DDI for the preservation of statistical data

DDI 2 or DDI 3?

Michiel de Brieder
Daidalos B.V. | DANS
14-05-2008

“Where is wisdom we have lost in knowledge? Where is the knowledge we have lost in information?”
~ T.S. Elliot (1888 – 1965)

“The information in the world doubles everyday. What they don't tell us ,is that our wisdom is cut in half at the same time.”
~ Joey Novick

How much DDI does one need for preservation purposes?

Preservation of digital data (or information) is a thoroughly discussed item these days. Man has created and gathered such an enormous amount of data in the last decades that concern is rising about the preservation of this data.
Perhaps equally import is the expressed concern on the usability of information in the future. This concern is twofold:

- The digital era has introduced a new concept of volatility for data. In early days man feared the burning of scrolls, books and so forth, nowadays the fear has risen that data can no longer be accessed in the future because the platform on which it was created is no longer in use or accessible. Whatever method of data destruction is feared, it is the outcome that must be addressed, loss of information (information leading to knowledge, leading to wisdom) is unacceptable.
- When as much data is preserved as possible then the world will need to deal with enormous amounts of data. The only way to access and successfully search through the amounts of data to come is by tagging the data. In the world of information technology these descriptive tags are called metadata. Without metadata every data object in a data set will have to be analysed each time it is within a data set that is being queried. Analysing a full object or analysing some pre-defined tags means a world of difference in performance aspects.

This article will deal with the metadata aspect of preservation, particularly for the SDFP standard. The SDFP standard is a preservation standard in the form of an umbrella format that can be used to store different data kinds. One of these kinds is the statistical data kind, others are, for instance, database data and spreadsheet data. SDFP concerns itself with these data kinds mainly because they are often captured in proprietary formats of long lived and widely used software programs. Conversion from proprietary standards to an open standard is the key for the development of SDFP.
The article will focus on the statistical data kind: “How will metadata, created by extracting the model from a statistical data file, be stored”.

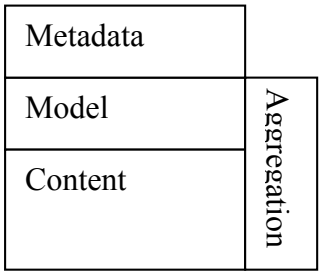
Index

Statistical data kind

What is regarded as statistical data kind? There is a concise answer to this question. To SDFP all data that is contained within a file that has been created by means of a statistical computer program is considered to belong to the statistical data kind.
There are several software programs that can be used to create a file that contains statistical information. A few examples of these programs are:

- SPSS
- SAS
- R

Each of these programs rely on the same structure. This structure is the following:



Each file contains a model with that consists of variables, where each variable is described by several attributes. The content is given within the model, each variable can have a value for an entry within the file. Together the model and content can be aggregated for research purposes using statistical analysis. The file also contains some metadata on itself (usually parameters like: date of creation, creator etc).
To preserve a statistical file in a non-proprietary format the document data initiative has been brought into life. A standard has been created with the name of DDI. DDI is in its second version at the moment of writing (DDI 2.1) and an expanded version has been created with the name DDI 3.0, however DDI 3.0 does **not** supersede DDI 2.1. According to the DDI alliance both versions will have their own path of progress.
The following chapters will provide insight in both versions.

DDI standards

This article will deal with the DDI 2 and DDI 3 standard, the following chapters describe each respectively and a comparison follows.

The DDI 2 standard

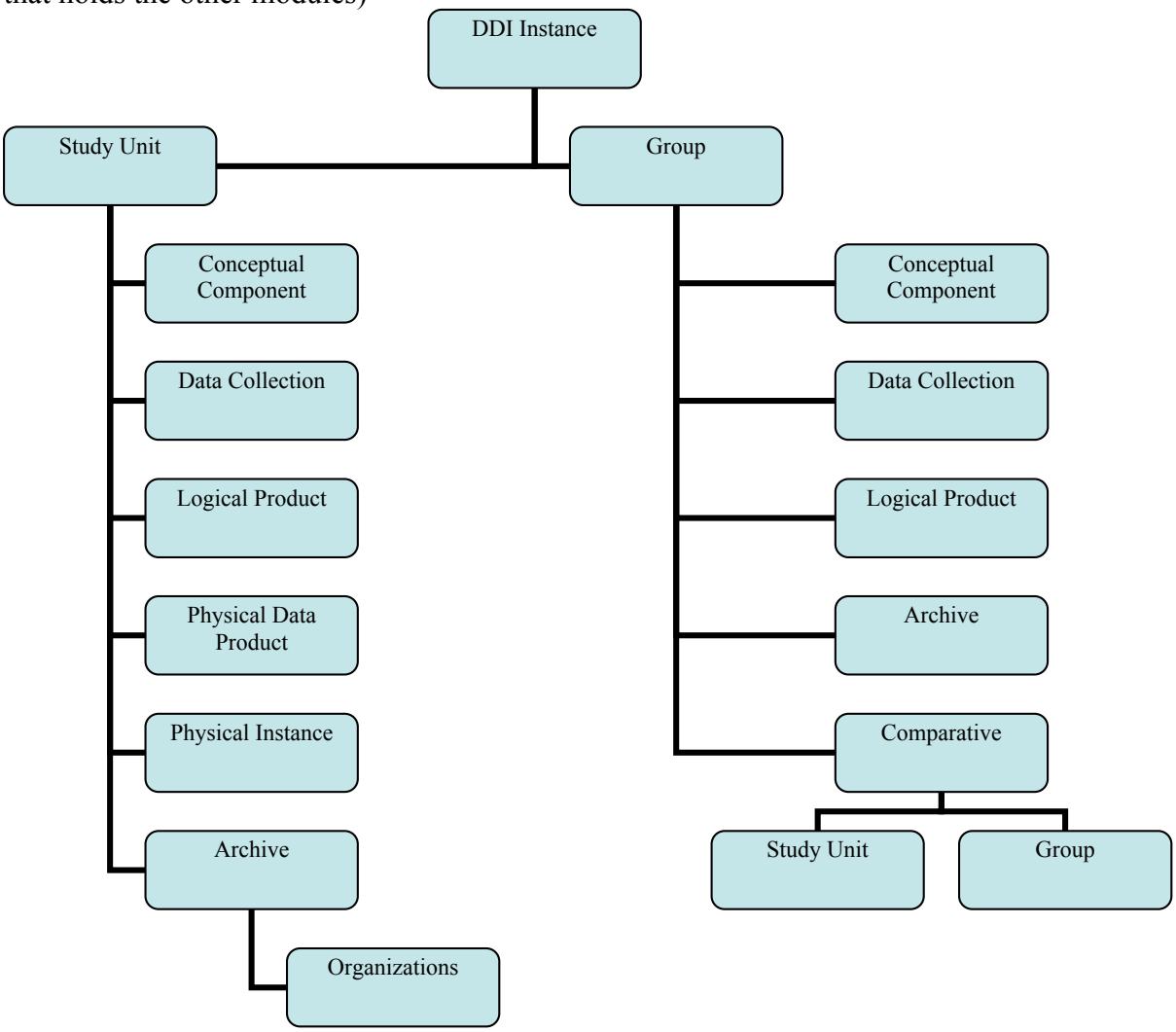
DDI 2 is described as a storage solution for microdata surveys with aggregate tabular data. The structure of a DDI 2 file consists of the following elements:

Document description	The information stored in the document description is the metadata on the DDI file itself. Creation and bibliographical citation is included in this metadata set
Study Description	The context of the study, this includes: • Creators • Methodology of research • The abstract • Keywords • How the data is produced • How the data is distributed • Etc
Data files description	One or more data files may be included within the dataset. This element holds information such as format(s), size(s), number of cases etc)
Variable Description	The model of the statistical data file (the variables with their attributes) are kept in the variable description. (the model may be compared to that of a database, however, the columns in a statistical data file may contain more attributes) @TODO check!
Other study materials	Additional materials may be included within the statistical file, either as inline information or external references. For instance: citations to publications, coding schemes, thesauri etc)

One of the many strengths of DDI 2 is the ‘readability’ of its structure. A human can decipher its meaning fairly easy which makes it a practical standard for preservation (everything that can be done will be done for preserving data, however, if all else fails it is good to be able to rely on something that may even in the future be decoded by hand)

The DDI 3 standard

DDI 3 can be described as a modular life cycle model with complex and comparative data files.
The modularity of DDI 3 can easily be recognized by the fact that its structure consists of referenced files. The following model shows modularity of a DDI 3 instance (since it consists of several files it is preferable to refer to ‘instance’ because ‘instance’ is the top level module that holds the other modules)



Comparing the above model with the DDI 2 standard quickly leads to the conclusion that DDI 3 is much broader than DDI 2. One might say that the DDI 2 concept is more oriented on the storage of content and model along with some metadata and the DDI 3 concept is more a study description or a collection of study descriptions with a layer of metadata to describe the relations between the studies.
In short: DDI 3 has the same functionality as DDI 2, but it is far more versatile and has far more functionality to describe different aspects of a study and the grouping of studies.

Comparing DDI 2 and DDI 3

The following table compares DDI 2 to DDI 3.

DDI 2[1]	DDI 3[1]	Relevance to SDFP
Single study	Several study units	Whether combined studies or separate studies are stored within the standard is irrelevant. If studies should be kept together then the archiver will be responsible.
Inadequate representation of complex / hierarchical data	Detailed documentation for complex / hierarchical data	The inadequacy of DDI 2 towards representation of complex / hierarchical data does not lie in scope of the model or the content but in the metadata description thereof. This is not within the scope of SDFP.
No instrument coverage	Full description of instrument as a separate entity	The instrument with which the study has been executed is not part of the scope of SDFP
Question text appears only as part of variable description	Compatible with Computer Assisted Interviewing (CAI) software	SDFP is meant to be a standard on its own, utilising other open standards. CAI is not an open standard and not supported within the format
No documentation for question flow / Conditions	Documents specific use of questions: flow, conditions, loops	Currently flow is not a functionality of major statistical software applications
Initially designed for microdata only	Adds support for tabular, spreadsheet-type representation of aggregate data	This is irrelevant, aggregation is not a focus area of SDFP
Aggregate data section added in V 2.1 to support limited representation (Census-type data, delimited files)	Aggregate data transport option: cell content may be included inline with the data item description	This is irrelevant, aggregation is not a focus area of SDFP
No data transport function	Inline inclusion enabled for both aggregate data and microdata	This is irrelevant, aggregation is not a focus area of SDFP
No longitudinal / time series / cross-national data comparability	Grouping structure documents studies related on one or several dimensions (time, geography, language etc.) as well as their comparability	Irrelevant since grouping and comparing studies is a form of aggregation
Limited multilingual support	Support for multiple language use and translations	Language support is not relevant for SDFP

Single file, hierarchical design	Modular design: facilitates reuse, facilitates versioning and maintenance, supports life cycle model, allows flexibility in organizing the DDI Instance, supports grouping and comparing of studies, supports creation of metadata registries	DDI 3 offers a modular design of reusability, versioning and maintenance. This modularity is useful when the lifecycle of the study is ongoing. The moment of archivation however, marks the end of the life cycle.

Life cycle or not?

The main difference between DDI 3 and DDI 2 is the fact that the life cycle of information can be stored within the DDI 3 format. The life cycle of information can be a very valuable asset to preserve, as it can give insight in a lot of factors surrounding a survey or research study.
However, the main concern of SDFP is the long term preservation of data as-is: research data that may be studied in the future without aggregation information of the previous analyst.

Different viewpoints

The storage of information can be viewed with different mindsets. In this case the point of view of an archiver and the point of view of a statistical analyst/researcher will be taken into account.

The archivers' point of view

Preserving data has to be practical. There should be no excess, but there should not be any loss of important information. With SDFP in mind the following statement will be the centre point for the archivers' point of view:
Preservation of data should be focussed on preserving the content and model of data as delivered to the archiver.

The statistical analysts' point of view

The researcher (or statistical analyst) has a different view on his/her data than that of the archiver. The more information on the study is preserved, the happier the researcher will be. The full preservation of the life cycle of a study is therefore very interesting from a researchers point of view. Moreover, the aggregation of data that leads to a conclusion put forth by the researcher is, to the researcher, of as much or maybe even more value than the data itself.
One could argue though, that conclusions and aggregation can be kept in reports and papers which is outside the scope of SDFP for the moment.

Conclusion

SDFP is a standard that focuses on the preservation of data in the form of content and model. Aggregation, interpretation and lifecycle documentation are not within the scope of the SDFP standard. At first sight one might say that DDI 2 is the standard on which SDFP will base its storage possibilities for the model and content of statistical data because it is human readable and in line with the thoughts behind SDFP.
At a second glance it must be mentioned that the influence of DDI 3 may increase over time. The SDFP standard may be updated in the future to support this in some way.

For now, the basis for the preservation of the model of statistical data files has been found in DDI 2.

References

The following references have been used in the creation of this article.

Internet references

The DDI (Data Documentation Initiative) website, http://www.ddialliance.org (May 14, 2008)

Green, A., Feb 5-6 2008, Data Documentation Initiative (Data Share Project Meeting, University of Edinburgh), http://www.disc-uk.org/docs/DDI_Green.pdf (May 14, 2008)

Martinez, L., Feb 28 2008, The Data Documentation Initiative (DDI) and Institutional Repositories, http://www.disc-uk.org/docs/DDI_and_IRs.pdf (May 15, 2008)