

While Jani is certainly right, DDI does not document the ways the data are analyzed, but rather the way they are presented to the analyst (who then chooses what to do with them). And qualitative vs. quantitative are terms that have been traditionally used by archives to indicate presence of “quantifiable” or “unquantifiable” data, so to speak. So people will want to use them.

The attribute type, with choices: quantitative, qualitative, mixed, is already hard-coded in DDI 3 and we do not want to take it out.

Its established existence leads us to decide that the second alternative offered for the KindOfData list is really not viable. Sanda makes a note of this in the KindOfData column.

Group: will review the cleaned-up copy of definitions and will be ready to comment.

Next meeting: February 29th, 2012, at 13:00 GMT

We will take advantage of this extra day made available to us as a last chance to catch up on all of the unfinished business accumulated in the past four years!

**DDI-CVG meeting, January 11, 2012**

Minutes

Updating TimeMethod: : Jani has agreed to help with publishing CVs and has requested the tools from Achim. Achim has not responded yet (he has been on vacation).

CV versioning info on the Web – Sanda will try to revisit this text before the next meeting (did not have time to do this so far).

Incorporating qualitative data terms in our CVs: we remain open to suggestions from people who work with qualitative data as long as we can incorporate them in our normal workflow and projected timelines.

ModeOfCollection list:

We remove “recordings” and “notes” as subcategories of face-to-face and telephone interviews; these are modes of recording data rather than modes of data collection.

Sanda’s proposal to add “notes” as a data collection mode (same level as recordings, simulation, etc. – i.e., top level) was voted down on the grounds that it could be confusing, and observation as a more general concept covers for the case in which a researcher would be taking notes about some event (not during an interview)

Subcategories email, paper and web-based will \*not\* be dropped for “self administered writings”, since in this case they do describe the way the data were collected. These are media used by respondents to create data, therefore may have an impact on the actual data content, while recordings and notes as described in a. above are really means by which the researcher records the responses given by the respondent through a different medium (say, telephone)

ACASI is added as a subcategory under self-administered questionnaire, Interactive. This mode of collection was mentioned in an issue of the Journal of Public Health, found by Hilde (<http://jpubhealth.oxfordjournals.org/content/27/3/281/T1.expansion.html>). The definition is added by Sanda in the ModeOfCollection definitions document.

Adding subcategories “video” and “audio” under “recording” was considered, but eventually voted down on the grounds that the type of recording would be specified under KindOfData.

Adding types of focus groups as subcategories – the consensus was that they are not really needed. However, we decided to add subcategories “face-to-face”, “telephone”, and “online” to be consistent with the ‘medium’ being mentioned under Interview and Self-administered questionnaire. Sharon will provide definitions for these.

ModeOfCollection definitions:

Definition for self-administered writings: amended (how collected was deleted, to make it more general)

Definition for interview discussed and approved: we include group interviews in interview with the specification that Focus group should be used when group interaction is part of the method.

New definition for focus group discussed and approved: we basically get rid of the words “group interview” and emphasize the interaction among participants.

Issues we did not have time to discuss;

Point raised by Sanda regarding the subcategories in Observation. *(Observation – field and laboratory refer to a place, whereas “participant” refers to the researcher’s role. This is a “conceptual” problem as well. We probably need participant/non-participant for “field” and – not sure about “laboratory”: perhaps we need the participant/non-participant subcategories there as well? And, to a recent comment in one of the emailed contributions – yes, observation can also be in a laboratory, it is not necessarily done in the field (i.e. natural environment)*

Compilation and synthesis: (*Definitely keep both. Delete textual from synthesis (may use other media, like video?). Rewrite in the light of Chryssa’s research and Sanda’s notes.*

*(I think the distinguishing factor here is what happens to the constituent parts:*

*Compile – syn. Collect, assemble. Constituent pieces retain their identity: example collection of poems, list of names. Synthesize – a new entity with a new identity is created, while the identity (individuality) of the pieces is lost. Syn: combine, blend, fusion. Best example is from chemistry – synthesis produces a new  substance, with different properties than the individual components.)*

Next meeting: Wednesday, February 1st, 13:00 GMT.

**DDI-CVG Meeting Minutes: November 9, 2011**

TimeMethod: Sanda has found a definition for “cross-sectional” data that seems to answer the concern that this term with its current definition might not be useful for describing certain types of qualitative data that are not longitudinal, but have multiple data collection events.

The idea is that longitudinal data attempts to capture / analyze change, by means of comparison (differences over time). The new definition for cross-sectional addresses this issue, and has been approved in the following form:” *Cross-sectional data refers to data collected by observing subjects at the same point in time, or without regard to differences in time. Analysis of cross-sectional data usually consists in comparing the differences among subjects. (adapted from Wikipedia)”*

Sanda will try to produce a new version of the TimeMethod vocabulary and publish it on the Web.

External contributions to controlled vocabularies: this is still not possible, as the tools for creating and publishing vocabularies are not made available via the Alliance website. TIC / The Alliance have to decide first how we credit / document external contributions and then this kind of information has to be added to the input form (spreadsheet) for vocabularies. (Currently, it is only possible to credit the DDI Alliance as author).

Covering for qualitative data in our controlled vocabularies: we would be interested to have the qualitative data working group review our vocabularies before publication, to find out if they can be used to describe qualitative data as well as quantitative. Sanda will email the group with this request.

Mode of Collection : list of terms. We change Online/Internet from Self-administered writings to Web-based to be consistent with Self-administered questionnaire. We also add “web-based” type to interview.

Mode of Collection, definitions. We review the new or edited ones, including the comments made by John Southall from Qualidata (UKDA). Decisions are marked in bold type in the document ModeOfCollection CV 2011\_11\_09, to be circulated and posted as a working document by Taina on the thesauri page. It is an excellent idea to preserve the history of our discussions; however, it is also time consuming, and therefore might not be feasible in the long run. Perhaps preserving such “working documents” – i.e. annotated versions will help in documenting how we made decisions.

Tasks for next meeting:

Taina will try to produce a definition for our “new” entry – Interview—Web –based.

In addition to the tasks outlined above, Sanda will try to circulate a proposal regarding “Instrument” (replacing Interview with something like Interview guides/protocols with two subdivisions – semistructured and unstructured (actually, this *is* the proposal, but we will discuss it next time)

All members will spend some “quality time” with, and get ready to comment on the definitions for ModeOfCollection as well as for DataSource:SourceType. Particularly useful would be a definition for Self-administered questionnaire: Interactive.

Next videoconference: Wednesday, December 7, 13:00 GMT. Full attendance is highly desirable :-) so that we end the year on a high note, and also wish Happy Holidays to everyone!

**DDI-CVG videoconference September 14, 2011**

Minutes

Present: Hilde, Sharon, Taina, Sanda.

Michaela Olde, Assistant Director for Data Management at the Australian Bureau of Statistics (ABS) has requested to be added to our mailing list. She is interested in finding out more about our work, and has promised to contribute whenever possible by sending in comments and/or suggestions. She will not be able to participate in our meetings, but we welcome her interest and virtual presence in our discussions.

The Qualitative Data Working Group has expressed interest in looking more closely at our work and giving us ideas / concrete proposals on how we might incorporate terms that are specific to qualitative research. We certainly welcome a dialogue and are also interested in finding out more about their work, especially after their December meeting in Gothenburg. Taina can serve as a “liaison” since she works together with Arja Kuula from the Qualitative Data Group, and Arja herself is welcome to participate in our meetings as appropriate.

Today’s discussion centers on some proposals / comments originating from Arja on how to incorporate some terms specific to qualitative data in our lists.

DataSource: SourceType: We will keep processes on our list. We are also inclined to keep subcategories, as the term processes on its own may be too general/all encompassing. We will try to find good definitions for the existing subcategories, and as we define them we hope to be able to ascertain which ones should stay. Taina has added Hilde’s definition of natural processes on version 11 of the DataSource Definitions, which is now uploaded on the thesauri site. Definitions for all subcategories are still up for discussion.

TimeMethod: We agree that the term “longitudinal” cannot be taken out from the list. We find that Arja’s examples can actually be defined as cross-sectional with multiple data collection events or simply longitudinal:panel if the time frame was long enough.

Sanda emphasizes that the term “panel” itself refers to the sample rather than the time dimension. It is used in this list only to differentiate between longitudinal studies: some use a panel-type sample, and others a sample that is representative of the population, but composed of different individuals (an entirely new sample is drawn with every iteration of the data collection, usually a “wave”). Taina suggests we rework the definition of “cross-section” to clarify that the length of the time period is of the essence here. Participants agree with changing the current wording to :” Data about a population are obtained only once, or over a short period of time”. Need consent from the absentees.

ModeOfCollection: Participants agree that a revised list of terms should be submitted for evaluation and approval, including by today’s absentees. The list runs as follows and basically eliminates the word survey so that it can cover qualitative data in addition to quantitative:

Interview

Face-to-face

CAPI

PAPI

Recording(s)

Notes

Telephone

CATI

Recording(s)

Notes

E-mail

Self-administered questionnaire

Fixed form

E-mail

Fax

Paper (SAQ)

SMS/MMS

Web-based

Interactive

CASI

CAWI

Self-administered writings

E-mail

Paper

Online (Internet ?)

Coding

Transcription

Compilation

Synthesis

Aggregation

Recording

Simulation

Observation

Field observation

Laboratory observation

Participant observation

Experiment

Laboratory experiment

Field experiment/Intervention

Focus group

Physical measurement

Other

When we have everybody’s approval, we also need to reconsider the existing definitions.

ResearchInstrument:Type – we came up with the following revised proposal:

Questionnaire:

structured

semi-structured

unstructured

(Open?) Interview

semi-structured

unstructured

Instructions;

written

verbal

Is this a satisfactory list?

For next time, we need to consider the proposed definitions for the list above.

Next meeting: November 9, 13:00 GMT

**DDI-CVG meeting minutes, August 10, 2011.**

Present: Hilde, Sharon, Taina, Sanda.

Controlled vocabularies Web page:

Some minor changes to the main page were made by Sanda as a result of our July meeting, including a special tab for download on the right hand side of the page.

More detailed text on contributions, translations and how we create vocabularies is in the works, but not yet in stable draft form. Sanda needs to do more work on this. August is a vacation month for many, so TIC and Mary will be consulted starting September.

Data Source Definitions:

Taina’s proposal for amending the Court proceedings definition (change case to plural) was approved, and the definition was edited.

Events/transactions: Are they really a data source? Yes, they can be, when events are directly observed and/or recorded by researchers as they happen. Added a sentence to definition to reflect this, and re-worked the examples. Final draft of definition in SourceType\_definitions\_V9, needs to be looked at by today’s absentees, and approved.

Events vs. Processes – there is a need to clearly differentiate between the two. In this light, natural phenomena can also be events (earthquake) or processes (erosion). Therefore it seems appropriate to change natural phenomena to natural events and make it a “child” of events.   
This makes sense.

Definition of natural events revised and final draft provided for group’s approval.

As a result of the discussion about natural phenomena, natural processes is added under “processes”. Hilde might look into providing a draft definition, as appropriate.

Revising the data source list:

We agreed to make natural events a child of events.

Processes: again, the question is – can they really be sources for data? This is rather unlikely for social, economic, political processes because normally one becomes aware of them (and decides to study them) after they start taking place, sometimes even after they “end”. It is much harder / quite unlikely to observe them directly. But that is not true of all processes – biological and behavioral may be observed directly in controlled environments, experiments. How about natural processes ? Can something like “decay’ or “erosion” be observed directly, or only through measurements, and then they are not a source. Although they may be “videotaped” (direct recording) it is unlikely that they would be studied that way?

These are the questions that need to be discussed and decided next time: do we drop all processes from the list? Do we keep some (biological, behavioral) – which ones? (natural, yes or no?) Input requested from entire group.

Revisiting Mode of Collection in conjunction with thoughts about Instrument Type

In her work at FSD, Taina has come across some types of questionnaires that do not seem to be covered in the current Mode of Collection list. In consultation with Hilde, she will provide a proposal for modifying (adding to) the current Mode of Collection list, with some draft definitions.

They will also produce some thoughts on how we might start to build the Instrument Type list.

Task for all: make sure you can attend our next meeting – There’s power in numbers!

Next meeting: September 14, 13:00 GMT

**DDI-CVG July 14, 2011**

**Minutes**

Present: Chryssa, Hilde, Meinhard, Sanda, Taina

Controlled Vocabularies Web pages:

On the main page, we will keep the first two paragraphs as they are.

We will try to move the download link to the left hand side of the page (Sanda, ICPSR Web).

Usage : will keep the current link, while adjusting for content (Sanda, Taina?).

Publication, Maintenance and Management – short paragraph on main page, with link to more developed content. Sanda will work on the content, and then will submit to Achim/TIC for consultation and approval. Will include versioning policy here rather than on a different page.

Translation – Taina will try to produce a draft, based on the texts she has already produced.

How we work – Taina will send over notes, based on which Sanda will start a draft, then circulate it for others to add/improve.

Data Source Definitions:

Definitions through “voting results” were finalized.

All group members will focus on all remaining definitions, as well as the possibility of reordering some terms or restructuring the classification (this item was on the agenda, but we did not have time to discuss it).

Next meeting: Wednesday August 10, 2011, 13:00 GMT

**DDI-CVG MEETING June 16, 2011**

Minutes

Present: Hilde, Meinhard, Sharon, Sanda.

Published vocabularies: Web presentation/documentation, CVs management, creation of new vocabularies outside the DDI Alliance, publication of tools

DDI-CVG will manage the published vocabularies (make changes and publish new versions according to type of change) and publish new ones on behalf of the DDI Alliance.

Requests for changes/edits/corrections to existing vocabularies should be submitted to DDI-CVG. That is already stated on the website. Perhaps stronger/clearer wording is needed?

In conjunction with a. above, need to publish versioning policy on the web site. (Where?)

When the tools package is made available for download, we should also publish a clear policy regarding the creation of new vocabularies (in addition to Achim’s instruction on *how* to create them). Suggestions: New vocabularies may also be created by other organizations/groups with the option of using the tools provided by the Alliance. In this case, the creators can choose to submit their lists to the Alliance for approval/adoption, or not. If the vocabularies are adopted by the Alliance, they would be published along with the others (?) on the DDI site (will they appear as co-authors? how do we credit them?). If they don’t submit them for adoption by the Alliance, they cannot reference the Alliance in the agency attributes. They should reference themselves, but we still want to know about the existence of such vocabularies. Perhaps we will want to reference them from our site (?) although they cannot sit on our site. All of the above has to be discussed and decided with Achim, TIC, and Mary Vardigan.

Creation of new vocabularies for other research fields (medicine, biology, etc.) in conjunction with the projected DDI 4. This is an area where other organizations might indeed bring valuable contributions, however we should not publish anything to this effect on the website until we have a clearer plan about DDI 4 – both concerning its contents and the actual timeline for its development.

According to TIC’s report at the Expert Committee meeting the publication of DDI 3.2 will be delayed due to some technical problems that still need ironing out, primarily regarding identification.

Some information on translating the vocabularies should be published on the website.

We would say that while we are not translating them we would coordinate such an effort, therefore we should be notified of anyone’s intention to produce translations.

Translations should be ultimately submitted to CVG for publication on the Alliance site along with the US-English version. The goal is to have only one authoritative translation for a particular language, to support interoperability.

Definitions for Source Type:

Historical records – Sharon will incorporate in the definition Randy’s suggestions about how to distinguish these from “administrative” ; Meinhard will try to add some examples pertaining to Germany. So will Sanda, if she finds any that would be specific to the US.

Clinical records – move Sanda’s definition to “final”. We think it is a good idea to stay away from references to anything psychological because we would then have to explain the difference between mental illness (covered in clinical records) and psychological assessments, which might not be for medical purposes. Not sure if there is agreement on this however. Might have to raise the question again.

Assessment results – while we like Sharon’s definition, we find that there are other types of assessments (see Sanda’s document with term definitions). Sharon will re-write this definition trying to incorporate financial (tax as example?) assessments, risk assessments, and psychological.

Court/legal proceedings – move to final.

Voting results. – move to final. Exit polls are not really results, but just another type of survey.

Event/transactions – financial transactions included in examples, but still need to talk about computing (clarify how/if they qualify as events?)

Next meeting : Thursday, July 14, at 13:00 GMT.

Hopefully Taina will be able to participate.

**DDI-CVG minutes, May 4, 2011**

Present: Chryssa, Hilde, Randy, Sanda, Sharon, Taina

The first set of CVs should be published by the end of the week (May 6, 2011). When they are on the DDI site, the CVG will have a chance to look at them, browse the related web pages, and then offer comments, feedback and ideas on how to improve the web presentation at our next meeting (June 16) and beyond. Taina has created a folder, called “Web pages and publication” on the Thesauri site related to this topic. It will be useful to consult it in conjunction with what we actually see on the site. We can work at leisure at this task over the summer.

Contributor@role: Instigator will be changed to initiator. Data Manager will be added, with Sanda’s definition.

SourceType definitions: Definitions for Administrative records and Historical records almost finalized in meeting. Sharon will try to provide examples for “Historical records” as well as updated definitions for clinical records, and assessment results (to better reflect the fact that the first are associated with health-care issues, while the second apply rather to an educational or professional environment). Organization proceedings deleted as entry, found similar to administrative records. More examples will be added to administrative records to reflect broadening of scope (Randy).

Next meeting – June 16 (this is a Thursday!) at 13:00 GMT.

**DDI-CVG meeting minutes**

April 13, 2011

Present: Chryssa, Hilde, Taina, Sanda.

Update on publication of first finalized CVs: We hope to publish by the end of April. Sanda and Mary are working with Achim to fix errors, and iron out Web design and display issues. A realistic goal for now would be to publish the CVs with some information on downloads, usage, and contacts for questions and feedback. The Web content can be further developed after publication. One thing we need to sort out are details about maintenance, publishing new CVs, as well as editing and translations. We will ask for input from TIC, and discuss these issues in our future meetings.

When draft pages are set up on “staging” both the CVG and TIC will be asked to review prior to public release.

DataType list: “Character” was deleted per advice from Ornulf and Achim. Sanda will add “normalized string” with definition as requested by Achim, and will then send the list to Achim for review.

We think that when Achim has approved, we can close this list.

CategoryStatistic: closed.

Adding DataManager to Contributor@role: group is still undecided. Sanda will attempt a definition, perhaps with Mary’s help, and will re-submit to the group.

AggregationMethod: Definition for “sum” changed to: “Value obtained as a result of adding two or more numbers”. Definitions for “arithmetic mean” and “count” seem okay. We compare the definitions of “sum” and “mean” for aggregation method with those given to the same terms in the summary statistics list. Participants agree to change the definition of “mean” in summary statistics to the same as the one given in the aggregation method list. This seems reasonable, and it will be done before initial publication of the summary statistics CV, which should happen soon.

Definitions for DataSource: we will have a more detailed discussion next time. Until then, to facilitate a productive discussion, all members will review Sharon’s and Hilde’s definitions, and will add comments/questions/alternative definitions/examples on the same document, using different colors, and signing with their first name.

Meinhard has sent in a few comments, Sanda will add them to Sharon’s table and redistribute both Sharon’s and Hilde’s definitions. Other comments made during the meeting: “voting/election results”: do these include parliament votes? “Administrative records” – ICPSR definition deemed helpful (better?)

In view of the latest comments/clarifications, Randy will attempt definitions for his part of the list.

Next meeting: May 4.

Tentative date for June meeting: June 15. Members are kindly asked to indicate availability.

DDI-CVG minutes

March 2nd, 2011

Atle finds he is too busy to participate in the meetings, but will stay on the email list so that he is up-to-date with our progress. Hilde Orten, who has been working at NSD since 2002, with a focus on cross-national surveys, especially the European Social Survey, has kindly agreed to join the group as Atle’s successor. We are delighted to welcome her, as well as Sharon Bolton from the UKDA and Randy Raphael from the University of Utah, who joined since the beginning of this year.

DataType: keep both STRING and CHAR or just STRING? Meinhard has asked Achim. Achim’s answer: Achim's answer regarding DataType. In short and in English: The distinction between character and string has its historical origin > in the programming language C. Today "character" is simply a string of > one character, i.e. "string" is sufficient! In addition Achim guesses to check titles and definitions with reference to the Primitive Data Types in the XML schema: <http://www.w3.org/TR/xmlschema-2/#built-in-datatypes>

Why did we drop "normalized string"? Achim guesses to include it (if we did not already do so in the meantime) ... He would like to see the final (current) list before publication.

We need more input from IT people on this. In the meantime, Hilde has asked Ornulf - he says: “ In modern computing, there isn't really much differentiation, and I can in fact not recall having dealt with the type "Character" in my 10 years at NSD. As you say, Strings with one character is still a String, so unless there are compelling reasons to differentiate, I wouldn't bother; I'd use String all over.” So, it looks like we can delete Char from the list, but we’ll wait until next meeting to do this.

CategoryStatisticType : new members Sharon, Hilde, Randy will review, and if they don’t have any problems, we will close the list next time.

Contributor@role (Data Manager): Sanda will ask for Mary’s and Wendy’s input on this.

AggregationMethod: Definition of mean: Sanda will change ... The average is calculated by .. to read: The mean is calculated by...

Newer members will carefully review entire list

Not so recent members (nobody is ‘old’ in the group!) will focus on the definitions of sum, arithmetic mean, and count.

CVs for DataSource, ModeOfCollection, Instrument and KindOfData.

History and background are briefly reviewed, all of this is detailed in the “KindOfDataType\_v..” document on the thesauri site, as well as the “DataKind discussion 2008..” document that was distributed to the group.

ModeOfCollection needs to be made into a codevaluetype in the next DDI version (in order to support a CV). Wendy has confirmed that this request has already been entered under bug 253.

Sanda will delete the quantitative data and qualitative data entries from the SourceType list (they will be too confusing, since we have those two types on the KindOfData vocabulary, and they seem more appropriate for describing the finished (data) product, as opposed to the source for the data.

KindOfData attempts to describe the physical format of the (published) data rather than the specific content. Sharon might have a chat with Louise Corti to see if our list seems sufficient for documenting qualitative data for now. (Bearing in mind that there is an on-going effort to add a special section/module to DDI dedicated to describing qualitative data in more detail)

Hilde raises a good point that oftentimes the ModeOfCollection is also described as single vs. mixed. While this is a valid point, the two terms would not be consistent with the rest of the list, so we might need to investigate another way to incorporate this kind of information in DDI. A new subelement Type (on ModeOfCollection)? Sanda will try to look at the DDI 3 schema to identify a possible solution, but this will not be a priority for next time.

Some of the terms included in the current table already have definitions. Sanda will try to identify them and put them in a more structured format so that they are more accessible for review and discussion.

The idea is to create individual spreadsheets for each element from now on (rather than one single document for all) but Achim is still revising the Excel template so we will start using it when he thinks he has a final version (should happen soon – Achim was contacted about this)

Finally, and most importantly, Sharon will try to put together some draft definitions for the terms running from “administrative records” through “natural phenomena”, Hilde has agreed to try to define “processes” and “interactions” (with their “children”) while Randy will attempt to define from “personal communications” through the end of the SourceType list. Trying to define these terms might also help us all decide on the actual content of the list.

- Definition of elements in DDI 3 schema: TIC is agreeable to changing the documentation on certain elements to better reflect the CV contents. We need to consider whether it might be more efficient for our group to propose new definitions, than to let TIC do this, since TIC is too busy doing too many things as it is (we need to talk about this next time).

- Next meeting: April 13 at 13:00 GMT

**DDI-CVG Minutes 2/9/2011**

Present: Chryssa, Meinhard, Randy, Taina, Sanda.

A first set of controlled vocabs has been finalized and submitted for publication at the end of November 2010. They will be published soon, perhaps within a month or so. Achim is taking care of this.

A number of other vocabularies are almost finalized. It should be possible to publish these as soon as DDI 3.2 is released – the new DDI version will include structural changes that are needed before these vocabularies can be published.

Sanda will email Wendy to confirm that these changes are actually being made. These changes are:

**-change configuration of Date to incorporate CV for calendar.**

**-add subelement Role to element Publisher (make CodeValueType)**

**-change attribute Role on Contributor to element (makeCodeValueType)**

**-change AggregationMethod from attribute to element (make CodeValueType)**

**-clarify structure of CategoryStatisticType.**

The group proceeded to review the “almost finalized” vocabularies:

-DataType: we agreed to get rid of the “complex” part in Other/Complex for consistency and because we really can’t provide a definition for Complex. Other will be used for any type that is not included in the list.

Inclusion of Char is discussed. Sanda thinks both Char and String should be included, as they describe different types in computing languages. String can be used to describe Char but in programming defining one or the other involves allocating different sizes of memory.

Fredy not available to comment.

Meinhard will contact Achim to ask him whether it is ok to have both Char and String. Also perhaps ask Achim to give a quick review to entire list.

-CategoryStatisticType:  
Sanda will change “percent of total sum” to “percent of total.

Group needs to review entire list for next time, to make sure everything looks fine and we can approve.

-Date@calendar

List is okay, no changes needed.

-Publisher@role

Reviewed and approved

Contributor@Role:

“Researcher” is too vague, it is agreed to drop that term.

Agreed to have both Data Manager and Data Producer as they may have quite distinct roles.

Chryssa will try to produce better/clearer/more complete definitions to both terms, based on discussions and Randy’s contribution:

**Randy Raphael: The data manager designs (and possibly personally implements) the database (depending on resources available), loads data from case report forms (essentially questionnaires, now more often via electronic data capture), handles queries about data problems, and eventually locks the database prior to analysis. At that point, the data manager might move on to another project. So, I would say that manager and producer can be quite distinct.**

Group will review definitions next time.

-AggregationMethod:

Sanda will look at this list and try to bring some improvements to be discussed next time.

Next meeting: Wednesday March 2nd, 13:00 GMT.

**CVG Agenda 22 September**

1) Finalising the CVs in the Proposed controlled vocabularies 20100914\_most finalised

- Taina will add a description from DDI2 tag library to Category Statistic. (However, this was not helpful, so Sanda provided a definition.)

- New descriptions have been added to all CVs. Taina will send them separately in a short document to the group and ask for a review, giving a week deadline.

- Definitions and some new terms added to LifeCycleEventType (following a decision last time that we would see what GBSPM terms used by statistical institutes might be useful for our CV).

- Definitions added to CommonalityTypeCoded accepted.

- Definition of “Time Series” in TimeMethod was amended to clarify concept.

- CharacterSet: Unicode- 5.1 dropped as it is always encoded, and the encodings should appear as the terms. UFT-32 added. ASCII will be called ASCII since no-one knows its ISO-code and we have tried to use as the code the most commonly used name.

- FiscalYear dropped from Date@Calendar

- Definitions of “Families”, “Household families” and “Households” in AnalysisUnit: As there are many definitions of “Household”, we decided to cite as an example definition one which is often used by Eurostat.

- AnalysisUnit: Decided to keep the term “Events/Processes” rather than change it to “Events” as one reviewer suggested, as we thought we saw a difference. Sanda amends the definition to clarify the difference between events and processes

2) DataType

- ABS recommends that we explain the difference between the CV and W3C data types and asks why all W3C types are not included.

Decision: We add a note explaining that the CVs are based on the W3C data types but we have included only the types relevant for documentation of data.

- Some types are missing from our CV. The missing ones are QName, Notation, NormalizedString, Token, Language, Name, NMToken, NCName, ID, IDRef, Entity, and what seem to be to be subtypes of nonNegativeInteger: unsignedLong, unsignedInt, unsignedShort, unsignedByte. If we add them, we also need definitions.

<http://www.w3.org/TR/xmlschema-2/>

- We decided to include unsignedLong, unsignedInt, unsignedShort, unsignedByte but not the others since they are metalanguage for specifying XML language and not relevant for descriptions of data. They are at the same level as integer.

3) ModeOfDataCollection

- Decided not to add ‘Data integration’ as a term, as it is a LifeCycleEvent rather than data collection method. “Data integration” is already in the LifeCycleEventType CV.

- Changed the definition of “Compilation” to clarify difference to “Data integration”.

4) StudyUnit: KindOfData

- See KindOfDataType\_v6\_20100524\_tj

- KindOfData has an attribute KindOfDataTypeType which provides a description of the types of data described by the KindOfData element. It has a CV “Qualitative”, “Quantitative” and “Mixed” which we asked TIC to implement into the DDI3.

- Now we are trying to make a CV for the element KindOfData itself which is x:string(?)

- Our draft CV for KindOfData was:

Numeric data

Text

Still image (i.e. photos, drawings/graphical representations)

Maps

Sound (or Audio?)

Moving image (mute or sound, film/video, animation)

Programme source code

Other

However, for quantitative data we would then enter the Type attribute as “quantitative” and KindOfData as “numeric data”.

This seems somewhat superfluous and offers no new information. Do we need two lists? We could specify what kind of numeric data (micro, aggregate) but it seems that the information about whether data is micro or aggregate is entered somewhere in the Logical product.

- We discussed this briefly but made no decisions.

5) Next videoconference:

Wednesday 6 October

**DDI CVG MINUTES**

7 July 2010

1) More definitions to CV terms

- ABS would like to see all terms have definitions.

- Most terms have definitions, the LifeCycleEventType and AggregationMethod seem most lacking in this respect.

- Meinhard will provide definitions for AggregationMethod and Sanda and Taina for LifeCycleEventType

2) CategoryStatistic

- Relative Standard Error (RSE) will be added, and defined as:

“The relative standard error is calculated by dividing the standard error (SE) of the estimate by the estimate itself, then multiplying that result by 100. Relative standard error is expressed as a percent of the estimate”.

3) CharacterSet

- Shall we change ASCII and Unicode terms to ISO versions like John W Shepherdson from the UKDA suggests?

ASCII - ISO 14962

Unicode - ISO 8859

- Shall ask Fredy what he thinks. ASCII is very commonly used, IT people do not necessarily know that ISO14962 is ASCII?

4) Date@calendar

- FiscalYear: John W.S says US and UK fiscal year is different so need to distinguish but should we rather add a note that Fiscal Years differ from country to country (cannot distinguish them all…)? Sanda will take a look whether this information can be entered in some other field. Best option might be to drop FiscalYear altogether from the calendar (it is not a calendar).

5) Intended Frequency

- Are the new definitions clearer (Proposed contolled vocs 20100706\_\_most finalised)

- TIC is still reviewing whether changing (N) can be used. What does Fredy say about the use of N, from the IT point of view, is this possible?

6) CommonalityTypeCoded

- Reviewers are always asking why ‘none’ is there, so Sanda will provide definitions for all terms in this hierarchy.

7) LifeCycleEventType

- Can we try to improve the CV with GSBPM and SDMX terms, asks ABS.

- Sanda and Taina will provide definitions and review the terms over the summer. Best option seems to be that while we cannot produce total similarity of terms due to the fact that DDI3 has to cover all types of data, not just quantitative data, we tried to see that Level 1 and 2 terms of Metadata Management of the Generic Statistical Business Model can be described using some LifeCycleEventType term - although the GBSPM term may be more numerous - and try to define the CV terms so that this becomes clear.

8) TimeMethod

- Trend/RepeatedCrossSection and Discrete Time Series definition will be changed to make them clearer (following ABS comments)

- Hierarchy becomes clear in the published version.

9) AnalysisUnit

- More specific definition/term for families, “household families”?

http://www.abs.gov.au/AUSSTATS/abs@.nsf/Latestproducts/93E468A5C48B343ECA25703C0082B0D6?opendocument

- John W S says analysis unit can be combinatory (e.g. individual and household), but will the fact that the element is repeatable solve the problem?

10) ModeOfCollection (in the web folder)

- New definitions are OK.

- Changed terms for interviews and questionnaire main terms are OK.

- Change of ‘Compilation’ to ‘Data integration’? Decision: No, as they are two different methods.

11) DataType (did not have time to discuss this)

- ABS recommends that we add a note saying that these are based on the W3C types. And should explain the difference between the CV and W3C data types.

- So we could add a note and say that we have shorter and easier-to-understand definitions.

- There are some types that are missing. Some of them seem very marginal, subtypes of subtypes but not sure about the others.

<http://www.w3.org/TR/xmlschema-2/>

12) Data Collection: SoftwarePackage CV (this is copied from the prior minutes just as a reminder to Taina of what to answer John W S )

- Fredy has taken a preliminary look.

- Achim asked why we cannot use the same list as for SoftwarePackage.

- We have three options:

1) Separate lists for the two SoftwarePackage CVs

2) One common list used for both

3) One common list but divided in two groups: statistical packages and data collection software (some software, like SPSS, may appear in both)

- We thought option 3 sounds too complicated and option 2 would produce a horribly long list, so though we would stick with option 1 since in many cases data collection software is not the same as the statistical software used.

- It has to be remembered that our software package CVs are not exhaustive but only contain some of the most used ones. We cannot promise to maintain an exhaustive list.

13) Next videoconference in August/September

**DDI CVG MINUTES**

19 May 2010

Present: Fredy, Sanda, Chryssa, Meinhard, Taina

1) No ‘Other’ in three vocabularies

- There is no sense in adding ‘Other’ to CommonalityTypeCoded

- Taina will add ‘Other’ to Publisher: Role and TimeMethod CVs

2) DataType

- Achim asks whether we take definitions for the terms from:

<http://www.w3.org/TR/xmlschema-2/>

- But we already have definitions which are shorter and easier to understand, so we would not like to change them.

3) XPaths

- Achim asks whether XPaths of DDI elements make sense in the documentation of the CVs?

- Considering how hard it has sometimes been to actually find the element we are creating the CV for, we think this information is relevant.

- Even more important might be to have the description of the DDI element in the documentation of the specific CV because the CVs are published separately from the DDI documentation. If we have this information, the XPaths may not be needed.

- Achim said “For each usage of a CV, a section will exist in the documentation of the CV, like DDI2, DDI3, XY (in future). The DDI sections can have links to the related DDI documentation. This way the documentation is not doubled”.

4) Synonyms for terms?

- We decided no synonyms. However, for CESSDA it might be a good idea to introduce at least the methodological CVs into ELSST where synonyms can be added in all languages.

5) KindOfDataType

- While KindOfDataType, ModeOfDataCollection, DataSourceType and Instrument: Type will have to be finalised together, we created a draft for KindOfDataType,

- Looked at QuDEx specification also, but decided to use “Moving image” instead of “Video” since video seems to refer to tapes which have sound whereas moving images can be animations, film etc. with or without sound.

- Draft v1 terms:

Numeric data

Text

Still image (i.e. photos, drawings/graphical representations)

Maps

Sound (or Audio?)

Moving image (mute or sound, film/video, animation)

Programme source code

Other

- Taina will make the draft CV with definitions where needed and send it out for discussion.

6) ModeOfCollection

- Added Aggregation and Physical measurement as terms.

- Sanda will provide a definition for Physical measurement.

- We need to make the captions/terms clearer and more understandable in this CV. Might not be so self-evident to the user if he sees in a data catalogue mode of collections described as “Fixed form”. Taina will make a draft for term changes.

7) DataSource: SourceType

- Sanda has provided draft definitions for processes and interactions etc.

- Can historical records be administrative records from old times? How to differentiate?

- This is an optional element in DDI3

- Did not have time to discuss this, next time.

8) Instrument: (Type)

- Is not repeatable but if different types in a dataset, one repeats the Instrument element (and not just the type).

9) Data Collection: SoftwarePackage CV

- Fredy has taken a preliminary look.

- Achim asked why we cannot use the same list as for SoftwarePackage.

- We have three options:

1) Separate lists for the two SoftwarePackage CVs

2) One common list used for both

3) One common list but divided in two groups: statistical packages and data collection software (some software, like SPSS, may appear in both)

- We thought option 3 sounds too complicated and option 2 would produce a horribly long list, so though we would stick with option 1 since in many cases data collection software is not the same as the statistical software used.

- It has to be remembered that our software package CVs are not exhaustive but only contain some of the most used ones. We cannot promise to maintain an exhaustive list.

10) Next videoconference

Sometime in the first half of July.

**DDI CVG VIDEOCONFERENCE MINUTES**

28 April 2010

Present: Atle, Chryssa, Meinhard, Sanda, Taina

1) KindOfDataType

We have two draft CVs that have overlapping terms and refer to different aspects/facets. One seems to refer to the kind of data source (administrative data, aggregate/census data, measurement/observational data, survey data, personal communication data, public communication data etc.), and the other to the format/type of the data we are disseminating, particularly in the case of qualitative data (still images, moving images, maps, text etc.)

CVG members will try to think of terms to include in the latter list (format/type), particularly if their organization holds non-quantitative data. Meinhard will consult the Historical department of GESIS. However, this will probably need a contribution from the group working on qualitative data.

Sanda will discuss with TIC members where the information contained in the first list could go, in their meeting next week, and do some further thinking on the issue. Some of the information in our lists may be conveyed through other elements, so she will look at the DDI3 as a whole.

2) Data Collection: SoftwarePackage CV

- Fredy will make a draft CV for next time

3) Other issues with TIC

Taina will send an e-mail to TIC members of issues the CVG wishes they would discuss in their meeting. The e-mail text is attached below.

4) Next videoconference

Wednesday 19 May at the usual time.

*Dear TIC members,   
  
the Controlled Vocs Group has now amended the controlled vocabularies, based on the comments received from CESSDA archives and elsewhere.   
We are still working on some vocabularies (KindOfDataType, and question and variable group types, for example).   
  
You will find the latest and pretty final versions of the vocabularies in the document "Proposed controlled vocabularies 20100429".   
The folder 'CVS in spreadsheet, html and genericode with notes' contains the CVs in these formats. (The XPaths may be incorrect.)   
  
These documents are available in the CVG web folder:*[*http://www.fsd.uta.fi/jemma/thesauri/*](http://www.fsd.uta.fi/jemma/thesauri/) *thesauri   
t9oI2vFm   
  
  
The CVG is hoping that you can find time to discuss the following issues at some point of your meeting next week:   
  
- Many of the vocabularies could also function as recommended CVs for corresponding DDI2 elements. If you agree, we suggest that the vocabularies be added to the DDI2 documentation. I have made tentative mappings between DDI2 and DDI3 elements, Sanda will work more on this.   
  
- Language versions of the CVs. We have understood that in documentation it is the code (for a term) that will be used. However, catalogues of data repositories will display the CV terms in the local language, using some kind of tool to transform the code into the local term. It would promote consistency and make the life of IT people and data seekers easier, if everybody used the same translations for terms. For example, if Canadians, the Swiss and the French catalogues would display the same French term for, say, 'self-administered questionnaire' instead of each organisation having their own translation. That would mean that these language versions would have to be produced and published somewhere in DDI3 documentation. (CESSDA will certainly provide versions in several languages, so that may help). Even if using an 'authorised' language version is not obligatory, having them published would increase the likelihood of them being adopted. How would this sound to you?   
  
- The IT people in the CVG group are saying that it would make programming plus harvesting data from different sources much easier if the CVs were implemented in the DDI3 instead of being just recommended vocabularies. The lesson we have learned in CESSDA is that if CVs are not obligatory, they will not be used.   
  
  
- We hope that Sanda may be present when you discuss these issues, as she knows what has been discussed and concluded in the CVG.   
  
- KindOfDataType: we found that there may be a need for two elements, one for data source (for example, administrative data, survey data, personal communication data, public communication data etc.) and other for what kind/format of data the organisation is disseminating (for example, sound, still image, moving image, text). Sanda can fill you in on this issue.   
  
  
Best regards,   
Taina*

**DDI CVG VIDEOCONFERENCE MINUTES**

24 March 2010

Present: Atle, Chryssa, Fredy, Sanda, Taina

1) Historical Date/Calendar terms

- Sanda has told by a web expert that good search interfaces allow searching for terms even when slashes are used, so having “Alexandrian/Coptic” as a term would not be a problem. So we have slashed terms whenever there is more than one possible name of the calendar. Alternative names suggested by GESIS are added to the names

- “Hindus” is changed to “Hindu”. “Bahá'í” will be changed into “Bahai/Badi” (as Badi seems to be alternative naming).

- A definition will be added to ISO week date.

2) TimeMethod

- In the same manner, we will keep the slashed term "Trend/Repeated cross-section" in TimeMethod.

3) DataTypes

- Wendy has recorded as bug ( in Mantis as bug 330) that DefaultDataType (in PhysicalStructure, for both microdata and NCubes) is STRINGType - should be

CodeValueType.

- We will leave the DataType CV as it is as it contains the same terms that appear in built-in datatypes in W3C list Wendy referred to:

<http://www.w3.org/TR/xmlschema-2/#datatype>

4) CharacterSet

- The DDI 3.1 documentation for the element says: If the character set is not known, for example for some proprietary data files, enter "unspecified".

- The term “Unspecified” will be added to the CV with this definition. “Other” is used when the character set is known but not mentioned in the list.

5) Other CVs

- No changes to AnalysisUnit, ContributorType: Role or Publisher: Role.

- We agreed that KindOfDataType has too specific terms which are hard to define. We need to drop some narrower ones, e.g. have only one broader term for Surveys and for Voting data etc. Program source code might be added. Chryssa will make a draft CV for discussion next time.

6) Data Collection: SoftwarePackage CV

- Fredy will make a draft CV for next time

7) DataSource

- Again, we discussed whether we need a CV for DataSource. Since we ended up in a dead alley the last time we tried, I’m copying the conclusion from that time from my notes:

*“As for Data Source, after a series of more or less confusing discussions we seemed to have come to the conclusion that Sanda asks TIC to:*

*Modify element <Source> in <Data Collection> to include just two subelements :<Description> and <Citation>. Content of <source> will be free text, uncontrolled.”*

Situation 20100325: DDI3 still seems to include a SourceType so TIC has either rejected our proposal or has not got round to changing the specification yet.

However, this element refers to published data sources, and may not be relevant for the CV we discussed in connection with KindOfDataType.

8) Next videoconferences

TIC will deal with CV issues in their meeting in May 3-7. They need our finalised CVs by then, and Taina has to produce them in xml and Genericode formats before that – and this will take some time.

So we will have meetings both on 21 April and 28 April.

**DDI CVG VIDEOCONFERENCE**

17 March 2010

Present: Atle, Fredy, Meinhard, Sanda, Taina

1) CommonalityTypeCoded

- Decided to have the terms originally suggested: Identical , Some and None.

- ‘None’ was left to make the CV exhaustive and because we could not be 100% sure that it is never needed for machine actionability.

- Note: There is an element CommonalityWeightType which can be used to indicate the degree of commonality.

2) ResponseUnit

- No changes on the basis of comments received

3) DataType

- Wendy commented that she used the standard W3C list and is not sure it is worthwhile to change that. Taina will ask Wendy where the original list she refers to is to be found in DDI documentation and is it "integer, real, string" or something else.

- And explain to Wendy that CVG would like this CV to be embedded/incorporated into DDI3 to allow machine actionability (ei vielä tehty).

-StorageFormat (in RecordLayout:DataItem:PhysicalLocation) is CodeValueType, no CV incorporated in the schema, in-line documentation says:

" An explicit definition of the data type. This field is necessary in

the case of some numeric data formats where the format definition would

allow real values, but the values are integer values. Allowed values

are: integer (default), real, string."

-DefaultDataType (in PhysicalStructure, for both microdata and NCubes) is STRINGType, no attached CV, in-line documentation says:

"An explicit definition of the data type that is applied to the majority

of the dataitems reducing the amount of repetitive markup required. It

can be overridden at the dataitem level. This field is necessary in the

case of some numeric data formats where the format definition would

allow real values, but the values are integer values. Allowed values

are: integer (default), real, string."

3) Software: Name

- This CV is really for the SoftwarePackage element in Proprietary Module, and not the Name element.

- The CV does not need to have terms for different versions of software packages, as there is an element Version in SoftwareType where this information can be entered.

- DIF, DBase and Statistica will be added to the list

- There should be a different list for the other SofwarePackage element in Data Collection (e.g. Blaise, CSPro, LimeSurvey, in-house solution)

- Everyone will try to find more information about software used for data collection for next time.

- I sent an e-mail to the Association for Survey Computing to ask when their software register will be online. No response yet.

4) SummaryStatisticsCodedType

- Pascal asked whether there should be a term “Percentile”

- There are already many percentile terms in the CV but the list is not exhaustive. However, we do not want to make a hierarchy (i.e.’percentile’ with narrower terms) so decided to adopt an additional term "Other percentile" instead. Term added with the definition: A percentile not covered by any of the other percentile terms.

5) HistoricalDate/@Calendar

- Discussion about whether can use multiple terms with slashes between them: how will these function in search interfaces? If we have "Islamic/Hijri/Muslim" will search systems find Hijri?

- Sanda said she would look further into the terms suggested by GESIS historical data staff member.

6) Codes, captions and search interfaces

- We discussed whether, considering the slash issue above, Time Method should have two separate terms: Event-based and Cohort instead of Cohort/Event-based but concluded that best to have just one (Cohort) with the present definition. If anyone objects, they will inform others before the next videoconference.

- Lively discussion on Trend/Repeated cross-section, which term to use. No conclusions as yet.

- However, Taina remembered after the videoconference that Joachim has said that it will not be the Caption (i.e. Trend/Repeated cross-section) that will be entered in the metadata but what will be entered as the value of the element is the **code** (i.e. Longitudinal.EventbasedOrCohort where Longitudinal comes from the broader term in the hierarchy and EventbasedOrCohort is the code for the narrower term). The caption will only appear in tables and such, he said.

- What does this mean for search? Maybe best discuss this search issue with Achim before we change any captions in TimeMethods, for example. Unfortunately, he is not in the office before 13 April.

- The codes I have given each term can be seen in the document Finalised controlled documents 20091123 with codes (though the CVs are outdates as we have changed them since then, and I have not given new terms codes yet and may need to amend some old one depending on the result of the search issue discussion with Achim).

6. Next videoconference Wednesday 24 March

- Because of Easter holidays, after the conference on 24 March, the next one will be at the end of April.

**DDI CVG VIDEOCONFERENCE**

24 February 2010

Present: Meinhard, Fredy, Sanda, Taina

ModeOfDataCollection reviewed.

We reviewed and amended the structure of ModeOfDataCollection. Dropped some terms and added others. Changed the name of "Self-completed questionnaire" to "Self-administered questionnaire" since this is a more commonly used term, and decided to have two main subgroups under it: Fixed form and Interactive.

Decided to have to two subgroups under Experiments: Laboratory and Field/intervention.

Sanda and Taina will provide new definitions.

Taina will send an e-mail to the DDI Alliance Survey Design and Implementation Working Group at some point, sending them the latest version of ModeOfDataCollection CV for their information and saying that “Invitation mode” sampling (through e-mail, SMS, web banners, web pop-ups, web links etc.) came up in the discussion and we though this is more an issue of sampling than mode of collection as mode refers to how and in what form information is collected, not how respondents are reached.

Next videoconference Wednesday 17 March at the usual time.

**DDI CVG VIDEOCONFERENCE AGENDA 28 October 2009**

Present: Atle, Fredy, Meinhard, Sanda, Taina

1. **LifeCycleEvent CV**

Decided to leave the CV as it is, for pragmatic reasons.

2. **CommonalityTypeCoded**

Finalised the draft CV as *Identical, Some, None.*

3. **Question and variable group types**

- We need to finalise Question response group types (see the document Variable\_Question\_groups\_v5)

- Sanda suggests that we keep “Randomised” group type for questions and maybe not for variables (?) since by the time variables are generated the actual order in which a group of questions were asked is no longer preserved. See her e-mails on 14 and 15 October.

- Any new comments on group types from GESIS?

Decision: Will be discussed next time after Uwe informs us of discussions with WP9 on this.

4. **Household in AnalysisUnit CV**

Possible definitions:

1. A household is a small group of persons who share the same living accommodation, who pool some, or all, of their income and wealth and who consume certain types of goods and services collectively, mainly housing and food. (United Nations, Systems of National Accounts)

*- Does not seem to include one-person households, nor refers to the fact that in some censuses/surveys household is taken to be all people living together.*

2. A household comprises either one person living alone or a group of persons, not necessarily related, living at the same address with common house-keeping - i.e. sharing a meal on most days or sharing a living or sitting room. Persons currently residing in the household, persons temporarily institutionalised (health home, full-time education, military service) or absent for work or travel are included in the definition. (EU)

*- Does not refer to the fact that in some censuses/surveys household is taken to be all people living together.*

3. Either a one-person household, defined as an arrangement in which one person makes provision for his or her own food or other essentials for living without combining with any other person to form part of multi-person household or a multi-person household, defined as a group of two or more persons living together who make common provision for food or other essentials for living. The persons in the group may pool their incomes and have a related or unrelated persons or a combination of persons both related and unrelated. This arrangement exemplifies the

housekeeping concept. In an alternative definition used in many

countries exemplifying the so-called household-dwelling concept, a household consists of all persons living together in a housing unit.

(UN Principles and Recommendations for Population and Housing Censuses)

- *Long but seems to cover everything except the specification of including members temporarily away – but that is probably not relevant here (for AnalysisUnit, it is always household)*

4. "A household is either one person living alone or a group of persons, not necessarily related, living at the same address with common house-keeping - i.e. sharing a meal on most days or sharing a living or sitting room, and pooling some, or all, of their income and wealth (Housekeeping concept). An alternative definition used sometimes is that a household consists of all persons living together in a housing unit (household-dwelling concept). The definition of a household may include persons temporarily institutionalised (health home, full-time education, military service or absent for work or travel."

*- This is mix of the above, written by Taina. However, is not an “official”, so should we go for 3?*

Decision: After a long discussion, we concluded that it is not good to make too specific a definition as each study will have defined ‘household’ in its own way and this may be problematic if DDI3 CV definition is specific and differs from that.

However, need to differentiate from the terms ‘Families’ and ‘Housing units’ in CV, so decided to add to the present definition of household in the CV the words: “sharing all or some of their resources”.

5. **PhysicalDataProduct/ProprietaryModule/ProprietaryRecordLayoutType/r: Software/Name**

- Pascal Heus from TIC says that it would be valuable to try to maintain en exhaustive list of software names but that TIC would certainly also publish a ‘commonly used software’ list.

Decision: CVG will not maintain an exhaustive list – it would be a promise we cannot keep. Taina has added some software names to the list, on the basis on information received from Pascal.

6. **Reusable/Coverage/CoverageType/SpatialCoverage**

- From Ken’s spreadsheet I see that we had felt the following distinction should be documented somewhere in the DDI3:

Whether the coverage/universe is - Sub-national, National, or Cross-National.

- A CV for an optional attribute to SpatialCoverage . Or an attribute suggested to Universe? Neither of these elements has an attribute at the moment, so would need to suggest an attribute to TIC, if so decided

If a combined datafile of, say, ISSP (=surveys in different countries): cross-national.

If a datafile of, say, Greek data of ISSP survey: national.

If a survey of customs officers in Finland and Russia: cross-national.

If a survey of inhabitants in Frankfurt-am-Main: subnational.

Decision: There may be studies the coverage of which is both sub-national and cross-national (e.g. national in some countries and a city/region in others), so the terms discussed are non-exclusive. After a discussion we did not find a good element to suggest this CV as attribute, either (SpatialCoverage has two mandatory elements: upper-most level and lower-most level). So decided to drop this idea altogether.

7. **Potential CVs**

Mari suggests we look at the following elements and consider whether they need a CV or not:

attribute type in r:otherMaterial (probably difficult)

d:DataCollectio/Instrument/d:type

attribute type in l:VariableScheme/Variable/Label (reusable, so difficult)

l:VariableScheme/Variable/Representation/Role

p: PhysicalDataProduct/PhysicalStructure/Format

Decision: Taina will take a look at some point.

8. **Next videoconference**

Wednesday 18 November at the usual time

**DDI CVG VIDEOCONFERENCE MINUTES 14 October 2009**

Present: Atle, Christina, Chryssa, Fredy, Meinhard, Sanda, Taina

1. **Reusable: Role CV for element “Contributor”**

CV for Role (Contributor):

Data collector

(Person, corporate body, or agency responsible for collecting data via instrument)

Data producer

(Person, corporate body, or agency responsible for producing data after collection)

Depositor

(Person, corporate body, or agency responsible for depositing object with publisher)

Metadata producer

(Person, corporate body, or agency primarily responsible for compiling and maintaining

the original description of a metadata set)

Research Instigator

(Research instigator other than a funder or the creator. For example, Parliament

expressing a wish to have a phenomena studied.)

Other

2. **Reusable:** **Publisher, attributes.**

- We felt that there should be a way to distinguish between Publisher/Distributor within the Publisher element, citing a following example: Eurostat publishes a dataset which is distributed by national statistical institutes, with national data archives providing metadata about it (without distributing the dataset themselves).

- DDI CVG is going to suggest to TIC that there should be a Role also for the element Publisher (to refine the element), with a CV. Role should be optional, and the Publisher element repeatable. This proposal will be added as a note to our draft CV.

Draft Role CV for element Publisher:

Publisher

(Person or organization responsible for making the resource available in its present form)

Distributor

(Person or organization responsible for distributing the resource to the public or to a

specific population. If not publisher as well, the distributor distributes the resource in an

unaltered form.)

Other

Publisher and Distributor are refinements (=roles) to Publisher element also in Dublin Core.

3. **“Other” in CVs**

All other CVs will have “Other” except CommonalityTypeCoded and TimeMethod.

4. **Question and variable types**

- Decided to drop Iteration (meaning unclear) and Grid (can be documented through other elements). See the latest version of the document in the web folder.

- There is some confusion about how to use ELSST thesaurus terms for documentation. According to TIC, concepts are more specific than terms in a thesaurus (e.g. in thesaurus we have “television viewing”. However, this term covers several concepts like “time of television viewing in a day/week”, “at what time of day watches television”, “what kind of programs” “viewing of TV news”, “viewing of current affairs programs” and so on). In DDI3, variables are linked to concepts, and to concepts schemes, and there is no keyword/subject for variables. Thesaurus terms can be used for variables only through grouping variables for Subject or Keyword via group types. So CESSDA may have a problem here for comparison and question bank unless decides to use ELSST for Concept.

- Randomised: Sanda will see whether this can be documented in some other way. To enquiry, Wendy answered:

“Create a Sequence including ControlConstructReferences to each of the question constructs involved in the randomization. Then use ConstructSequence/QuestionSequenceType.

Options are: InOrderOfAppearance/Random/Rotate/Other

If Other is used explain the sequence to follow in AlternateSequenceType.

Note that in the final version of 3.1 this will be optional (was inadvertaintely listed as required in the review copy).”

See also Sanda’s e-mails on 15 October.

- Chryssa has provided info on projective types for Question Response types (see document: Question response types \_ projective 20091010 in web folder). Will be discussed the next time.

5. **Next videoconference**: Wednesday 28 October 2009 at 13:00 UK time.

**DDI CVG VIDEOCONFERENCE MINUTES, 23 SEMPTEMBER 2009**

Present: Atle, Chryssa, Fredy, Taina

1. “Other” in our proposed CVs

Decision: Decided to have “Other” element in relevant CVs. Taina will review the CVs to see which should have Other and which not, and present findings at next conference. It is supposed most CVs will have “Other”.

2. Intended Frequency type

Amended CV was approved with one addition: Hpp(N) for hours.

3. LifeCycleEvent CV

We decided last time that secondary analysis is not in itself a life cycle event meant by this CV whereas reports are. So in the new version produced by Sanda, “Primary analysis reports” and “Secondary analysis reports” have been dropped, and “Data analysis reports” used instead. No other changes. See the amended CV in the web folder.

Decision: We made one amendment to Sanda’s draft CV: Data analysis reports was placed above New Version Release/Publication.

Atle will review this CV and think whether the Dissemination (delivering resource for further use) part of the process needs any new terms to be added, and e-mail the group before the next videoconference.

4. AnalysisUnit CV

Decision: Participants felt that Unit of Analysis elements should really be Unit of Observation elements to avoid ambiguity (for example, if organizations are units of analysis but information about them is asked from individuals, what to fill in the elements). So while the CV is OK, Taina adds a note of this issue for TIC.

5. Reusable: Role CV.

Ken’s spreadsheet for finalised CVs (2008?) contained a CV for Role. It seems this has not been discussed by TIC yet, though Ken may have submitted the spreadsheet.

About the Role element element : "Role" on "contributor" NOT "role" on "FundingInformationType" which has a different function.

**(Contributor:** The name of a contributing author or creator, who worked in support of the primary creator given above). Role is String type element

**Draft CV:**

Data collector

(Person, corporate body, or agency responsible for collecting data via instrument)

Data producer

(Person, corporate body, or agency responsible for producing data after collection)

Depositor

(Person, corporate body, or agency responsible for depositing object with publisher)

**Later additions to the CV:**

Distributor

Wendy’s comment: According to Dublin Core this should be listed in publisher as refinements of Publisher, and not here as a publisher or distributor may or may not also be a “contributor” to the resource. TIC wants to make sure these elements map cleanly "The entity responsible for making the resource available."

Copyright holder

Wendy says this should be listed under Copyright. It maps to Dublin Core: rights.

Research Instigator

(Research instigator other than a funder or the creator. For example, Parliament

expressing a wish to gain information about a phenomena.)

Other

However, please note that there are different elements in DDI3 for:

**Creator** : Person, corporate body, or agency responsible for the substantive and intellectual content of the described object.

**Publisher**: Person or organization responsible for making the resource available in its present form

Wendy’s comment about the Distributor/Publisher issue:

“First be aware that we will be expanding the use of native dublin core in DDI to include DCTERMS which allows for more detail and definition of the elements, plus some that are used in archives. We simply hadn't nailed down the appropriate scheme and reference mechanism for 3.1.   
Distributor - according to Dublin Core this should be listed in publisher as refinements of Publisher as a publisher or distributor may or may not also be a "contributor" to the resource. We want to make sure these elements map cleanly   
"The entity responsible for making the resource available."

Decision: Will finalise the CV next time. Decided first 3 terms are OK. Regarding later additions:

ask Wendy more info regarding Distributor/Publisher, remove the Copyright holder and keep the Research Instigator.

6. Next videoconference: 14 October 2009, 13:00 UK time

Where we will discuss Question and Variable Groups (with comments by Meinhard, Sanda’s added info of what is covered by other elements in DDI3 and Chryssa will look closer at Projective type, maybe with definitions)

**MINUTES FOR DDI CVG VIDEOCONFERENCE 9 SEMPTEMBER 2009**

Present: Atle, Meinhard, Chryssa, Taina, Christina, Sanda

1.“Other” in our proposed CVs

Two options:

1) Add Other to all the lists

2) Remove all instances of "Other" and recommend that if people want to use a term that's not on our lists (or other lists), they should just include it as content and not reference anything? Because CodeValueType provides the possibility of using a term that is not on a CV list. In that case no list would be referenced in the attributes.

If we choose this option, we need to ask TIC to remove the *OtherValue* attribute from the DDI specification.

Decision: Decided to go for option 2. We wondered whether there are other lists (than the DDI3 lists) which include “Other”. However, felt the same solution should be recommended for those lists as well.

Sanda will ask Wendy how to submit the request to remove the *OtherValue* attribute*.*

*As for CESSDA, we will provide some kind of CV revision handling procedure or maybe have an application that picks up what is used as content instead of the CV terms, and then revise those.*

2. IntendedFrequency type

Taina’s question: We have Ypp(N) in this CV (Years per period, where N indicates the number of years.

Can’t we make the same for minute, seconds, and milliseconds?

Mpp(N) – Minutes per period, where n indicates the number of minutes.

Spp(N) – Seconds per period, where n indicates the number of seconds

Mspp(N) – Milliseconds per period, where n indicates the number of milliseconds

Decision: Yes, will add those, and in fact thought the whole vocabulary should work like this.

Taina will revise the list accordingly by the next meeting.

3. LifeCycleEvent CV

In the previous meeting, we decided to change “Secondary Analysis reports” to “Secondary analysis” with narrower terms under it. Sanda suggested at least Data discovery, Subsetting, Reports) and wanted to hear suggestions from Atle and Meinhard what other narrower terms to add.

UKDA’s former director defined *secondary analysis* as: Any further analysis of an existing data-set which presents interpretations, conclusions or knowledge additional to, or different from, those presented in the first report on the data collection and its results. Secondary analysis is carried out a later point in time than primary analysis, usually by researchers who did not themselves collect the data.

Source: Catherine Hakim: Secondary Analysis and the Relationship between Official and Academic Social Research. Sociology, Vol 16, NO. 1, 12-28 (1982)

Decision: We revised the decision of the last meeting, and decided to have only “analysis reports” and remove “Primary analysis reports” and “Secondary analysis reports” from the CV. In fact, after a long discussion, we thought that secondary analysis is not in itself a life cycle event meant by this CV whereas reports are.

Discussed whether one can use this element for tracking the use of a dataset for secondary analysis, and concluded no, this is not what the element is meant for.

4. Question response types

Approved Chryssa’s proposed response types. For programmers of software, Close-ended b and d types are the same and will be treated as such. We decided to need both types (b and d) as people will probably want to group Attitude scale responses for other reasons.

Chryssa will produce a definition for the Projective type.

5. Question and variable types

- Minutes of the last meeting said: Meinhard has promised to comment the latest version (see Web Folder). Sanda said she’d do more research on certain types which may be covered by other elements in DDI3. After all, the original types lists we started with were based on the DDI2 list.

Decision: Did not have time to discuss this. Meinhard will send his comments by e-mail before Sanda goes on holiday, and we will discuss this CV only after she comes back.

6. Proposal for systematic testing of CVs by CESSDA archives

Taina will present a test plan in the CESSDA Expert seminar in Ljubljana. It is in the interests of the archives to test the CVs as many of the CVs will be part of the metadata model CESSDA-ERIC is moving towards.

7. Next videoconference: 23 September 13:00 UK time.

**DDI CVG VIDEOCONFERENCE** 19 AUGUST 2009

Present: Chryssa, Fredy, Sanda, Taina

1. Pascal's suggestions (in cursive) for CVs

LifeCycleEvent:

*- Add "Other".*

*- I would suggest using "Secondary Analysis" (drop 'report' suffix).*

*- How well does this cover the case of a researcher doing it's own data*

*collection or performing secondary analysis (Data Discovery,*

*Reshape/subset, etc.)?*

Decision: Decided could add “Other”. However, whether we will actually add “Other” depends on the decision made on “Other” in CVs on the whole.

Decided to drop ‘report’ from “Secondary Analysis” and add narrower terms under it (e.g. Data discovery, Subsetting, Reports).

We seemed to remember that Atle had defined Secondary analysis in some meeting or e-mail. We would like to hear/read that definition again. We would also like suggestions from Atle and Meinhard of what kind narrower terms to add.

FrequencyType

*- I would include smaller units in such as minute, seconds and*

*milliseconds to support scientific datasets*

Decision: Seemed too difficult, considering how we had constructed the CV (hourly, weekly monthly etc).

Still, on second thoughts, Taina asks: We have Ypp(N) in the CV (Years per period, where N indicates the number of years. Can’t we make the same for minute, seconds, and milliseconds?

Mpp(N) – Minutes per period, where n indicates the number of minutes.

Spp(N) – Seconds per period, where n indicates the number of seconds

Mspp(N) – Milliseconds per period, where n indicates the number of milliseconds.

AggregationMethod:

*- suggest adding "Other"*

Decision: “Other” added (if any Others will be used).

r: Software:

*- Pascal says this is missing several packages*

*- "For the vocabulary, we're looking for generic package name (like spss,*

*sas, stata) as the version specific information in under the Software*

*element"*

*- at least the following are not on our CV:*

*Blaise*

*CASES*

*CSPro*

*Fame*

*Limdep*

*Matlab*

*Nesstar*

*Shazam*

*Troll*

*TSP*

Decision: Sanda found the element this list is meant for:

“PhysicalDataProduct\_Proprietary/ProprietaryRecordLayout/Software:Name

[Reference] (r:Software)

Software used by this record layout.

[Reference] (Name) - min. 0 - max. unbounded

The name of the software package, including its producer

These are part of a separate module -PhysicalDataProduct\_Proprietary- which may be used instead of the regular PhysicalDataProduct to describe data that are in a Proprietary format (for instance, in an SPSS portable file, rather than ASCII data)”

Sanda points out that the element is not constructed as a CodeValueType in DDI3. So even though we created a CV, it is not necessary to take the software from a list.

Taina's question:

What was the intention with our list? Just cite the most common ones as examples and to harmonise naming? Complete list seems not possible and would be too difficult

to maintain? Decision next time.

SummaryStatistics:

*Add Percentile?*

Sanda’s e-mail:

“I just talked to one of our researchers and I have confirmation that the word "Percentile" as such has no meaning as a summary statistic.

Rather, a "certain" percentile will indicate the percent of a distribution that is equal to, or below it, therefore a number from 1 to 100 has to be added to the word percentile in order for this to make sense as a statistic - i.e. 4th percentile, 35th percentile, etc.

-How do we want to account for this in a CV? Don't know, food for thought. Do we want to include 100 different "rows"? Is there a way to just include one entry suggesting that any number from 1 to 100 (including decimals of course) may be attached to the word "percentile"

- like, "nth percentile"? (Atle? Meinhard? Please comment)”

2. Question and variable types

- Meinhard has promised to comment the latest version (see Web Folder).

- Sanda will do more research on certain types which may be covered by other elements in DDI3. After all, the original types lists we started with were based on the DDI2 list.

3. Question response types

- Chryssa will work more on this, and provide definitions. We decided that we need to have three types in the first section:1) dichotomous (only two possible answer alternatives, fixed, one choice), 2) fixed – one choice out of many, and 3) fixed – multiple choice. The terms may not necessarily be these, though.

4. Next videoconference: 9 September 13:00 UK time.

- - - - - - - - - - - - - - - - - -

Developments after the videoconference:

**DISCUSSION ABOUT OTHER IN CVs**

Copy of *Sanda’s e*-mail:

“I looked at the schema - thanks, Fredy - and here's my interpretation of the intended usage of "Other":

-"Other" is actually included as an entry in the **\*external list\***

-The attributes of CodeValueType **\*reference\*** the external list (name,

agency, ID, URN) and **\*if\*** locally a certain term is used in place of

"Other" that particular term is entered in the otherValue attribute. For example, the external FrequencyType list includes "Other". I'm using this list to mark up in DDI a study where the frequency is "every 10 minutes". Then I would enter "Every 10 minutes" in the otherValue attribute, but that would not replace the "Other" entry in the main, external list. Somebody else would use "every 25 seconds" instead of "other", then again they would enter that in the otherValue attribute, while the main list would always include only "Other".

Do we need to check with Achim or Wendy (not available this week) if my interpretation is correct?

-If it is correct, then I think we need to go ahead and include "Other" in all of our lists except: CommonalityTypeCoded, and TimeMethod (from latest spreadsheet)”

*Fredy to Sanda:*

Regarding (2) - "other" attribute   
I think you are right. I interpreted it in the same way, but didn't remember it today.   
  
But i don't like so much to clutter all the lists with 'other' values. I think it complicates usage of ddi in software design, and introduces an additional dimension (like the 'other-what' structures in SPSS, which you have to analyze separately). It opens a backdoor which goes against the basic idea of CVs. And finally, OtherValue is probably not language-aware.   
  
If we are interested to catch a maximum of metainformation, it would make sense. From the point of view of standardization of information, it would be a step back.   
  
So do we really need all these 'other' values ?

*Atle:*

did not participate yesterday, but I am trying to follow the

discussion and I just want to state my agreement with Fredy's viewpoint. "Other"-possibilities should if possible be avoided,

that is a principle.

*Sanda to Fredy:*

I totally agree with your comments below.

So here's what I think:

-Adding "Other" to the list was a way for us to avoid trying to provide exhaustive lists, which would have been very hard to assemble, difficult to use, and to maintain.

Plus, philosophically, can we ever be sure that a list is exhaustive?

-Still, CodeValueType also provides for the possibility of using a term that is not on a CV list. (Then, no list would be referenced in the attributes)

Should we then remove all instances of "Other" and recommend that if

people want to use a term that's not on our lists (or other lists), they should just include it as content and not reference anything?

IN WHICH CASE THE otherValue ATTRIBUTE ALSO NEEDS TO BE REMOVED FROM THE ddi SPECIFICATION.

This seems a good solution to me.

We have to find out if everybody agrees.

*Fredy*:

Interesting philosophical question below :-P   .   
  
Yes, I agree with Sanda: i.e. to use simple terms, if a CV isn't appropriate. (This results in almost the same information as if we would use a CV entry 'other' and add text to 'OtherValue')   
  
Does this setting fit to user needs?   
  
YES, if a DDI-3 aware application supports such simple terms in addition to a CV. If not, users might be stuck, if a list does not cover 100% of its scope. Thus, applications should support simple terms (same is true if we would use CV-other in combination with OtherValue)   
  
Having this said, it might be usefull to inform the TIC, because it concerns implementation.   
  
In this context (Atle addressed the same?), it becomes important to think about CV revision handling, and allow users to suggest additions, as well as having an agency who takes care about this.   
  
  
P.S. In case we decide to keep 'other' in CVs, applications should be able to identify the 'other' element (an flag or similar)   
  
  
Fredy

*Atle:*

If we need to use "other", then we have a situation that could have been

solved by

a) "Other"

b) Leave it blank ?

c) Include (invent) content

a) or b) could be coupled with c) ?

Should this be evaluated in relation to revision potential/cycluses ?

How would Taina like to get info for revision/maintenance ?

*Sanda*:

I don't think we need to use "Other."

That was the point of my proposal below.

*Taina*:

Yes, we did add "Other" to avoid having to take into account every

possibility on earth (which is not possible anyway).

But "Other"-additions seems to cause trouble for IT design. And the

original purpose of the controlled vocs was to increase machine

actionability. So adding "Other" would seem contrary to one of the main purposes of CVs.

So personally I can live without "Others", particularly after Sanda

explained how people can work around not finding a good term in a CV.

We must make a decision on this at the next videoconference on 9

September, so send e-mails of your views on this issue.

<<Should this be evaluated in relation to revision potential/cycluses ?

<<< How would Taina like to get info for revision/maintenance?

Not sure what you mean, Atle.

*Atle*:

I only mean that if we let users "work around", we will potentially after some time have an idea about what is missing, we get info that potentially could lead to additions in a revision process.   
How and where we gather and store such info is another question. But since WP4 should think on maintenance, it would be some kind of a maintenance question.

*Taina*:

Good point, Atle. When CESSDA begins to use agreed CVs, maybe the portal could produce logs of text entered instead of using a CV term? Or does anyone have any other suggestions?

Irena has sent me a mail saying she would like to have Controlled vocs discussed in the CESSDA Expert Seminar.

I’m going to suggest that CESSDA members test the proposed vocabularies systematically, either by using them in documentation (in which case they have to be translated and this would be feasible only if they do not already use another, own CV for that field) or by testing their functionality by randomly selecting a part of their data and seeing whether they would function for those datasets. In the Expert seminar, I will give a quick overview of the proposed CV, then present a test plan with timetable for testing, and ask archives to name a contact person for this testing.

**DDI CVG VIDEOCONFERENCE 29 JULY 2009**

**Present: Chryssa, Fredy, Sanda and Taina**

1. Variable and Question Group Types CV

- Sanda will make a new draft of CVs and a list of group types that we decided to drop with explanations why.

- We established that unlike many other CVs, group types are not mutually exclusive.

- We thought that the group type Response should contain subgroups (suggested syntax Reponse: single choice) to be useful and machine actionable. Chryssa will look into possible response types and make a draft.

2. Would the same ResponseUnit CV for survey type data go for both:

LogicalProduct/VariableScheme/Variable/ResponseUnit

LogicalProduct/VariableScheme/NCube/ResponseUnit

- In short: yes.

3. Version control system for CVG

- Taina will enquire from FSD’s IT people if it would be possible to put up a web folder containing the latest versions of documents.

4. Sample markup of different types of questions on DDI3 pages

- Was considered to be a good idea

- DDI CVG should suggest use cases for the Daghstuhl seminar, e.g. how to tag in DDI3

- multiple item questions

- repeated questions (how to tag them in ControlConstruct)

- grouping a question/variable to multiple groups (for now, Fredy will try to do

this with a variable and see)

- etc.

5. Finalised CVs spreadsheet

- Achim confirmed receiving this and will get back to us regarding procedure.

6. Next videoconference: Wednesday 19 August

**MINUTES FOR DDI CVG VIDEOCONFERENCE 15 July 2009**

Present: Taina, Fredy, Sanda

1. Finalised controlled vocs spreadsheet

- Fredy will re-add the DateTime Format CV, and send the spreadsheet to Taina (though the need to re-add the Format CV may have been a misunderstanding)

- Taina will send the spreadsheet to Joachim asking him to submit the CVs for DDI Alliance Expert Group Review

- TIC would like us to add a code to all CV terms. We decided to add the codes only after the review (as it may result in changes to CVs)

2. Question/Variable Group Types

- The conclusion of the discussion: we will go on preparing the CVs for both group types even though Question Groups as GroupTypeCodedType is not included in the present version DDI3 or in the forthcoming DDI 3.1. as there had been some technical problems with the element.

- Wendy reported assuming that question group types would be in the next DDI update 3.2 or 3.1.1.

- We thought Grid should be a group type in both CVs. However, it is not clear how these would be grouped in practice (what do we group, main question + its subquestions, or all multiple item main questions in the study or what). Sanda will do mark-ups of multiple item questions, and Fredy will send her examples of them.

3. Response Types in DDI3

- Taina will contact Wendy and ask if one can tag response types in DDI3 (note, different from Response Domain in Data Collection Module), or should we consider adding 'Response type' as group type.

Examples of response types:

Open-ended text questions - accept only open-ended text responses

Single choice questions – only one response accepted from a predetermined group of choices

Multiple choice questions – more than one response accepted from a predetermined group of choices

Scale/rank choice questions

Image choice questions??

Numeric entry questions – accept only numeric numbers as responses

Wendy said consider adding Response type as Group type as not possible to tag all different types in DDI3, though for some response types can be tagged by other elements like Response Domain.

4. Would the same ResponseUnit CV for survey type data go for both:

LogicalProduct/VariableScheme/Variable/ResponseUnit

LogicalProduct/VariableScheme/NCube/ResponseUnit

- did not have time to discuss this

5. Date for the next videoconference: Wednesday 29 July at 13:00 UK time.

**DDI CVG Videoconference** 29 April 2009

Present: Ken, Sanda, Meinhard, Taina, Fredy

Concepts, subjects, keywords and variable/question grouping

- Ken will have a lot of questions to ask from Wendy, for example, can Topical coverage information be referred to, do Subject and Keyword have an ID, cannot structured thesauri terms be used for Keyword, how to build a question bank in practice, using which elements (if use ELSST terms for finding comparative questions), will there be GroupTypeCodedType for both questions and variables?

Data Collection: Time Methods

- Ken and Meinhard will ask their colleagues about “historical”, so that we can finalise the CV. **Deadline**: Thursday 7 May.

VariableGroup and QuestionGroup CVs

- Homework for everyone: as Wendy’s document suggest that there will be GroupType for both questions and variables, these need two separate CVs. Everyone will to try to draft these two lists. Meinhard’s draft CV for VariableGroups can be used as a starting point – should some of these terms be only in the Question Group CV or Variable Group CV, some in both? **Deadline**: Monday 11 May.

ResponseUnit

- Participants were satisfied with the terms Sanda proposed but new definitions are needed. It was agreed not to have an example list for Proxy.

- Definition of Proxy may include something like: Proxy replaces the sampling unit. If a respondent who **is** the sampling unit is asked mainly about another person/other persons, the unit of response is Informant. (Sanda, feel free to make changes, as this is only a note of the content of our discussion). **Deadline**: Monday 11 May.

Summary and Category Statistics

- Meinhard will provide definitions for the next time. **Deadline**: Monday 11 May.

Reusable/GeographicLevels/GeographicStructureScheme/GeographicStructure/

Geography/Level/Code:

- No time to look at the examples for UK and Germany that Ken and Meinhard prepared. Next time.

Next meeting: Wednesday 13 May at the usual time.

**DDI CVG Videoconference** 1 April 2009

Present: Ken, Sanda, Meinhardt, Taina, Chryssa

1. LogicalProduct/VariablesScheme/VariableGroup/GroupType:

- Meinhardt has sent the group possible additions to the group types mentioned in DDI 2 Tag Library.

- However, since Wendy’s document about changes in DDI seem to be relevant to this, we decided to discuss the whole issue next time when everyone has had time to read it.

- Homework for next time: Everyone will read Wendy’s document ConceptNameEtc bug correction. Meinhard will e-mail the group his comments on the document (deadline for his comments 8 April), and hopefully this will lead to an e-mail discussion before the meeting. Does DDI really need so many ways of grouping and describing concepts/subjects? Where can CESSDA put ELSST thesaurus terms to produce comparative variables according to this version of DDI 3.0?

2. Data Collection: Time Methods

- Sanda will produce a draft CV with definitions, deadline 9 April. Taina will send her definitions from a book and others provided helpful web site links. Particularly the difference between Trend study and Time series was discussed, and also if we have both these terms, do we really need ‘Cross-sectional regular study’ at all?

3. PhysicalDataProduct/ProprietaryModule/ProprietaryRecordLayoutType

- ProprietaryInfo or Software element???

- CV of Statistical Packages was finalised by adding Other to Ken’s list.

4. Reusable/DateTimeRepresentation/@format

- Was finalised in the form presented by Taina in the document

Locations with objects of type xs string.

- QUESTION 1 TO TIC: Should Week Date be added to the DateType list?

QUESTION 2 TO TIC: DataType CV for various elements (e.g. StorageType) has gYearMonth, gYear, gMonthDay and gMonth, referring to Gregorian calender. Should DateTypeCodeType also have 'g' in the beginning instead of just YearMonth, Year, MonthDay etc?

5. Summary and Category Statistics

- No feedback yet on these.

6. Reusable/GeographicLevels/GeographicStructureScheme/GeographicStructure/

Geography/Level/Code:

- Describes a single level in a geographical breakdown (eg, country, province, etc.)

- Preliminary discussion of Ken’s draft. Homework: CVG members show this to experts in their organization, to be discussed next time (to be found in Ken’s e-mail 1.4. with the misleading subject Variable Groups).

TAINA’s CONTRIBUTION HERE, as will not attend the next DDI CVG meeting:

There should be a term for first/second level administrative divisions which in Sweden is kommun (municipality), in Finland kunta (municipality, may be urban or rural, so no Cities or such: all are municipalities), in Denmark kommun (municipality), in Norway kommun (municipality, so even Oslo is a municipality). So, as far as I know, local elections in these countries mean electing Council members for the municipality. Suggestion: Add the term ‘Municipality’ to the list.

The administrative divisions seem to be very confusing and challenging, though. Browsing Wikipedia I only got more confused. In some countries first level administrative division seemed to provinces and in other they were second level division.

7. Next meeting: Wednesday 15 April at the usual time.

**DDI CVG Videoconference** 18 March 2009

Present: Ken, Sanda, Meinhardt, Taina, Fredy

1. We have now covered most entries in Wendy's CodeValueType list.

ProprietaryInfo

- No CV

GenericOutputFormat

- Will be discussed with other Format element later

Summary and Category Statistics

- Meinhardt has sent these for reviews, we will wait for feedback.

Software names and Geography name types:

- Ken will provide draft CV for the next time.

Variable Group Types

- Sanda will ask Wendy about questions with multiple responses.

- Meinhardt will send an e-mail to the group about possible additions to the list.

Analysis Unit:

- We were happy with our draft CV but hoped for a clearer definition for the element in the DDI documentation. Ken has contacted Wendy about this and will enquire what the situation is now.

Wendy has not yet had time to prepare a new version of the Best Practice document on Concept/Keywords/Subjects, so topical coverage issues were left for later.

2. DataCollection/Methodology/TimeMethod

- Mary Vardigan wished for us to see whether we could produce a CV for this element.

- Ken will see what lists there are already for this, and what ELSST says (but I’m making his task easier by copying two lists here).

CESSDA CV proposal a few years back:

Cross-sectional, repeated cross-sectional, panel, cohort, time-series, historiographical

ELSST

: :..TIME METHODS (RESEARCH)

¦ :..CROSS-SECTIONAL STUDY

¦ :..CROSS-SECTIONAL AD HOC FOLLOW-UP STUDY

¦ :..CROSS-SECTIONAL ONE-TIME STUDY

¦ :..CROSS-SECTIONAL REGULAR STUDY

¦ :..LONGITUDINAL STUDY

¦ :..COHORT STUDY

¦ :..PANEL STUDY

¦ :..TIME SERIES

3. Location with objects of type xs string

Archive/OrganizationScheme/Organization/Location/@name

Archive/OrganizationScheme/Individual/Location/@name

-No CV

Archive/OrganizationScheme/Individual/Address/@type

Archive/OrganizationScheme/Organization/Address/@type

-Discussion deferred to next time. Will we try to produce a CV or suggest that the most common types are cited as examples in DDI 3.0 documentation?

LogicalProduct/VariableScheme/Variable/ResponseUnit

- Conclusion of discussion: the CV mentioned in the DDI documentation (Respondent, proxy, interviewer) is OK for surveys.

- Sanda will e-mail Wendy that we think that ResponseUnit should only be used for survey-type data, and suggest that there should be a DataSource element at variable level. At the moment DataSource is only for the whole dataset.

LogicalProduct/VariableScheme/Variable/Representation/Role

- After a lively discussion of pros and cons of CV for this element, Ken volunteered to produce a draft list of the most common types. Doubts were expressed whether a comprehensive list could be produced but a list of most common types may turn out to be useful – or not. We will see.

Reusable/DateTimeRepresentation/@format

-Taina asked people to look at her draft list for next time.

4. WhereWeStand

- Ken will look at the possibility to keep a Wiki-type site where we could keep record of all decisions and agreed-upon CVs. We have been sending a lot of e-mails and documents around and keeping track is proving to be difficult.

5. Next meeting: Wednesday 1 April at the usual time.

**DDI CVG Videoconference** 4 March 2009

Present: Ken, Chryssa, Reiner, Sanda, Meinhardt, Taina, Atle, Fredy

VariableGroup/GroupType:

Everyone was asked (again) to consult the experts within their organization as to whether the list cited in DDI 2.1 tag library should be extended and how variable grouping is used. Wendy had said that this element definitely needs a CV. Discussion next time.

SummaryStatistics:

CategoryStatistics:

Reiner and Meinhard will send the latest version to the group who will try to find experts to review the CVs.

(As for SummaryStatistics, in the previous videoconference it was decided that Ken will send the list to Wendy and ask her to find an expert. Don’t know where we are re this)

Reusable/Representation/RecommendedDataType:

The CV will be the same as for StorageType.

However, we did not discuss whether GenericOutputFormat needs a CV.

Reusable/GeographicLevels/GeographicStructureScheme/GeographicStructure

/Geography/Level/Code:

Reusable/GeograpiciLocation/Values/GeographyValue/GeographyCode/Value:

Ken will look to see whether EuroStat has anything relevant regarding this.

The decision at the meeting two weeks before: Ken reviews whether the INSPIRE standard (Infrastructure for Spatial Information in the European Community) has any relevance for this, at least at CESSDA level.

<http://inspire.jrc.ec.europa.eu/> (guess not since Ken did not mention it).

Software names:

Ken will clean up the list.

Sanda will find the element where to put the CV. She sent an e-mail later saying that it is probably:

PhysicalDataProduct – Proprietary/ProprietaryRecordLayout/r: Software

However, we did not discuss then why Wendy had included ProprietaryInfo to the list needing review CV-wise. Does ProprietaryInfo need a CV and what kind?

Next meeting: Wednesday 18 March 2009 at 13:00 UK time.

**DDI CVG Videoconference** 18 February 2009

Present: Ken, Taina, Chryssa, Reiner, Sanda, Meinhardt

STUDYUNIT/KINDOFDATA EXTENDS R:CODEVALUETYPE BY ADDING @TYPE:

@type is KindOfDataTypeType

Qualitative

Quantitative

Mixed

Qualitative, Quantitative and Mixed are attributes, and our DataKind CV is used for content, so no problem, as far as I could understand.

ModeOfCollection

Taina will adapt the definition for coding as agreed, and make an agreed-on suggestion for tag library description of the element, and send the finalized version to Ken. This CV is thus finalized.

ControlConstruct/Type:

Sanda’s e-mail after the videoconference:

"Control Construct" is an **\*abstract\*** entity in DDI 3 meaning that it

does NOT appear as an element per se in DDI markup, but rather stands

for a number of "real" elements that are enabled in its place.

These elements are IfThenElse, RepeatUntil, RepeatWhile, etc. etc. (the

entire list of elements that describe flow logic).

Any or all of these elements may be part of a Control Construct scheme.

Controlled vocabularies are usually applied or recommended for the

CONTENT of an element, and clearly this is not the case here. The

content of these elements would be a reference to a question or some

instruction on how to sequence the questions (or at least that is my

understanding at this time).

If we wanted to make changes (deletions, additions) to the list of

"flow" elements we would be requesting a revision of the DDI schema

rather than suggest a CV.

Fredy seemed to agree, judging from his e-mail.

Decision: Ken will ask Wendy why there should be a CV as it is a list of elements.

VariableGroup/GroupType:

Atle said in his e-mail that the groups cited in 2.1 are the same as in the Nesstar publisher. Ken will ask an expert at the UKDA to comment on the types, and asked everyone else to do so, if their organization uses the grouping. Discussion next time.

SummaryStatistics:

It was decided that Ken will send the list provided by Reiner and Meinhardt to Wendy and ask her to find an expert to review the list.

CategoryStatistics:

Reiner and Meinhardt will review the list in the Wendy document for the next time.

Reusable/GeographicLevels/GeographicStructureScheme/GeographicStructure

/Geography/Level/Code:

Reusable/GeograpiciLocation/Values/GeographyValue/GeographyCode/Value:

Ken will look into these, and review whether the INSPIRE standard (Infrastructure for Spatial Information in the European Community) has any relevance for this, at least at CESSDA level.

<http://inspire.jrc.ec.europa.eu/>

Will be discussed next time, Ken got two week’s extension on this

Next meeting: Wednesday 4 March 2009 at 13:00 UK time (=at the usual time).

**DDI CVG Videoconference** 4 February 2009

Present: Ken, Atle, Taina, Fredy and Dimitra (briefly)

Unit of Analysis:

Ken will send an e-mail to the DDI 3.0 User group with our intended CV, and ask for a clearer definition of Unit of Analysis elements. The tag library of DDI 2.1 was clearer, saying: *Basic unit of analysis or observation that the file describes.*

The participants of the meeting felt that the CV the DDI CVG has developed is good as it is, and describes units that data are collected about well but the problem is somewhat unclear definition of Units of Analysis elements in DDI 3.0.

ControlConstruct/Type:

Both Fredy and Atle will look at this issue. Atle will be on holiday (in Cuba!) at the time of the next meeting but will collaborate with Fredy and they will provide information beforehand.

VariableGroup/GroupType:

Ken will send an e-mail to the group asking whether a CV will be useful/possible (though this would require specifying what the GroupType is needed for). Atle will check what group types there are in the Nesstar Publisher (it turned out that these were exactly the same as specified in the DDI 2.1 tag library). Ken will also ask an expert at the UKDA to comment on the types. Discussion next time.

CharacterSet:

was finalized in the form suggested by Taina.

SummaryStatistics:

It was decided to ask Meinhard/Reiner to review the CV suggested in Wendy’s document and look into the issue by the next meeting.

Reusable/GeographicLevels/GeographicStructureScheme/GeographicStructure

/Geography/Level/Code:

Reusable/GeograpiciLocation/Values/GeographyValue/GeographyCode/Value:

Ken will look into these, and review whether the INSPIRE standard (Infrastructure for Spatial Information in the European Community) has any relevance for this, at least at CESSDA level.

<http://inspire.jrc.ec.europa.eu/>

Next meeting: Wednesday 18 February 2009 at 13:00 UK time (=at the usual time).

Web page where it is easy to find information on DDI 3.0:

<http://www.ddialliance.org/DDI/documentation/ddi3.0/index.html>

**DDI CVG Minutes 20090121**

Next meeting on Wednesday 4 February at 13:00 GSM.

Aggregation methods: The original 10-point list accepted.(?)

Storagetype list was accepted with definitions – so is finalised.

CollMode: Was decided (or not)? Group and Focus Group are the same (forgot to ask how this was decided last time). Taina will provide a new list with collected definitions.

Experiments – Field experiment definition included in the overall definition, not as narrower term.

Observation – the three narrower terms adopted (participation, laboratory and …)

Unit of Analysis:

Taina will send the proposed list to the group, along with the latest version of r\_codevaluetype Wendy v5. Chryssa will provide Eurostat definitions of Units of Analysis and Sanda will see where this information will sit in the DDI 3.0.

CharacterSets: People will show the recommended list to their IT experts. (tama unohtui käsitellä, menee seuraavaan kertaan)

**DDI CVG Minutes 20090109**

Next meeting at Wednesday 21 January at 13:00 GSM.

Commonailtytypecoded: We will propose a three-point list IDENTICAL – SOME – NONE. Ken will tell the DDI people that we had some doubts whether any CV is possible. The five-point list was rejected.

CollMode: Was decided (or not)? Group and Focus Group are the same

Sanda will provide definitions for Compilation, Synthesis and Simulation.

Is Clinical testing (aina potilas, ei rottakoe, joka ) part of Experiments

What is the difference between Observation laboratory and experiments?

Atle will work on Aggregations methods (?) tsekkaa

CharacterSets: People will show the recommended list to their IT experts.

**DDI CVG Minutes 20080226 (Sanda)**

1. We have agreement on DataKind and ModeofCollection - I am attaching an updated list. When submitting, Ken will add some definitions/clarifications on use where necessary.

2. Role: Distributor, Copyright Holder, Research Instigator, Other - to be added to initial submission. All terms on Role, as well as Publisher, and Creator need definitions (see also attachment)

3. Micro/macro, etc. - still under consideration.

For next time - not sure if Ken had something on the agenda (?) but I'd like to ask if people are willing to discuss a potential CV for LifeCycleEvent Type (this would be a list of events describing stages in the life cycle of a data collection/dataset). As a starting point, I am attaching a tentative list - just to give you an idea of what is needed.

(Taina said re Role definitions:

just for clarification - my definitions for Contributor, Creator and Publisher came from the schema documentation of DDI 3.0. so they are there already.

I was just not sure to which stage of the production process the element Publisher refers to but as it seemed to be clear to others, it means that the definition is sufficient and I need to study DDI 3.0 a bit more.)

**DDI CVG Minutes/Notes CVG 2007-12-18**

SamplingProcedure

Ken will submit request to TIC to add attributes to allow distinction between:-

Sampling and No Sampling

Probability and Non-probability

universe

Ken will request TIC to add attributes to allow distinction between:-

sub-national, national and multination

after Ken has come up with definitions and circulated to group

AnalysisUnit

Group is happy with Sanda's list and definitions

With "Objects" added to the list

Sanda or someone to add definition

Then Ken will submit controlled vocabulary to TIC

DataKind

Sanda/Mary to request TIC to reininstate in StudyUnit

Should not map to DataSource

Existing examples in Tag Library starting point for controlled vocabulary

Do we need more?

survey data,

census/enumeration data,

aggregate data,

clinical data,

event/transaction data,

program source code,

machine-readable text,

administrative records data,

experimental data,

psychological test,

textual data,

coded textual,

coded documents,

time budget diaries,

observation data/ratings,

process-produced data

Do we need definitions? Volunteers?

DataType

(Indicates whether data is alphabetic, numberic or mixed)

Sanda will investigate (see below)

DataTypeCodeType: Provides an enumerated list of data types.

Restricts: xs:NMTOKEN

Code: String - A string datatype corresponding to W3C XML Schema's xs:string datatype.

Code: Boolean - A datatype corresponding to W3C XML Schema's xs:boolean datatype.

Code: URI - A datatype corresponding to W3C XML Schema's xs:anyURI datatype.

Code: InclusiveValueRange - This value indicates that the startValue and endValue attributes provide an inclusive numeric range of type xs:double.

Code: ExclusiveValueRange - This value indicates that the startValue and endValue attributes provide an exclusive numeric range, of type xs:double.

In the new year Ken will investigate conference facilities at Essex

And setting up a Google Group

Ken Miller - Associate Director - Head of Information Development, Programming and e-Social Science - Tel: + 44 1206 872146

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

On Mon, 17 Dec 2007, Sanda Ionescu wrote:

Wendy,

I have a user asking where he can enter information such as a variable

being in ISO 8601, or GEOS. format.

Please help?

Also, this is for me - I can't find where exactly the element DataType

(indicates whether data is alphabetic, numeric or mixed) is situated

within the Logical Product - probably in the Variable section, but

where? sorry - I must be brain dead.

Thanks

Sanda.

Sanda Ionescu

ICPSR

University of Michigan

P.O. Box 1248

Ann Arbor, MI 48106

Phone, Fax: 734-615-7890

-----Original Message-----

From: Sanda Ionescu [mailto:sandai@umich.edu]

Sent: 18 December 2007 14:07

To: Miller, Kenneth P; reiner.mauer@gesis.org; Chryssa Kappi; Taina Jääskeläinen; Meinhard Moschner; Atle Alvheim

Subject: dataType

-----Original Message-----

From: Wendy Thomas [mailto:wlt@pop.umn.edu]

Sent: Monday, December 17, 2007 4:03 PM

To: Sanda Ionescu

Subject: Re: quick questions

Sanda,

We are in the process of fixing the DataType issues. Right now there is no "DataType in Variable as the RepresentationType made this redundant (NumericRepresentationType, TextRepresentationType, DateTimeRepresentation, etc). For the ISO 8601 the DateTimeRepresentation has a type attribute with an alternat format to provide explicit formats such as YYYY/MM or MM-DD-YY (default is the type standard). Options for type correxpond to the W3C XML Schema's all of which are ISO 8601

xs:dateTime

xs:date

xs:time

xs:year

xs:month etc

We are providing the ability to define specific structural constraints to text to allow for constraining to numeric codes (to be treated as text strings rather than numbers) etc.

So for now: NumericRepresentation = numeric

Text with no constraints is mixed

CodeSchemes are treated as mixed

I'll copy this to the bug tracker so that Arofan has the info.

Wendy

**MINUTES Controlled Vocabularies Group Tuesday 6th November 2007**

Present: Sanda Ionescu; Meinhard Moschner; Taina Jaaskelainen; Miller, Kenneth P; Atle Alvheim; Reto Hadorn

Unable to get connection: Chryssa Kappi

Apologies: Mary Vardigan

It was agreed that the purpose of the controlled vocabularies within a DDI instance should be to make the instance more computer actionable.

It was agreed that Wendy's list was by no means comprehensive and that before the next teleconference we should try and identify other elements/attributes that require immediate attention.

Initial investigation had shown that some of the listed items already had controlled vocabularies and perhaps some others could be rejected straightaway. It was agreed that the whole group should look at these list and determine their relevance and whether external experts are required (eg in aggregate data). It was also considered important in most cases to have an "other" option.

In looking at those existing list within the reusable module, it was agreed that controlled vocabularies were not required for

a) affiliation (eg every organisation in the world).

[What might be required is to point to a used list, where for instance organisation name changes could be recorded]

b) HistoricalDateFormat - if this meant terms such as 16th Centuary, Middle Ages for example, then computer actionablity would not be helped.

[Having read the module again the word Historical might be misleading and NonISO8601Format could represent the attribute better. Then eg Little endian: dd/mm/yyyy dd.mm.yy Big endian: yyyy month dd Middle endian: mm-dd-yyyy could work]

Certain items in the list didn't have controlled vocabularies but did have examples. Such as "calendar" (Gregorian, Julian, Jewish). More investigation was required to determine whether this was comprehensive or needed additional terms.

Certain attributes need clarification. For example "role" as an attribute of FundingInformationType was obviously different than "role" as an attribute of ContributorType. Is the former describing the action funded rather than if Funding Agency had been described as a contributor where its role would have been sponsor.

This lead to a discussion of possible contributors, why "Creator", "Funder" and "Publisher" had been single out and therefore would not be part of any controlled vocabulary for contributor role. Initial listing of possible terms were Data Collector, Data Producer, Depositor, Distributor, Copyright Holder. This led to more clarification needed between Publisher, Distributor and Data Producer. Sanda stated that the adoption of Creator and Publisher was to bring the DDI more in line with Dublin Core. She will investigate DC definitions.

Sanda will investigate why Chryssa didn't get connected at all and why everyone apart from ken and Sanda got cut off just before the end of the teleconference.

Next call: Tuesday 20th November 2007 (also Little endian) same time 13:00 GMT (UK time)

Best regards Ken

UK Data Archive - a service provider for the Economic and Social Data Service (ESDS).