

## CROSS-NATIONAL AND DIACHRONIC DATASETS: DIAGRAMS

---

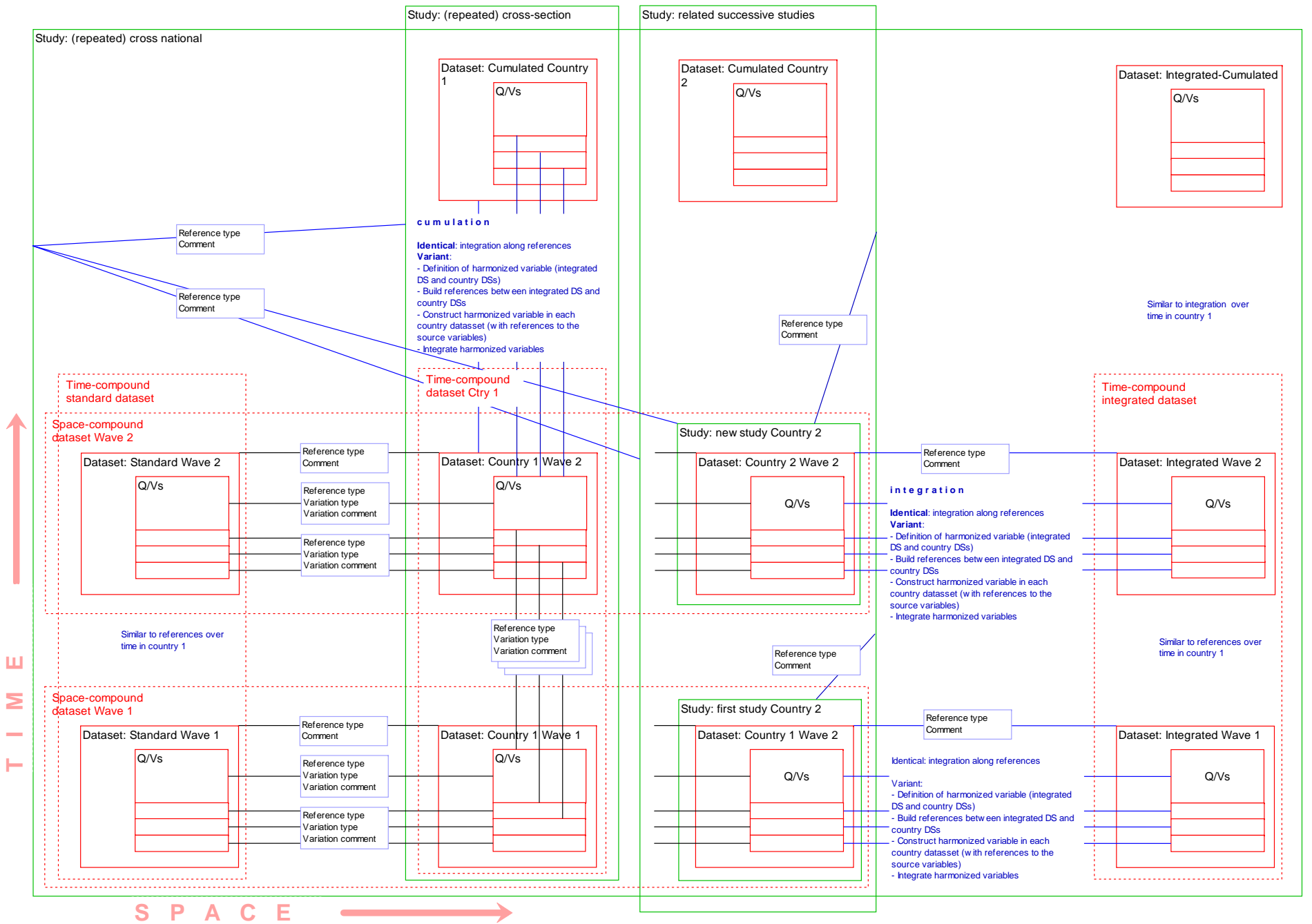
Author: Reto Hadorn, SIDOS

State: 17.3.2005

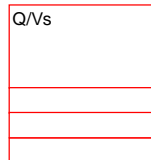
Status: draft

These diagrams illustrate the document "Handling the repeated cross-national datasets" by the same author.

Just print it on A3 Landscape for best reading.



## Legend



### Questions and variables

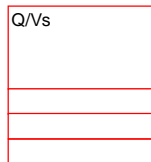
Questions are conceived only in relation with the variables they create.

Question structures, which organise the communication between interviewer and interviewee, may be of various degrees of complexity. The icon in this graph just suggests that several variables may be related to one single question unit.

Some variables are not related directly to questions. Constructed variables will be related to their source questions (eventually related to questions).

Extraneous variables must have a well documented source information.

Dataset: Country 1 Wave 1



### Simple dataset

A simple dataset is based on a single data collection operation, defined as the exhaustion of a sample in a limited time span.

The dataset may involve several files, hierarchical data, distinct instruments (FTF and written questionnaire). It contains many Questions and variables - only one being shown in the diagram. The questionnaire object is not shown, although it contributes to the organisation of Q/Vs.

In a database capable of managing repeated comparative datasets, any dataset must be characterised by a set of coordinates, which locate it in the time/space space (diagram on the right).

Space coordinates: Single;(Standard);Crete;Naxos;Zermatt;Florida;Compound;Standard/integrated  
Time coordinates: Single;2000;2001;2002;2003;compound;cumulated.

The standard definition for a comparative study can also be handled as a dataset, although it is virtual.

Integrated datasets are also simple datasets. Specificity: there is a space variable, which actually refers to distinct samples. The metadata incorporate information about differences between samples.

Cumulated files are also simple datasets. Specificity: The data include a time coordinate, which refers to distinct samples or at least distinct data collection operations.

Space-compound  
dataset Wave 1

### Compound dataset

A compound dataset is a set of related datasets within a comprehensive project, which defines the relationship between the datasets.

A space-compound dataset includes the standard definition and all the simple datasets collected in specific spaces and within a specific wave; as such, it is distinct from the integrated dataset.

In a database capable of describing repeated comparative studies, it must have variant space coordinate and a fixed time coordinate. A space-compound dataset will be defined for each wave of a repeated comparative study.

A time-compound dataset includes all the simple datasets collected in a specific space (sample base) but within successive waves. As such, it is distinct from the cumulated dataset. It will be defined for any space-specific longitudinal study/project. In a repeated cross-national study, there is no necessity for building in each distinct space the corresponding time-compound dataset. This is optional, since the time dimension is basically handled on the level of the compound standard dataset.

In a database capable of describing repeated comparative studies, it must have a fixed space coordinate and a variant time coordinate.

The successive standard definitions build a time-compound standard dataset. The successive integrated datasets build a time-compound dataset.

The space and time hyper-compound dataset is a combination of the time-compound standard dataset and the successive space-compound datasets related through the compound standard dataset.

Where time-compound datasets have been elaborated for some single countries, these are de facto part of the hyper-compound dataset, although they are not necessary components.

Study: first study Country 2

### Study

In this presentation, the Study has two components. An organisational one (who does what with which funds in which organization?)

and a methodological one (which methods are used on which sample defined on which universe?). In the Metadata metadata model, the first is on 'project' level, the second on study level.

The Projects object has been created for the case a same research project covers more than one study (defined from the methodological point of view).

A study may head a single simple dataset, a space-compound dataset, a time-compound dataset, or a space and time hyper-compound dataset. Yet:

- There will be a study definition for a single time space compound dataset. There will not be a study for each of the successive cross-national waves, just one encompassing the whole.

- There will be a study definition for a time-compound single sampling universe dataset. In a repeated cross-national study, a study definition will be given for each space, even where no compound dataset is constructed, since it stores information on the local organization of the data collection.

- If breaks occur in the organization or in the methodology of a longitudinal study (Diagram Country 2), successive studies will have to be defined ('first study' and 'new study' in the diagram).

In the diagram, the Study for the repeated cross-national dataset encompasses the successive integrated and the cumulated integrated datasets. This expresses a situation where integration and cumulation are done under the control of or within a mandate from the coordinating agency. If integration or cumulation are done fully independently, a new study may be defined, which will refer to the study for the repeated cross-national study.

Within a repeated cross-national study, it should be possible to select distinct sets for publication: a single space-compound dataset or all successive space-compound datasets

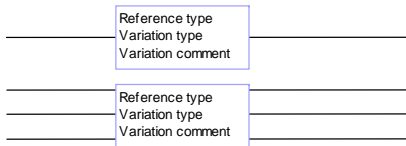
The study level information is in all cases taken from the description of the repeated cross-national study; the selection of the datasets is just a matter of defining the values of the coordinates for the compound dataset to be selected.

For the study description to be able to cover the integration and cumulation activities as well as the general coordination activities, several information elements should be marked to define the work context which they belong to; for example: citation, title, involved institutions and staff, abstract... to be continued.

## References

References are needed on several levels to wave the dataset fabric.

- On the lowest level, references from questions to question and variables to variables are necessary to establish the comparability over space and the continuity over time. Besides defining the relationships among questions, resp. variables, those relationships help building the whole metadata set economically, using existing information to define new elements. The same relationships can be used for making an overall diagnosis of the integrability or cumulativeness of sets of variables, an important step before harmonisation takes place. They are also the lines along which the variables from the single simple datasets are integrated or cumulated.
- On an intermediate level, references may be necessary from the single country datasets to the standard definition and from waves 2-n to wave 1 dataset. In very homogeneous studies, sharing the same time coordinate suffices to define a space-compound dataset and sharing the same space coordinate defines uniquely a longitudinal effort. Explicit links have yet the advantage that information can be added to the relationship, for example to characterise explicitly the type of relationship or to document some specifics about the relationship between the datasets concerned. These relationships will help combining the information of the single simple datasets into the metadata for the compound, integrated and cumulated datasets.
- On the highest level, references between studies must be defined to help the user understanding the relationships between those objects, which all will appear as a 'top entry' in the DB system. For example, if a country elaborates a time-compound dataset for its part of a repeated cross-national study, a reference must be built into the system, which would allow for automatic reference to related studies and navigation among the studies, since the accent will not be put on the same characteristics.



## References between questions, resp. variables

References are established separately for question elements and variable elements, as suggested in the diagram.

The exact structure of the relationships will depend on how the question structures are translated into database structures. In a VarInfo-like design, where multiple response and items are described on variable level, two levels of relationships do the whole job: one for questions, one for variables. With other designs, involving for example a 'subquestion' object as in the current MetaDater model (as of March 2005), additional linking elements may be necessary to document variations in question definition over space and time. If the elaboration of the typology of questions is continued and additional logical levels introduced, reference levels may also have to be added.

Besides the external keys (which must be directed, with a source and a target), the fields necessary are

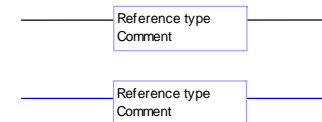
- a field defining the type of reference, so all kind of references can be stored in the same logical store (table)
- a field for a standard coded qualification of the variation (country specific deviation or variation over time)
- a field for a textual description or comment.

VarCode	Variation	CommentVariableiation
Id	Identical	Country question is wholly identical with question in reference questionnaire
Word	Var. in wording	Wording of question is different, aiming similar content
VarStruc	Var. in question structure	Subquestion list is eventually different or multiple where standard is simple etc.
Word+VarStruc	Var. in wording and quest. struct.	Both wording and subquestion list are different
NotAv	Not available in standard Q.	Country question is not available in reference questionnaire

Types of variations between questions, involving question wording and structure (e.g. number of items)

VarCode	Variation	CommentVariation
Id	Identical	Subquestion wording (if available) and value structure identical with question in reference questionnaire
VarWord	Var. wording subQ	Variation in subquestion wording
VarVal	Var. values structure	Variation in the structure of values
VarWord+Val	Var. wording and values	Variations both in subquestion wording and structure of values
NotAv	Not available in standard Q.	Subquestion not available in reference questionnaire, although question is.
NotAppropriate	Evaluation not appropriate	For example: the question is open, answers are postcoded

Types of variation between variables, involving item wordings and value domain structures



## References between simple datasets

The minimal requisite for this link is, besides the external keys, a file reference type and a textual comment. Experience will show whether or not information should be added here.

## References between Studies

The minimal requisite for this link is, besides the external keys, a file reference type and a textual comment. Experience will show whether or not information should be added here.

integration / Cumulation

Identical: integration along references

- Variant:
- Definition of harmonized variable (integrated DS and country DSs)
  - Build references between integrated DS and country DSs
  - Construct harmonized variable in each country dataset (with references to the source variables)
  - Integrate harmonized variables

The use of references in integration and cumulation processes

(Suppose all the metadata for the compound dataset in the database and the control on data available. An integrated or cumulated dataset definition has been generated automatically, using the Q/V definitions in the standard, incl. the anticipated harmonisation variables and the expected country or wave specific variables to be stored in the synthetic file).

- Screening: The program uses the coded information stored with the references (cf. tables above) to identify
- the strictly identical Q/Vs (over space or time, depending on the process), which can be readily integrated or cumulated
  - the Q/Vs where a change lets expect a change in the data properties.

Using the information stored with the references, the program can deliver a well informed warning for each Q/V in the standard or in the country datasets; the operator must use this information and some additional techniques (like selected cross-tabulations) to make a precise diagnosis, which can be stored with the standard Q/V concerned.

Depending on the specific diagnosis he makes, the operator adds an informative comment with the standard Q/V or decides to create a harmonized variable.

Documenting: the standard comment fields on the level of questions and wordings will be used (question level or value level).

Harmonizing:

- a harmonization variable is created and documented in the integrated or cumulated dataset, and
- copied to the single country or wave specific datasets to be integrated or cumulated;
- a reference is automatically created from the country or wave specific variable to the harmonized variable, with appropriate qualification.
- the operator defines the construction necessary in each single dataset to compute the dataset specific version of the harmonized variable;
- the construction is documented with the harmonized variable
- a reference link is created from the constructed harmonized variable to the 'parent' variables used in construction; as in other cas of construction, this reference can be used either to navigate the database and the published metadata or to draw information from the parent variables to inform the documentation of the constructed variable (for example, the precise question wording).

The harmonized variable is now ready for integration or cumulation, in the same way as are the rigorously identical variables.

In a cross-national study, if the need for harmonization has been anticipated and the harmonized variable defined in the standard, the process is somewhat different. The differences observed between the standard definition and the single dataset indicates a possible error on the part or the local partner. Formulation of original question. Distribution of values, undocumented codes. ... (wizard work)

For the sake of simplification, only variables are taken into account in the diagram below.

