To: Jane Fry, Carleton University, and Alex Cooper, Queen’s University:

Authors of the new edition of the Best Practices Document, Based on DDI 2.x. (hereafter referred to as the BPD)

From: Susan Mowers and Alexandre Paquet, GSG, U.Ottawa

Date: July 20, 2012

Comments on the BPD version 2 …

|  |
| --- |
| Alexandre:  RED is READY TO GO  WHAT DOES THIS MEAN? WOULD IT BE HELPFUL? IF SO, COMPOSE IN FRENCH?  BLUE means the changes have been made  SUSAN  TO CONFIRM |

**General:**

The BPD has been critical for training of new staff and is much appreciated.

Version 2 is focused primarily on social science survey microdata. Perhaps some changes for this next edition could also touch on aggregate social science data?

Looking further down the road, we’d like to mention that when creating DDI for the PCCF, Alexandre noticed that the BPD did not provide guidance. Perhaps this could be considered at some point (but not now)?

**1.1.2.1 Author Entity, p. 11 NOT A PRIORITY**

Background, during Nesstar 4.0 testing we noticed that any sublanguage records would not only indicate the authors of the sublanguage record, they also have to list all the default language record’s authors.  During testing (French sublanguage), we found more than 8 authors already for the default language record (from just two institutions).  NOT APPLICABLE CURRENTLY, COMMENT RE: MULTILINGUAL METADATA FEATURE OF NESSTAR 4: Tacking our author information after a lengthy list of English authors for a French record seems to me to be too open-ended a procedure (how long could or should the lists potentially be? Whether or not we use Nesstar multilingual metadata functions, could the BPD give any guidance on this?

**1.1.3.2 Copyright**

Would it be helpful or appropriate to add the example “Statistics Canada, Data Liberation Initiative”? Our Library uses a lot of their FR records from the DLI Nesstar server, partially, to a greater or lesser extent.

**1.1.3.3 Production Date, p. 13**

I’m still trying to understand this.

Does “Date the marked-up document was produced PUBLISBHED TO ODESI AS AN NDSstat file” mean the date of publication to Odesi?  If so, would it be correct to change the description to “Date the marked-up document was ~~produced~~ published to the archive” or words to that effect? The rationale is that the day the NSDstat file is completed being worked on, is usually the same day as it is published to ODESI.

Is there a one-to-one relationship between 1.1.3.1 and 1.1.3.3? If so, would it be appropriate to note this under 1.1.1.3?

**1.1.3.6 Funding agency, p. 14**

~~Add CIHR as an example of grant funding agencies – I have heard of CIHR funding the production of data deposit/archiving and metadata creation.~~

Could **Ontario Council of University Libraries (MarkIt! Program)** be an example?

And the Millennium Scholarship Foundation as another example? Or maybe that’s too outdated too?

Could Ontario Buys be removed from the guide since the program no longer funds DDI creation?

**1.1.4.1 Distributor**

What is the Gallup Canada referencing doing here in example 3, is it because Gallup Canada no longer exists and some history is required… and therefore the URL??? Which is now broken, now: Jane <http://www.library.carleton.ca/find/data/gallup-canada>

**1.1.6.1 Version** and **1.1.6.2 Version responsibility Statement, p.18**

Are these meant to be one-to-one relationships: e.g., 1.1.6.1 instance 1 and 1.1.6.2 instance 1 refer to the same update? If so, would it be appropriate to note this under 1.1.6.2.

Note: UOttawa and perhaps others would find this helpful. The names of these tags are general in nature and the relationship is not clear from the names.

**1.1.7 Bibliographic Citation, p. 19**

Could 2-3 examples for aggregate data tables be included please? One example might be. I can send a 1 or 2 more…

APA SUSAN … FRENCH EXAMPLES

Statistics Canada, Business Register Division. (2012). Table PRVNAIC4\_LOC December 2011 Location Counts by PROV/CAN, Industry Groups & Employment Size Ranges. [Table], Retrieved from http://odesi.scholarsportal.info/documentation/CBP/cbp-en.html. Beyond 20/20.

As an aside, these html pages are open to the public… and theoretically, these Open License tables could be open too… Technically in the BPD, we recommend reflecting this in a new example for Distributor below, please see note 5.

2.1.3.4 is interpreted as producers of the **data**, not of the work (as per 1.1.3.4) is this an error of copy/paste/versioning? JANE?

**2.1.4.1 Distributor** and **2.1.4.5 Date of Distribution, pp. 33 and 35**

1. Are these meant to be one-to-one relationships: e.g., 1.1.6.1 instance 1 and 1.1.6.2 instance 1 refer to the level of distribution? If so, would it be appropriate to note this under 2.1.4.5?

NOTE: UOttawa and perhaps others would find this helpful because of their separation in the standard by 4 other tags. We note that DLI is the distributor for most datasets.

1. The Note to Appendix C under 2.1.4.5 is not that helpful in the BPD as Appendix C only gives a very generic list of sources without distinguishing what the best sources are depending on the tag in question.
2. In the case of new STC microdata studies, we note that the [IMDB](http://www23.statcan.gc.ca/imdb-bmdi/pub/indexa-eng.htm) is a reliable source for the date of distribution. - Ajouté à l’annexe C.
3. Because 2.1.4.5 seems to us to be a more significant/obvious date, for DLI data, we use 2.1.4.5 and not 2.1.4.4 and we tend to declare DLI twice as applicable, for 2.1.4.1 and 2.1.4.3. Would the BPD have a better way? Could the BPD comment on why one might declare metadata for all these tags for usual social sciences microdata (from DLI)?
4. New example:   
   2.1.4.1 Distributor: Statistics Canada Open License Agreement

Entente de licence ouverte de Statistique Canada

**2.3.1.3 Sampling Procedure, p. 48**

1. Alexandre reports that the sampling procedures he comes across are usually more complicated than the examples. He provides something closer to [the IMDB description of sampling](http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3889&lang=en&db=imdb&adm=8&dis=2), along the lines of:

PLEASE COMPOSE …

*This is a sample survey with a cross-sectional design and a longitudinal follow-up.*

*The samples for SLID are selected from the [2012] monthly Labour Force Survey (LFS, record number 3701) and thus share the latter's sample design. The LFS sample is drawn from an area frame and is based on a stratified, multi-stage design that uses probability sampling. The total sample is composed of six independent samples, called rotation groups, because each month one sixth of the sample (or one rotation group) is replaced.*

*The SLID sample is composed of two panels. Each panel consists of two LFS rotation groups and includes roughly 17,000 households. A panel is surveyed for a period of six consecutive years. A new panel is introduced every three years, so two panels always overlap.*

… instead of, e.g., *Stratified, multi-stage design.*

Suggest naming 1 or 2 sources User Guide and IMDB (to help with a more simplified explanation of the methodology than the User Guide) – Ajouté à l’Annexe C

… In order to explain more about from what?, how and when

**2.3.1.8.1 Data Source, p.50**

We would suggest perhaps more representative examples such as:

*<titl> Survey of Labour and Income Dynamics</titl>*

*…*

*<dataSrc>The studies are based on samples taken from the Labour Force Survey.</dataSrc> Data source from IMDB…*

*Please revise both*

**We noted that the 2nd example is not that typical.**

**2.4.1.1. Location of Data Collection, p. 54**

Knowing when to apply this field can be confusing. Could a note be added to flag exclusions of which there any many such as microdata studies or other data sources which published for the data and the metadata using e.g., Nesstar editor?

URL FOR AGGREGATE DATA , e.g.,…

**2.4.1.2 Restrictions of Use, p. 57**

Since aggregate data is no longer be published to Odesi, with links to TDR, suggest deleting Examples 1 and 3. At some point, because we now have STC Open License data on Odesi, perhaps another example could be added along the lines of… *This data may be used by all members of the public, under the Open License.*

**2.4.1.4 Conditions of Use, p. 59**

Perhaps another example could be added?

<conditions>Statistics Canada Open Licence Agreement<ExtLink URI=” <http://www.statcan.gc.ca/reference/licence-eng.html>” title=” Statistics Canada Open Licence Agreement: /></conditions>

P.S. Did you notice the [exceptions](http://www.statcan.gc.ca/reference/exceptions-eng.html) and [FAQ](http://www.statcan.gc.ca/reference/licence-faq-eng.html)? I hadn’t.

QUESTION TEXT …

Questionnaire questions may be expressed as % % % . Example The use of these % symbols should be avoided where possible.

Proper French version

Codebook shorter and still meaningful

VARIABLE LABEL:

Some surveys, e.g., GSS use LONG questions … Instruct what you do with these?

**4.3.8.1 PreQuestion Text** and **4.3.8.3 PostQuestion Text, pp. 77-78**

Suggest either reversing the order of Notes 1 and 2 or simply dropping Note 1 as the Questionnaire / CODEBOOK seems to be the only (or best) source for this text. THIS USUALLY IS NOT IDENTIFIED AS PRE-POST question text.

In the case of 4.3.8.2 Literal Question, suggest reversing the order of Notes 1 and 2.

Discretion… Add pre-question text to the question itself as it’s often done this way in the codebook where the actual question is an incomplete text.

POST-question example: Weight can only be between 00 and 999 kg

pre-question text What about a preamble type …?

**4.3.12 Universe, p. 81**

As these examples sound somewhat like the survey universe, rather than the question universe, perhaps some other examples could be considered, such as :

*All respondents*

*Based on HUI\_01, HUI\_02, HUI\_03, HUI\_04, HUI\_05. System for measuring health status. (CHSMS-HUI3). See documentation on Derived Variables.*

*First-generation immigrants*

*Participating: Alberta, Yukon, Manitoba*

*BETTER AS NOTE OR UNIVERSE -- Alexandre??*

*ESCC – optional modules, e.g., Alberta, Yukon, Manitoba*

*Alexandre: Abstract??*

**Appendix C**

We recommend for consideration that for the following tags, that the [IMDB](http://www23.statcan.gc.ca/imdb-bmdi/pub/indexa-eng.htm) be used as the major source and in some cases (see asterisks), the IMDB should be used with the User Guide. Note, in these latter cases, the User Guide should be consulted, but may need to be significantly summarized, and the IMDB may help with this.:

|  |  |  |
| --- | --- | --- |
| Higher  Level tag | Specific tag |  |
| 1.1.5. |  | Series Statement |
|  | 2.1.4.5 | Date of distribution for new studies, not preceding cycles |
| 2.1.5 |  | Series Statement |
| 2.2 |  | Study scope - All and especially 2.2.2 and 2.2.3 and 2.2.3.9 |
| 2.3 |  | Methodology |

As there are some overlaps with the IMDB, we recommend for the following tags, that the User Guide also be used for

|  |  |
| --- | --- |
| 2.2.3.1 | Time Period covered |
| 2.2.3.2 | Date of Collection |
| 2.2.3.3 | Country |
| 2.2.3.4 | Geographic Coverage |
| 2.2.3.5 | Geographic Unit |
| **2.2.3.6** | **Unit of Analysis (user guide=major source)** |
| **2.2.3.8** | **Kind of Data (user guide=major source)** |
| **2.2.3.9** | **Universe** |
| 2.3.1.1-  2.3.1.4 and 2.3.1.6 | Data Collection … |
| **2.3.1.9** | **Characteristics of the Data Collection Situation (user guide=major source)** |
| 2.3.1.12 | Weighting |
| 2.3.3.1 | Response Rates |

*A shame that the Online Catalogue formerly on the STC site is no longer available for bibliographic citation, e.g., 1.1.1.5*