MetaDater Metadata Management and Production System for surveys in Empirical Socio-economic Research

Hold on, Jerry!

The great dataset movie

Screenplay and direction: Reto Hadorn
Production: SIDOS and EU (MetaDater project)
Co-production: DDI, CSDI, ICRC, Swiss household panel

First show: 2005 IASSIST festival
Edinburgh, 27 May 2005

Status: Conference draft

Plan

- Introduction
- The movie
- Issues and puzzles

Introduction

This show is part of a bundle of documents, which describe in abstract terms the handling of the single datasets in the context of a repeated comparative study. This bundle includes:

- text
- diagram
- this movie.

The stance is that the whole story is not a complicated one, as far as the definitions of the involved objects are precise and one problem is handled at a time. Yet the communication of abstract ideas is not unproblematic. Starting with a textual description of the construct, the presentation has been expanded to a diagram supposed to include all aspects and, since that diagram is unfortunately static, with the present movie, in which the diagram is built up step by step.

The movie has been conceived at the start for a presentation of the issue at the 2005 ISSIST Conference in Edinburgh. A meeting of the Expert group of the DDI-Alliance has take place just before the congress itself, where a 'grouping' concept was presented, which was discussed in depth in the 'comparative datasets' working group and… several informal and casual meetings. The discussions have shown how important a clear presentation of the relationships between the elements of the so called 'repeated comparative study' are. This was an additional motivation for making a more 'didactical' presentation of the whole story.

It remains that the ideas presented here are especially related to the MetaDater project, which needs a metadata model to put it in action, and to the SIDOS prototype for a variable level relational database, where relationships between questions and variables belonging to various country datasets (cross-national studies) or waves (panel study) were tested.

Efforts were made to keep all three documents consistent. Some inconsistencies may nevertheless subsist, since they have all their own 'story', being partly developed in distinct contexts.

To have a full view of the issue, it is recommended to refer to all three documents cited above and, in addition, to the following ones:

-Workflow diagram for new questions
- Description of dataset selection process
- description of question typology (2 documents).

Let's now turn back to the presentation.

The background

- **MetaDater project**: creating metadata structures accounting for the production of repeated cross-national datasets
- The **CSDI**: create better documentation of the cross-national programs
- The **DDI**: building metadata structures supporting the identification of comparable data

The challenges

- Support the **life-cycle** (long term)
  - Definition of the research program
  - Creation of the research instrument (multilingual)
  - Completing the field work
  - Processing data and metadata
  - Distributing data and metadata
  - Repurposing
  - …

The challenges

- **Economy** in metadata capture and management
  - **Normal** form
  - **Re-use** information by reference
  - **Copy** information for edition in the case it is just ‘varying’

The challenges

- Best **documentation** of:
  - Relationships between the studies involved
  - Series of questions and variables
  - Possible variations within series
  - The construction of new variables for harmonization

The challenges

- **Document by doing!**
  - Data are handled from within the metadata system
  - Documentation of the operations is largely a byproduct of doing them
    - Corrections on data file
    - Variable construction
    - Defining harmonized variables
    - Computation of harmonized variables

The challenges

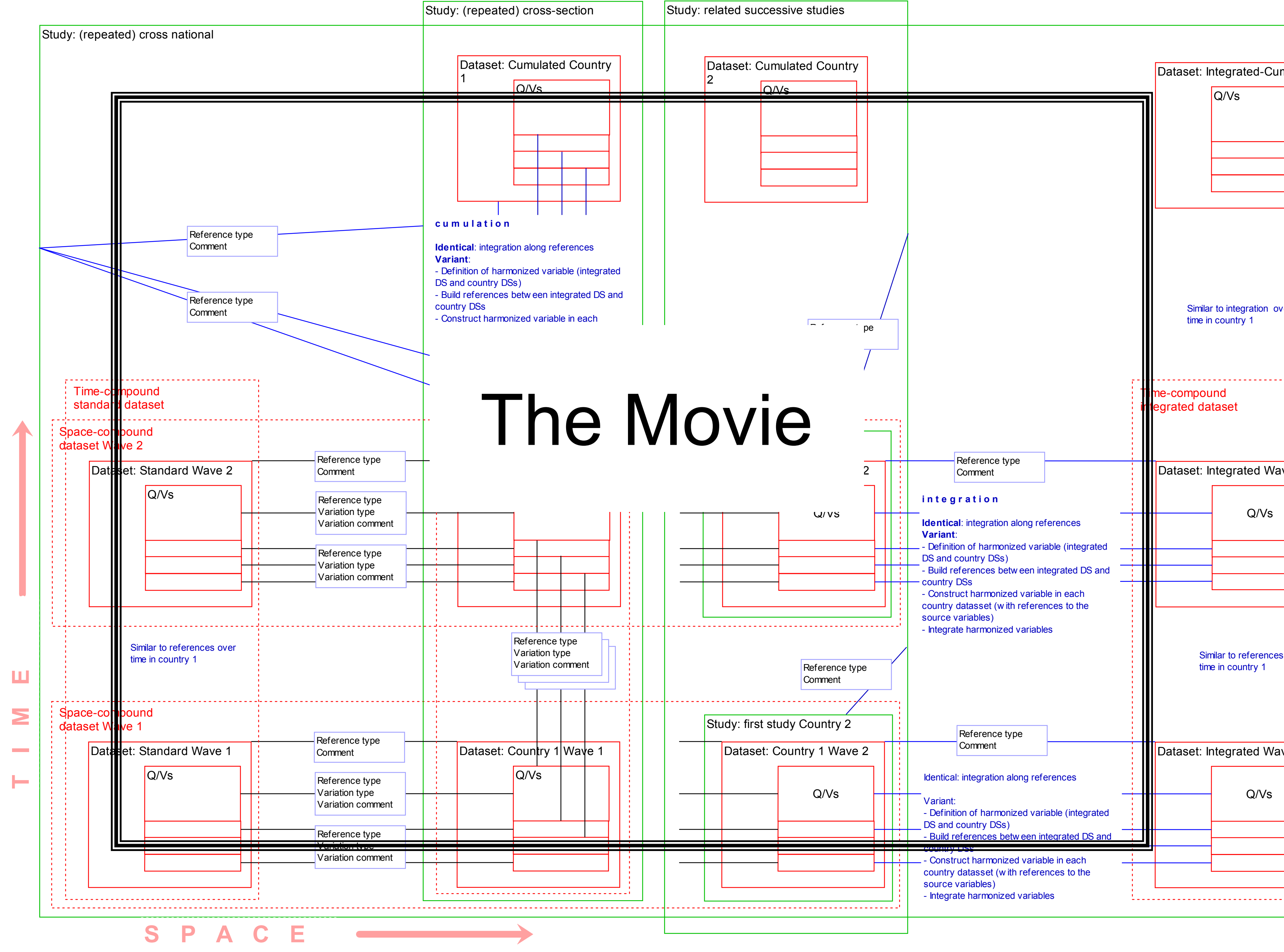
- To support…
  - **analysis** of the variations among simple datasets to be integrated or cumulated
  - **integration** of country datasets, **cumulation** of waves and cumulation of integrated datasets
  - **publication of metadata** at any stage

The challenges

- Make **several forms of metadata publication** possible:
  - – The sets of datasets (country DS or waves) as they are collected
  - – The integrated dataset (space), the cumulated dataset (time)

The challenges

- Make **several forms of metadata publication** possible:
  - The sets of datasets (country DS or waves) as they are collected
    - … full navigable documentation
  - The integrated dataset (space), the cumulated dataset (time)
    - … all documentation drawn from the single datasets and harmonization operations into a synthetic presentation



The Movie

Study: (repeated) cross national

Study: (repeated) cross-section

Study: related successive studies

Dataset: Cumulated Country 1

Q/Vs

Dataset: Cumulated Country 2

Q/Vs

Dataset: Integrated-Cu

Q/Vs

Reference type
Comment

Reference type
Comment

cumulation

Identical: integration along references
Variant:
- Definition of harmonized variable (integrated DS and country DSs)
- Build references betw een integrated DS and country DSs
- Construct harmonized variable in each

Re   pe

Similar to integration ov
time in country 1

Time-compound
standard dataset

Time-compound
integrated dataset

Space-compound
dataset Wave 2

Dataset: Standard Wave 2

Q/Vs

Reference type
Comment

Reference type
Variation type
Variation comment

Reference type
Variation type
Variation comment

2

Q/Vs

Reference type
Comment

Dataset: Integrated Wa

Q/Vs

integration

Identical: integration along references
Variant:
- Definition of harmonized variable (integrated DS and country DSs)
- Build references betw een integrated DS and country DSs
- Construct harmonized variable in each country dataset (w ith references to the source variables)
- Integrate harmonized variables

Similar to references over
time in country 1

Reference type
Variation type
Variation comment

Reference type
Comment

Similar to references
time in country 1

Space-compound
dataset Wave 1

Dataset: Standard Wave 1

Q/Vs

Reference type
Comment

Reference type
Variation type
Variation comment

Reference type
Variation type
Variation comment

Dataset: Country 1 Wave 1

Q/Vs

Study: first study Country 2

Dataset: Country 1 Wave 2

Q/Vs

Reference type
Comment

Dataset: Integrated Wa

Q/Vs

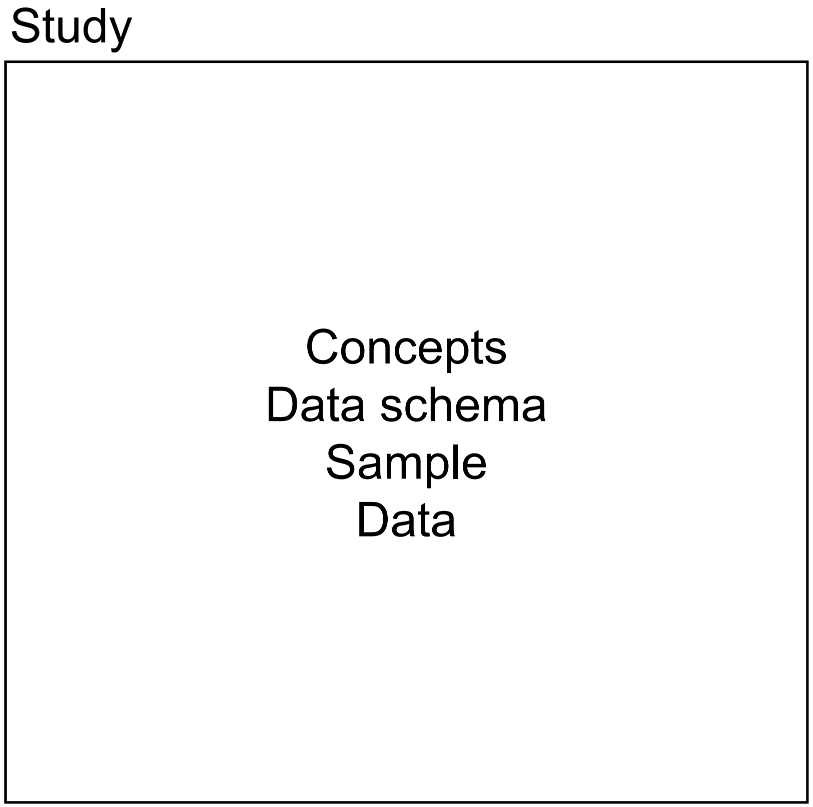
Identical: integration along references

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

TIME

SPACE

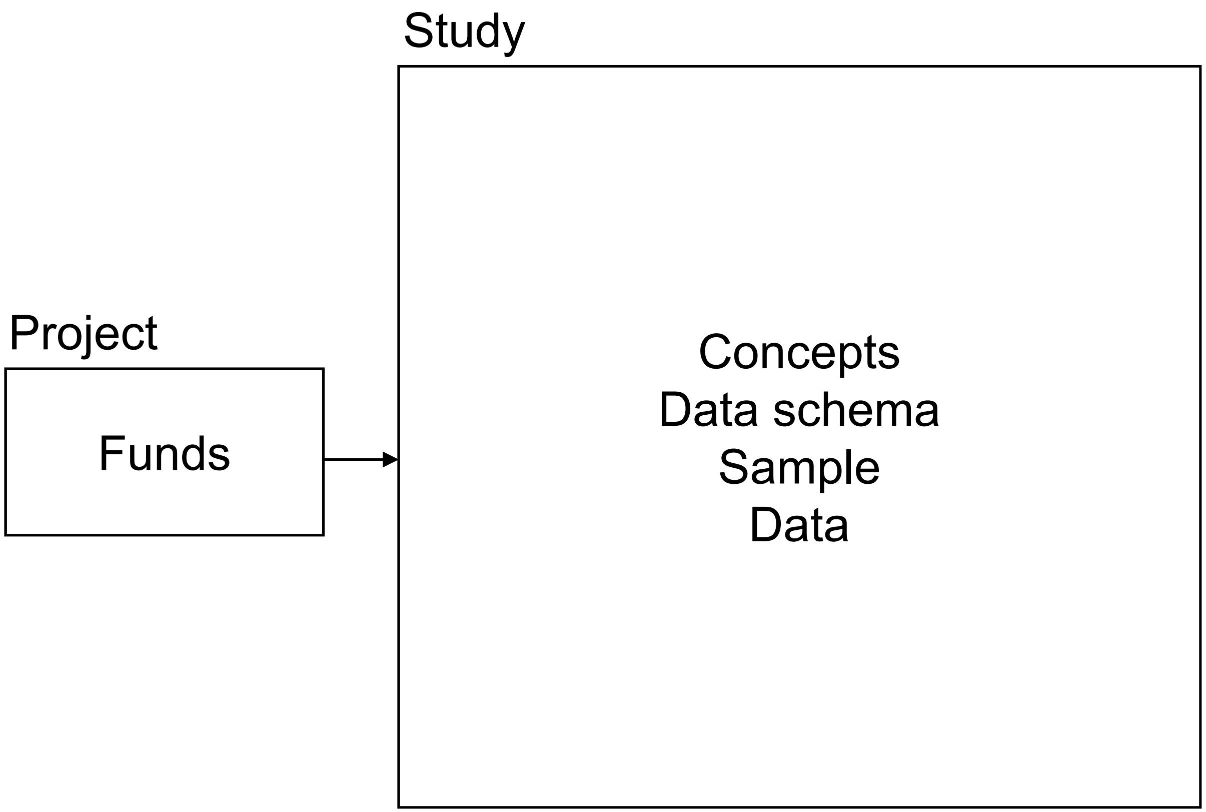
At the beginning was the study



Study

Concepts
Data schema
Sample
Data

A study can be funded successively under several research projects, identified by their title and described by a summary



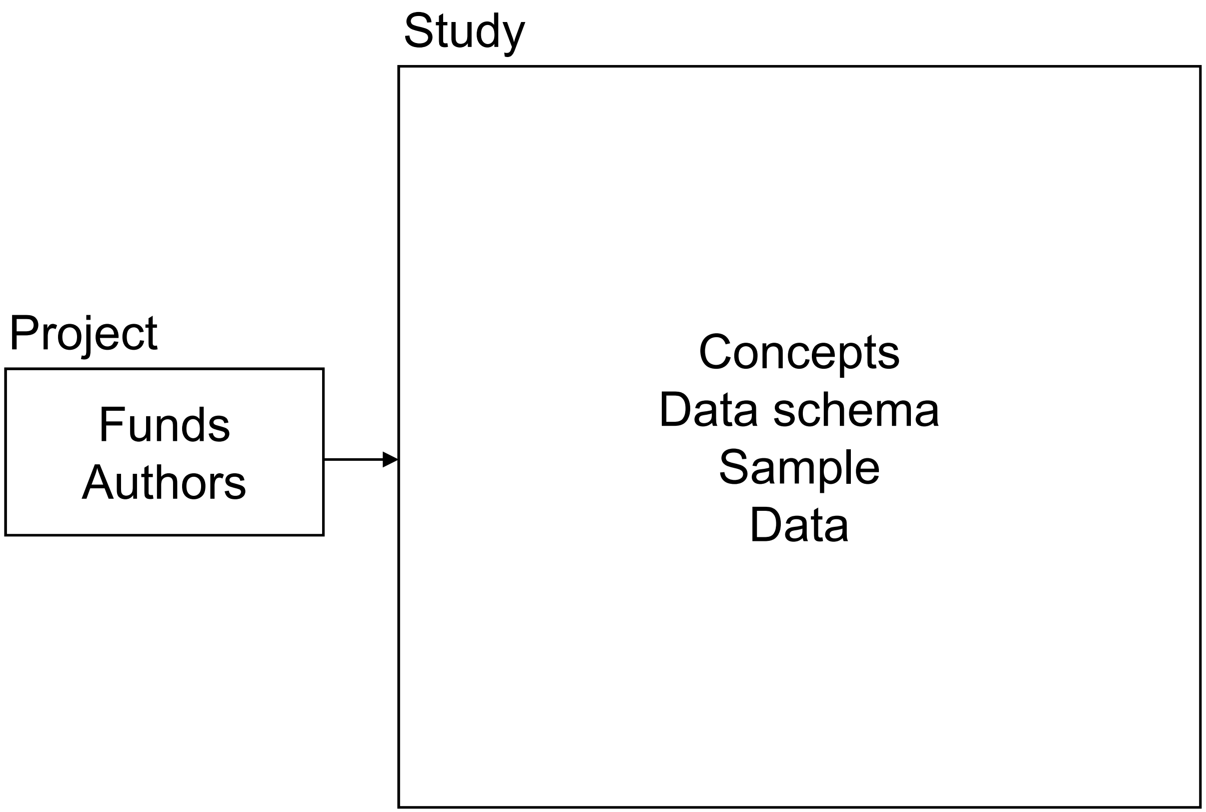
Study

Project

Funds

Concepts
Data schema
Sample
Data

Usually, the funding goes to services and authors

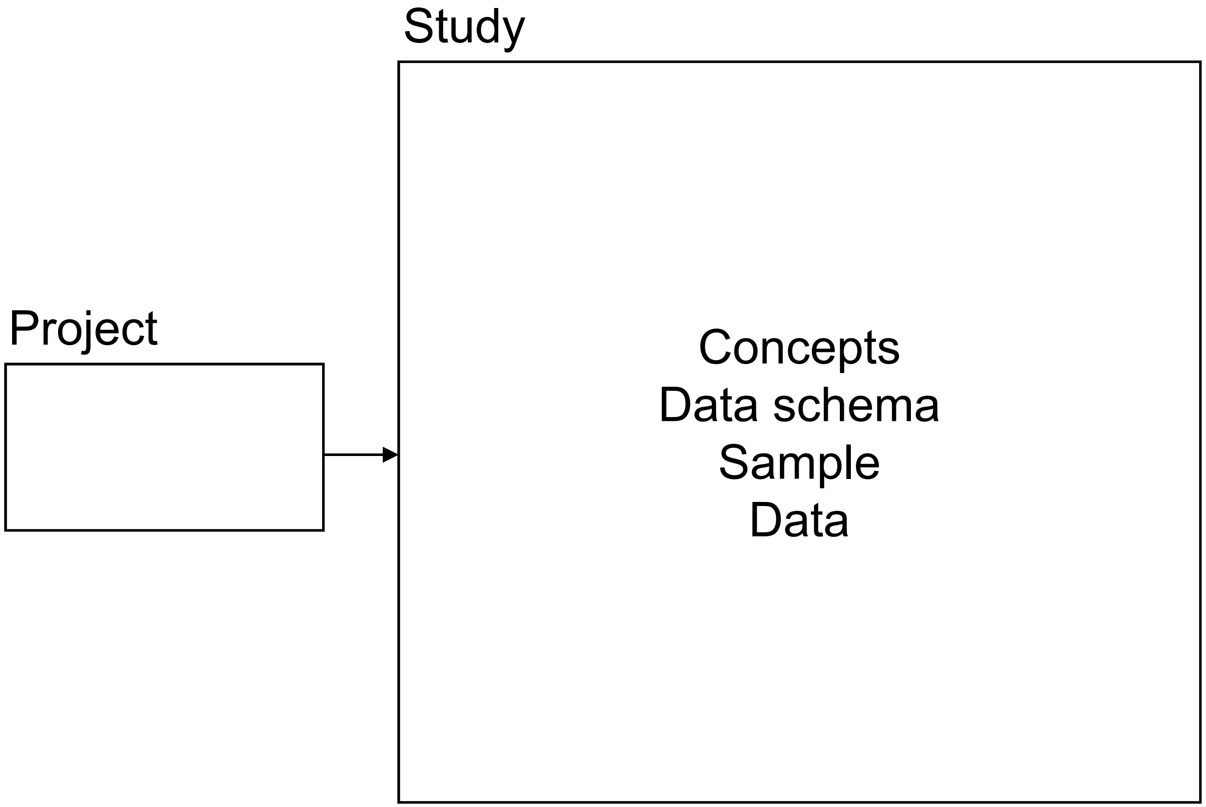

Study

Project

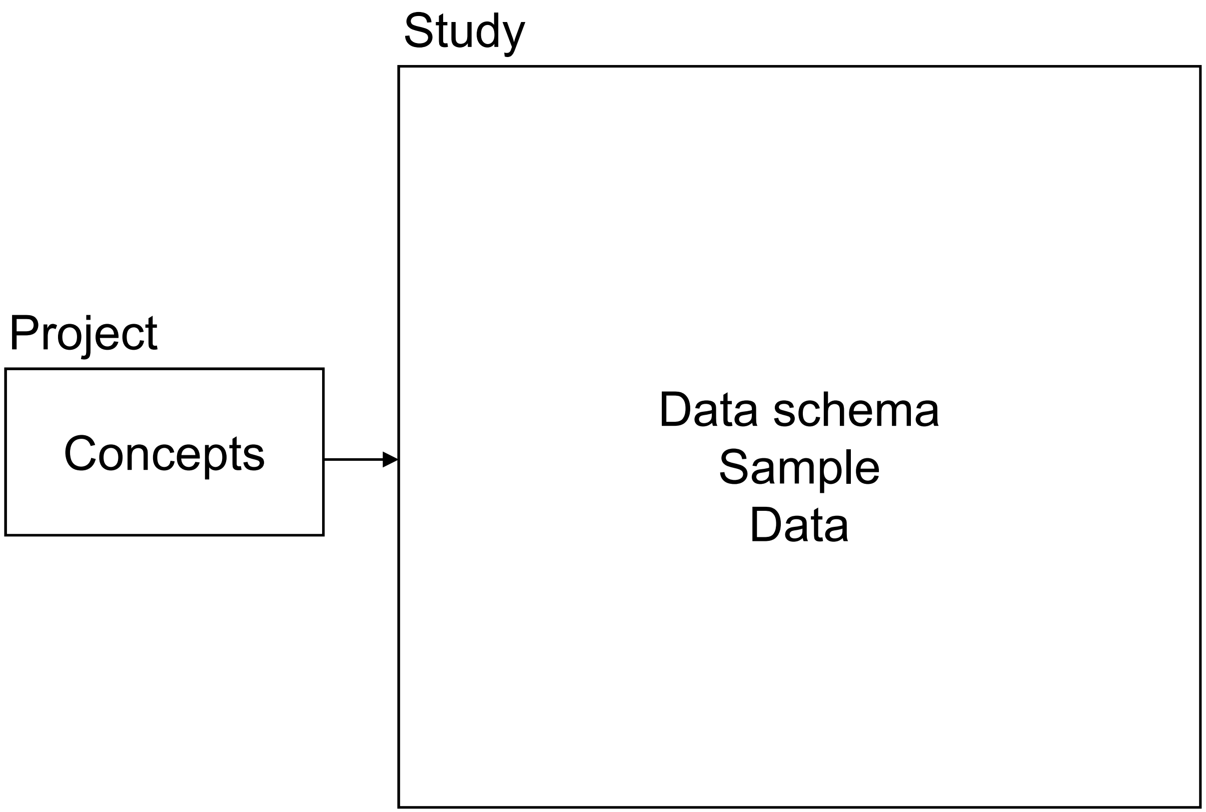
Funds
Authors

Concepts
Data schema
Sample
Data

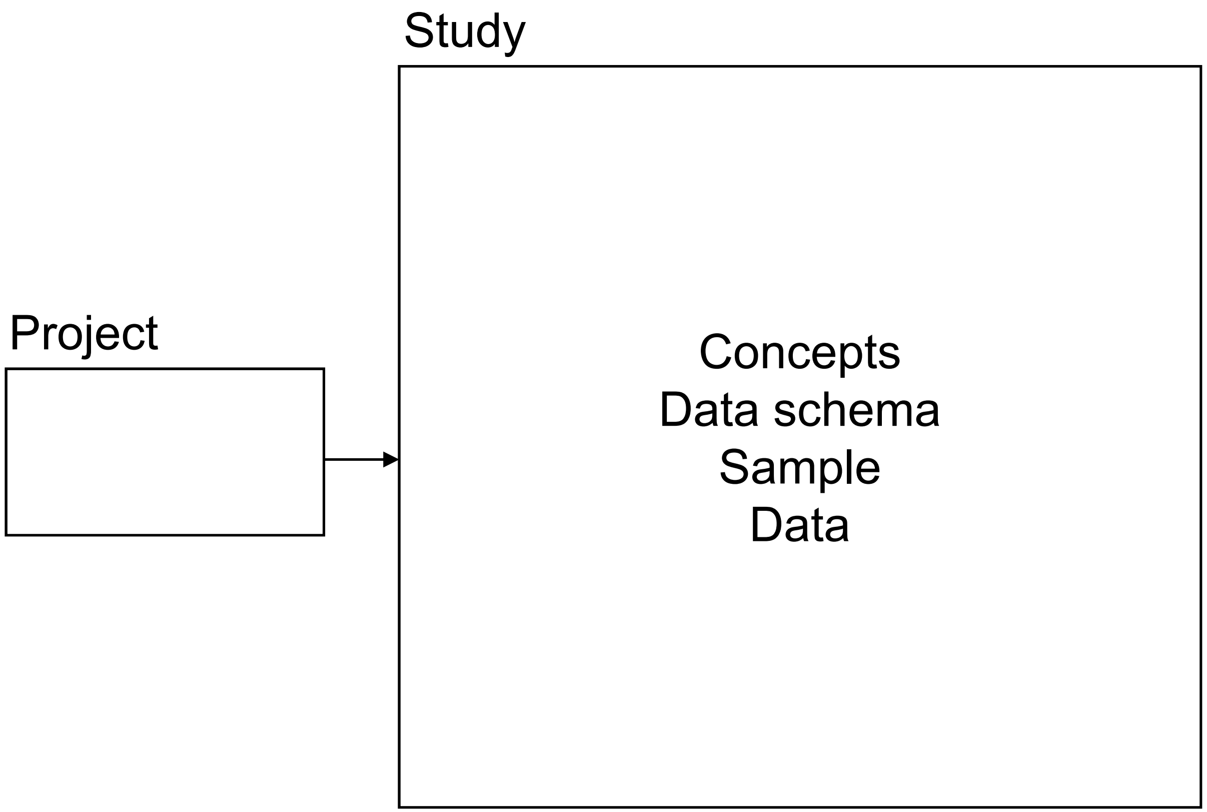
Let’s forget about them for the moment…



Maybe the Concepts should be migrated to the project; usually they are best expressed in terms of what researchers plan to do



Let this question open until we will have to distribute elements Among the main objects: here already project and Study, later perhaps more…

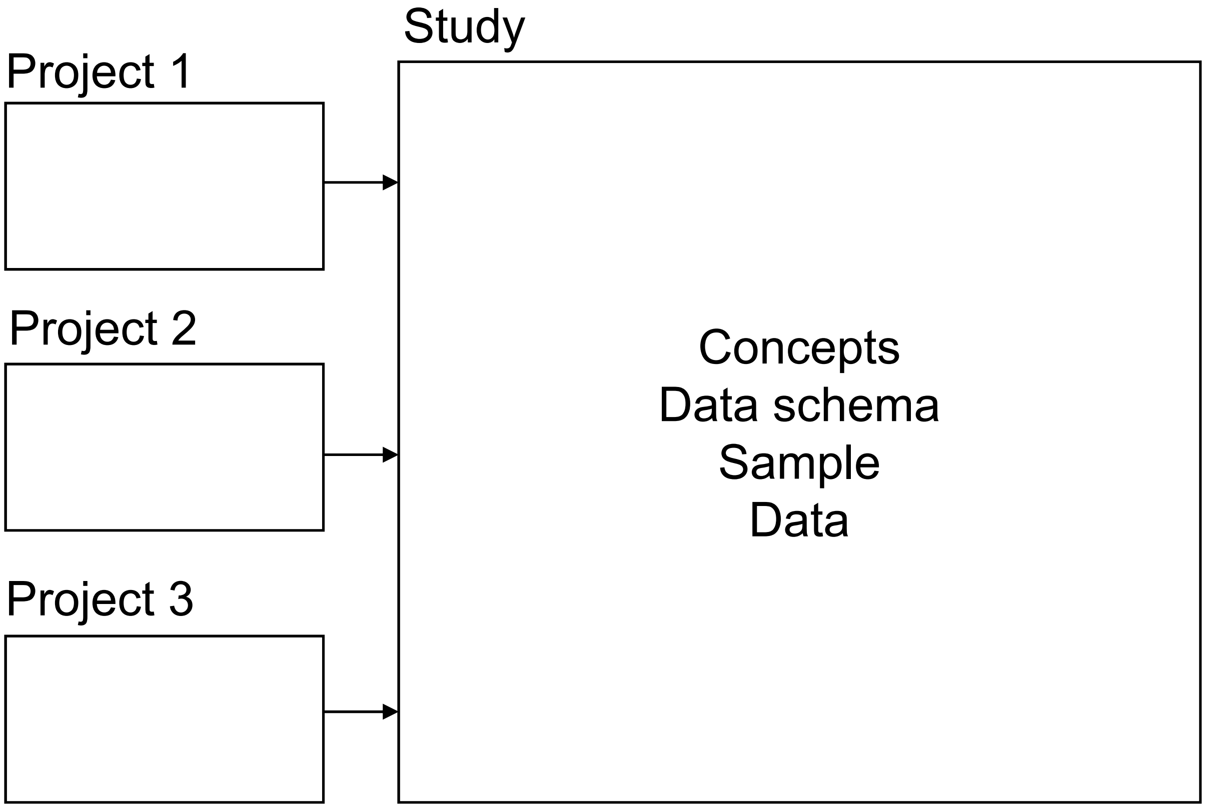


Study

Project

Concepts
Data schema
Sample
Data

A study can be funded successively under several research projects (association)



Study

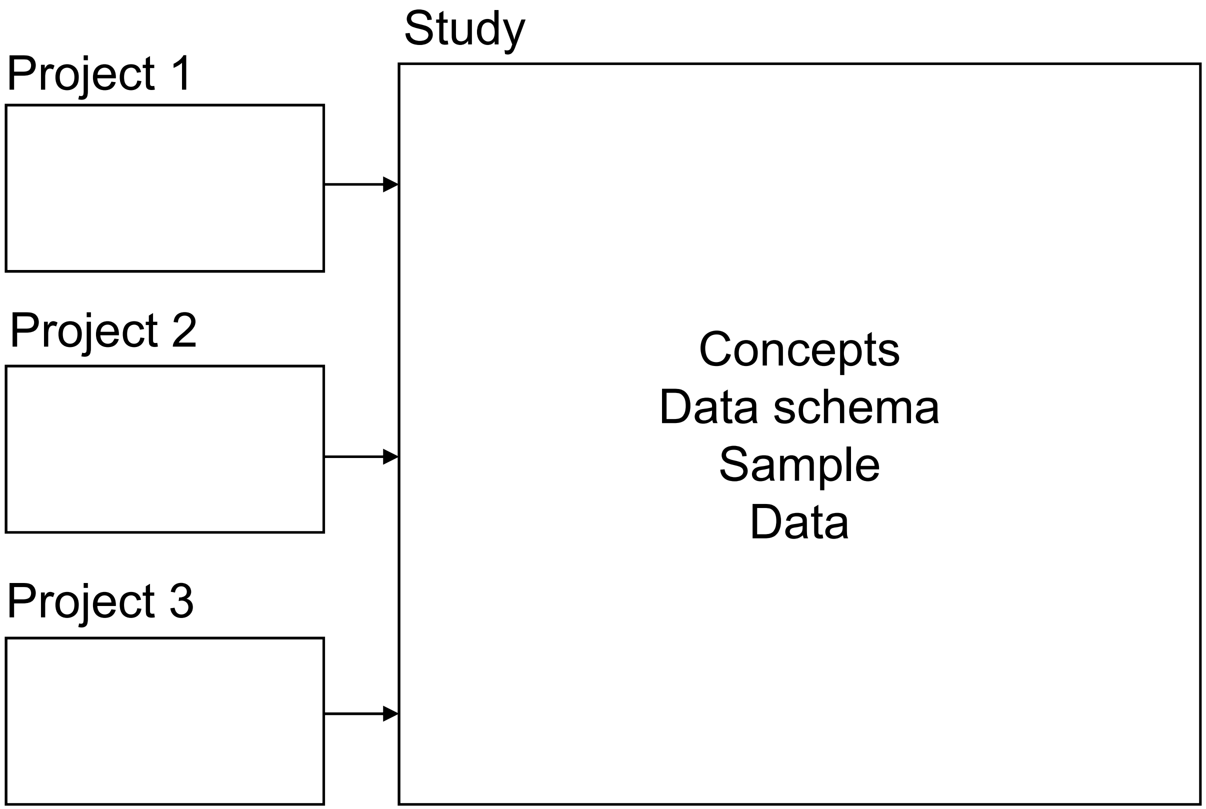
Project 1

Project 2

Project 3

Concepts
Data schema
Sample
Data

There are also activities related to the study, which depend on the stations the data go through over life:



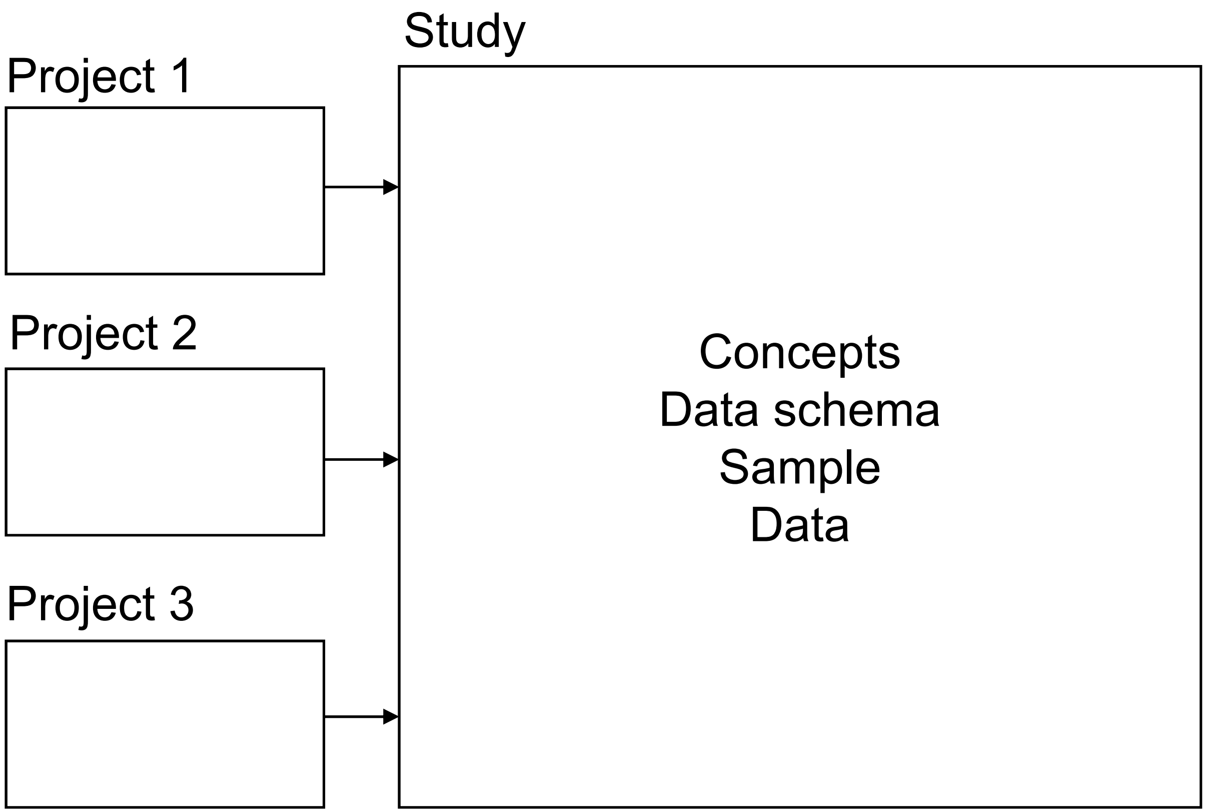
Activities:

Data collection
Data processing
Data publication
Data repurpousing
Data deposit
Data distribution
Etc.

Some repeated,
others not

Actors, Authority
Time, Content in
Various forms

Let’s forget the activities and turn back to the projects



Study

Project 1

Project 2

Project 3

Concepts
Data schema
Sample
Data

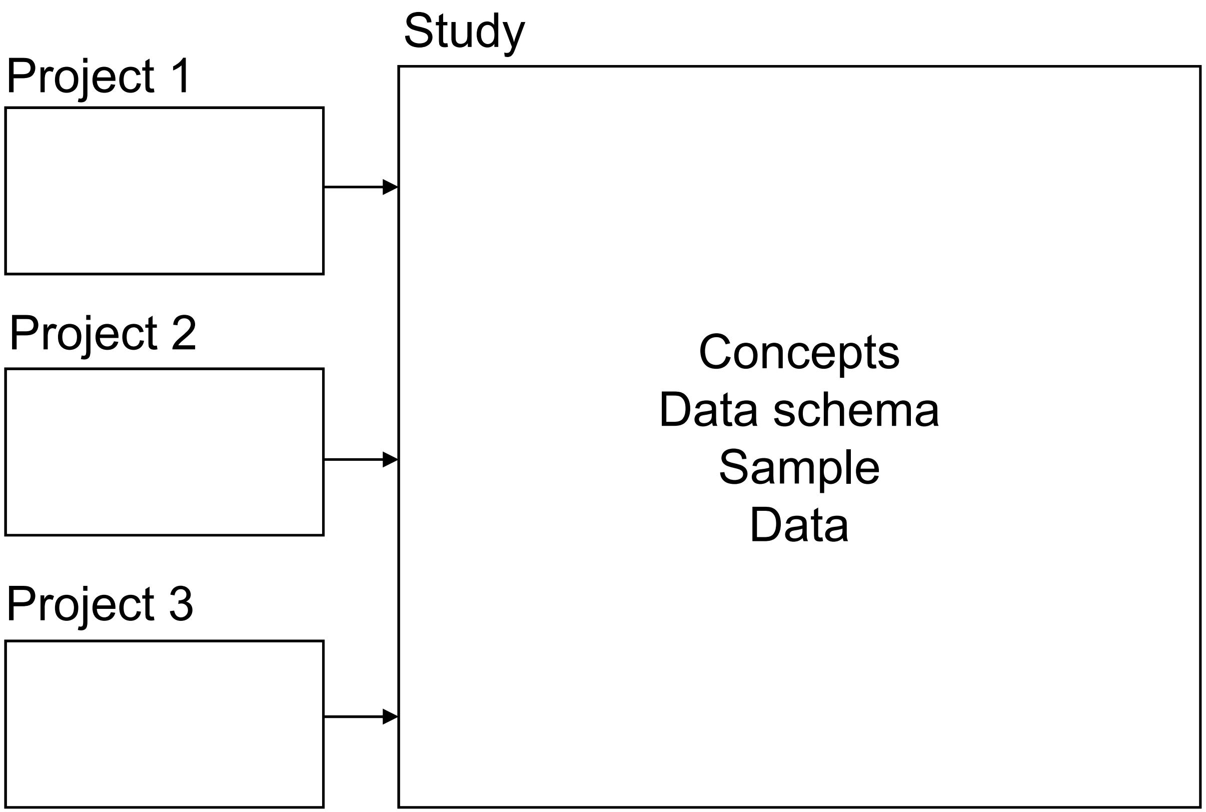
Activities:

Data collection
Data processing
Data publication
Data repurpousing
Data deposit
Data distribution
Etc.

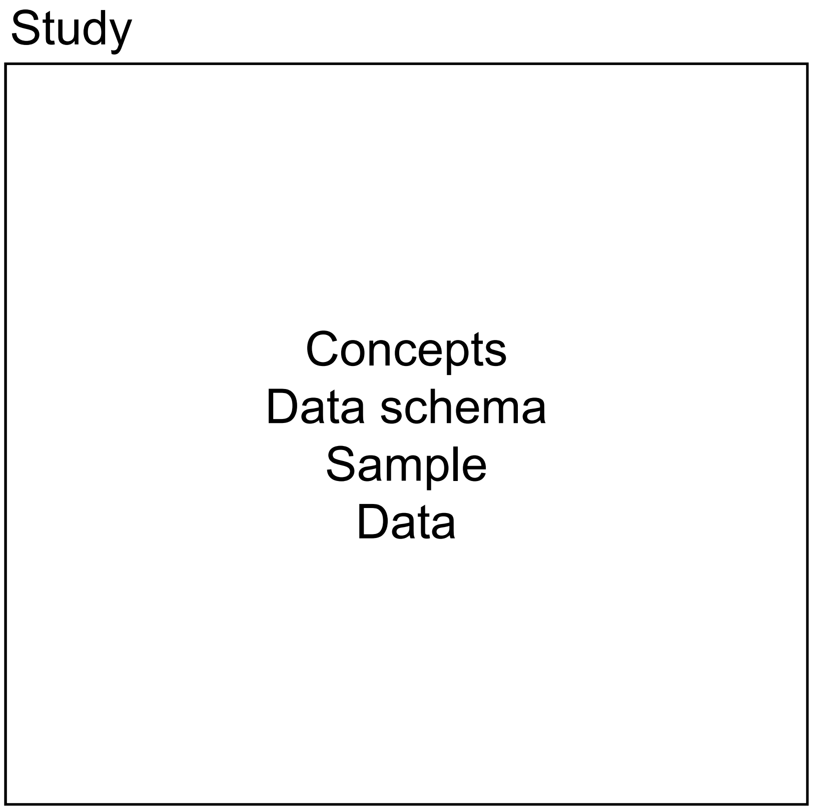
Some repeated,
others not

Actors, Authority
Time, Content in
Various forms

Let’s forget the activities and turn back to the projects



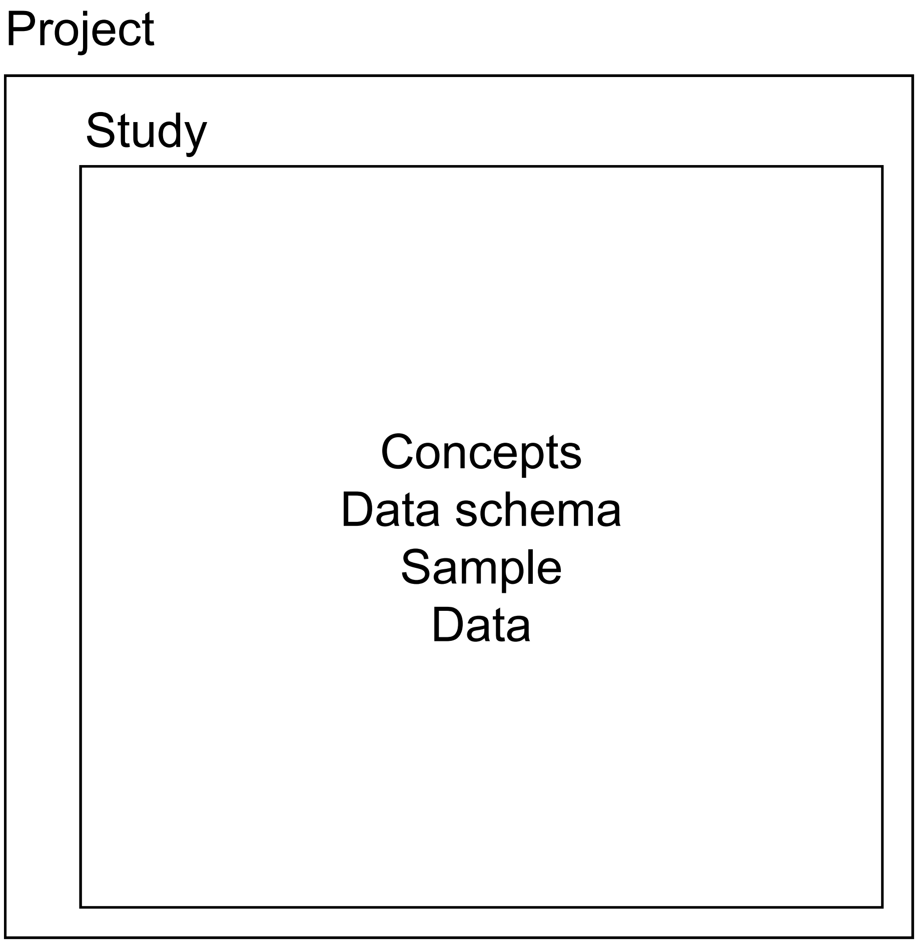
Let’s forget the activities and turn back to the projects



Study

Concepts
Data schema
Sample
Data

Most studies are actually produced within one single project:

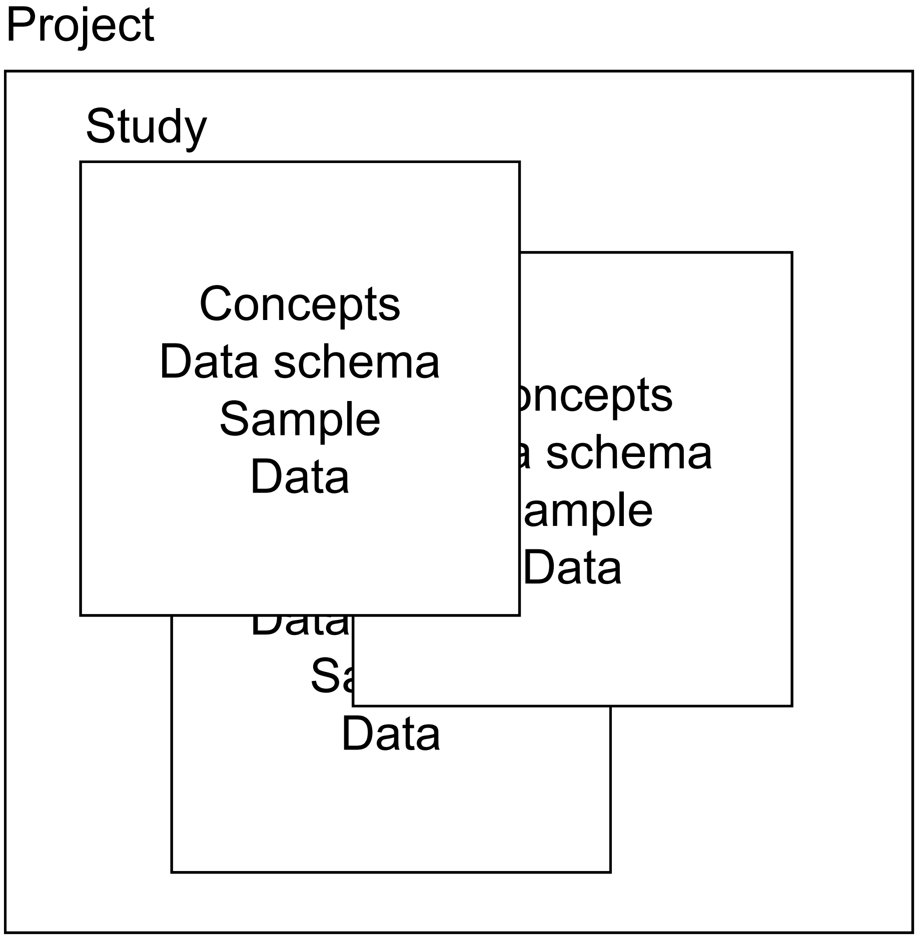


Project

Study

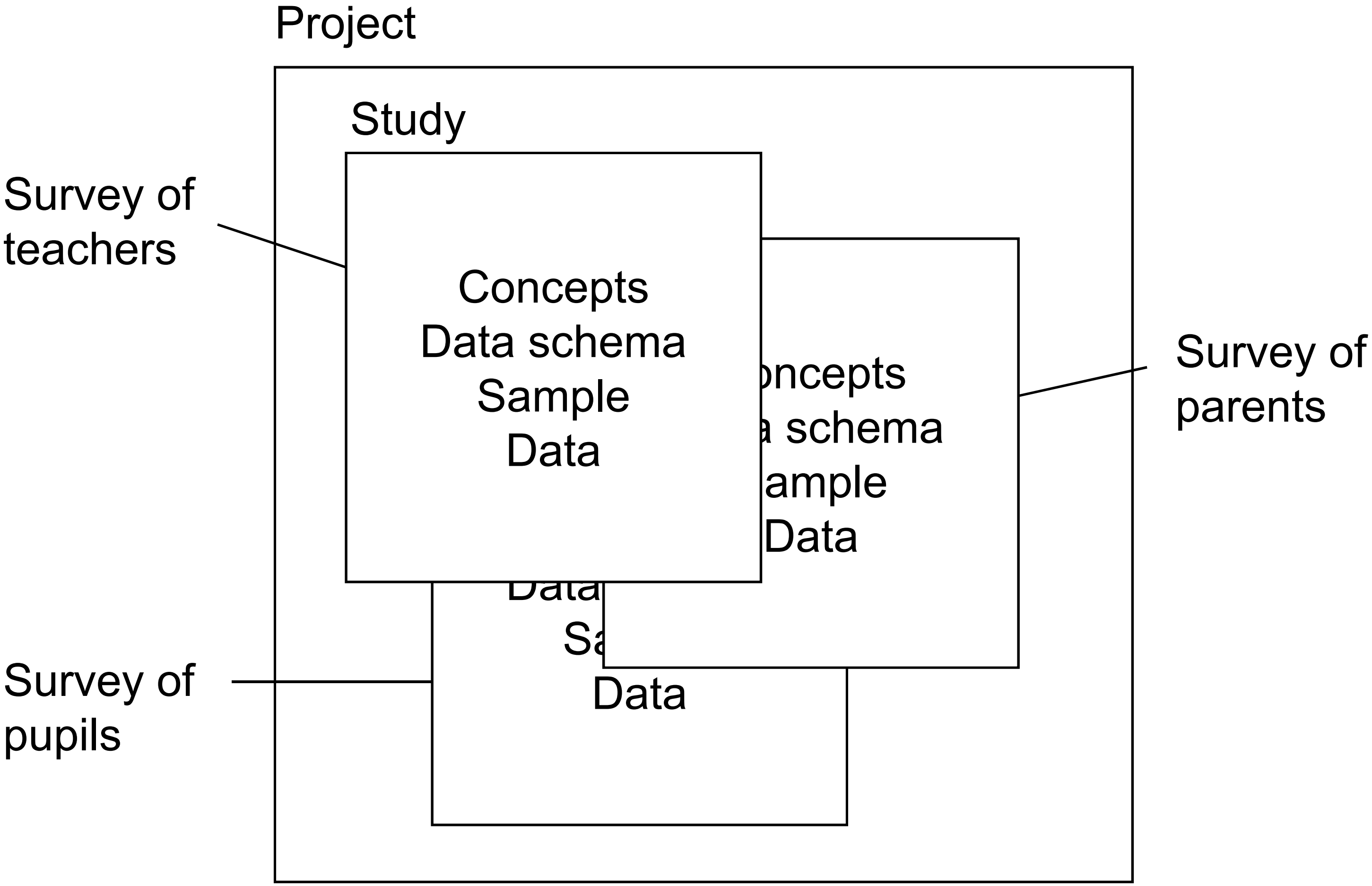
Concepts
Data schema
Sample
Data

In some research projects, you will actually fund more than one study



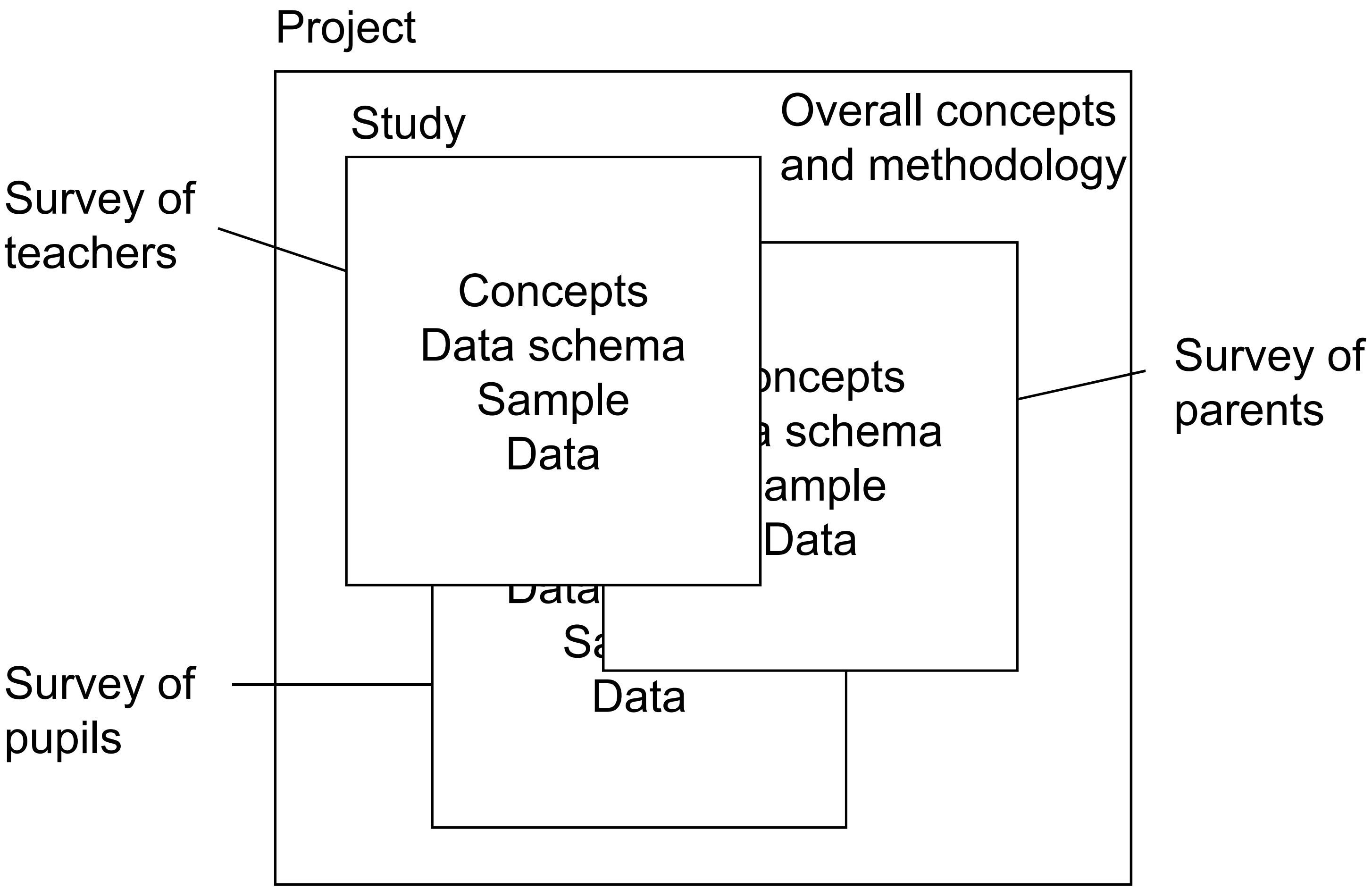
…so we need a good model for the project-study relationschip!

In some research projects, you will actually fund more than one study



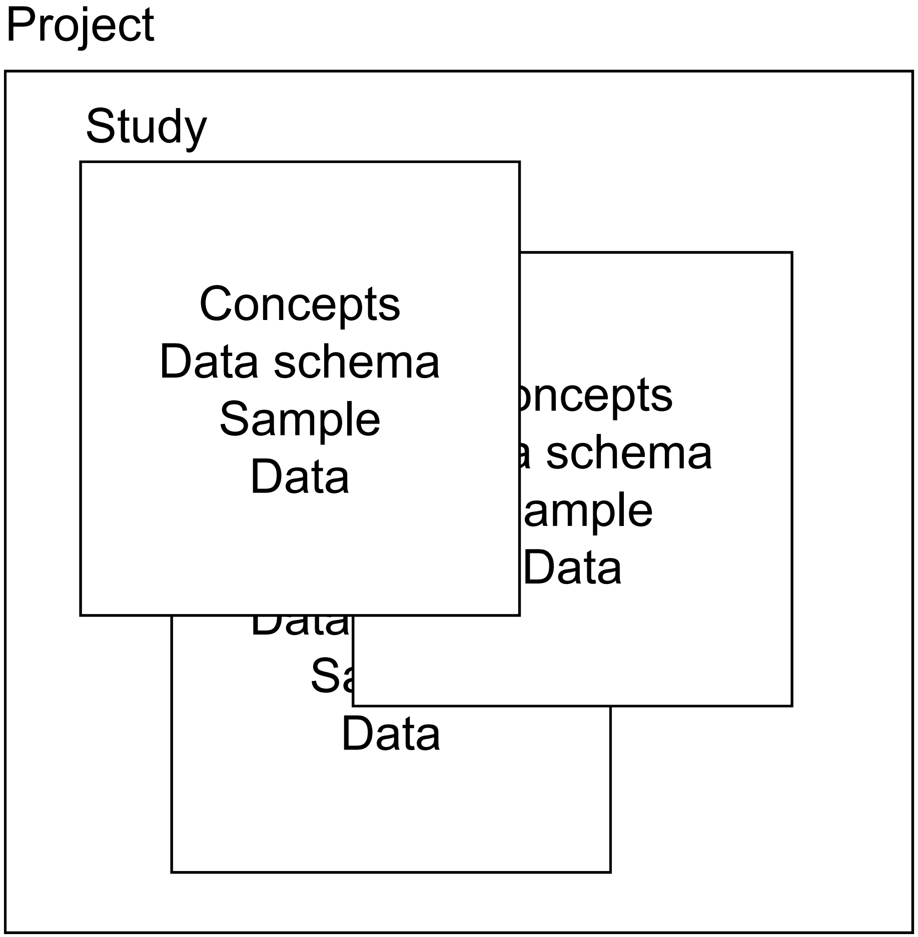
…so we need a good model for the project-study relationschip!

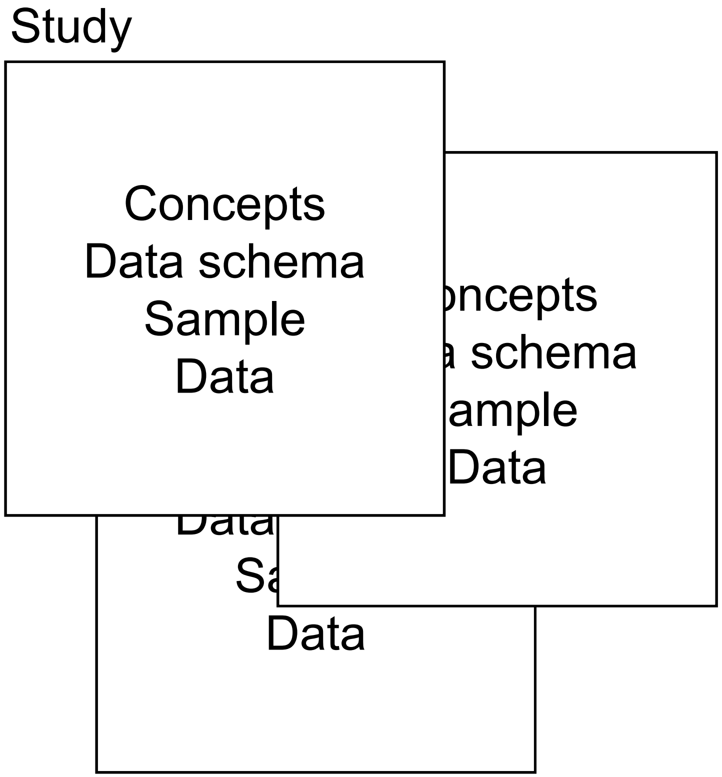
The overall concepts would be expressed on the level of the project



…so we need a good model for the project-study relationschip!

This is already a bit complicated, so let’s forget about the project..





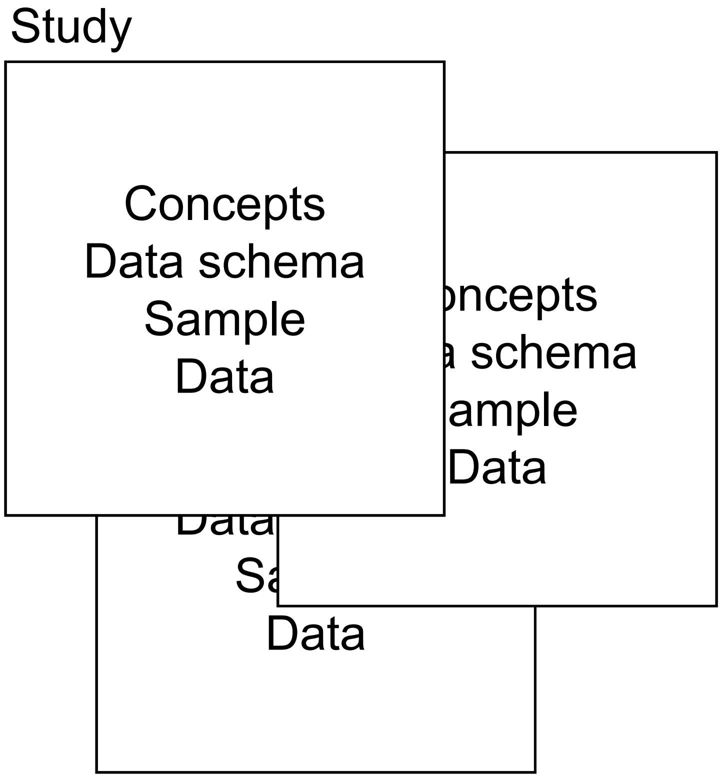
Study

Concepts
Data schema
Sample
Data

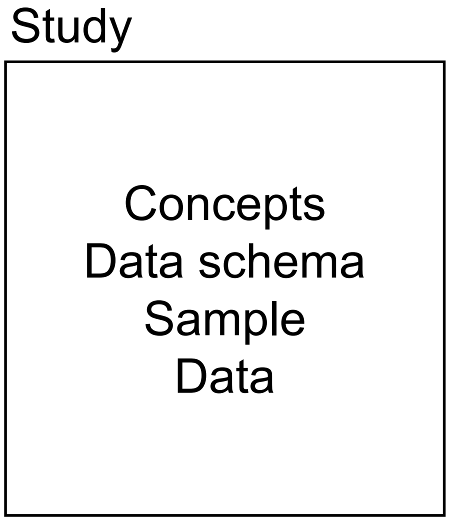
oncepts
a schema
ample
Data

Data
Sa
Data

Let's keep just one study…



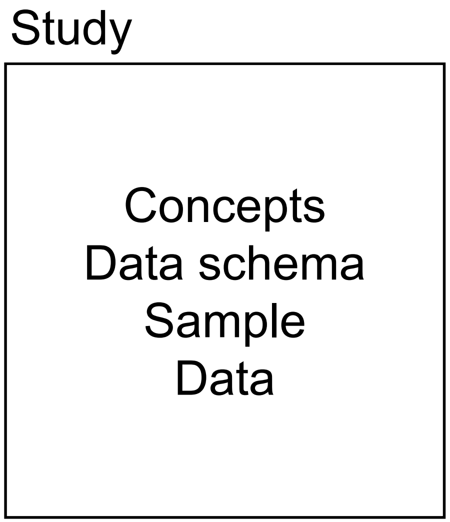
Let’s keep just one study…



Study

Concepts
Data schema
Sample
Data

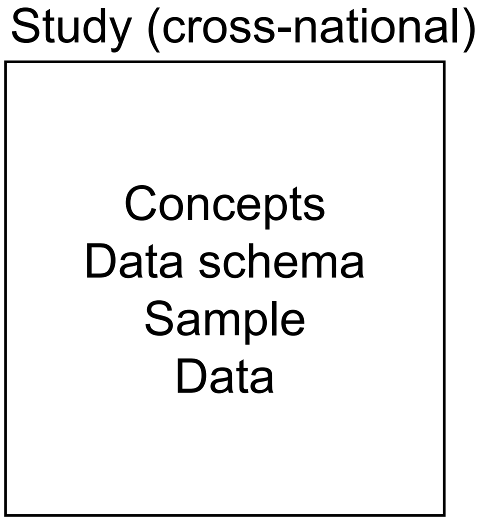
…but let it be a cross-national study:



Study

Concepts
Data schema
Sample
Data

…but let it be a cross-national study:



Study (cross-national)

Concepts
Data schema
Sample
Data

…we need more space!

…but let it be a cross-national study:

Cross-national study

Concepts
Data schema
Sample
Data

What happens to the contents?

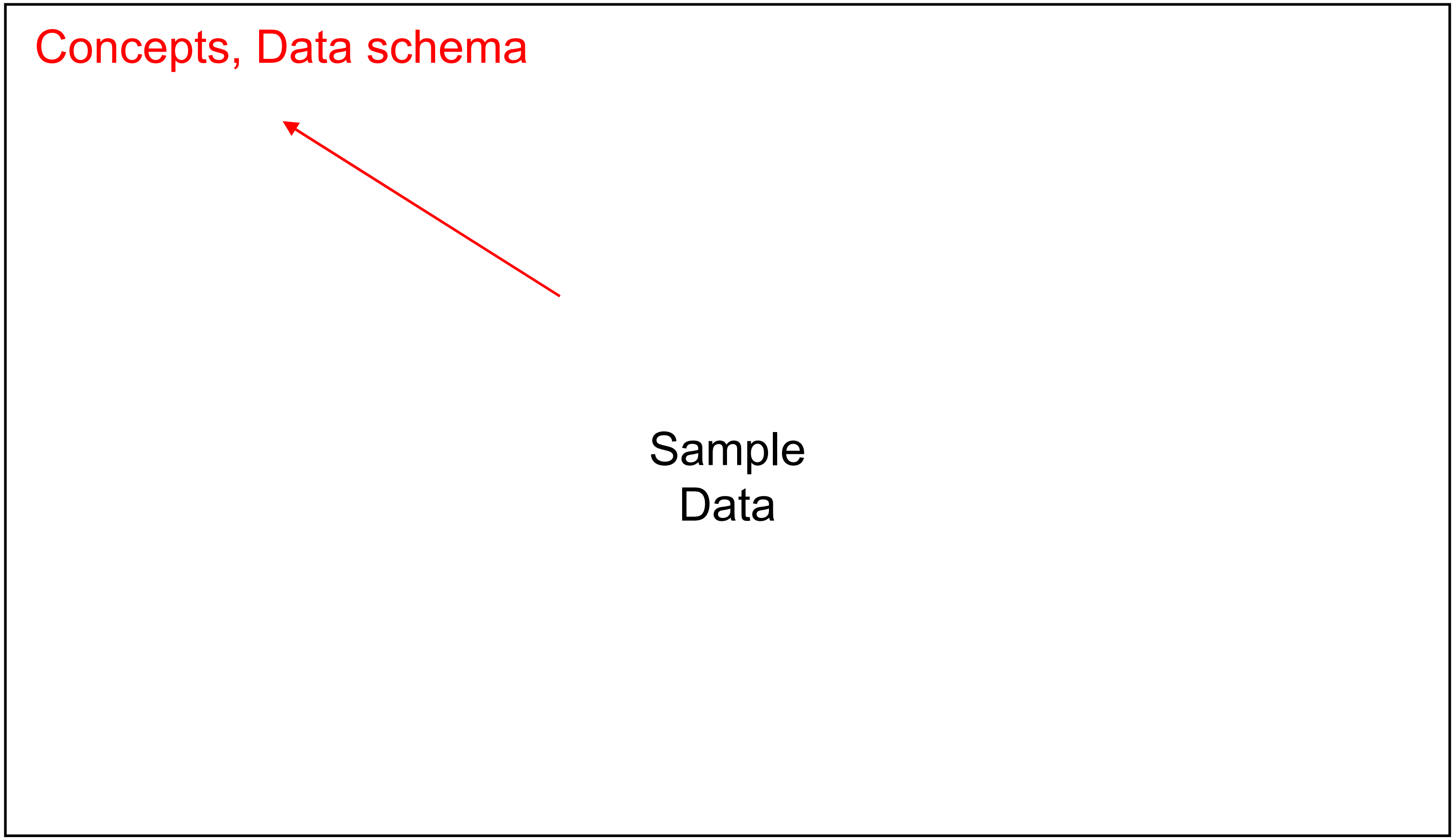
Cross-national study

?

Concepts
Data schema
Sample
Data

Concepts and data schema belong to the Study

Cross-national study



Concepts, Data schema

Sample
Data

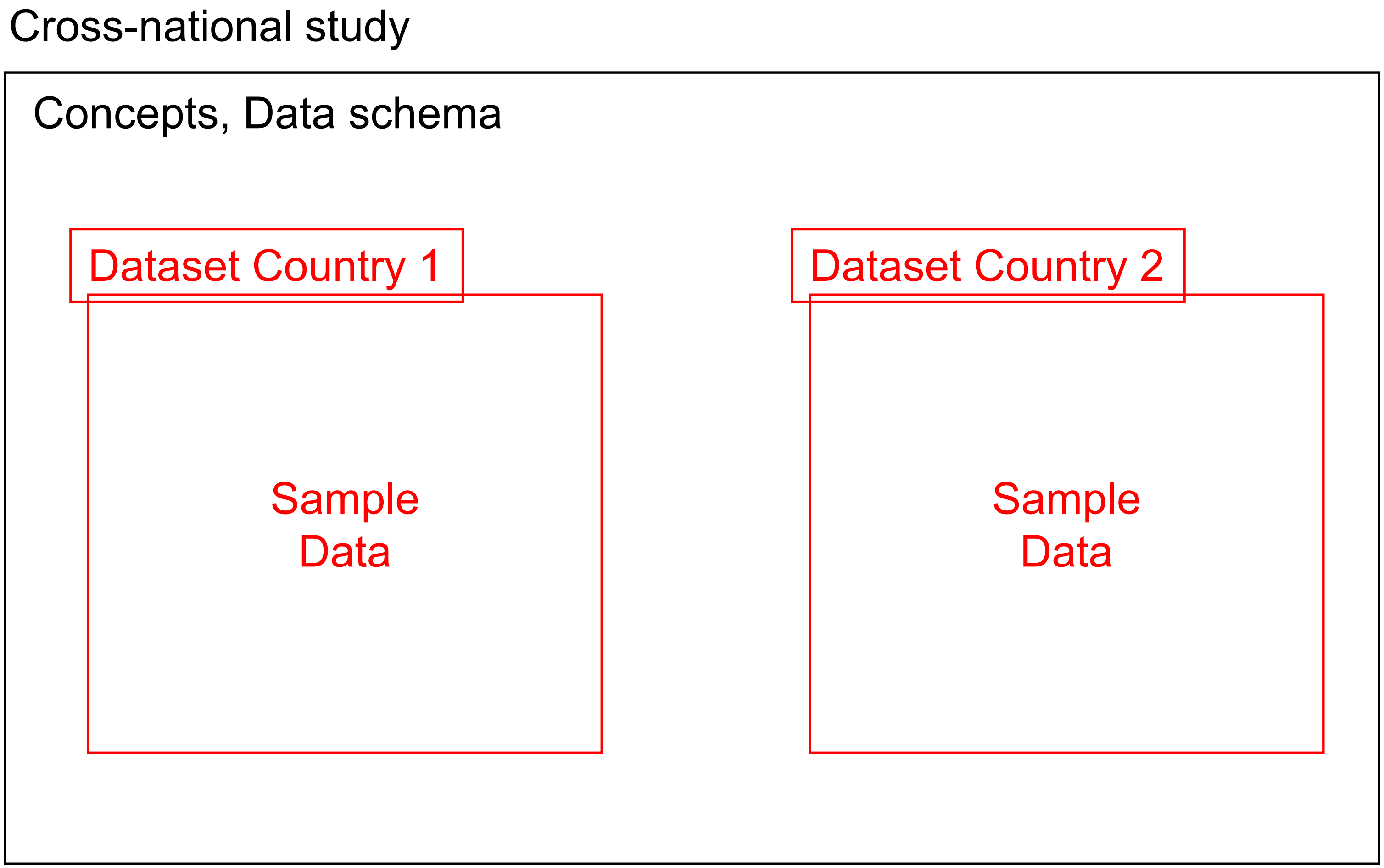
In a cross-national study, on compares sets of data collected on distinct samples, specific to each country

Cross-national study

Concepts, Data schema

Sample
Data

Fine. Let’s call those sets of data ‘datasets’



Cross-national study

Concepts, Data schema

Dataset Country 1

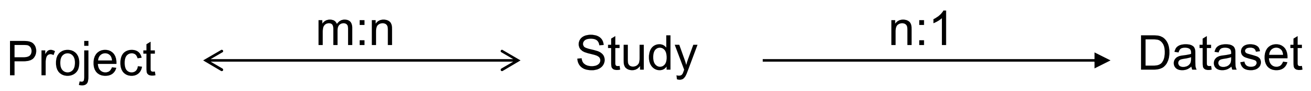
Sample Data

Dataset Country 2

Sample Data

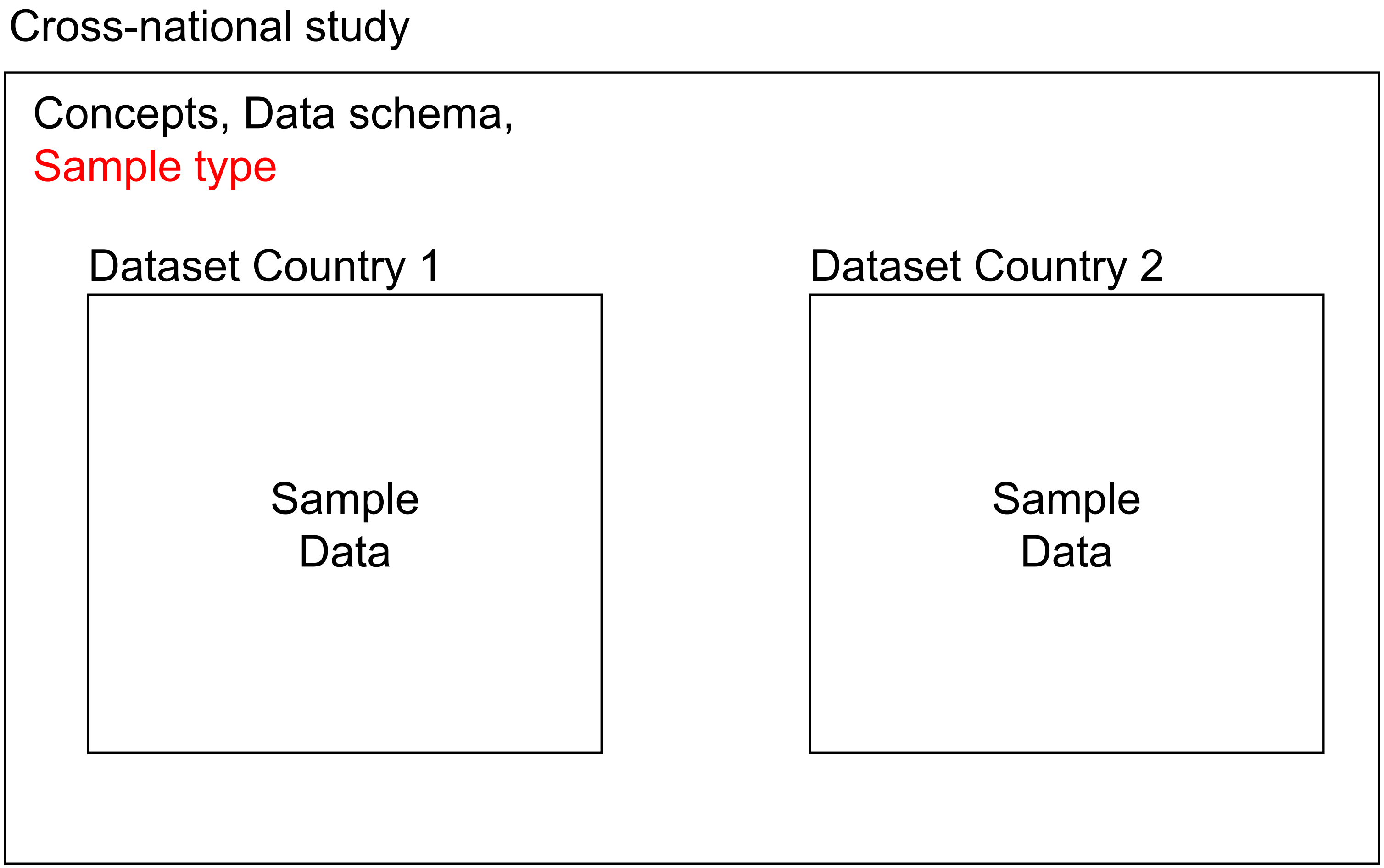
Now, the samples are distinct, but must be similar:

We have starte with the study, say, the DDI-study, and we have now three main objects, which must be desinguished because of the more complex relationships in a complex study:



Project m:n Study n:1 Dataset

So we need a sample type on study level, which prescribes what the samples should be:



Cross-national study

Concepts, Data schema,
Sample type

Dataset Country 1

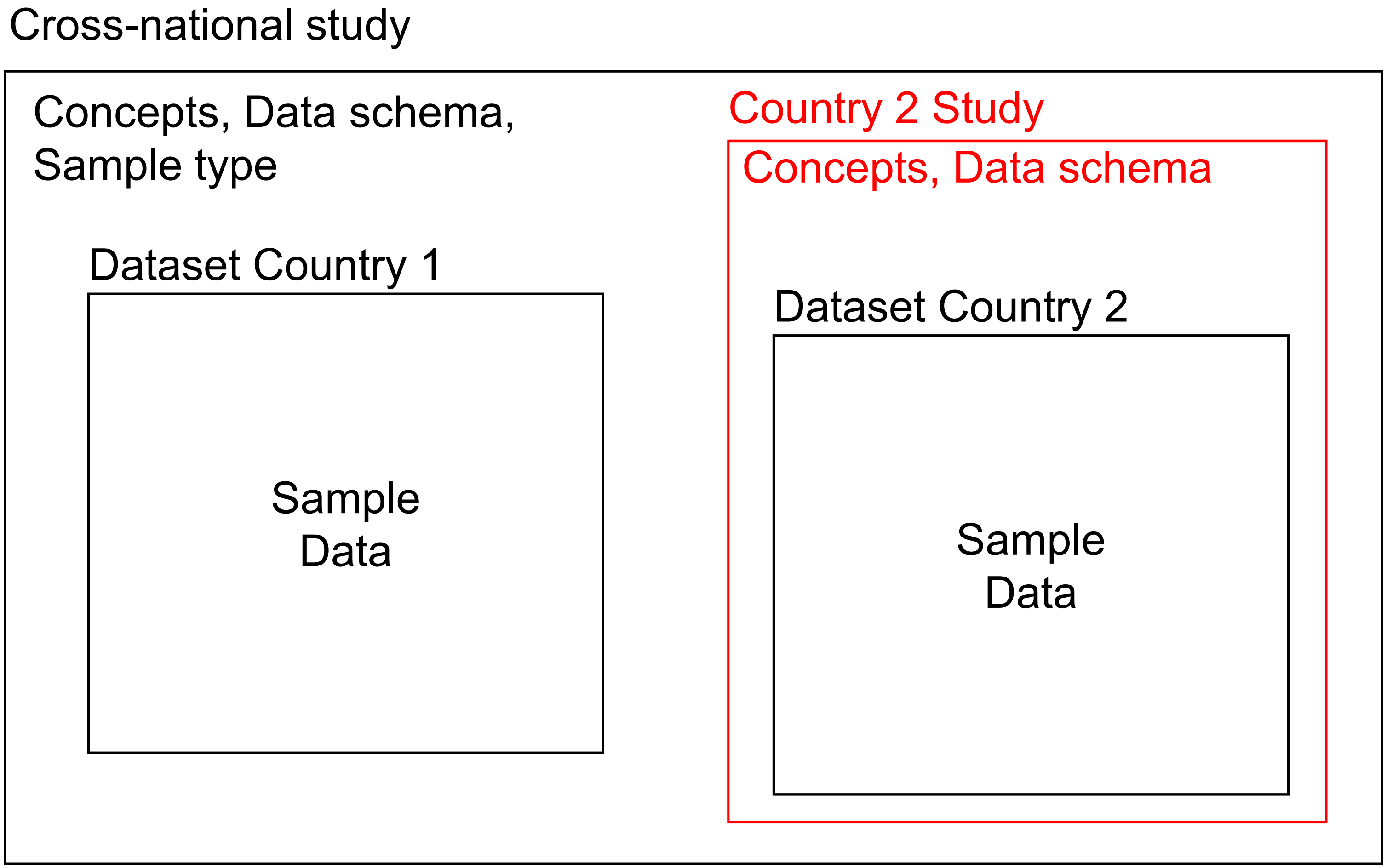
Sample
Data

Dataset Country 2

Sample
Data

Will country 2 really strictly conform to the data schema?

Well… more or less; sometimes less. They have their own view:



Cross-national study

Concepts, Data schema,
Sample type

Dataset Country 1

Sample
Data

Country 2 Study

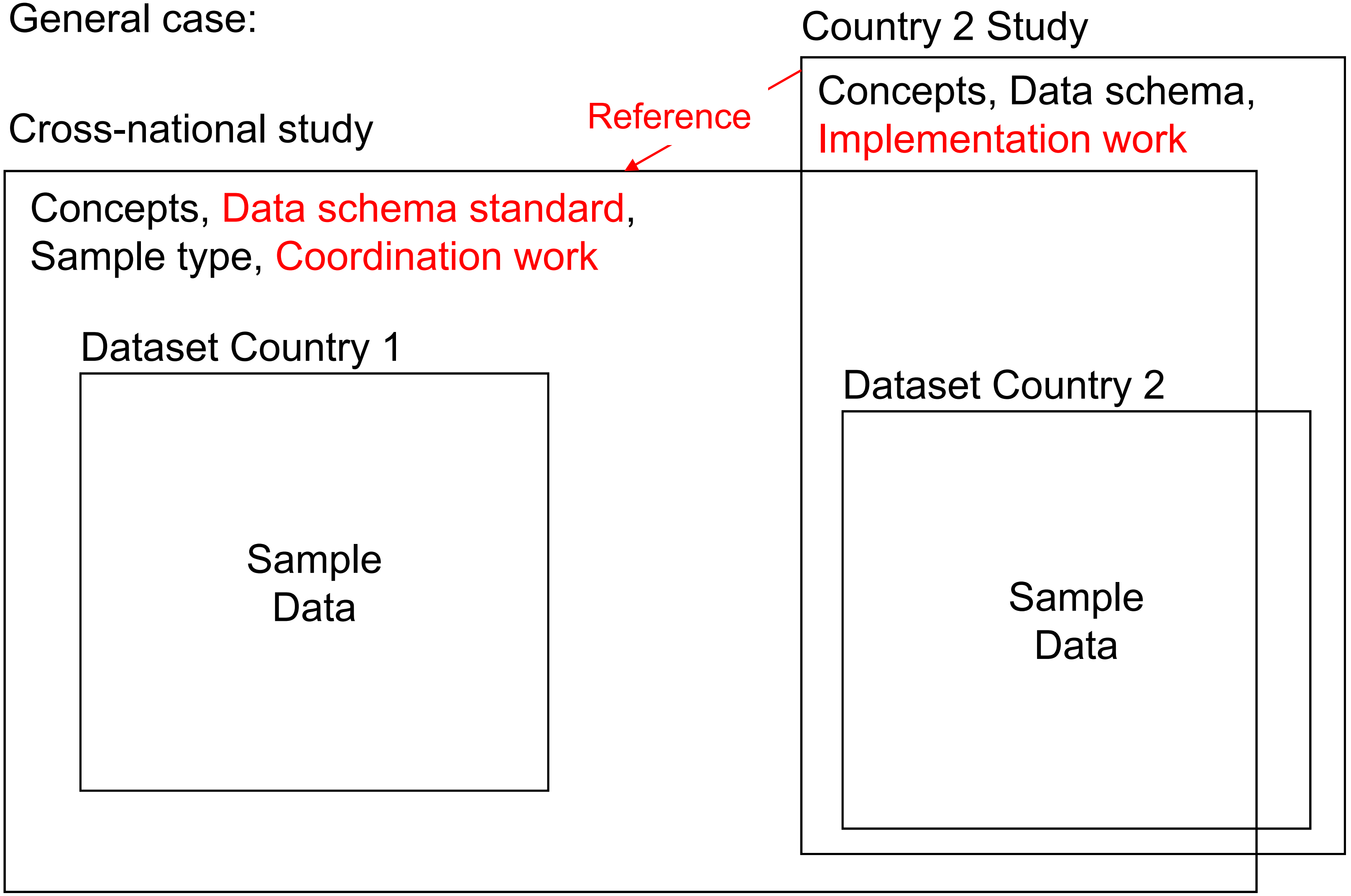
Concepts, Data schema

Dataset Country 2

Sample
Data

…which just overlaps with the overall data schema, so…

General case:



Country 2 Study

Concepts, Data schema, Implementation work

Reference

Cross-national study

Concepts, Data schema standard, Sample type, Coordination work

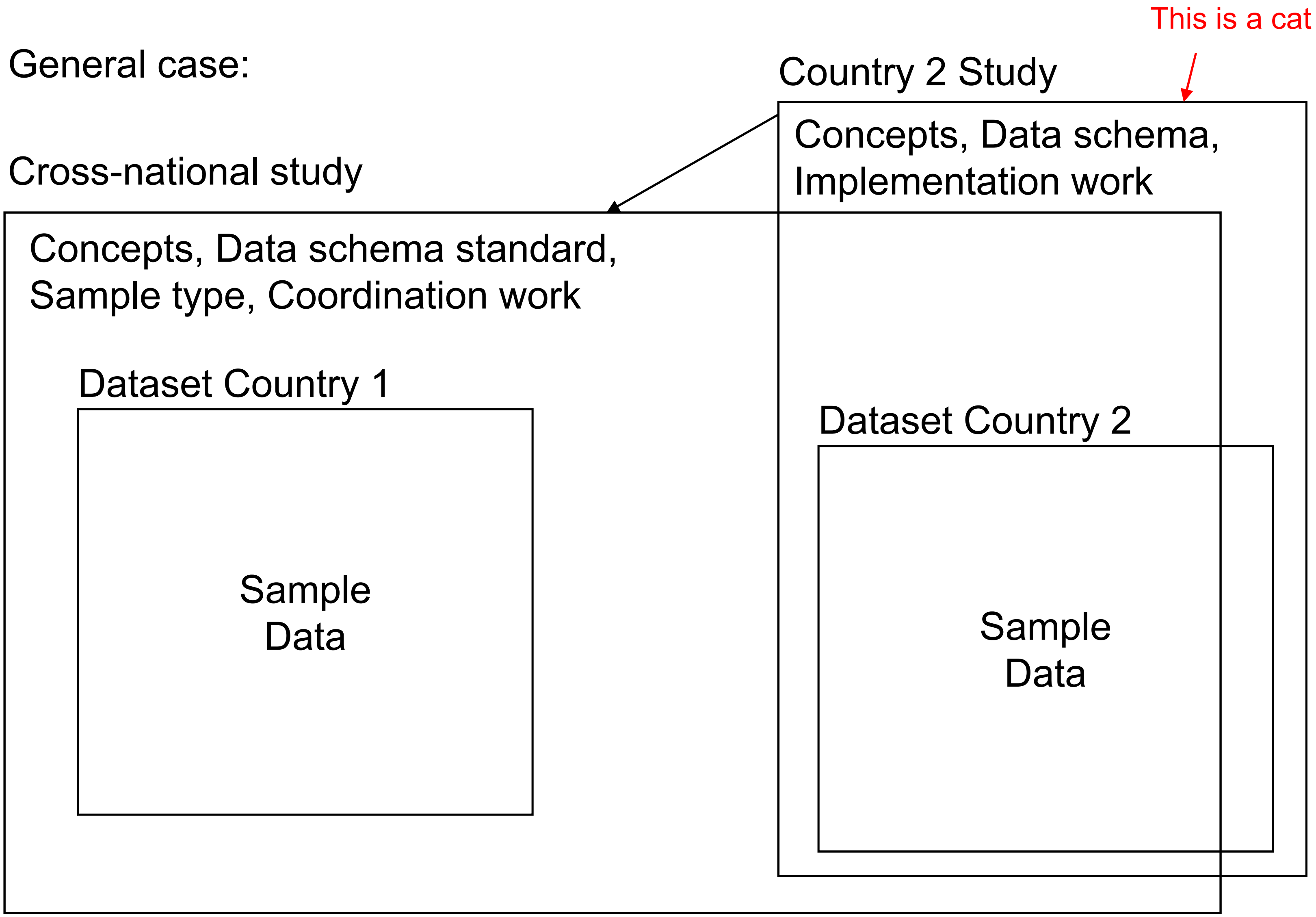
Dataset Country 1

Sample Data

Dataset Country 2

Sample Data

General case:



This is a cat

Country 2 Study

Concepts, Data schema, Implementation work

Cross-national study

Concepts, Data schema standard, Sample type, Coordination work

Dataset Country 1

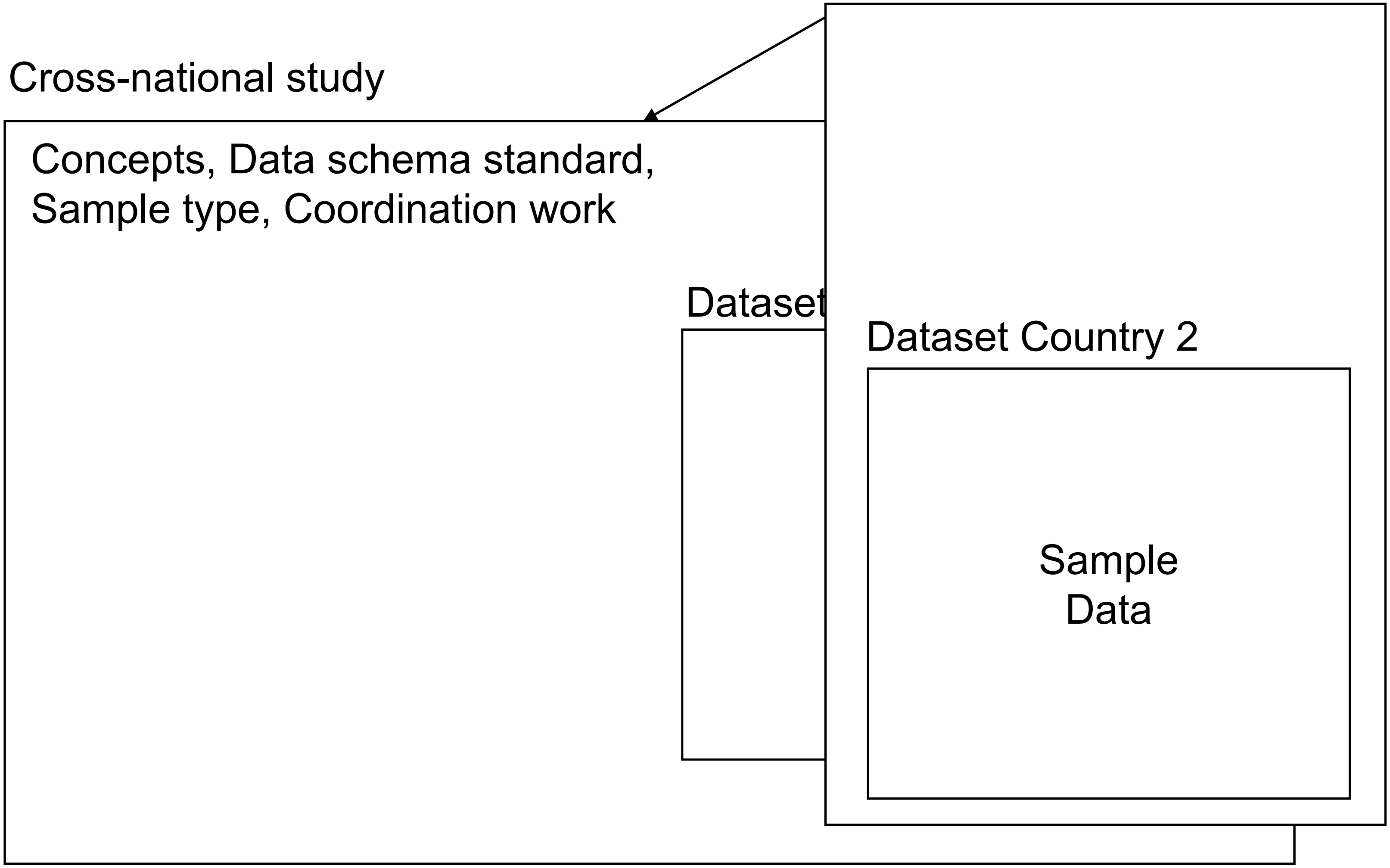
Sample Data

Dataset Country 2

Sample Data

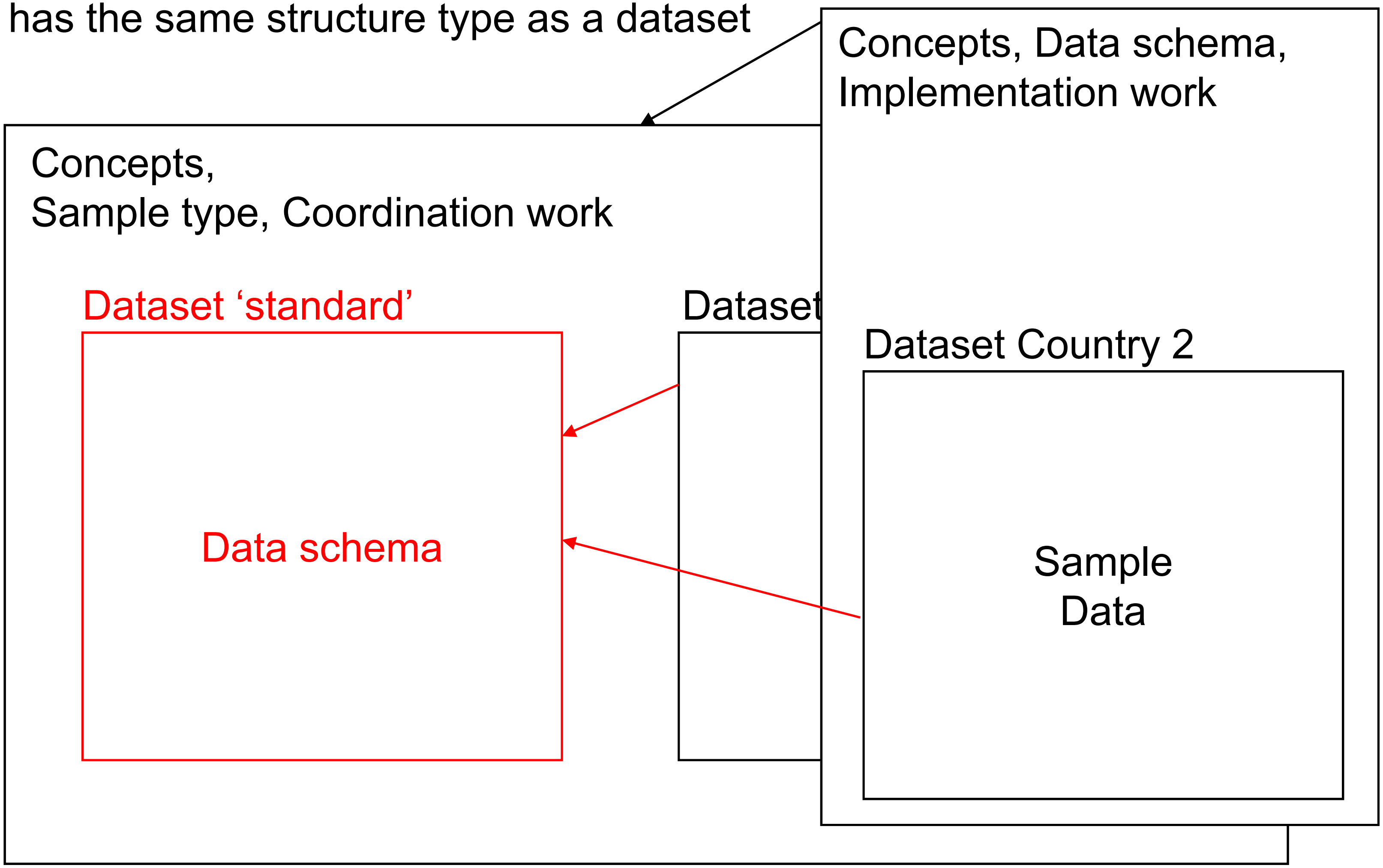
Represent the data schema standard as a dataset:

Clean some space to put the standard:



Represent the data schema standard as a dataset:

The definition of the standard data schema has the same structure type as a dataset



Concepts, Data schema, Implementation work

Concepts,
Sample type, Coordination work

Dataset ‘standard’

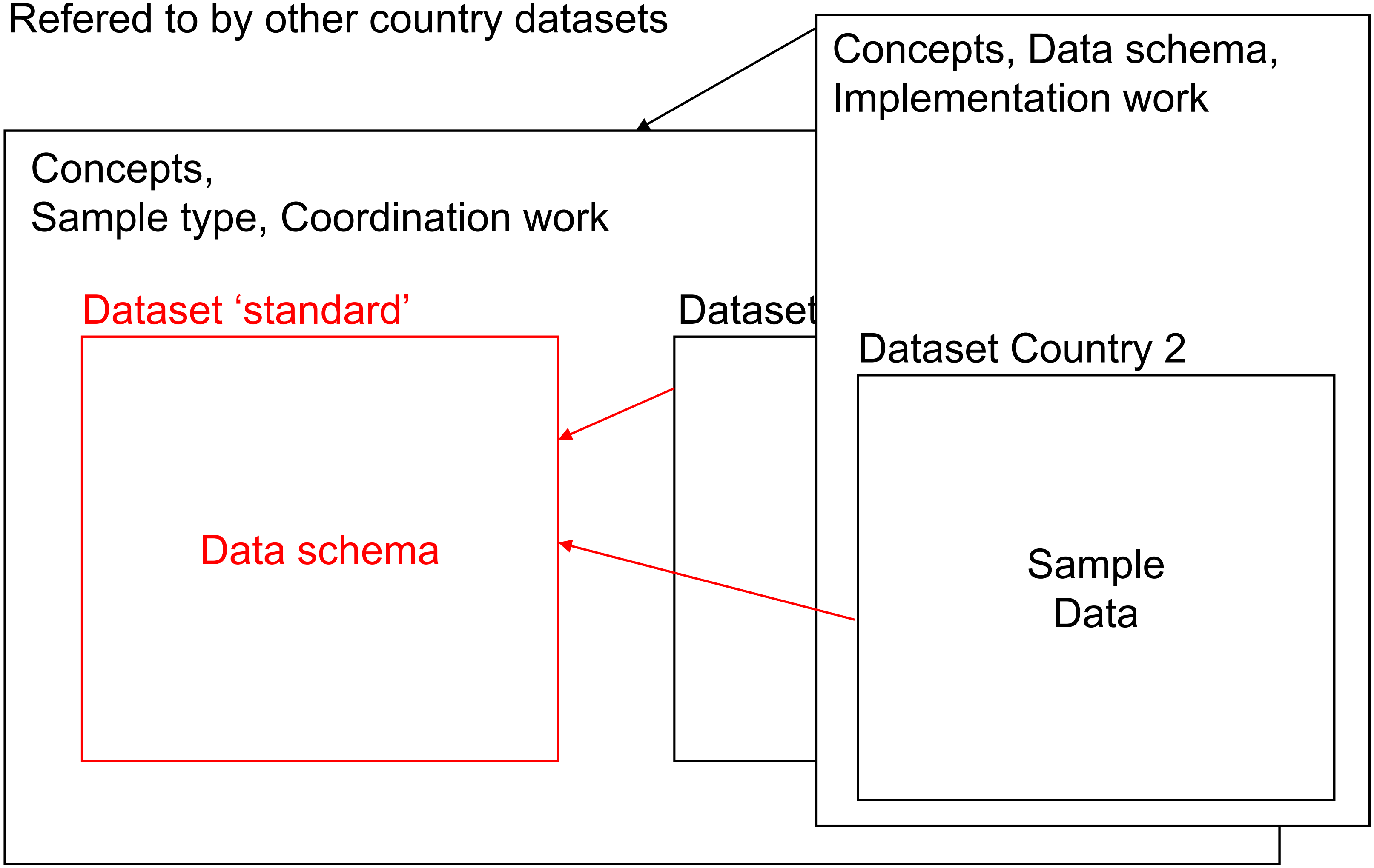
Data schema

Dataset

Dataset Country 2

Sample
Data

Characteristics:
Country name = « Standard »
Refered to by other country datasets



Concepts, Data schema, Implementation work

Concepts,
Sample type, Coordination work

Dataset ‘standard’

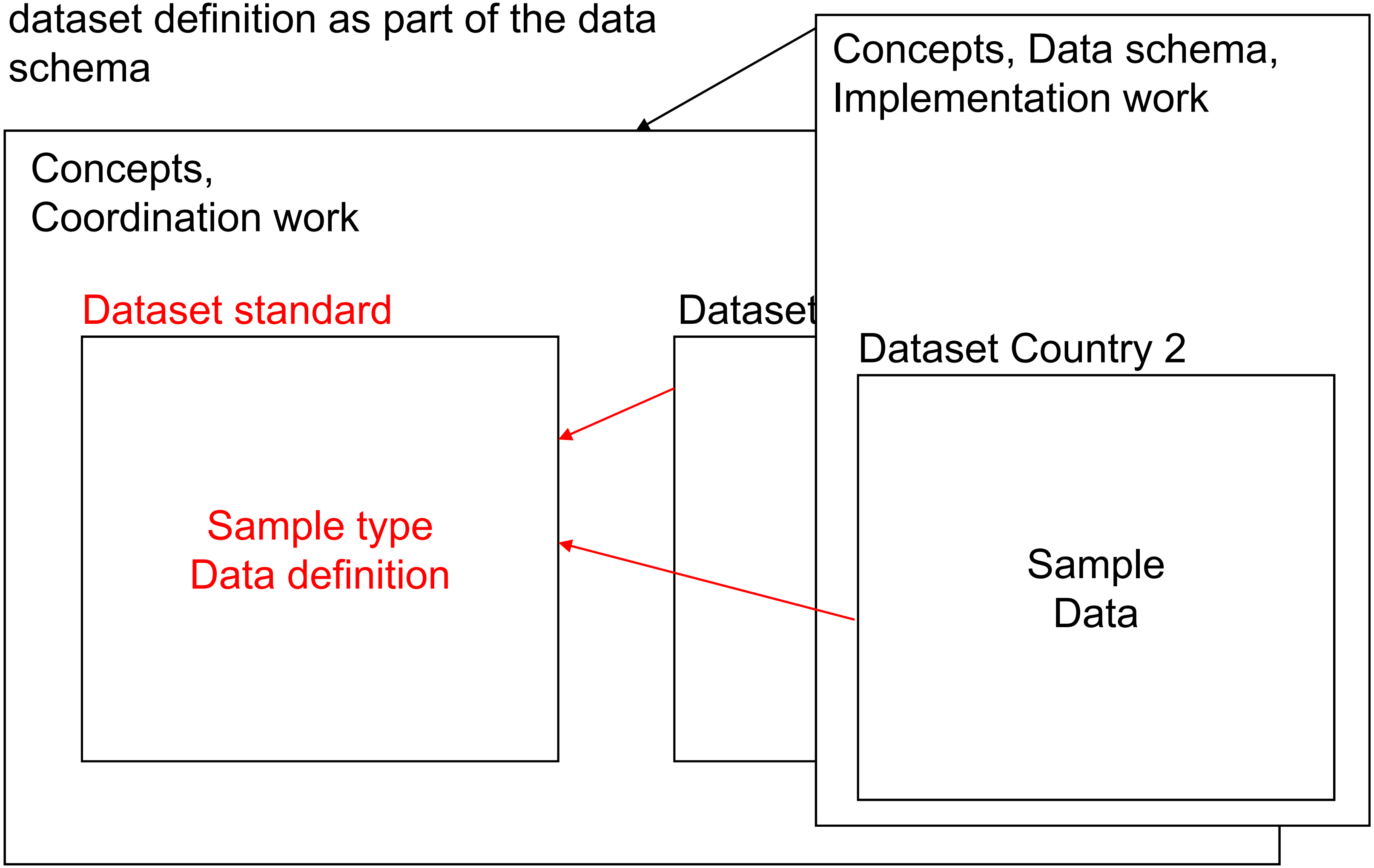
Data schema

Dataset

Dataset Country 2

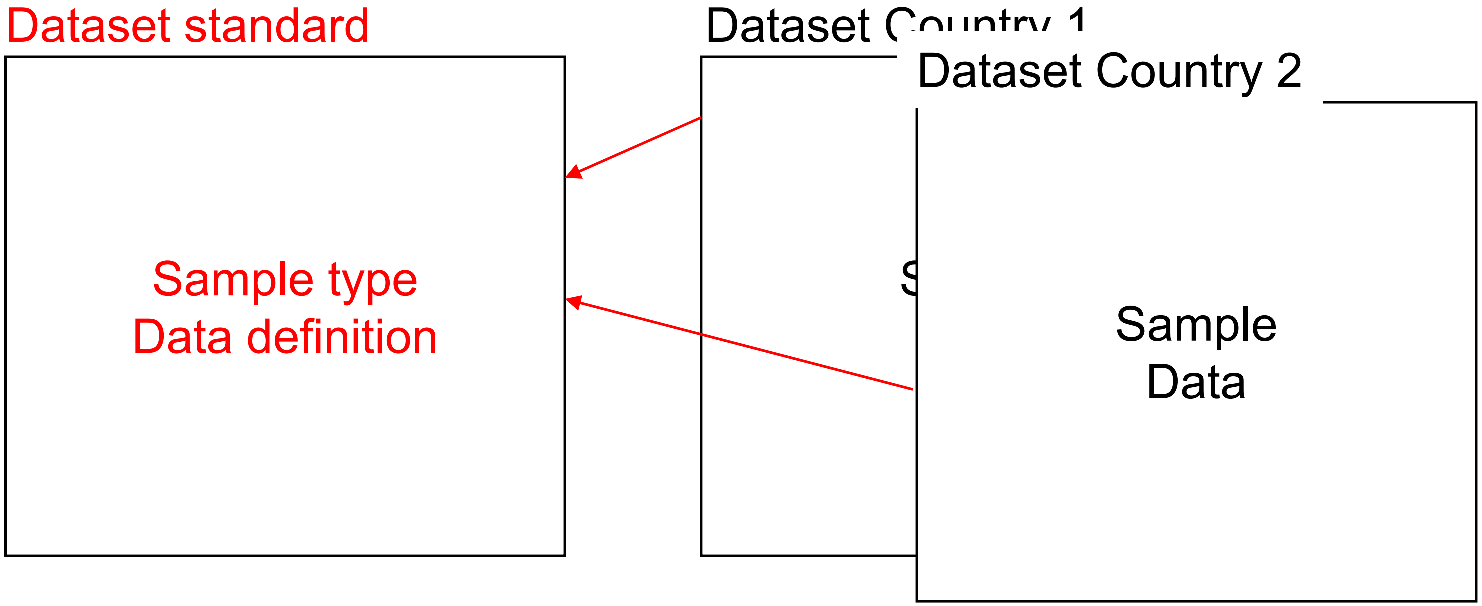
Sample
Data

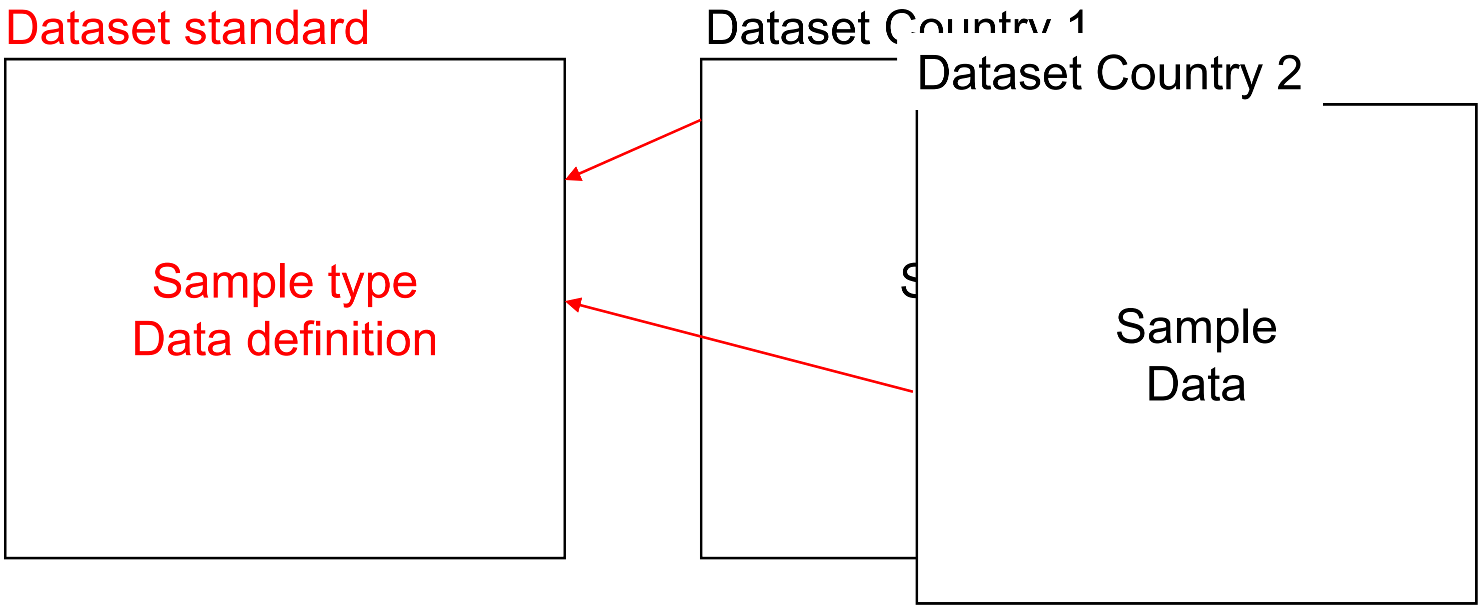
The sample type, previously defined on study level, migrates to the standard dataset definition as part of the data schema



Now, we still have to go deeper into it… forget of the studies!

The definition of the standard data schema
Has the same structure type as a dataset





Dataset standard

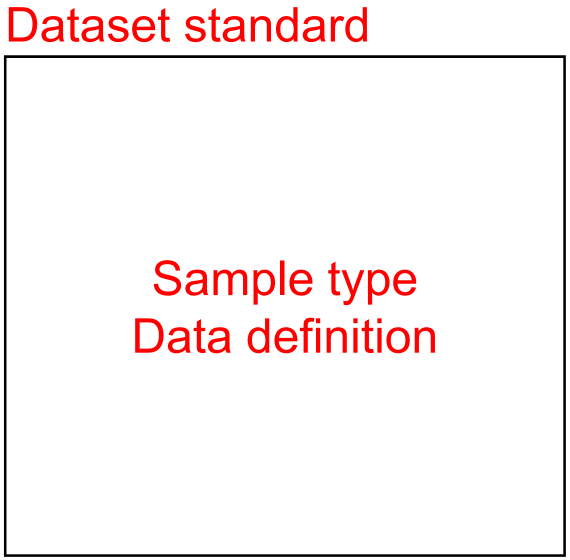
Sample type
Data definition

Dataset Country 1

Dataset Country 2

Sample
Data

Forget even of the country datasets….

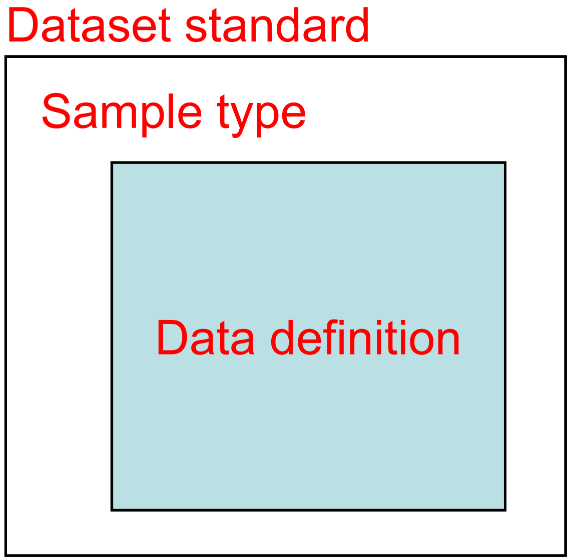


Dataset standard

Sample type
Data definition

The sample type belongs to the dataset standard as a framework:

The data definition appears to be the ‘content’ of the dataset standard

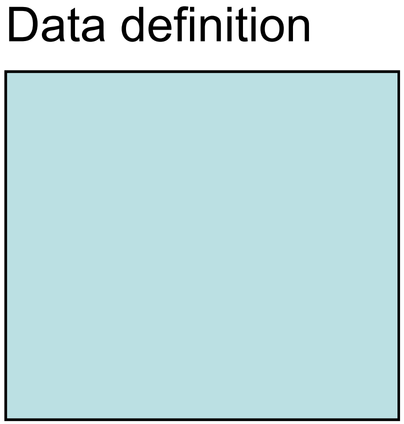


Now, forget of the dataset standard, just concentrate on the data definition:

The data definition appears to be the ‘content’ of the dataset standard

Data definition

Now, forget of the dataset standard, just concentrate on the data definition:

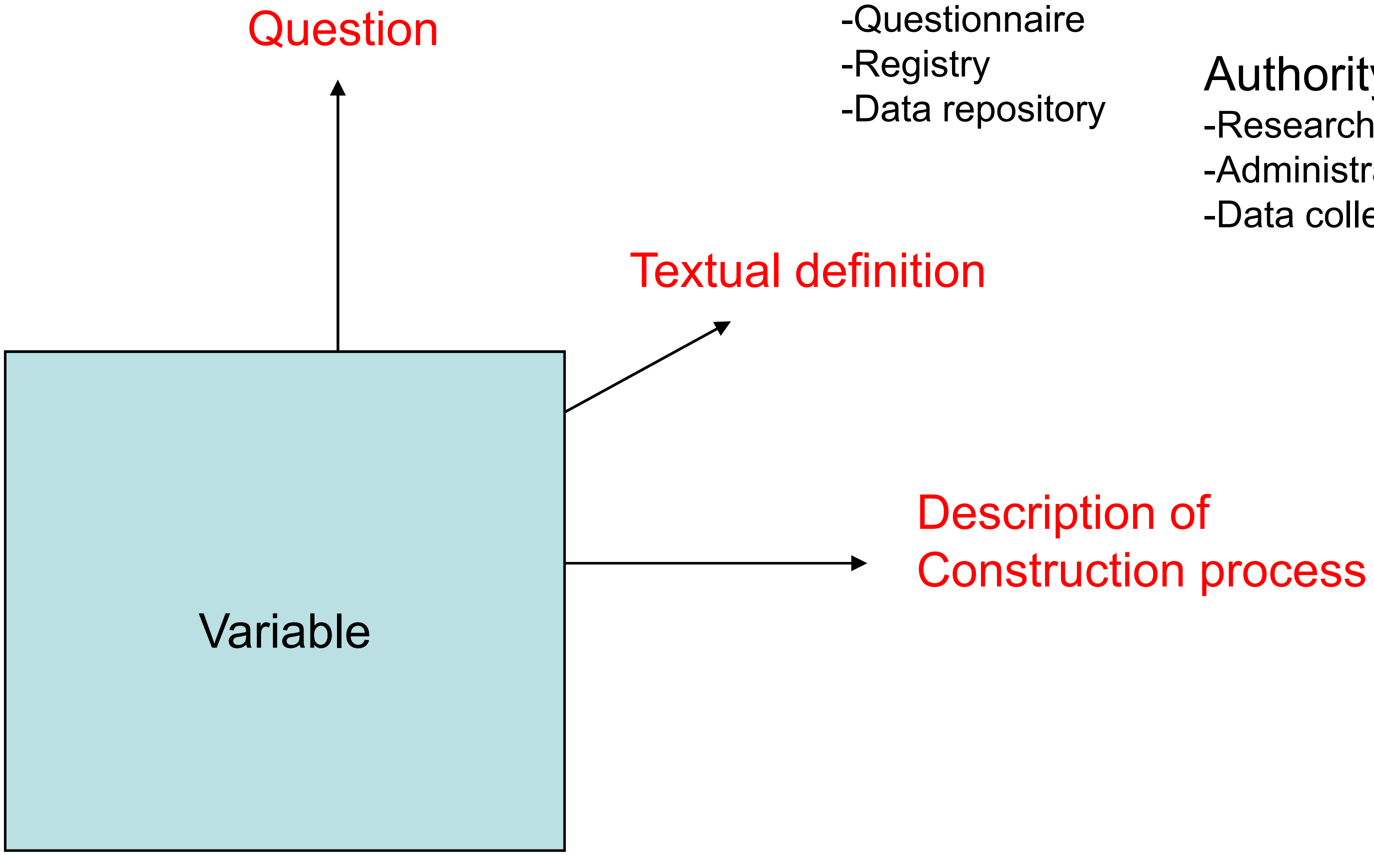


Data definition

… and look into it:

Data definition

A variable needs a definition



Question

Source
-Questionnaire
-Registry
-Data repository

Authority
-Researcher
-Administration
-Data collector

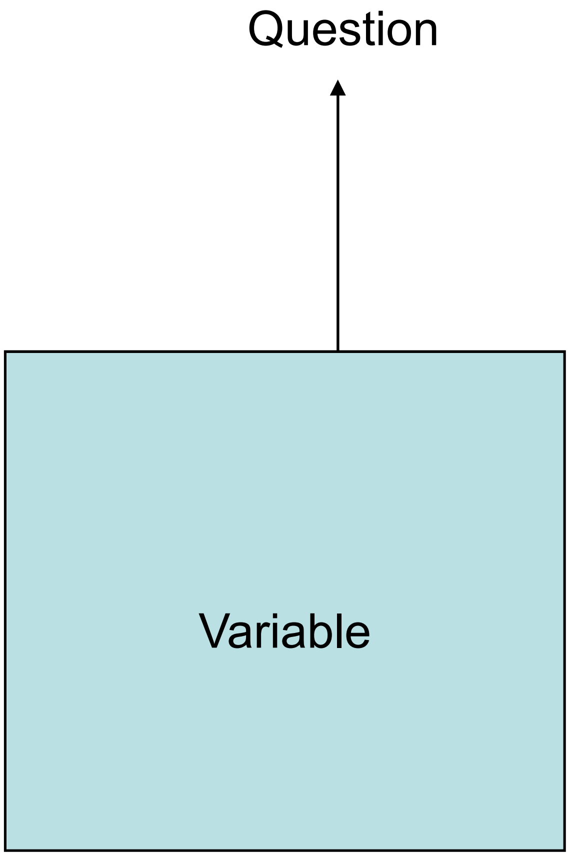
Textual definition

Description of
Construction process

Variable

In a survey, the most basic definition is the question.
Let’s concentrate on it.

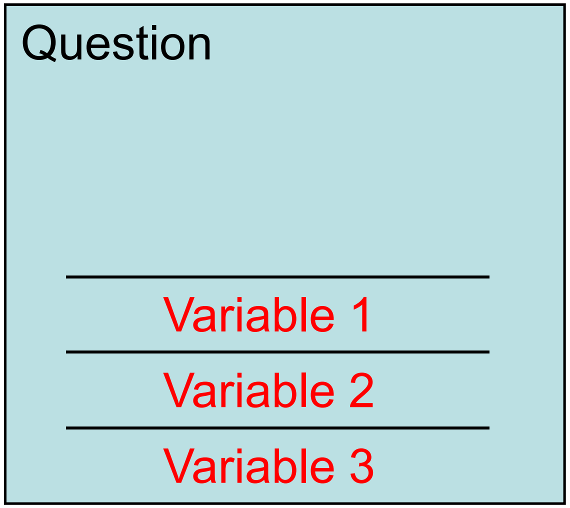
Questions may have various structures, which must be replicated in the database structure



Simple question
Items question
Multiple response
-By answer instance
-By answer category
Grid question

The most complex question structure is the generic case, from which Simpler forms can be obtained in a process of simplification:

So let’s take a representation of a question with some complexity to represent the generic question



Question

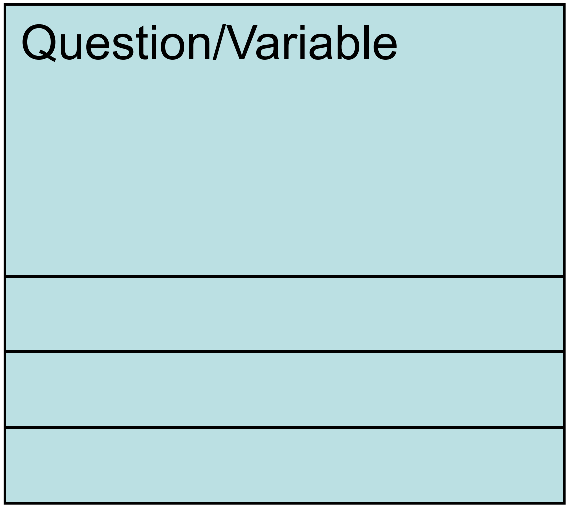
Variable 1

Variable 2

Variable 3

Items question
Multiple response
-By answer instance
-By answer category

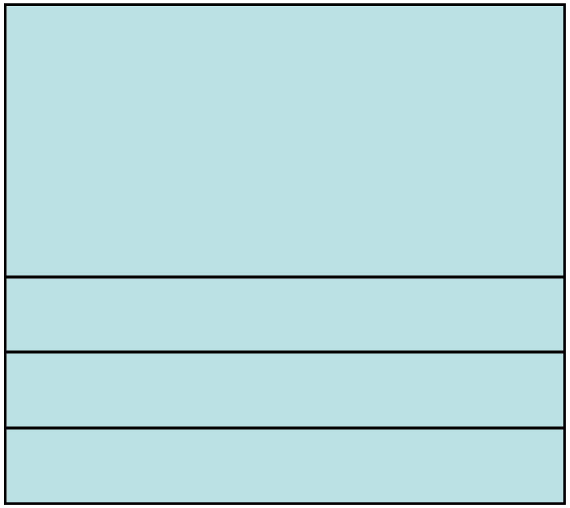
…and make it a symbol:



Question/Variable

…and even more simple:

Questions may have various structures, which must be replicated in the database structure



Keep it small:

Make it the standard definition:

Now, we have a good representation of the standard definition
In terms of standard questions and variables

Examples:

ISSP questionnaire + ‘standard setup’

ESS questionnaire + standard file

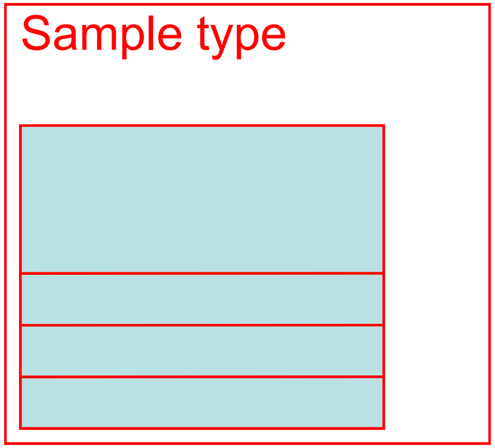
Standard

Remember, on higher levels we have the wrapping dataset…

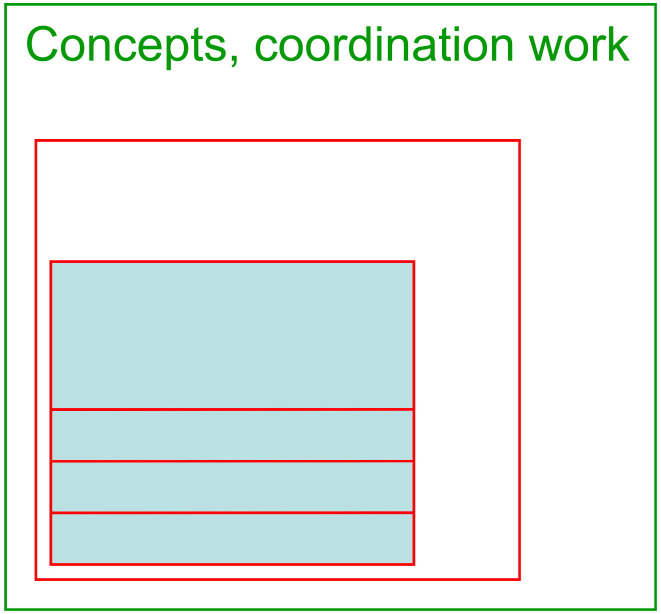
Examples:

ISSP questionnaire + ‘standard setup’

ESS questionnaire + standard file



… and even some kind of cross-national study…

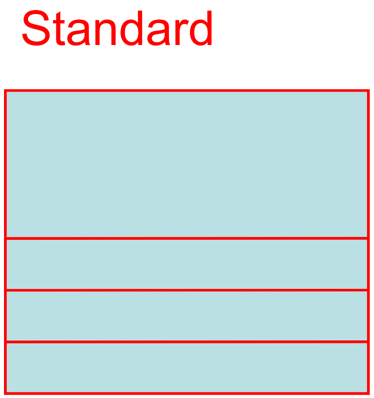


Examples:

ISSP questionnaire + ‘standard setup’

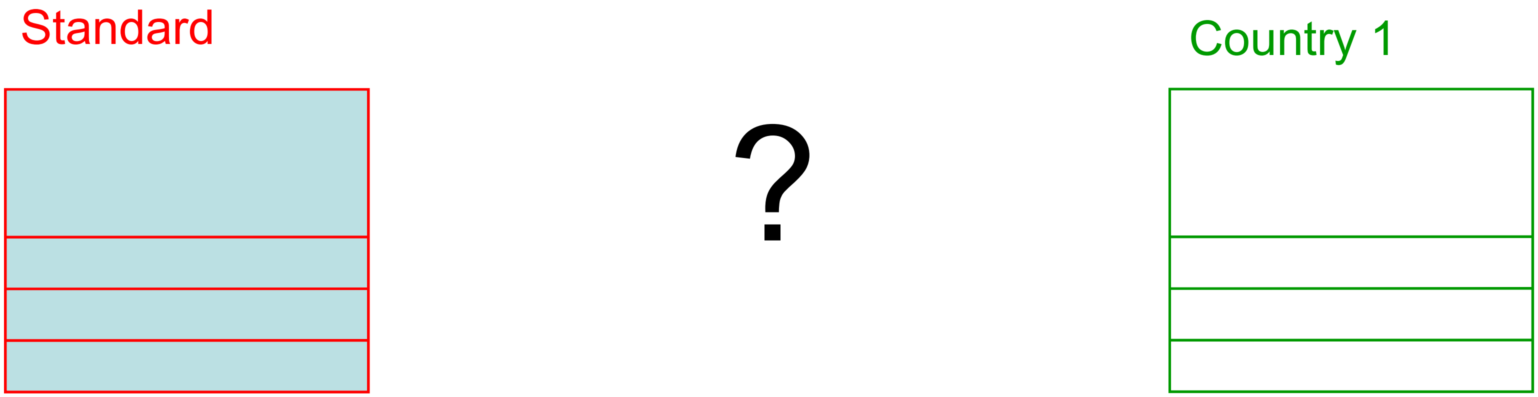
ESS questionnaire + standard file

Now, having the standard definition in our database, how would we Most economically enter the country definitions?



Standard

Create the country metadata using the standard:



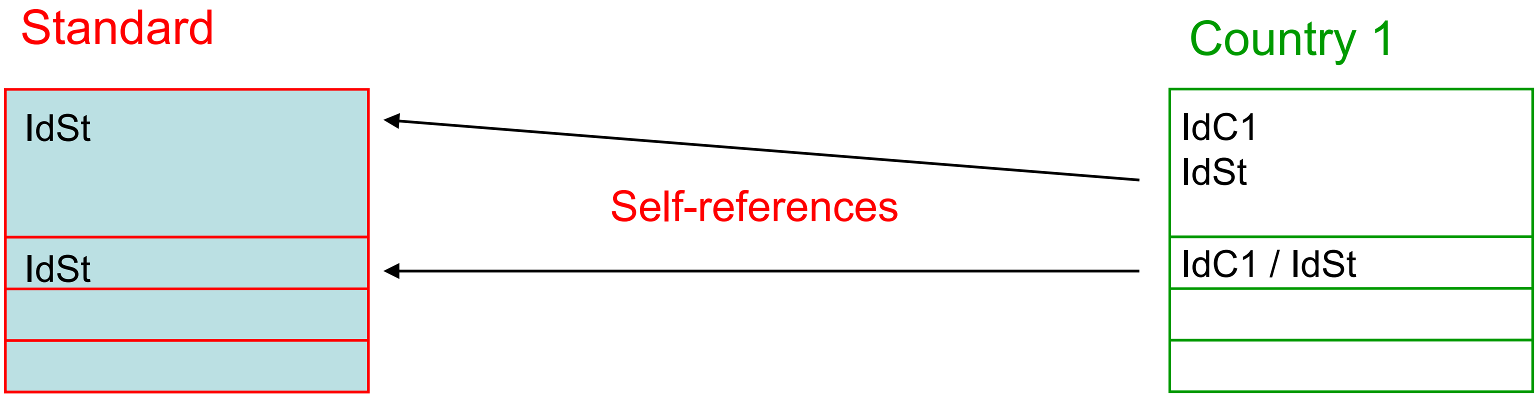
Standard

?

Country 1

Create the country metadata using the standard:
Derive!

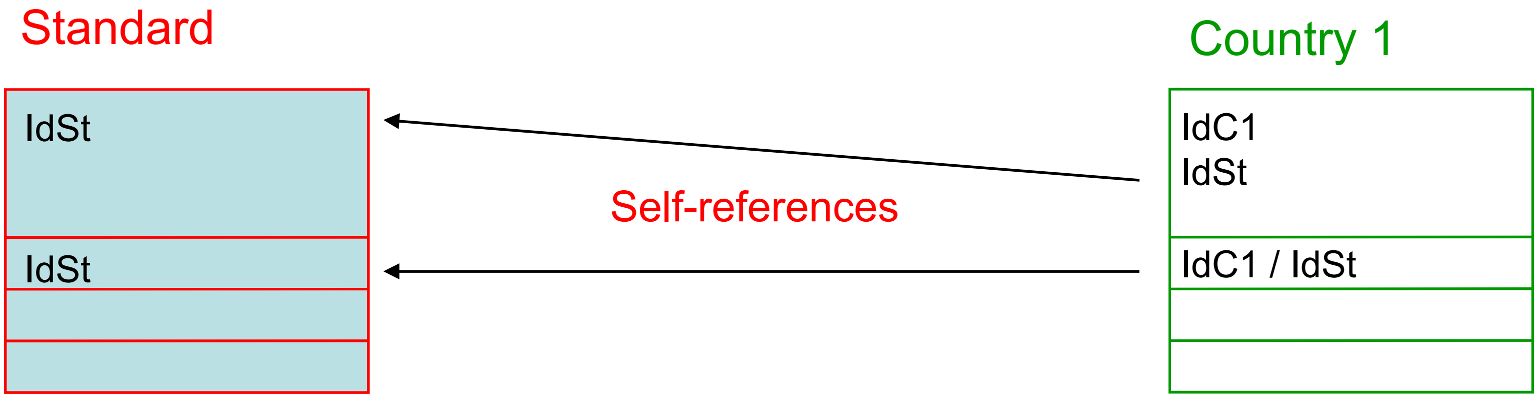
A self-reference of tables Question, rsp. Variable on themselves
Makes selected information inheritable:
- Wordings (multilingual)
- Value domains (multilingual)
- Descriptors (indexes)
- … and some other details



Create the country metadata using the standard:
Derive?

Question structures are complex; the higher the complexity, the higher is the probability for a small change somewhere in the structure
- What if a wording changes in one single language?
- What if an Interviewer instruction changes because of a change in the structure of the questionnaire?

Another structure is necessary for changing Q/Vs; multiple structures are a factor of complexity in all processes to be programmed.



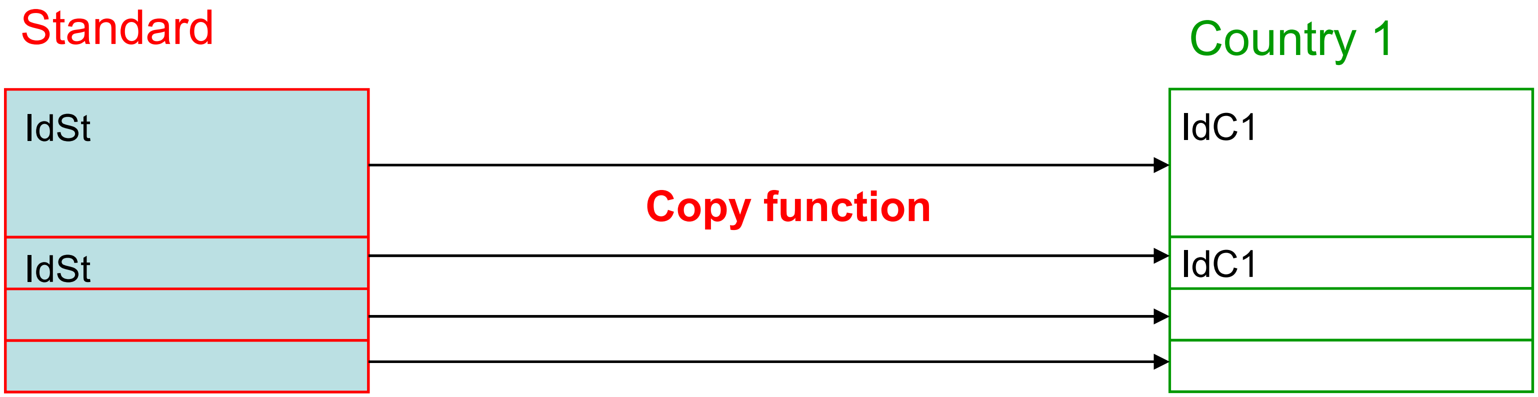
Create the country metadata using the standard:
Copy! You may also edit the copy (general solution)

All information for the standard is copied, even the information, which remains constant.

Redundance

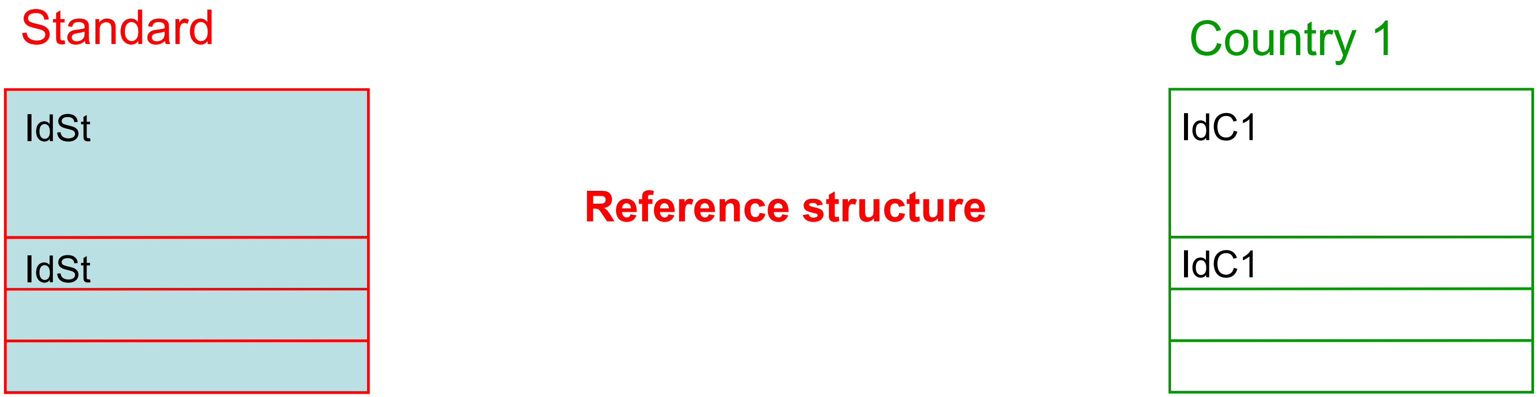
Independent versions

Copes with variations!



Where is the reference from Country 1 to the Standard?

Create reference link!



Standard

IdSt

IdSt

Reference structure

Country 1

IdC1

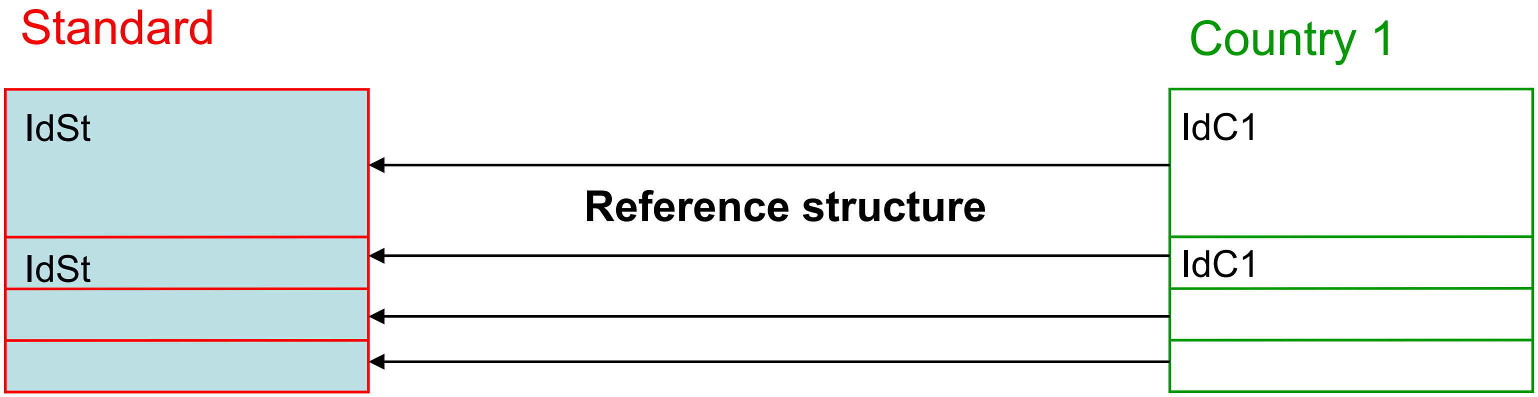
IdC1

The data element used for reference:

IdentTarget
IdentSource

LinkType

VariationType
VariationComment



Standard

IdSt

IdSt

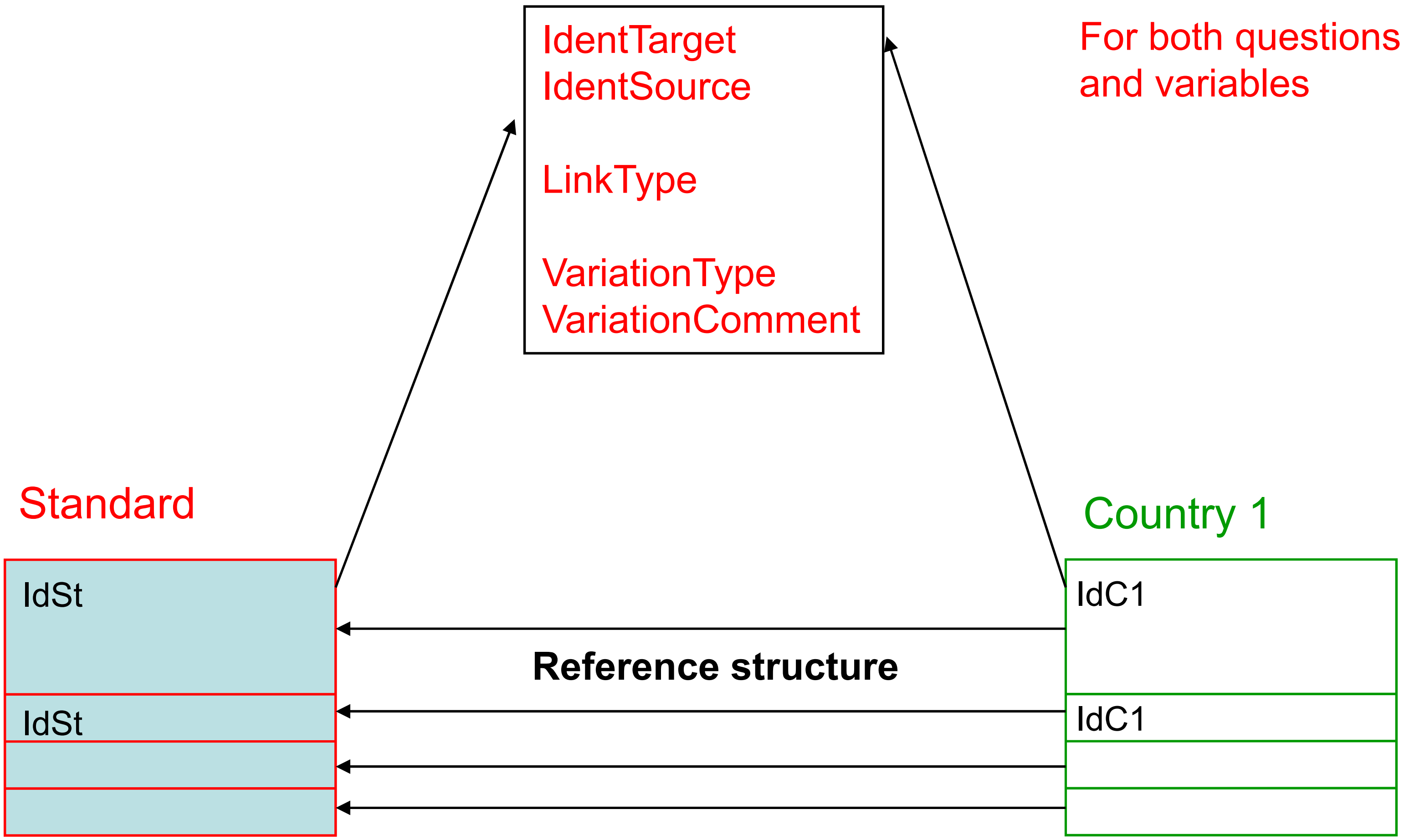
Reference structure

Country 1

IdC1

IdC1

The data element used for reference:



IdentTarget
IdentSource

LinkType

VariationType
VariationComment

For both questions and variables

Standard

IdSt

IdSt

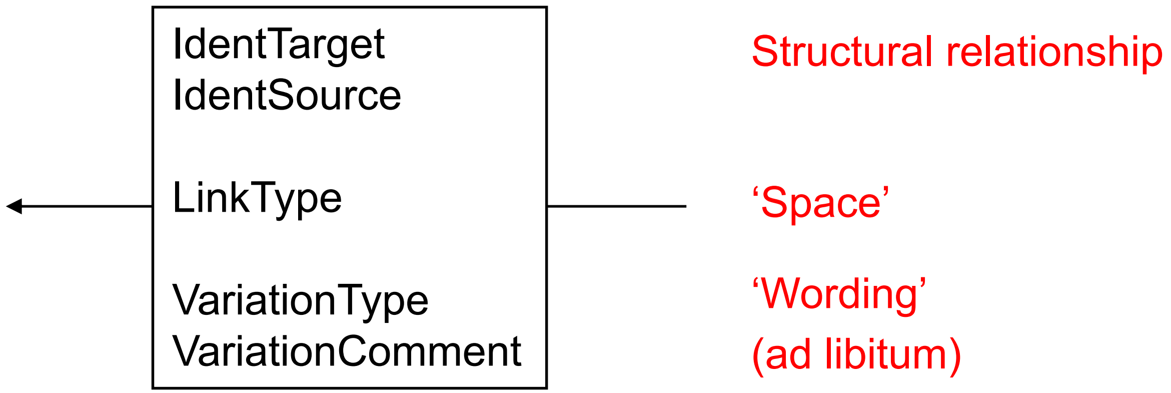
Country 1

IdC1

IdC1

Reference structure

The data element used for reference:



IdentTarget
IdentSource

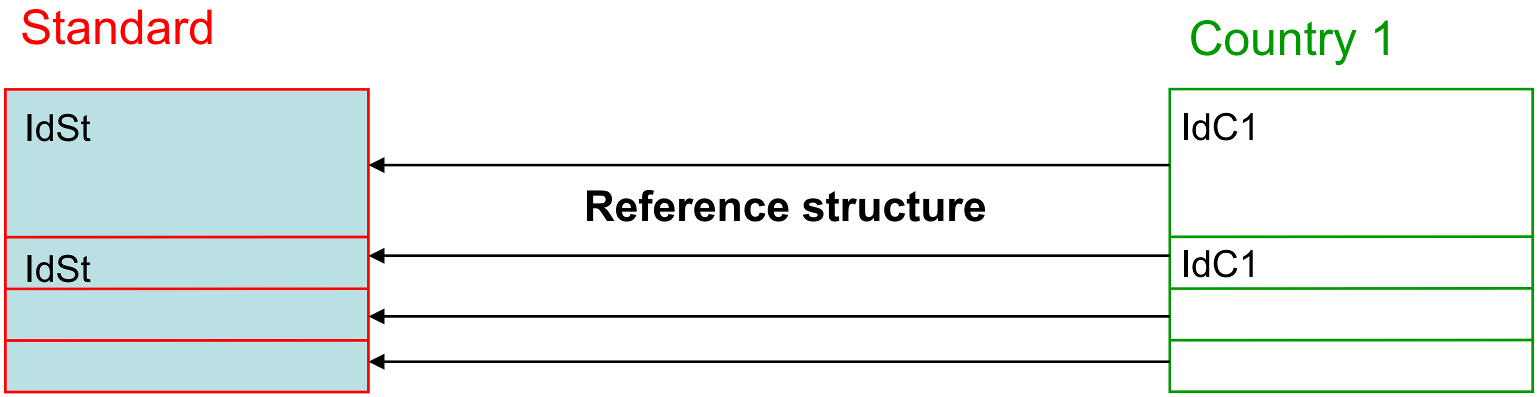
LinkType

VariationType
VariationComment

Structural relationship

‘Space’

‘Wording’
(ad libitum)



Standard

Country 1

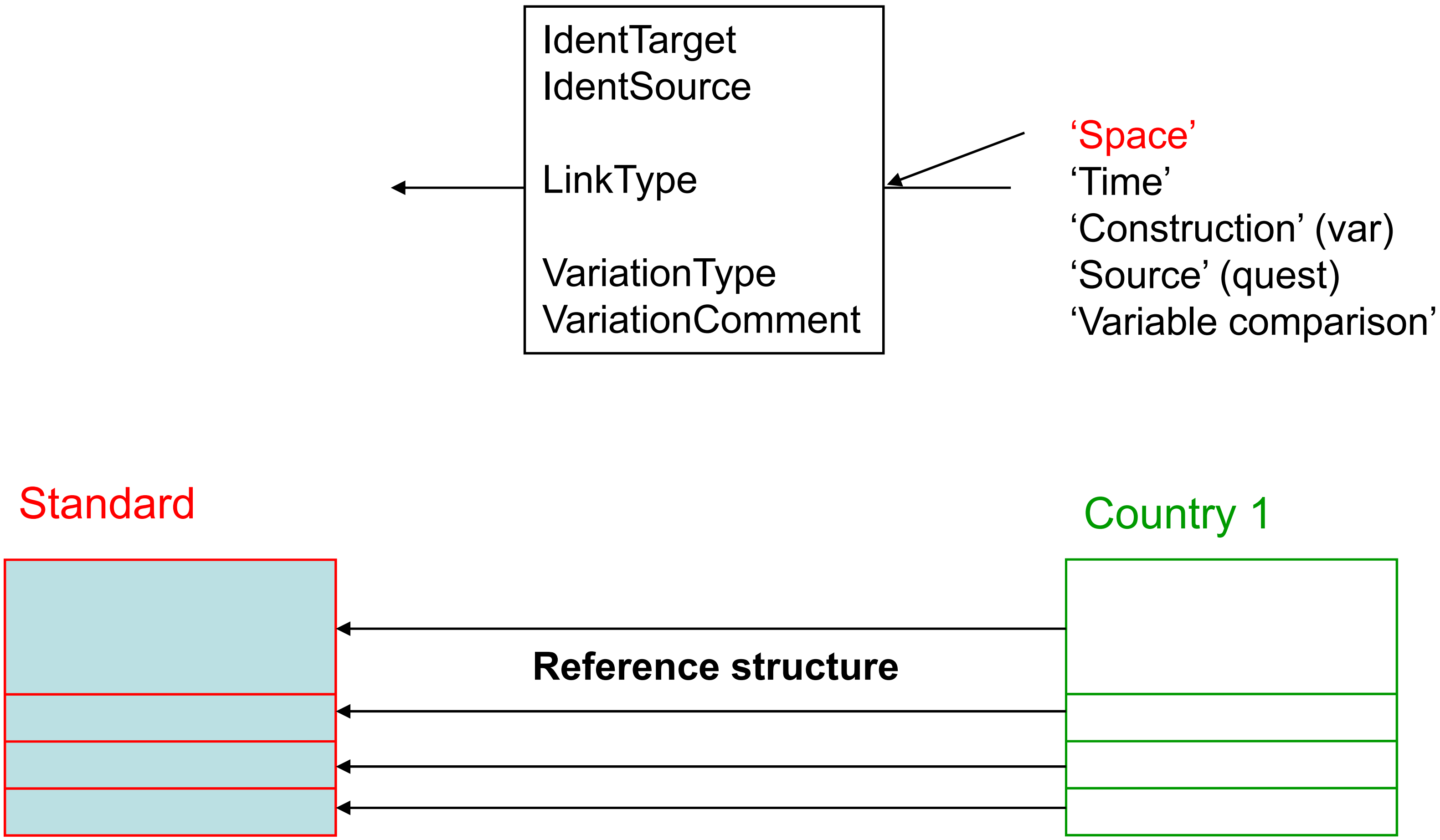
IdSt

IdSt

IdC1

IdC1

Reference structure



IdentTarget
IdentSource

LinkType

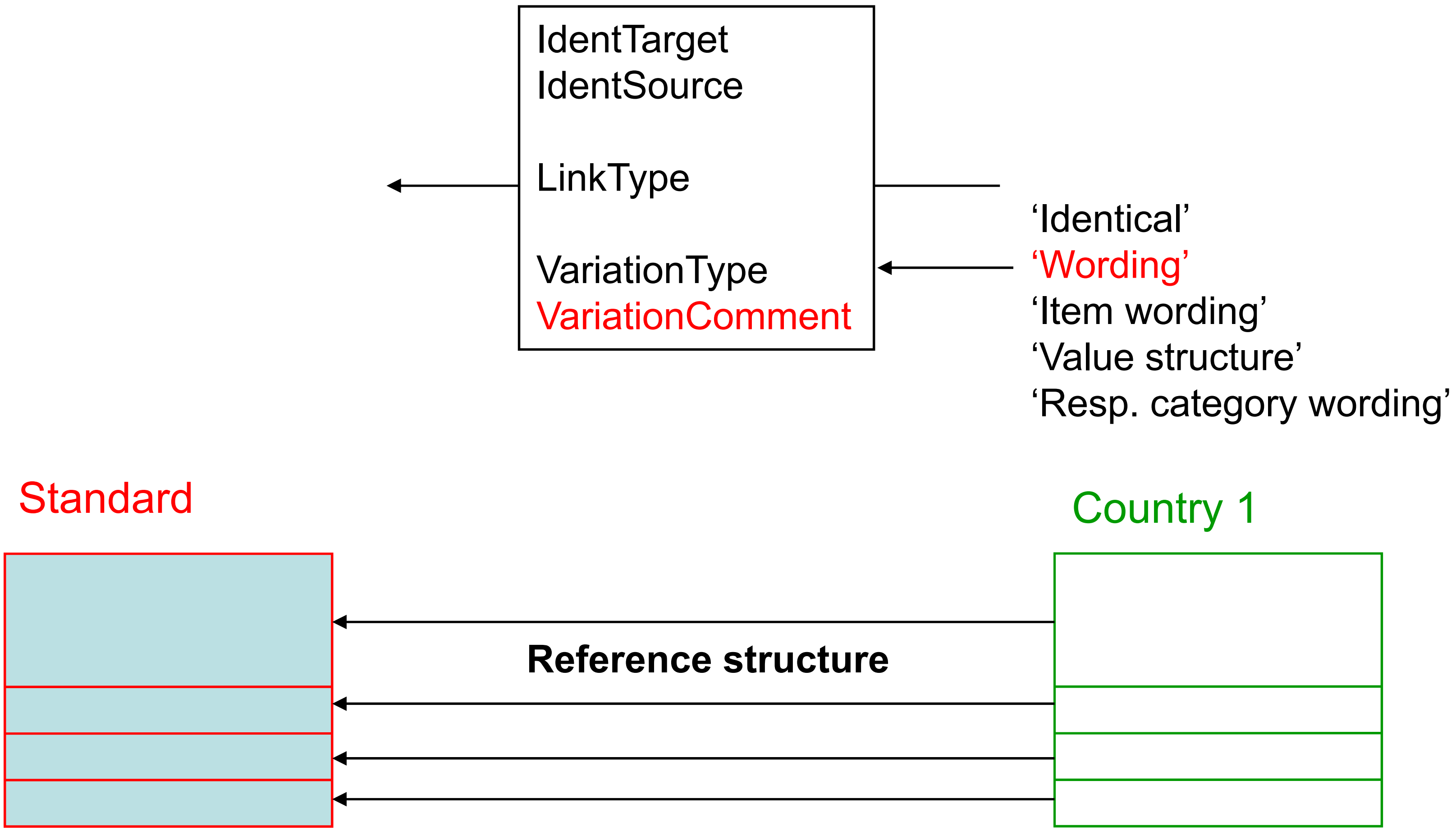
VariationType
VariationComment

‘Space’
‘Time’
‘Construction’ (var)
‘Source’ (quest)
‘Variable comparison’

Standard

Country 1

Reference structure



IdentTarget
IdentSource

LinkType

VariationType
VariationComment

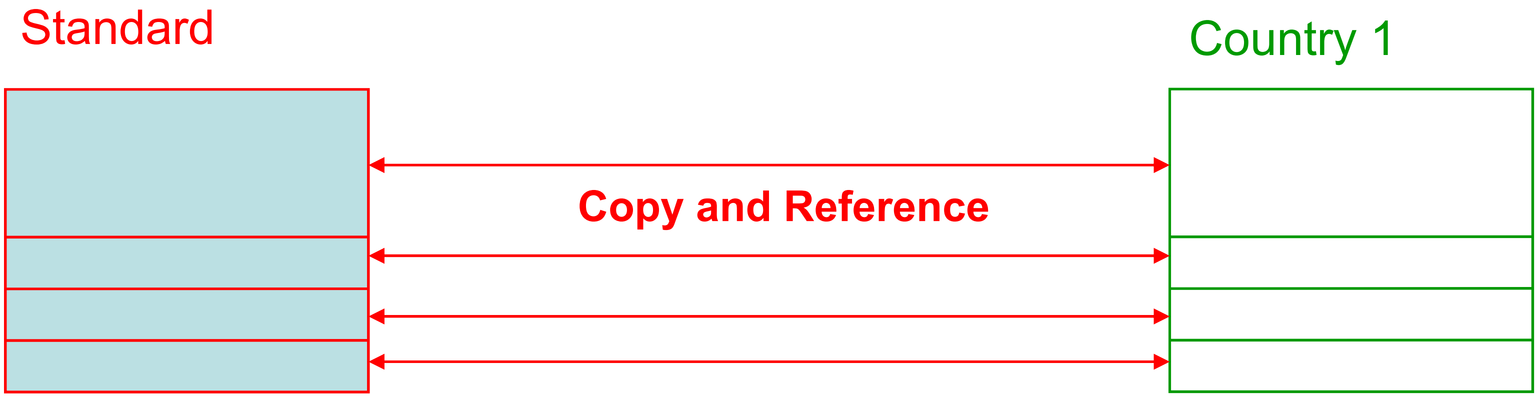
‘Identical’
‘Wording’
‘Item wording’
‘Value structure’
‘Resp. category wording’

Standard

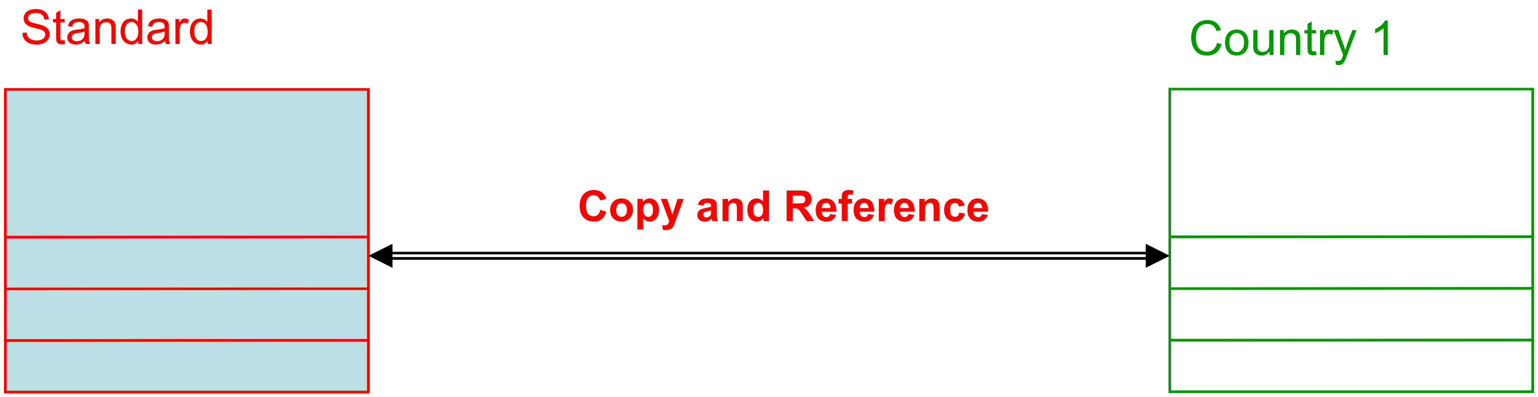
Country 1

Reference structure

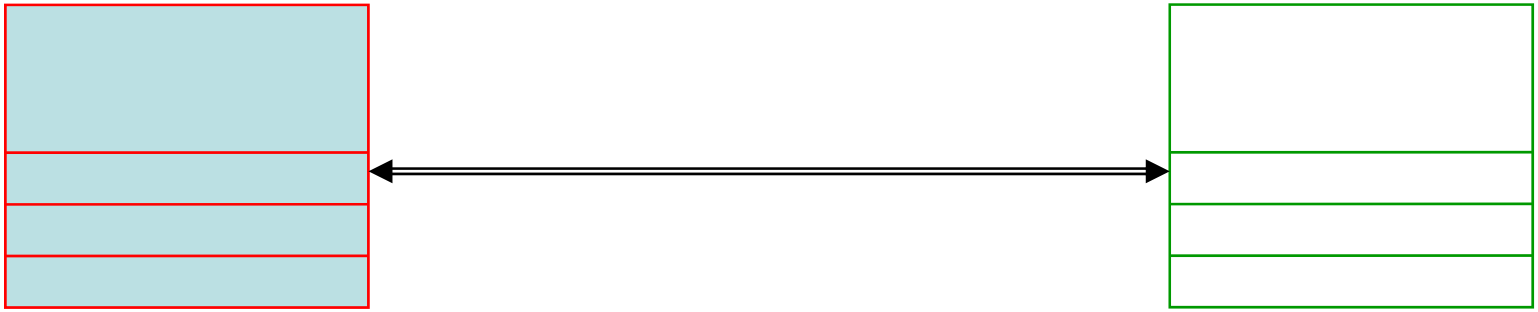
Synthetic view of copy function and reference structure:

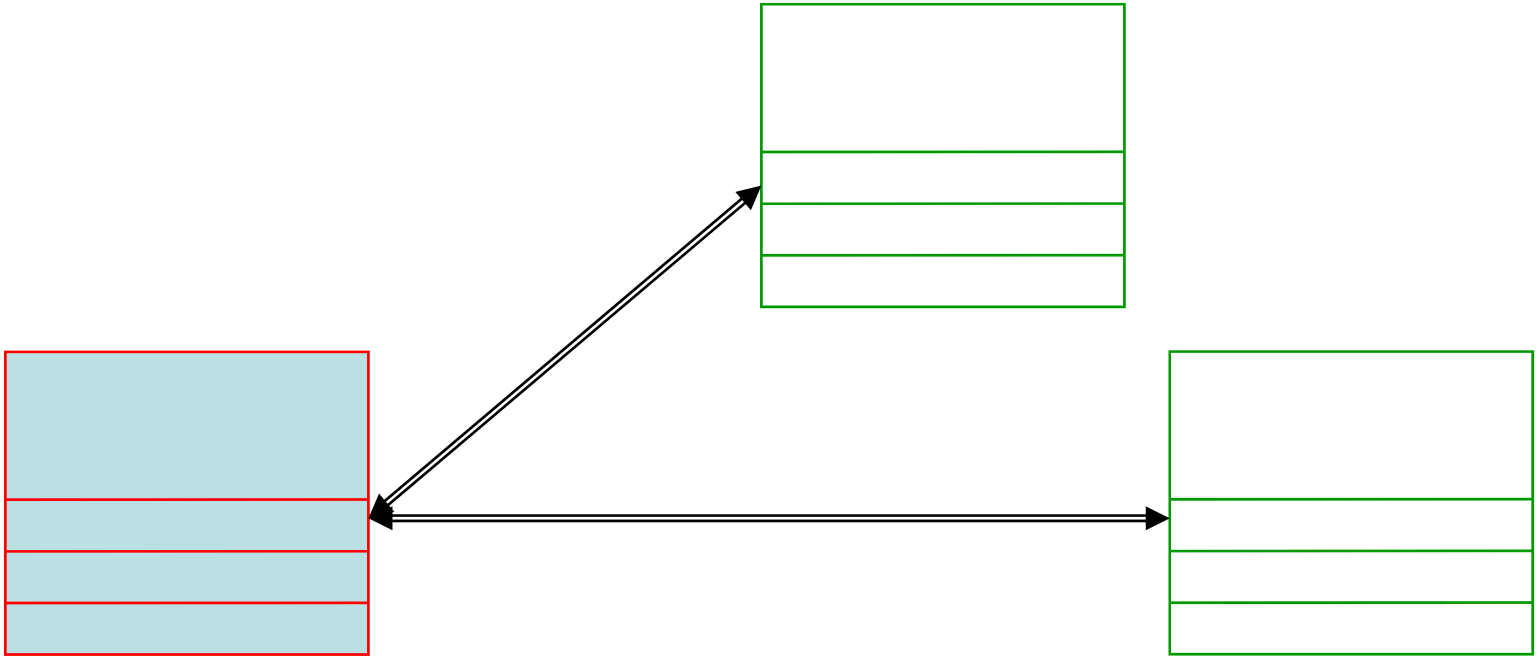


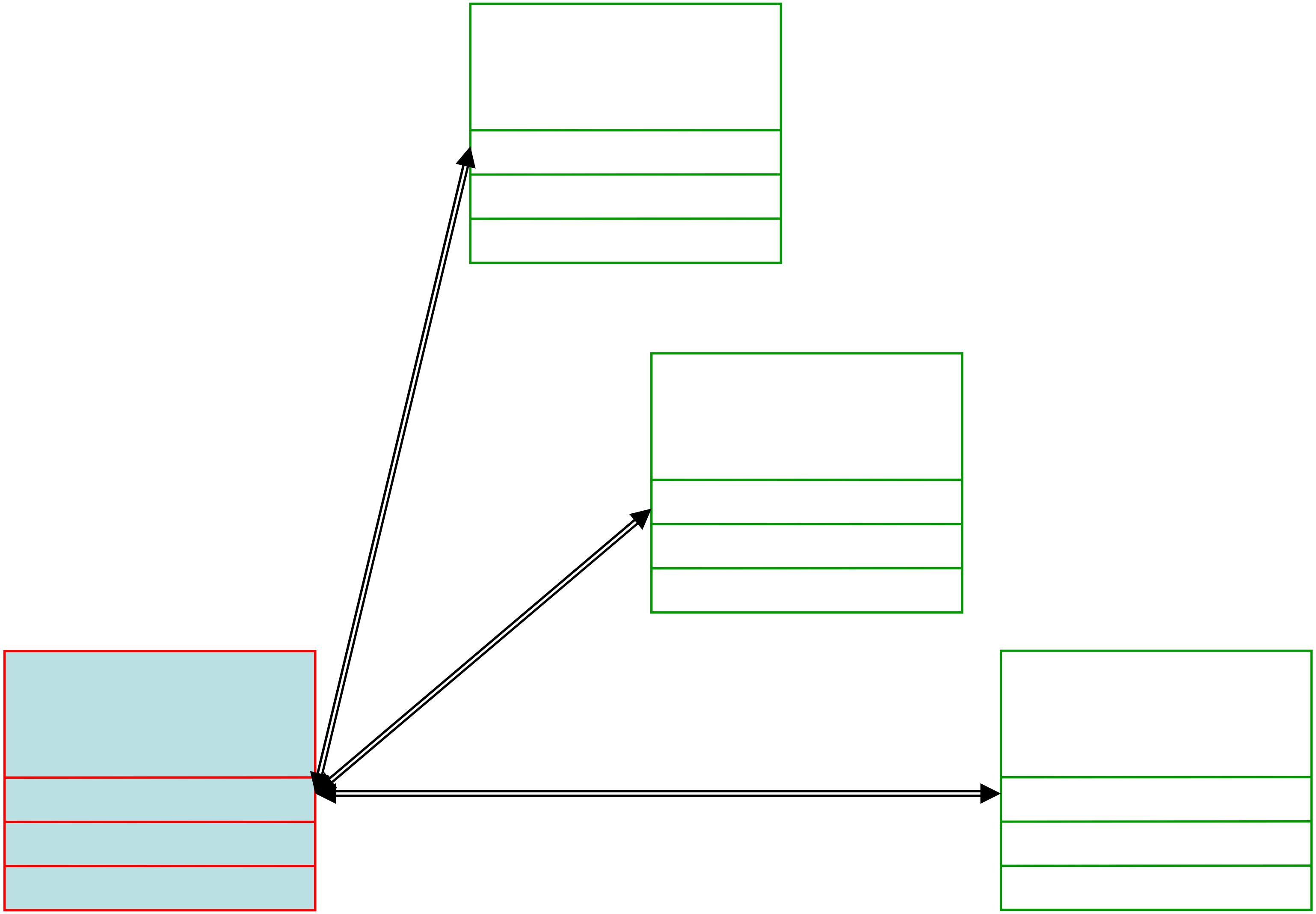
Synthetic view of copy function and reference structure:



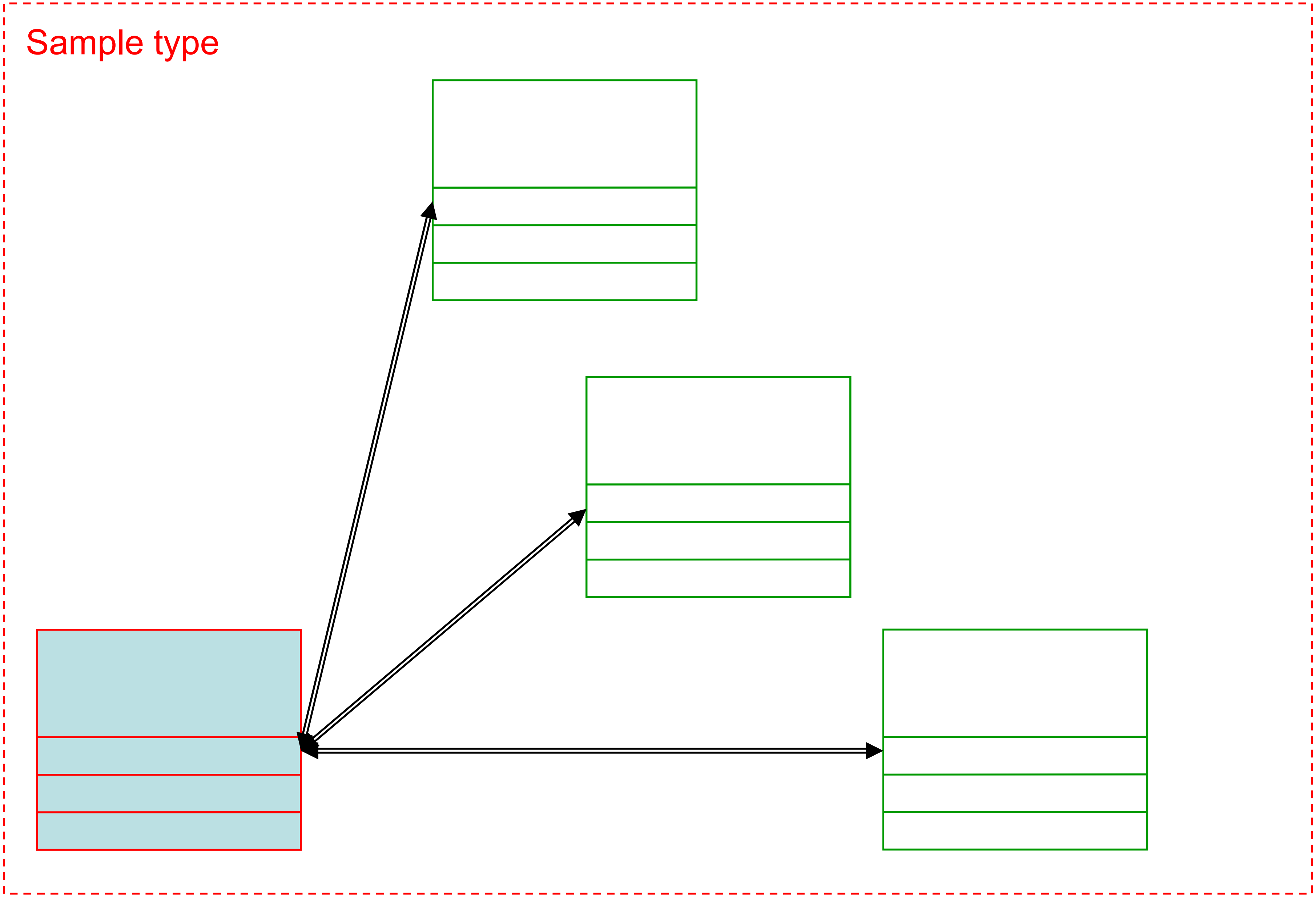
Now, let’s look what happens with more than one country dataset!





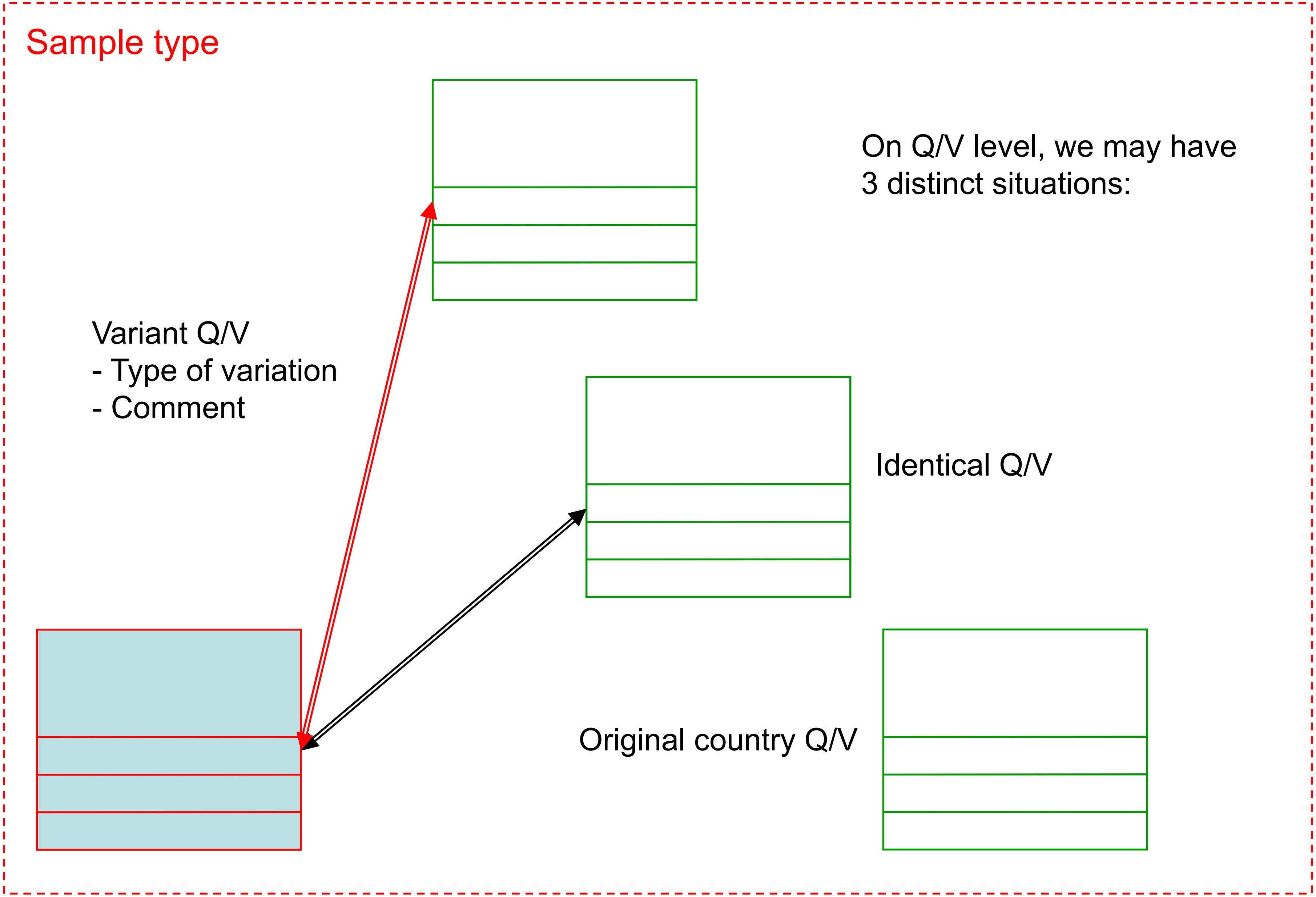


Space-compound dataset (higher level dataset, incl. Documentation of variations)



Sample type

Space-compound dataset (higher level dataset, incl. Documentation of variations)



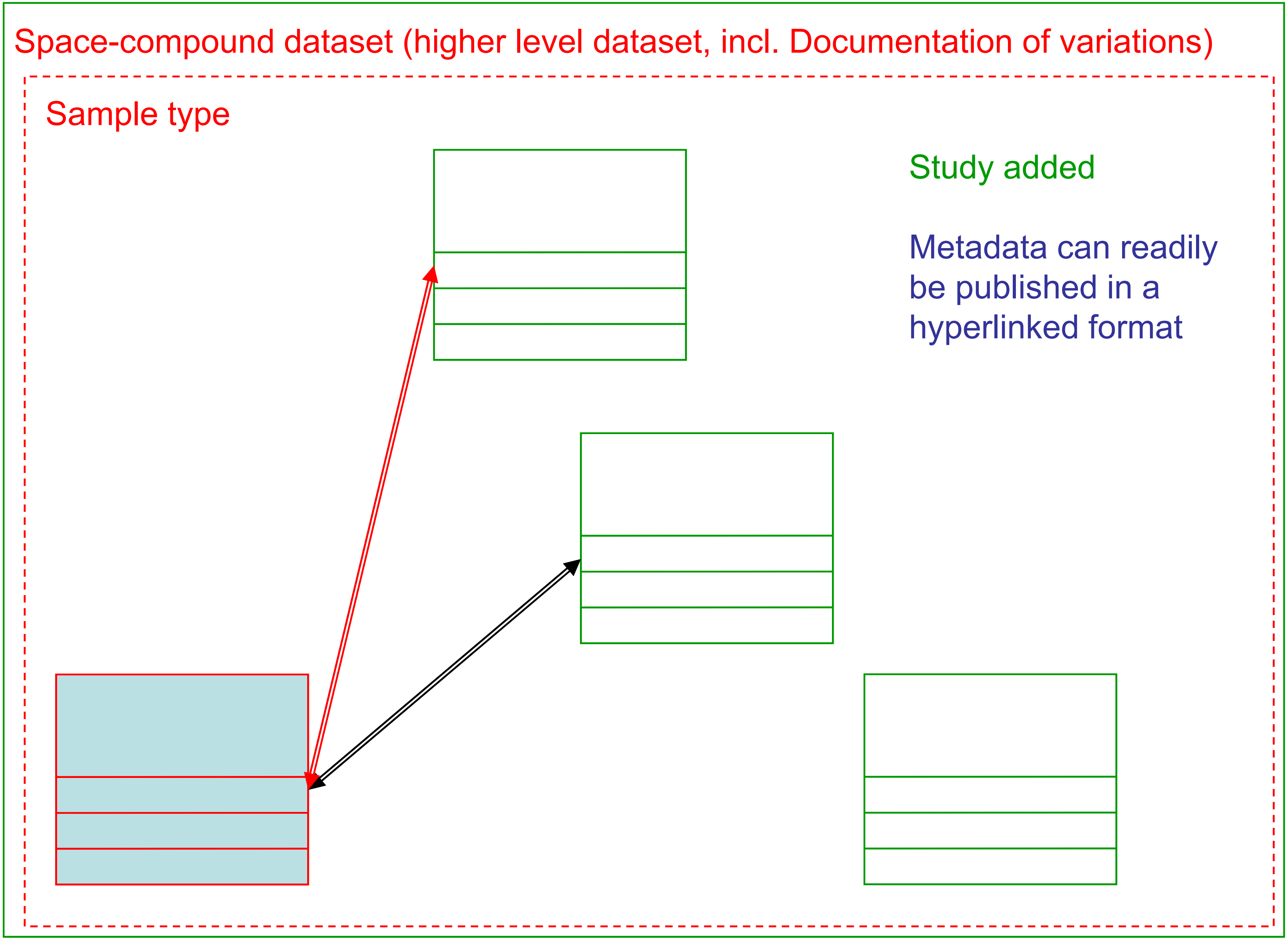
Sample type

On Q/V level, we may have
3 distinct situations:

Variant Q/V
- Type of variation
- Comment

Identical Q/V

Original country Q/V



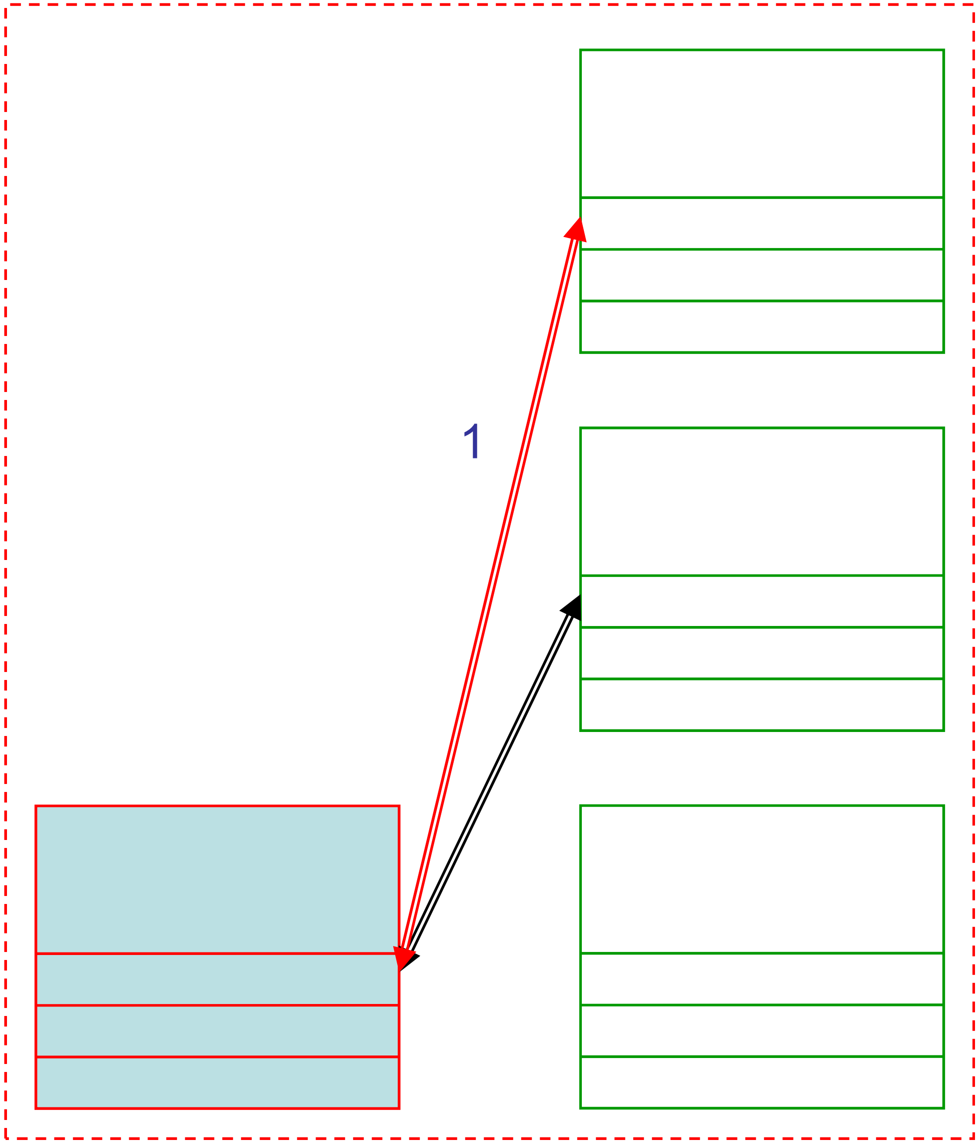
Space-compound dataset (higher level dataset, incl. Documentation of variations)

Sample type

Study added

Metadata can readily be published in a hyperlinked format

Now, look at the integration process

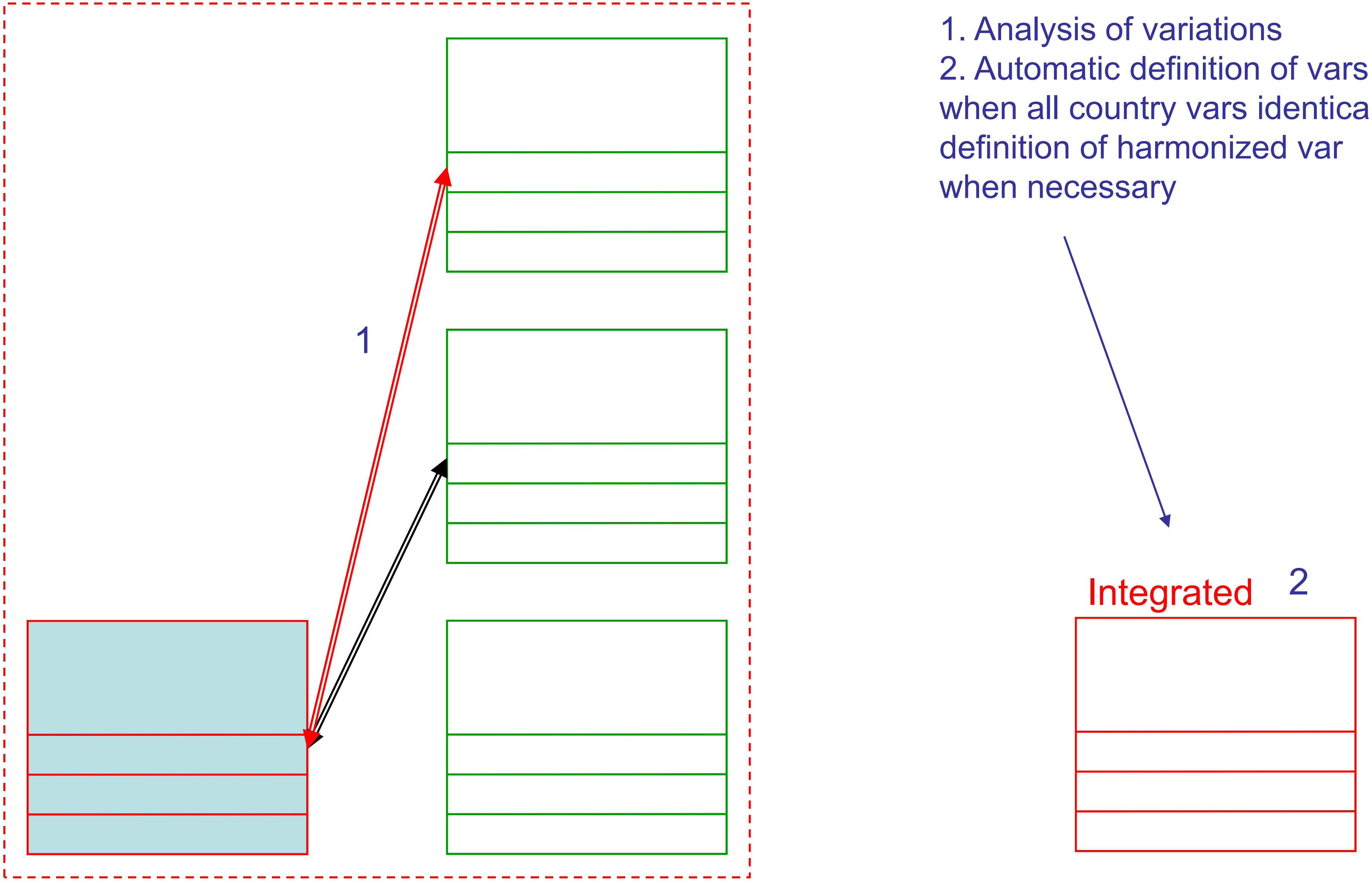


1. Analysis of variations

A country Q/V may be:

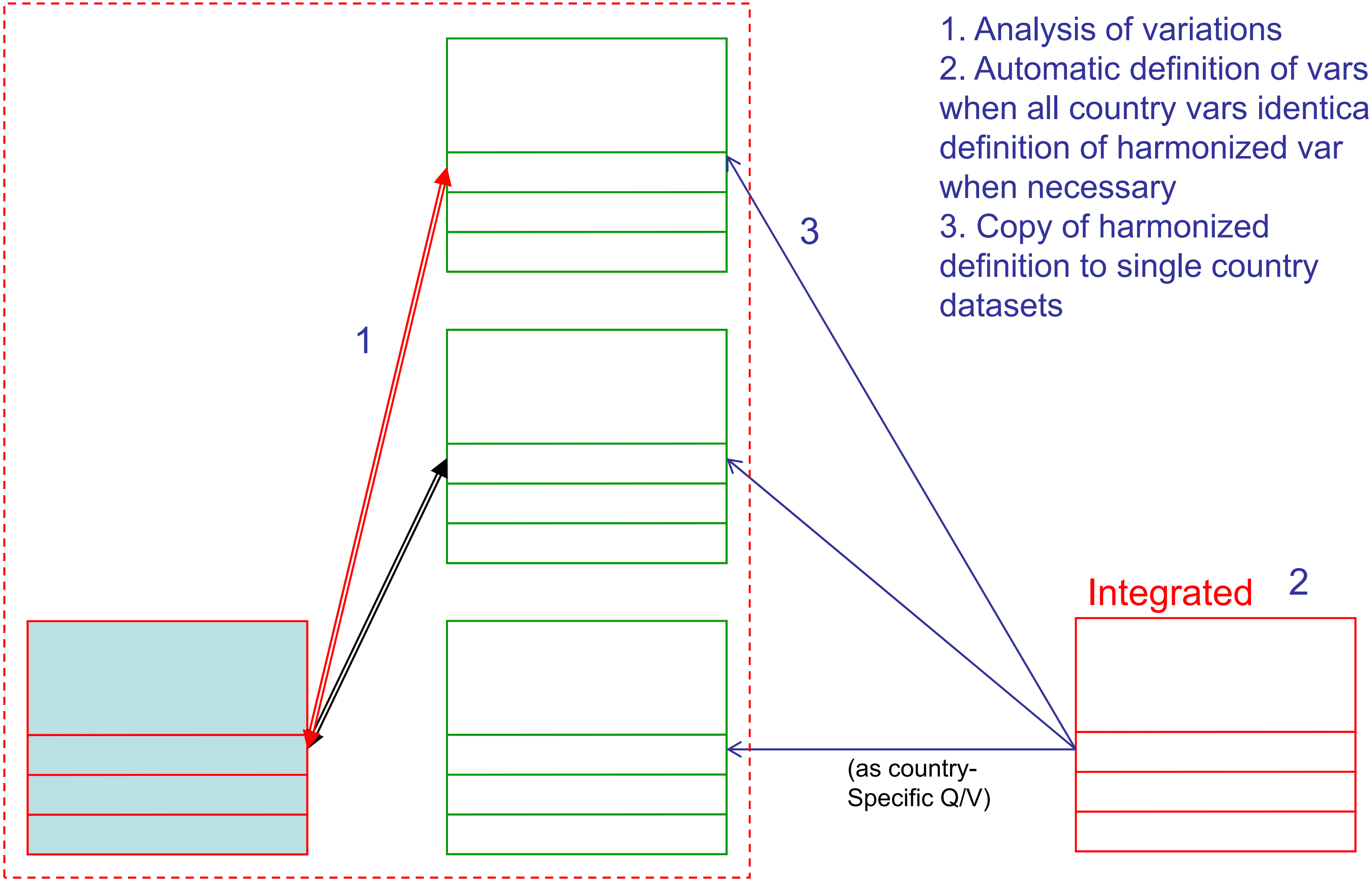
- identical to the standard
- varying on the standard
- specific to the country

Now, look at the integration process



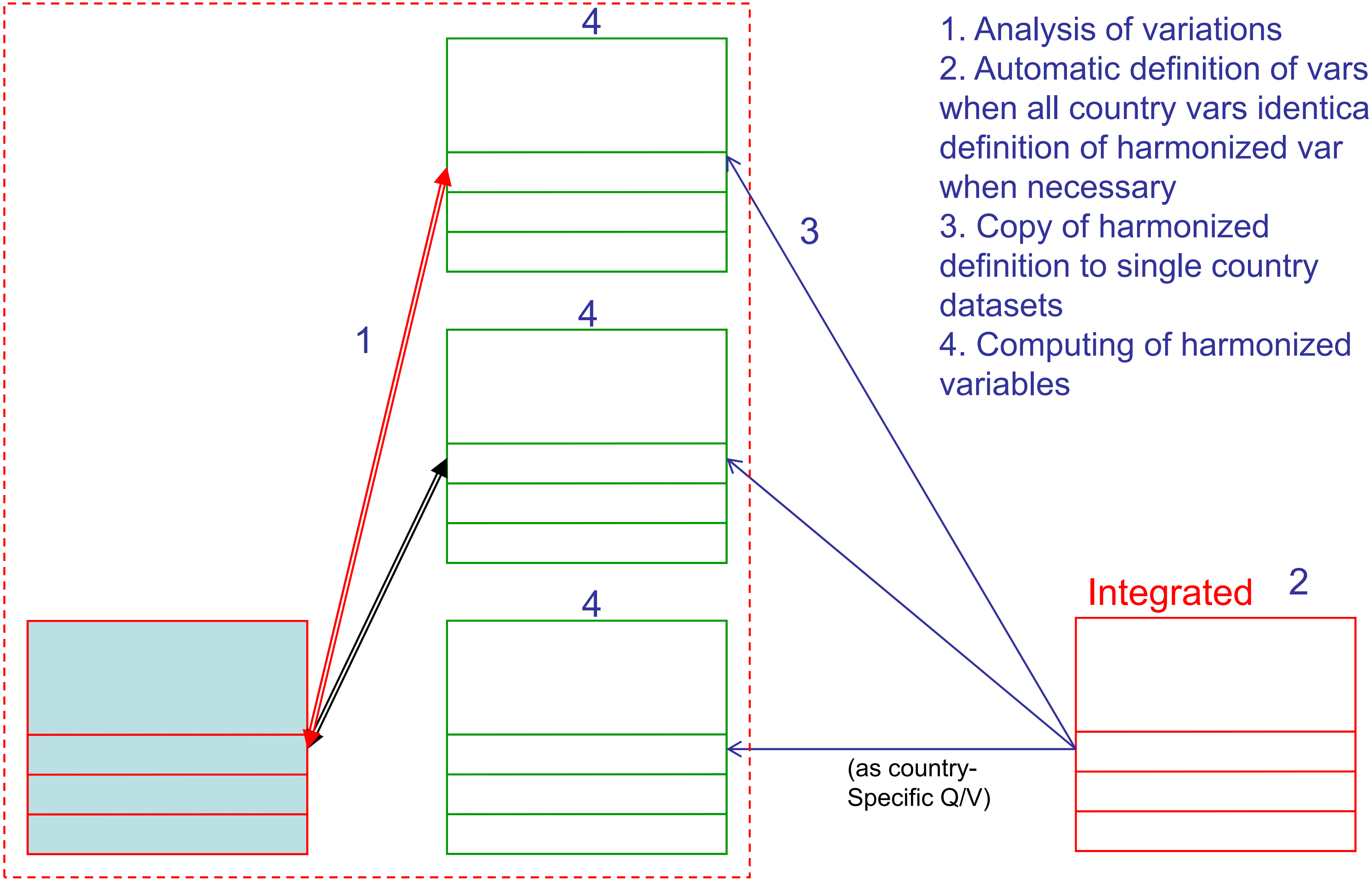
1. Analysis of variations
2. Automatic definition of vars when all country vars identical, definition of harmonized var when necessary

Now, look at the integration process



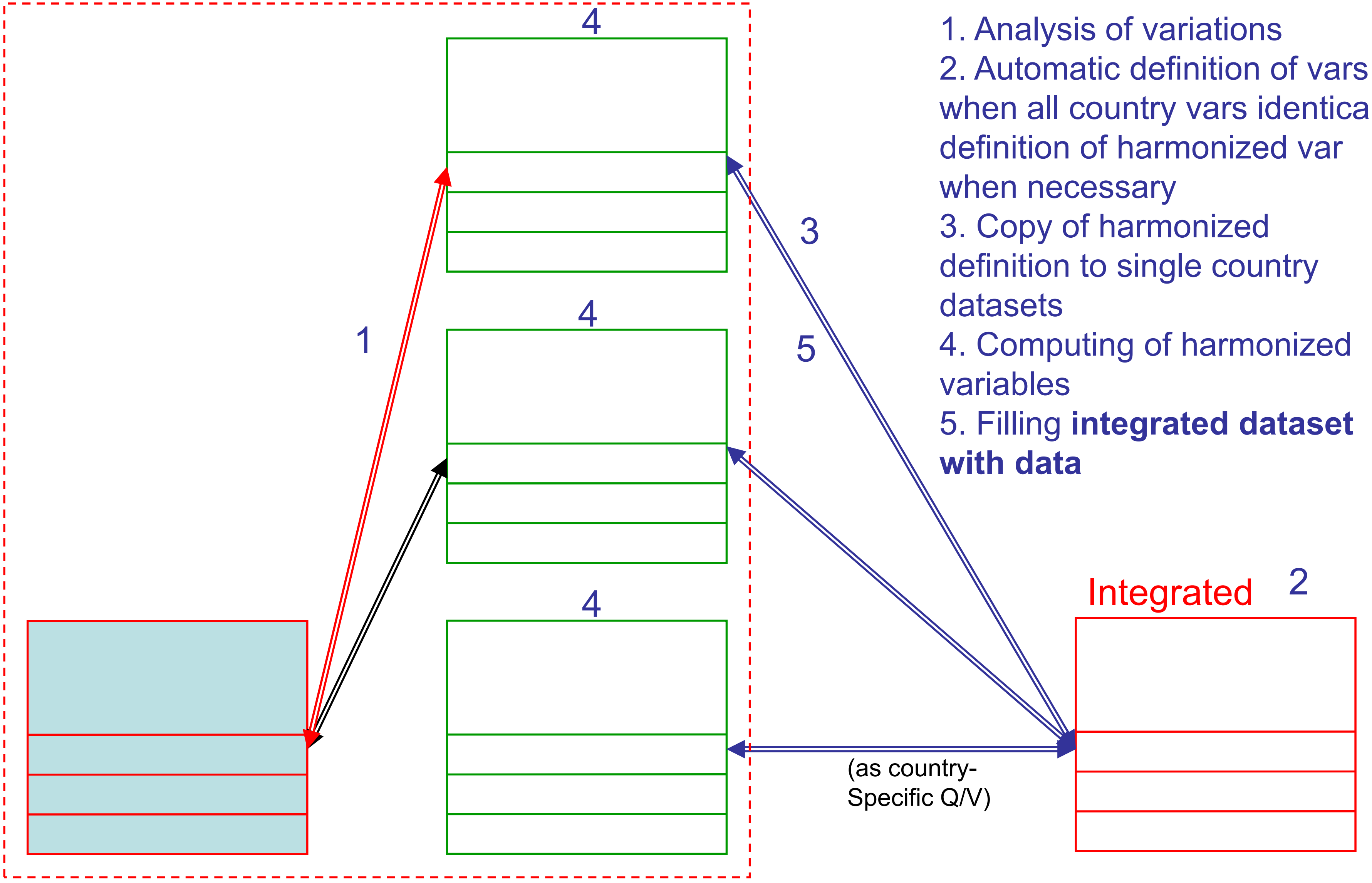
1. Analysis of variations
2. Automatic definition of vars when all country vars identical, definition of harmonized var when necessary
3. Copy of harmonized definition to single country datasets

Now, look at the integration process



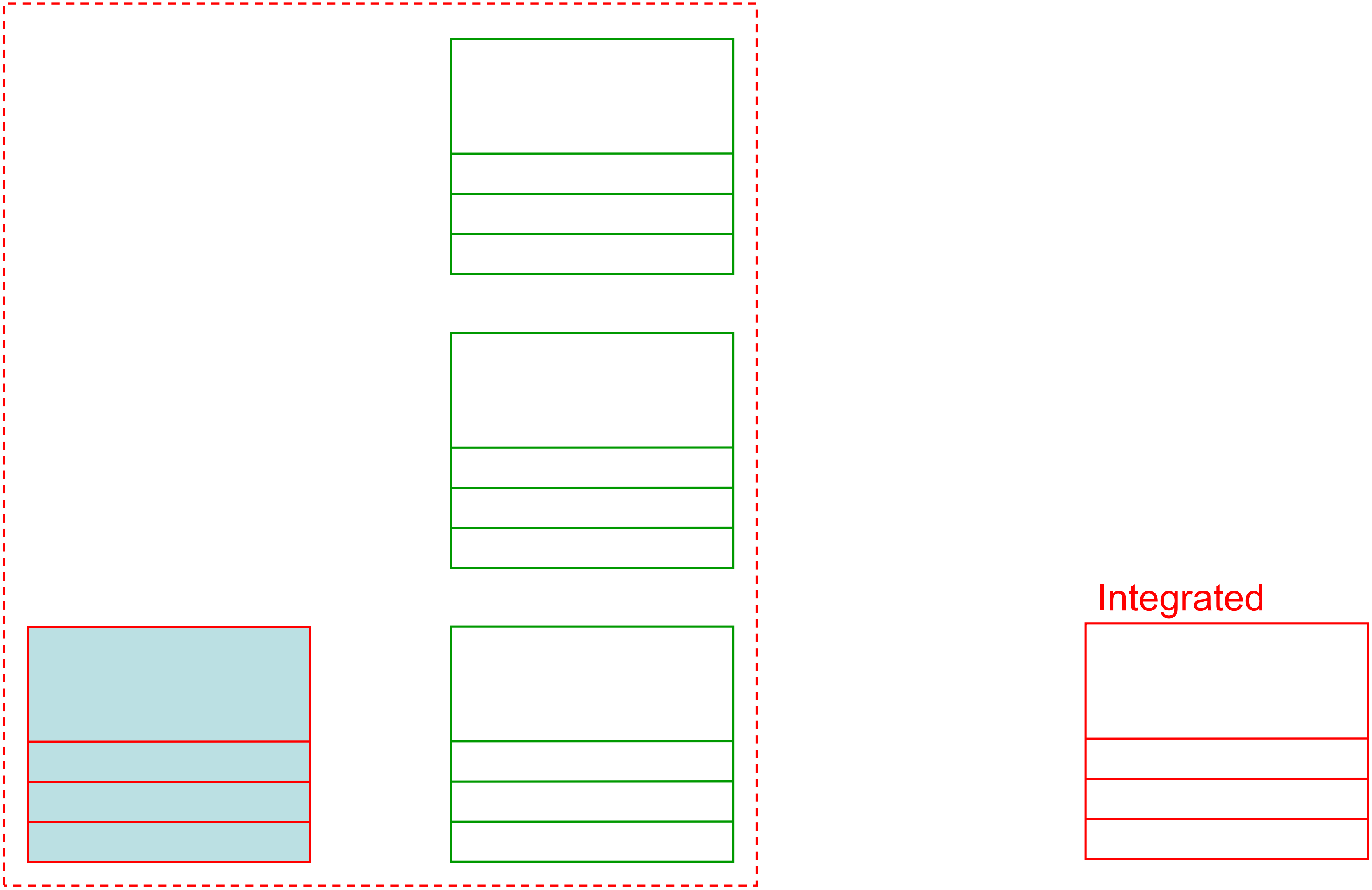
1. Analysis of variations
2. Automatic definition of vars when all country vars identical, definition of harmonized var when necessary
3. Copy of harmonized definition to single country datasets
4. Computing of harmonized variables

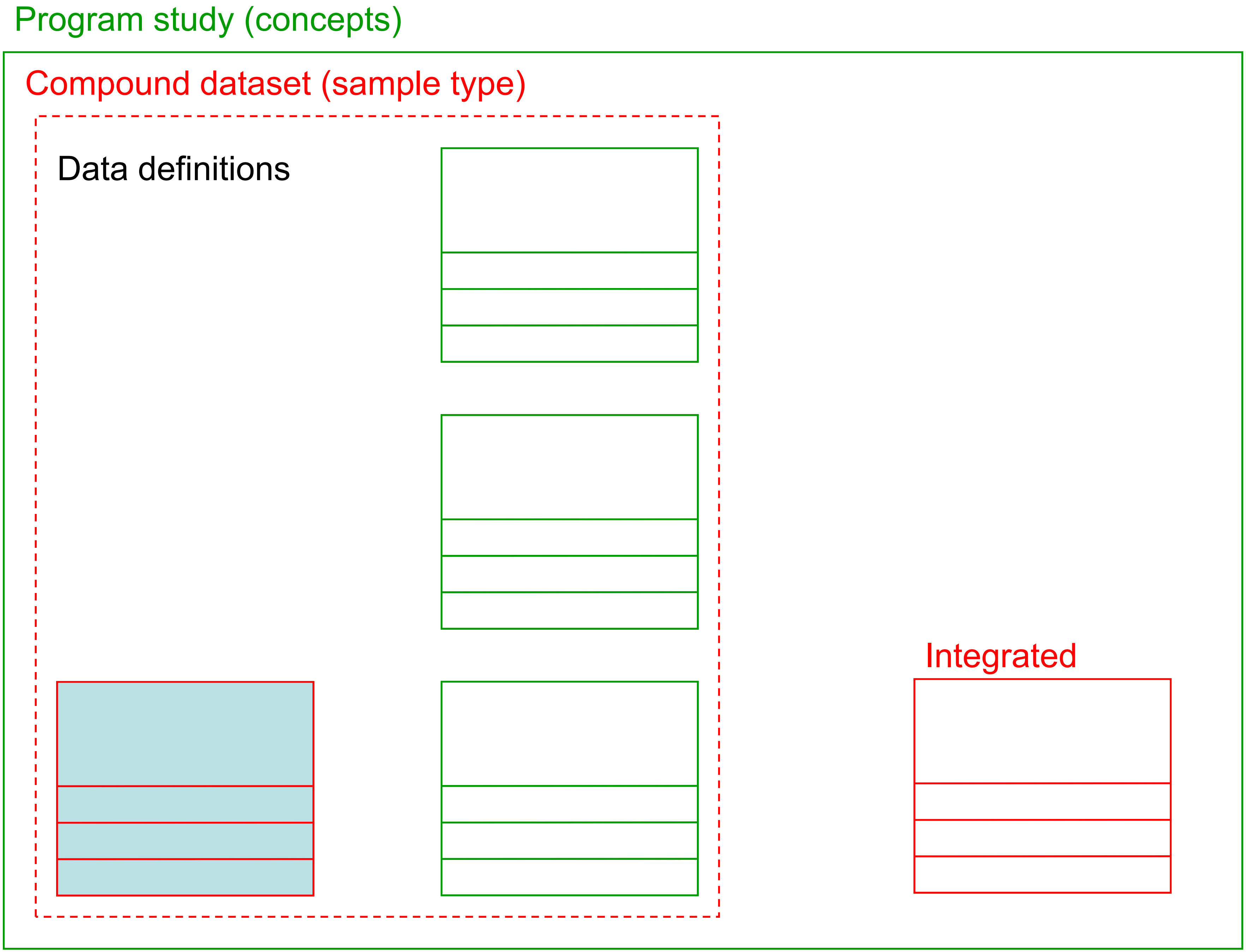
Now, look at the integration process



1. Analysis of variations
2. Automatic definition of vars when all country vars identical, definition of harmonized var when necessary
3. Copy of harmonized definition to single country datasets
4. Computing of harmonized variables
5. Filling **integrated dataset with data**

Let’s come back to the studies involved



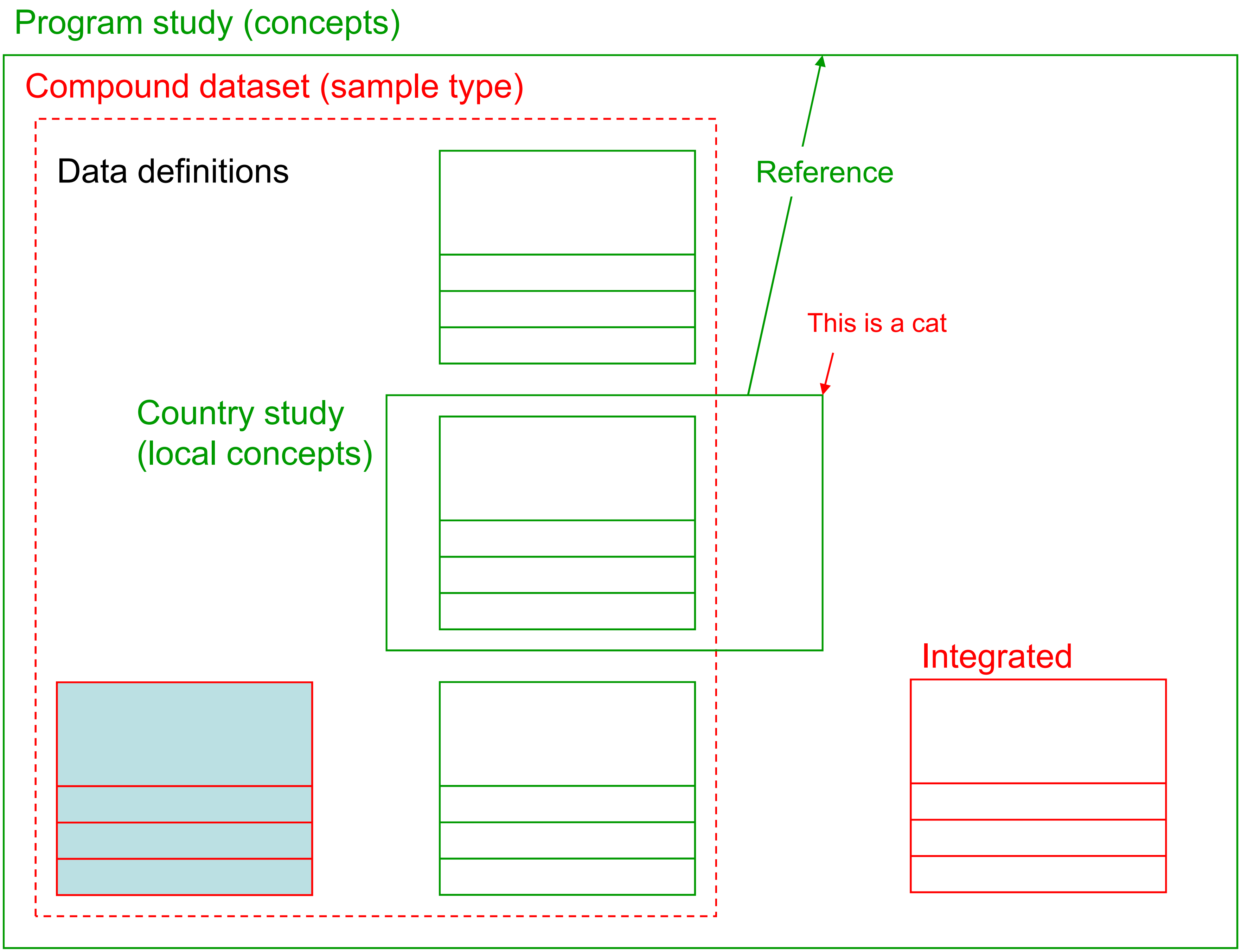


Program study (concepts)

Compound dataset (sample type)

Data definitions

Integrated



Program study (concepts)

Compound dataset (sample type)

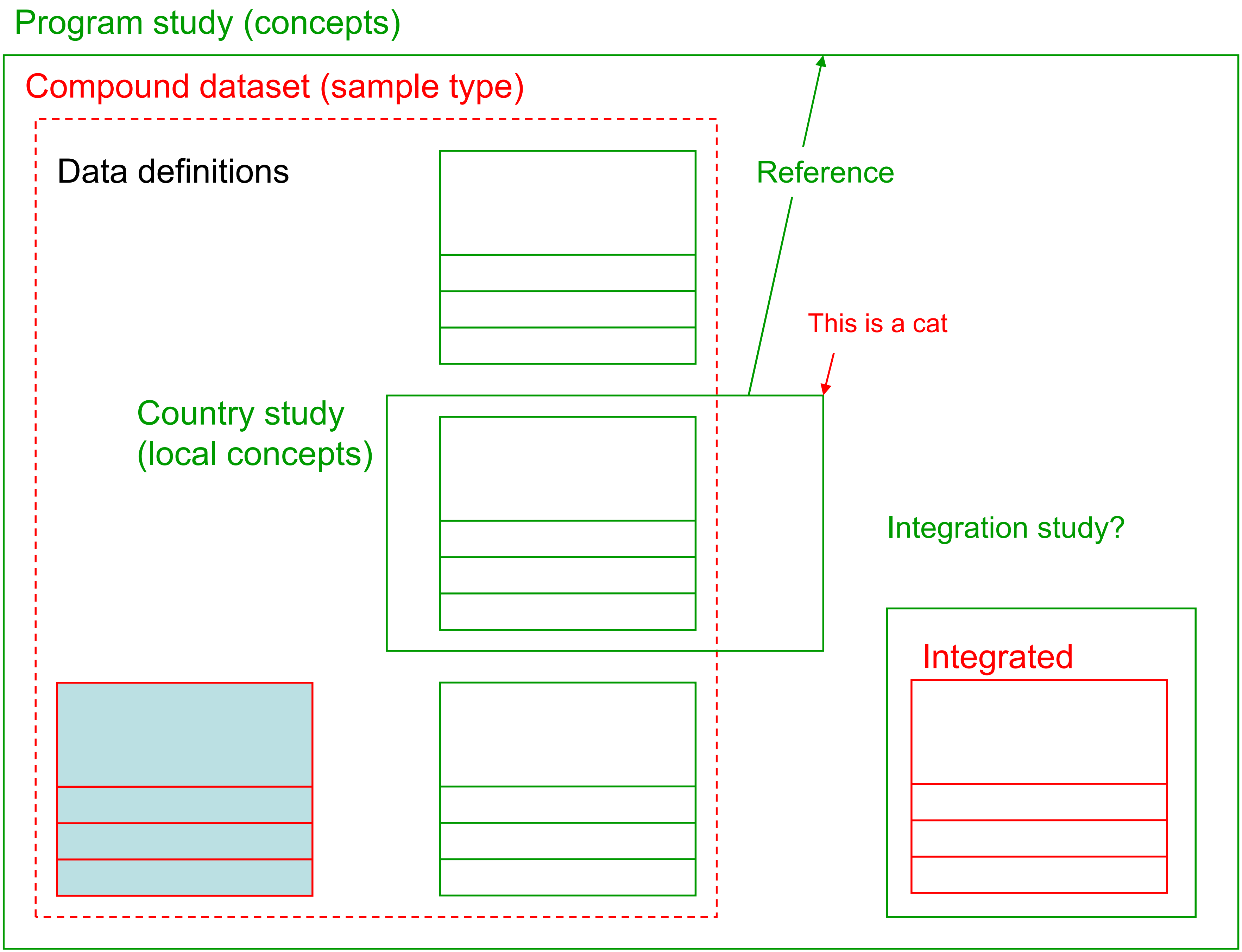
Data definitions

Reference

This is a cat

Country study (local concepts)

Integrated



Program study (concepts)

Compound dataset (sample type)

Data definitions

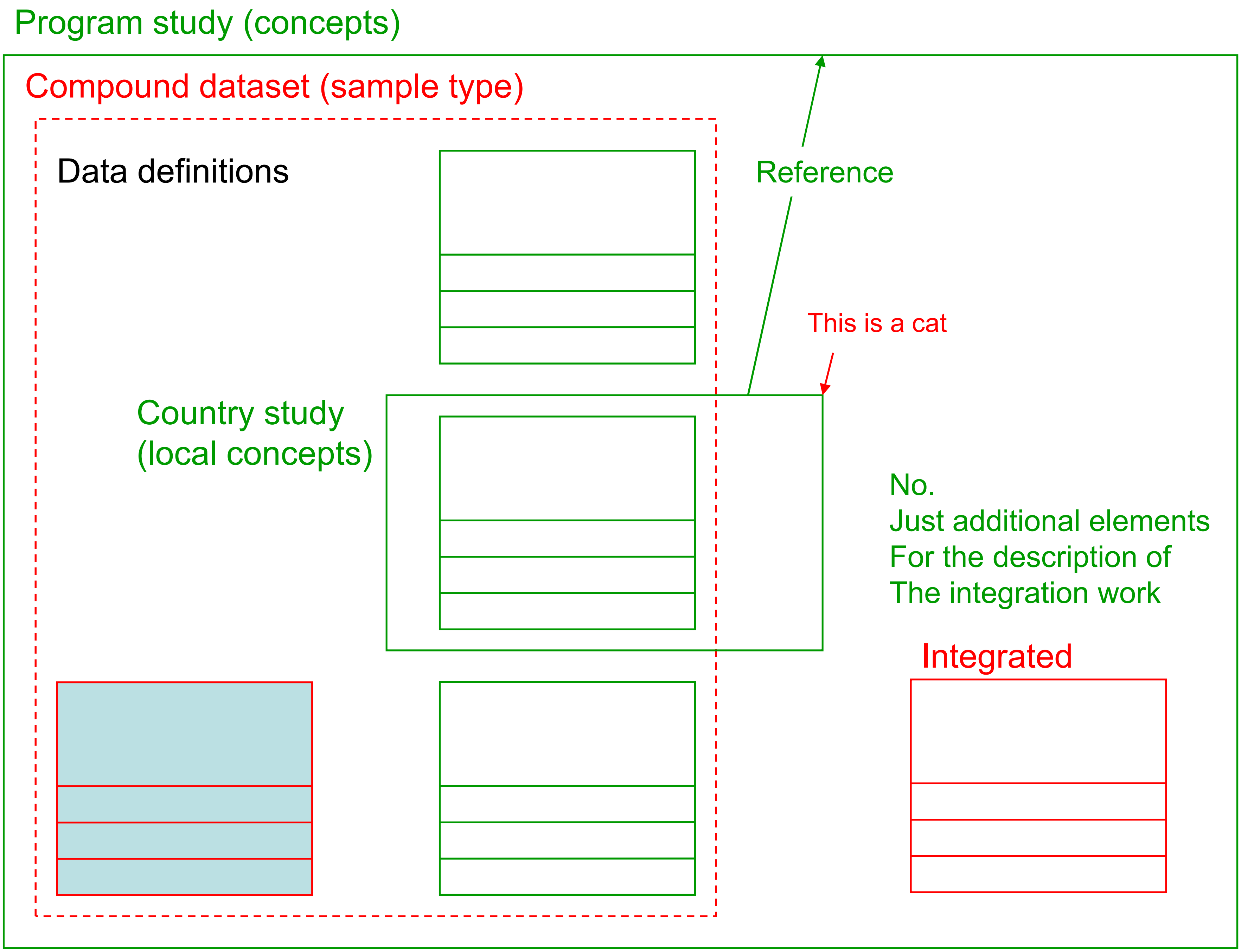
Reference

This is a cat

Country study
(local concepts)

Integration study?

Integrated



Program study (concepts)

Compound dataset (sample type)

Data definitions

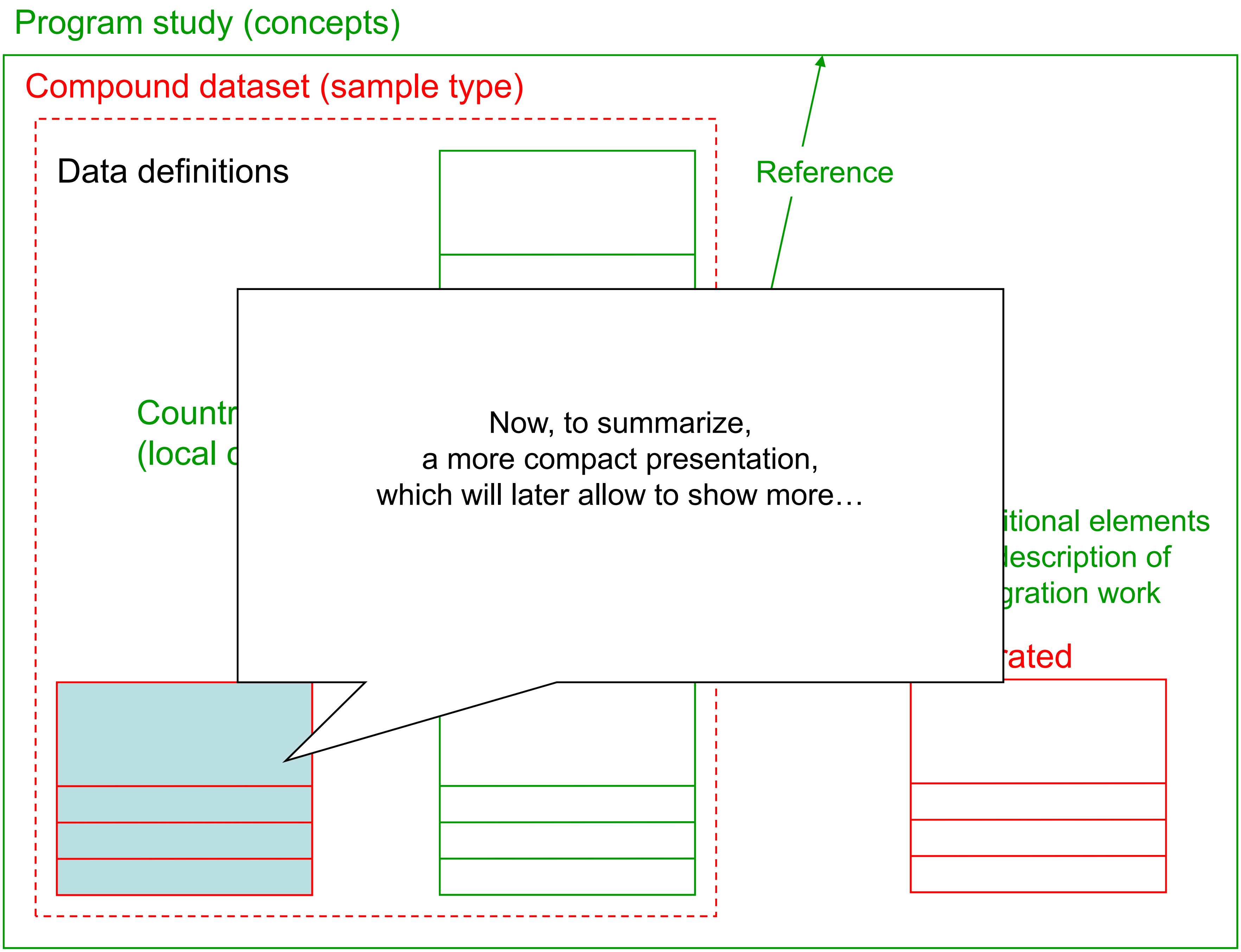
Reference

This is a cat

Country study
(local concepts)

No.
Just additional elements
For the description of
The integration work

Integrated



Program study (concepts)

Compound dataset (sample type)

Data definitions

Reference

Countr

(local c

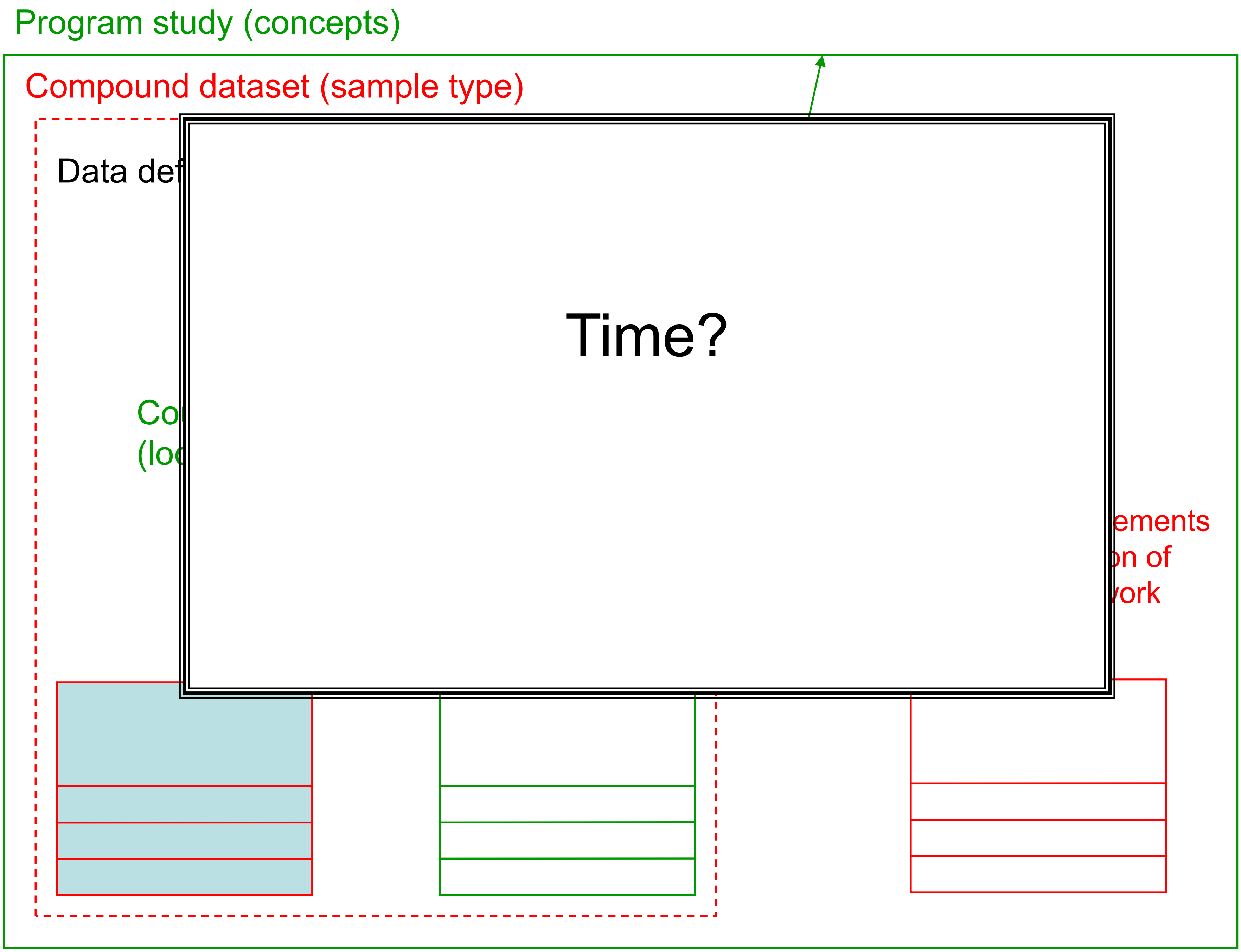
itional elements

lescription of

gration work

rated

Now, to summarize,
a more compact presentation,
which will later allow to show more…



Program study (concepts)

Compound dataset (sample type)

Data def

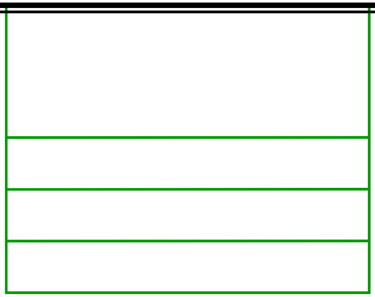
Co
(loc

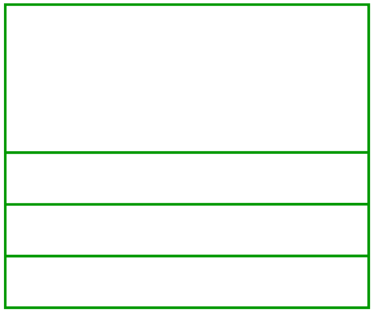
ements
on of
ork

Time?

Time!

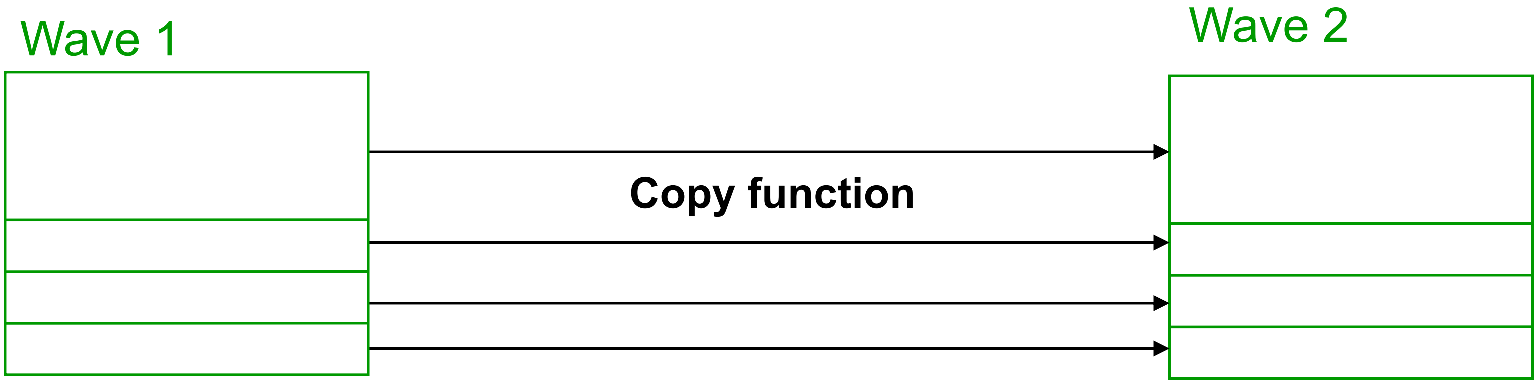
Just keep one dataset to start with

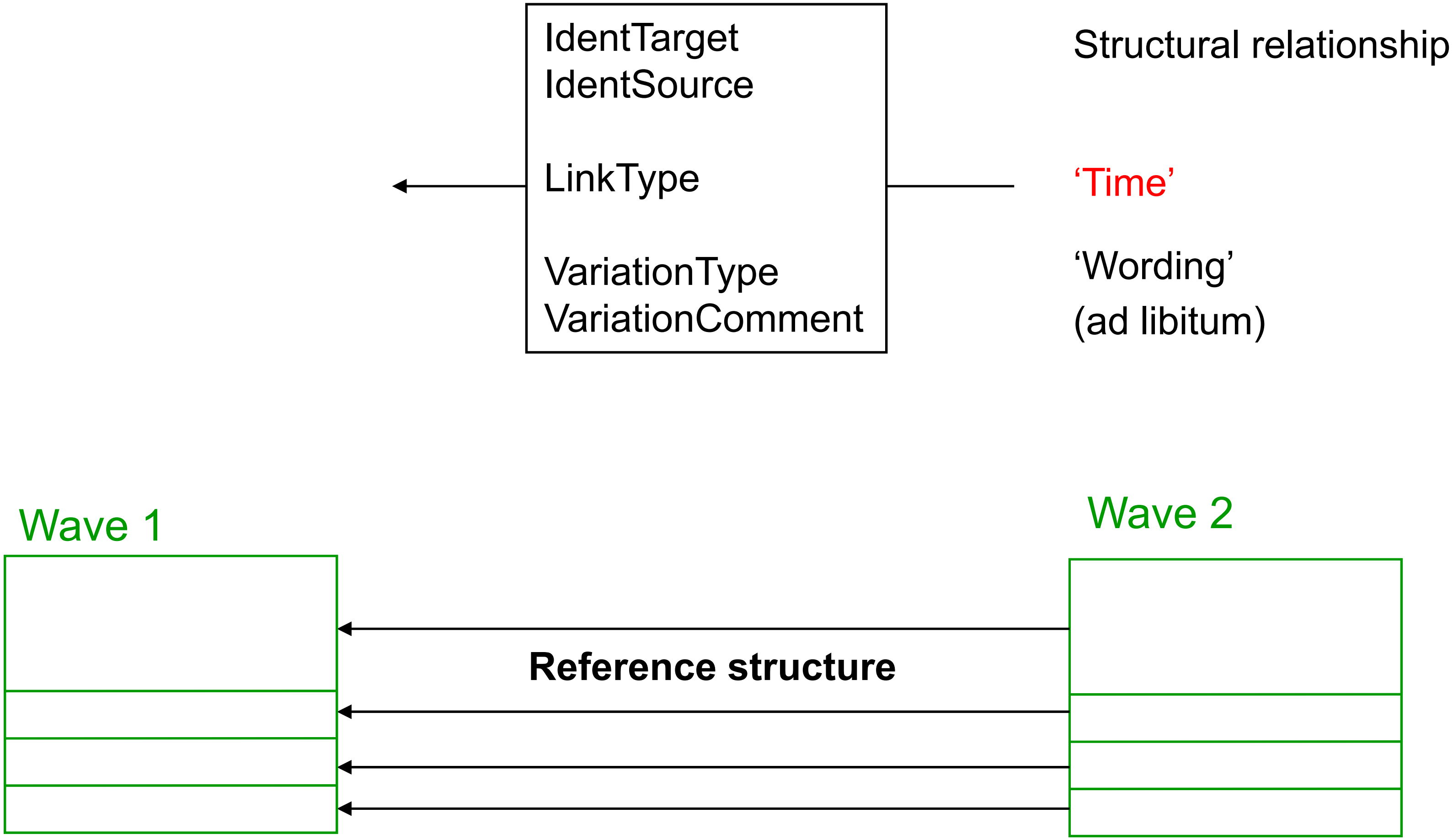




Wave 1

Use the information already in the database to create the metadata for wave 2





IdentTarget
IdentSource

LinkType

VariationType
VariationComment

Structural relationship

‘Time’

‘Wording’

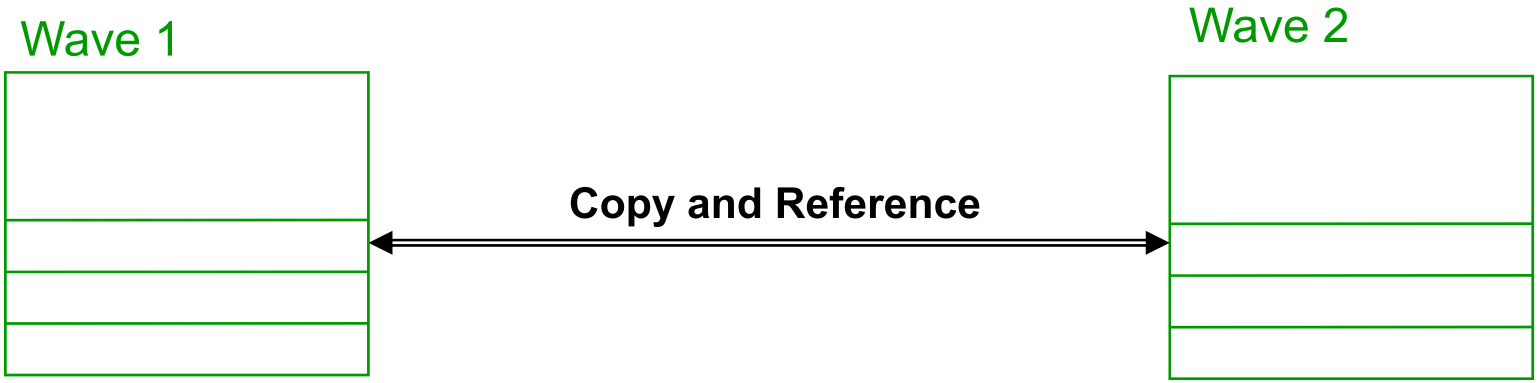
(ad libitum)

Wave 1

Wave 2

Reference structure

Synthetic view of copy function and reference structure:

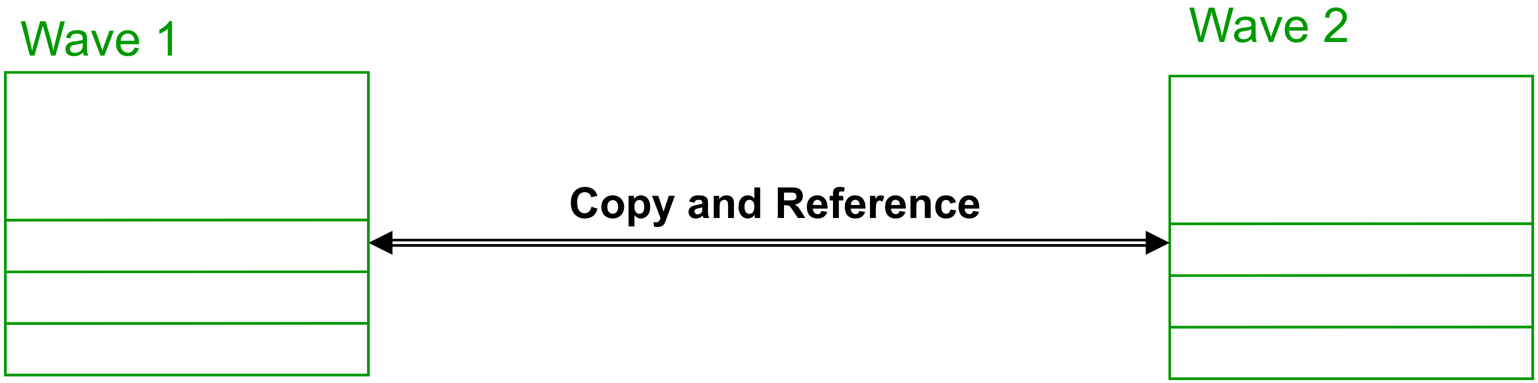


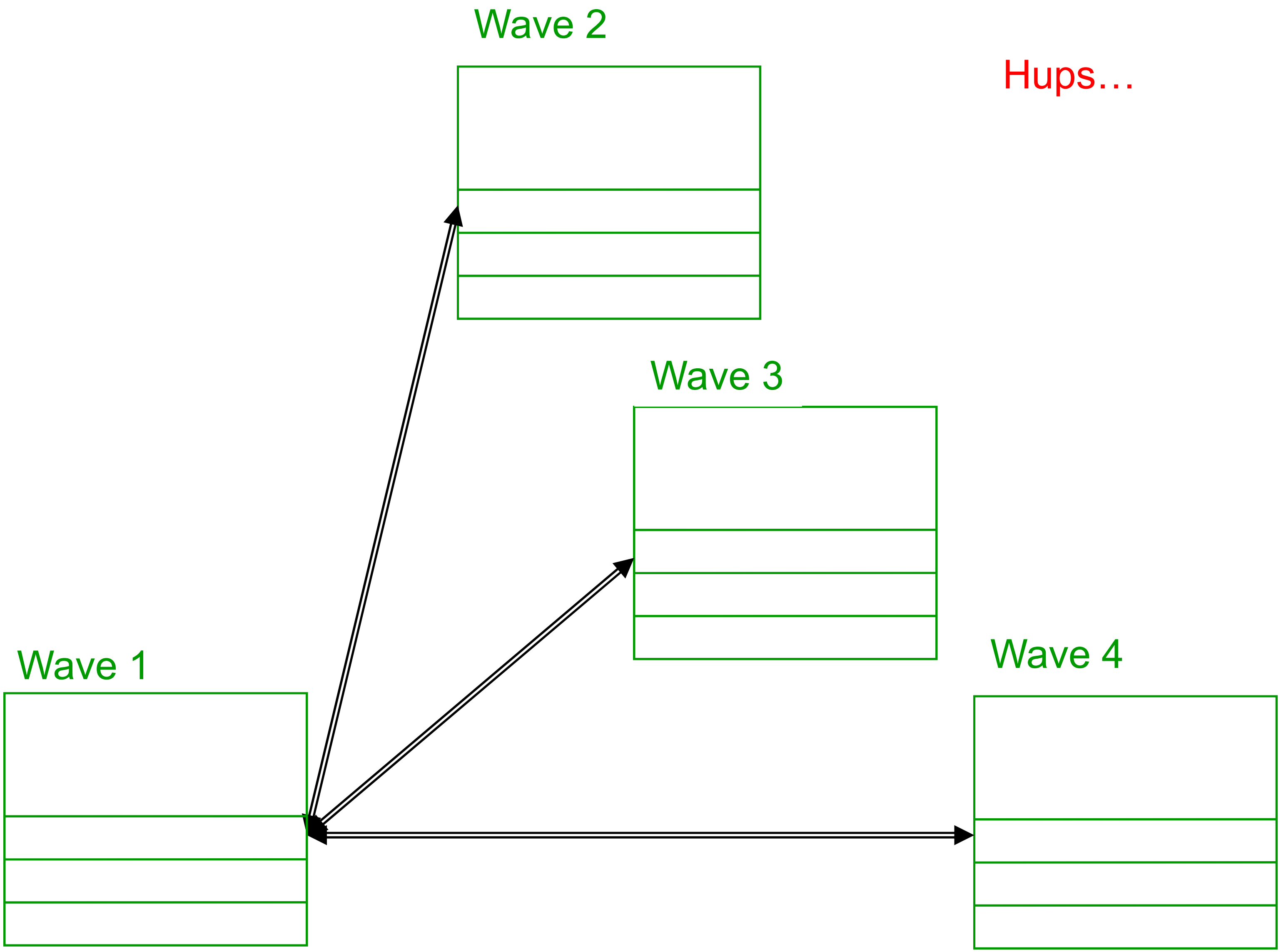
Wave 1

Copy and Reference

Wave 2

Let’s add waves:





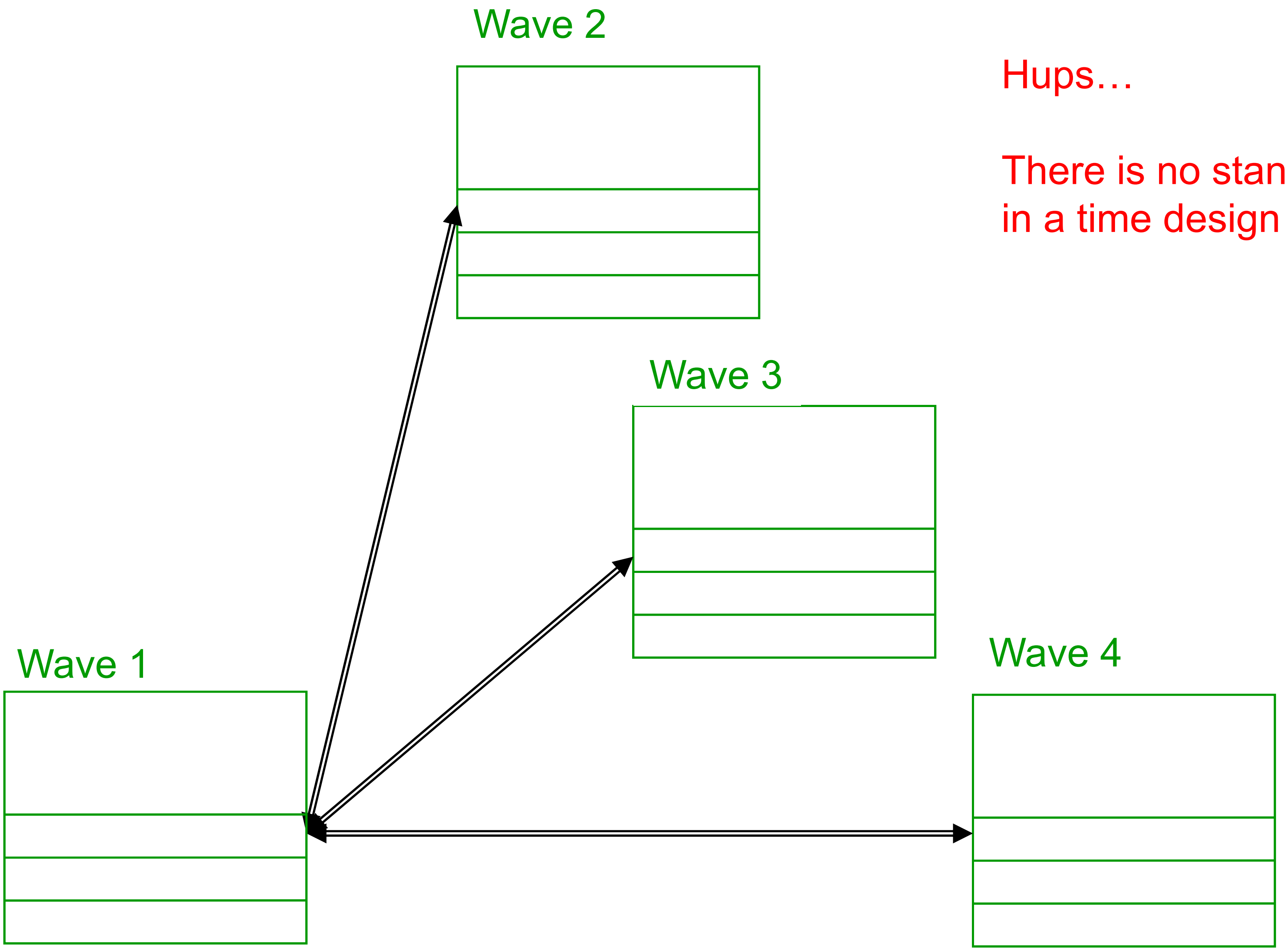
Wave 2

Wave 3

Wave 1

Wave 4

Hups…



Wave 2

Wave 3

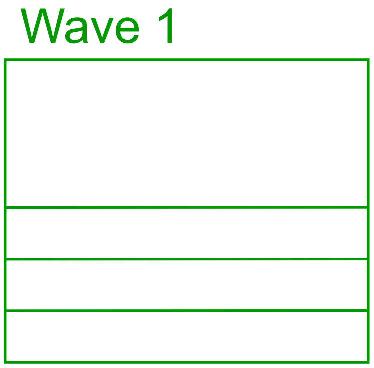
Wave 1

Wave 4

Hups…

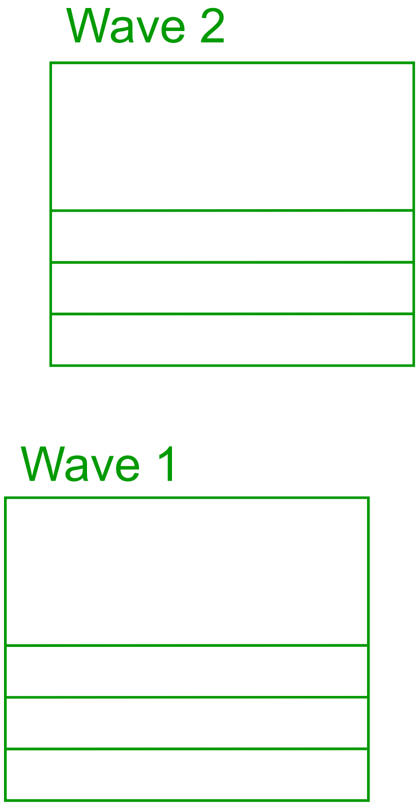
There is no standard
in a time design

Let’s start…



Wave 1

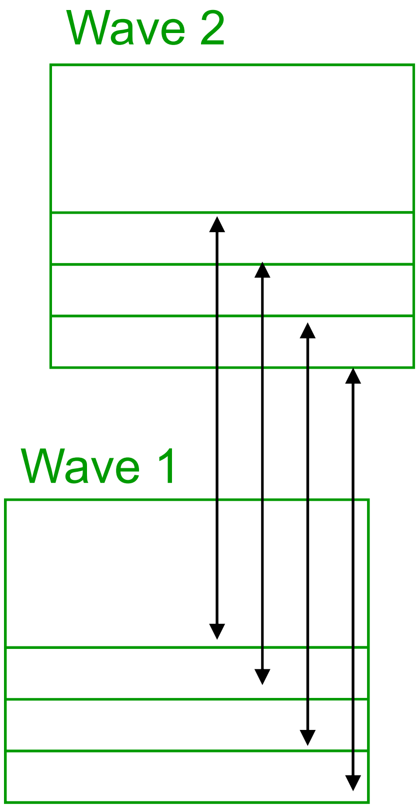
… and repeat



Wave 2

Wave 1

Some questions and variables will be repeated exactly in the same form

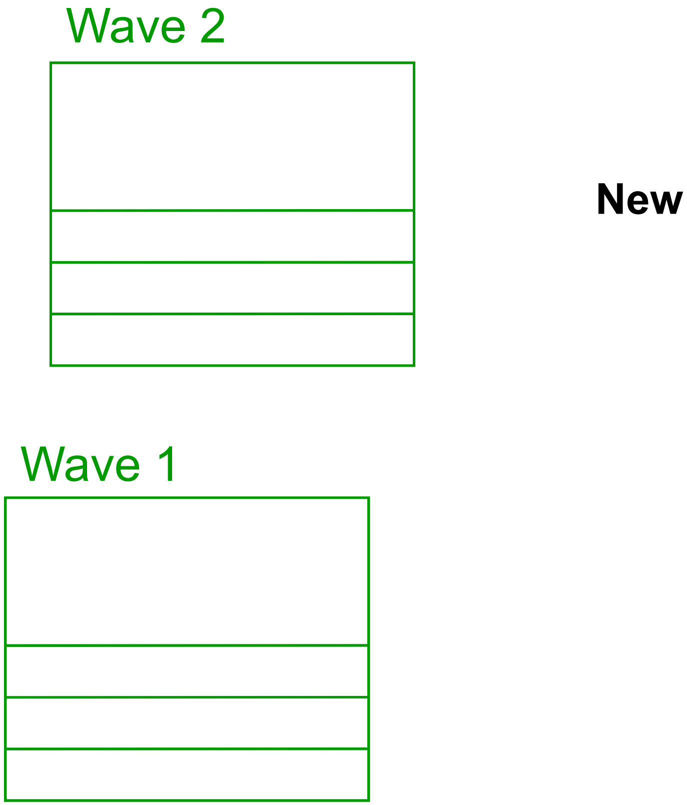


Repeated

Variation type = ‘Identical’

No comment

…others will be new

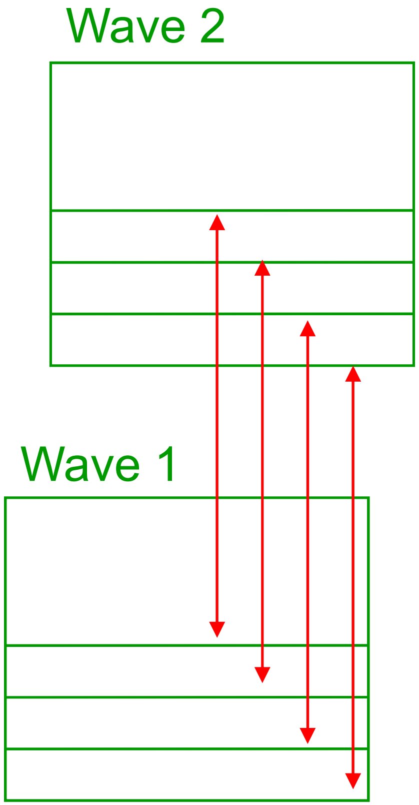


Wave 2

New

Wave 1

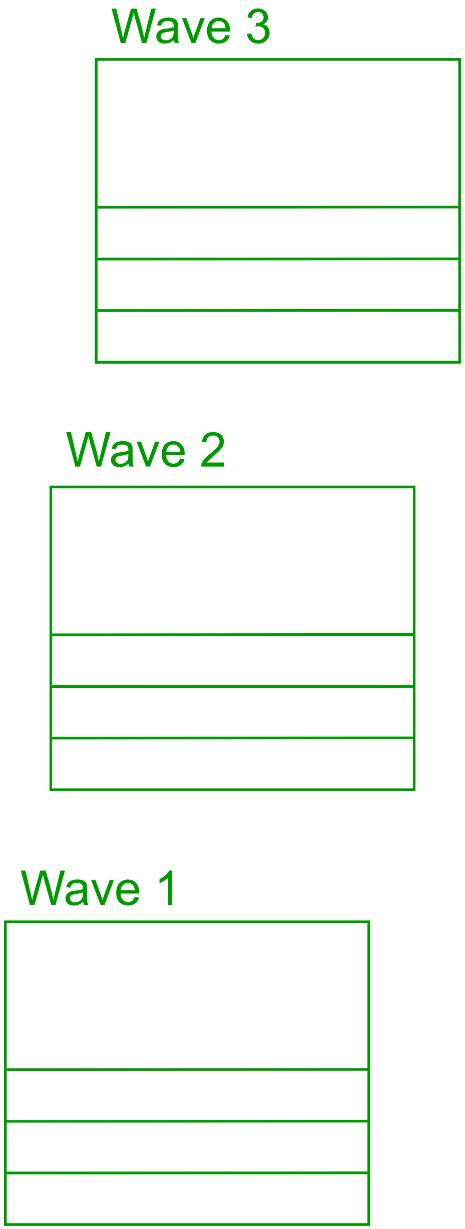
…still others will present variations, without fully breaking up the series



Variant

Variation type = ‘Wording’

Comment on the impact on meaning of the change in the wording

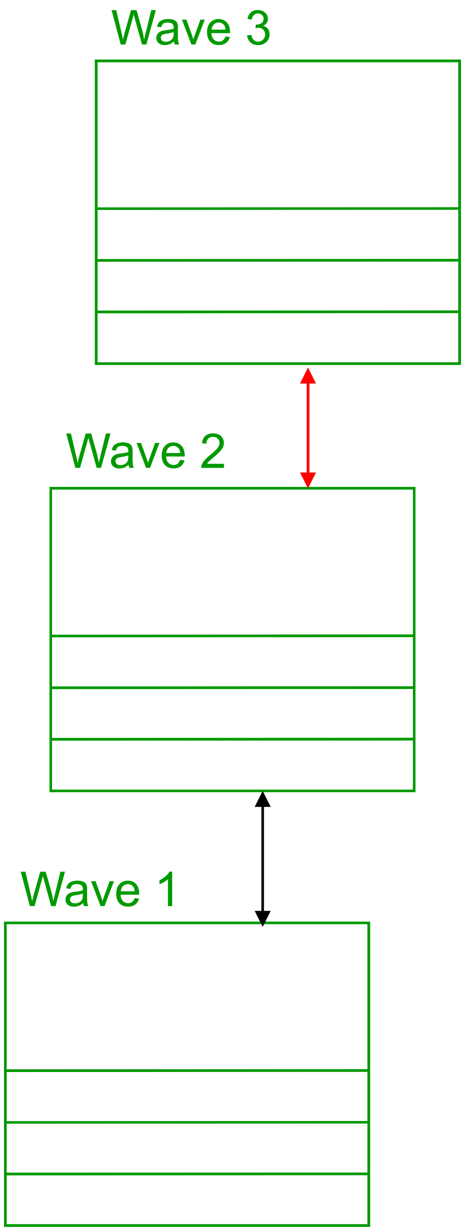


Wave 3

Wave 2

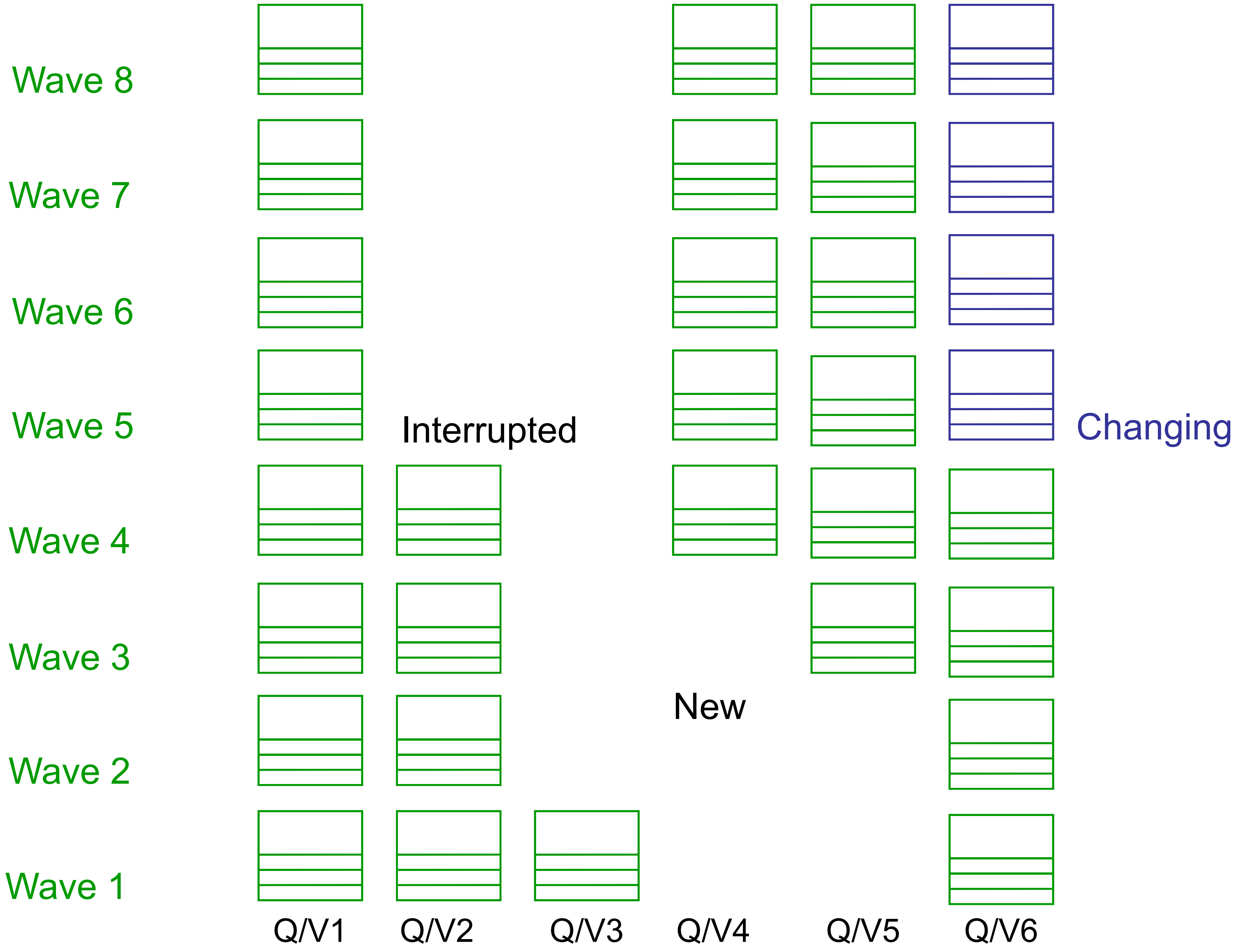
Wave 1

Add more waves



Simplify the presentation of the references

Series of questions/variables



Wave 8

Wave 7

Wave 6

Wave 5

Wave 4

Wave 3

Wave 2

Wave 1

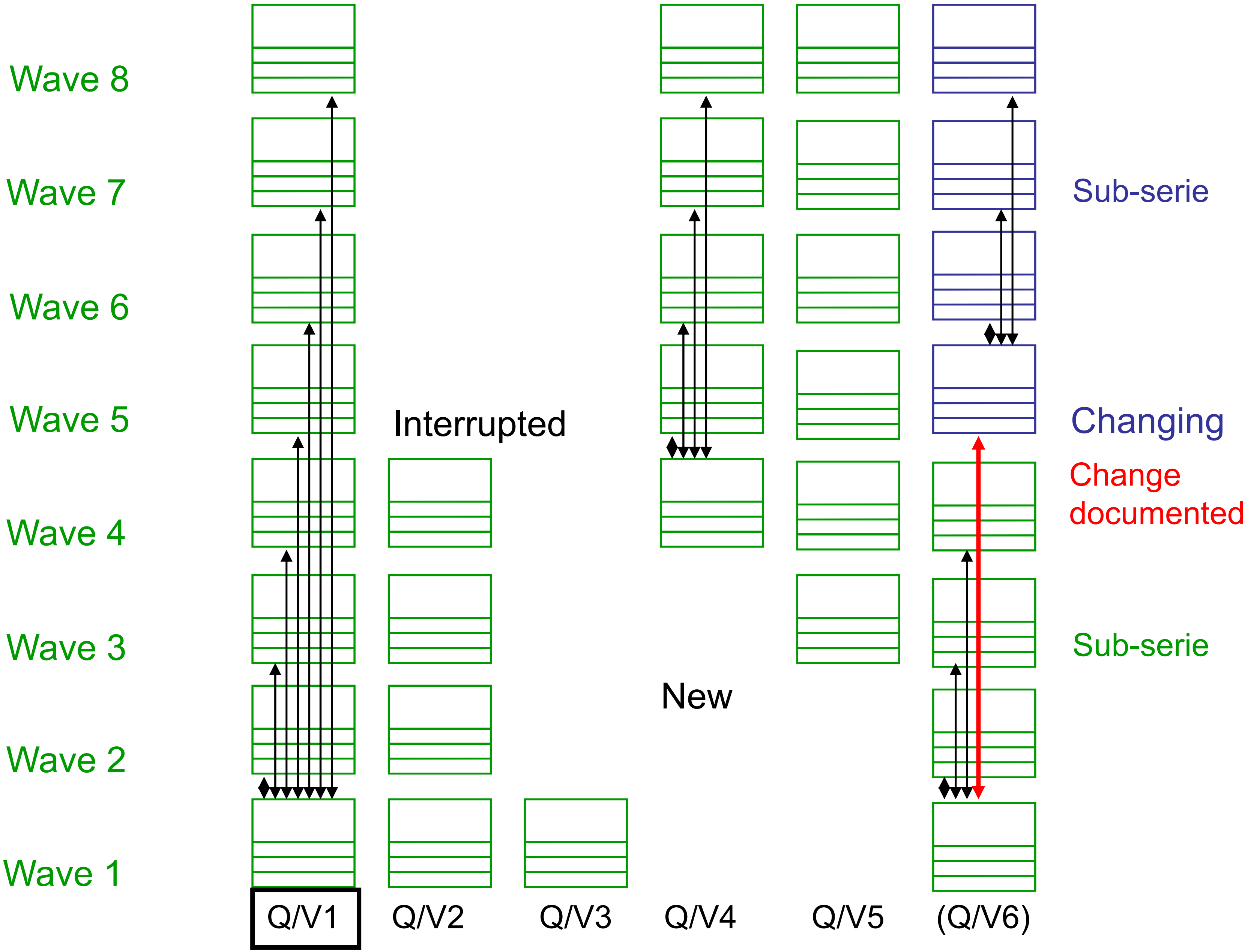
Interrupted

New

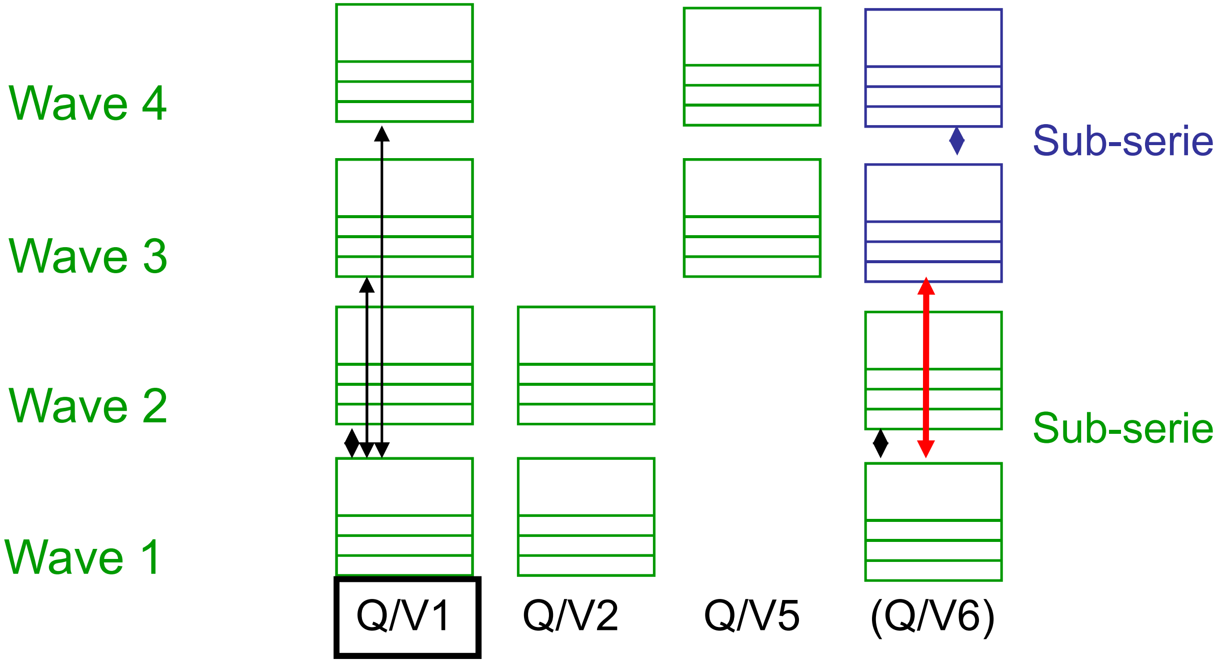
Changing

Q/V1 Q/V2 Q/V3 Q/V4 Q/V5 Q/V6

Series of questions/variables

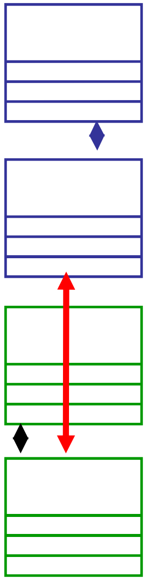


Simplifying the presentation of the longitudinal study



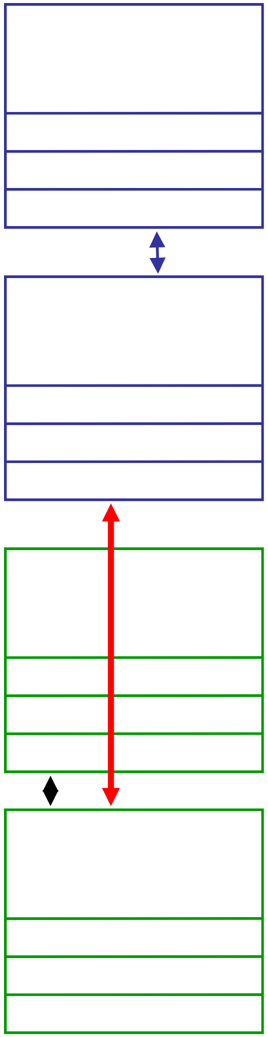
Simplifying still more…

Let this series of variables be a general representation of a series of datasets over time

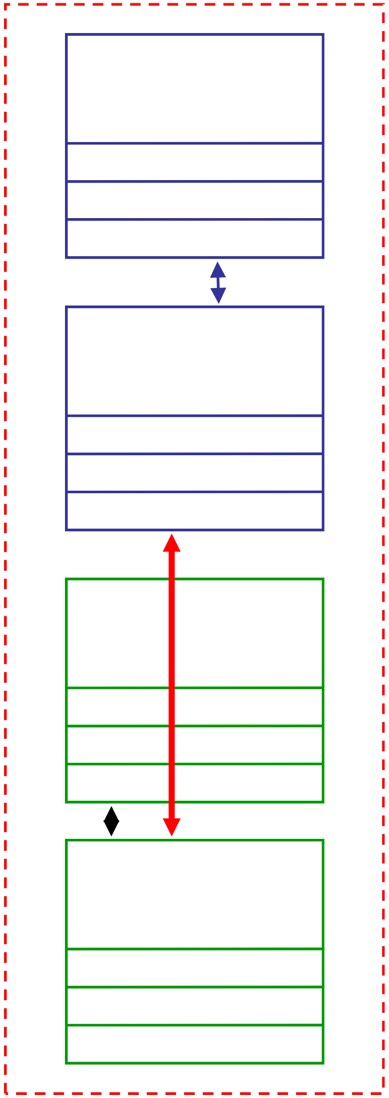


(Q/V6)

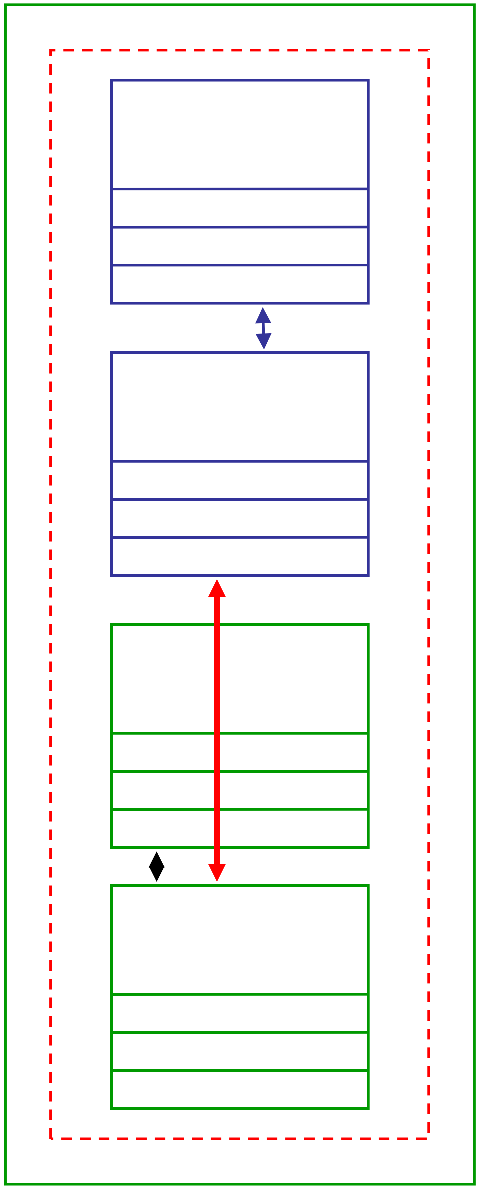
…enlarge it



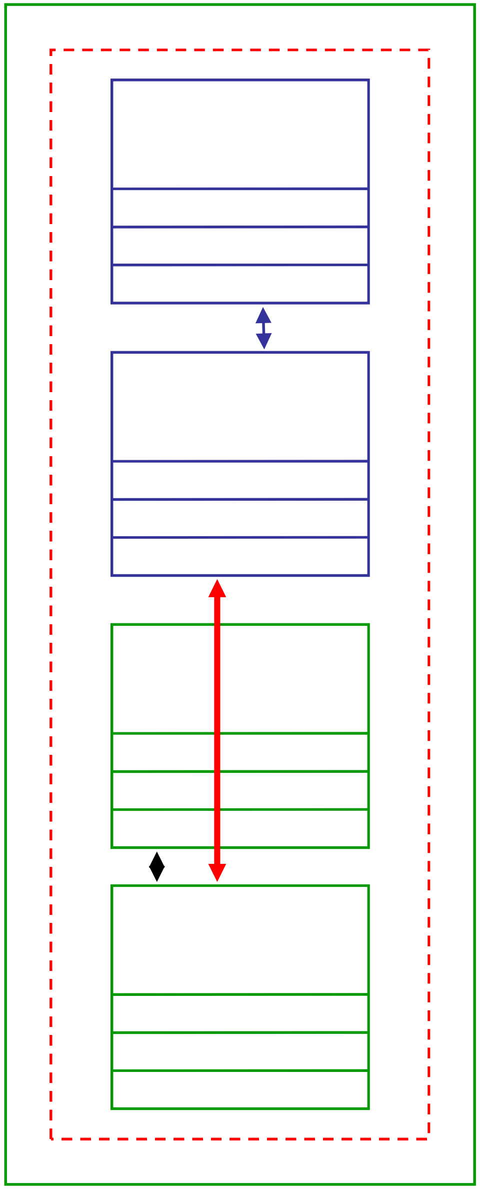
This is actually a compound dataset, a time-compound dataset
(higher level dataset)



…which may be included in the longitudinal study



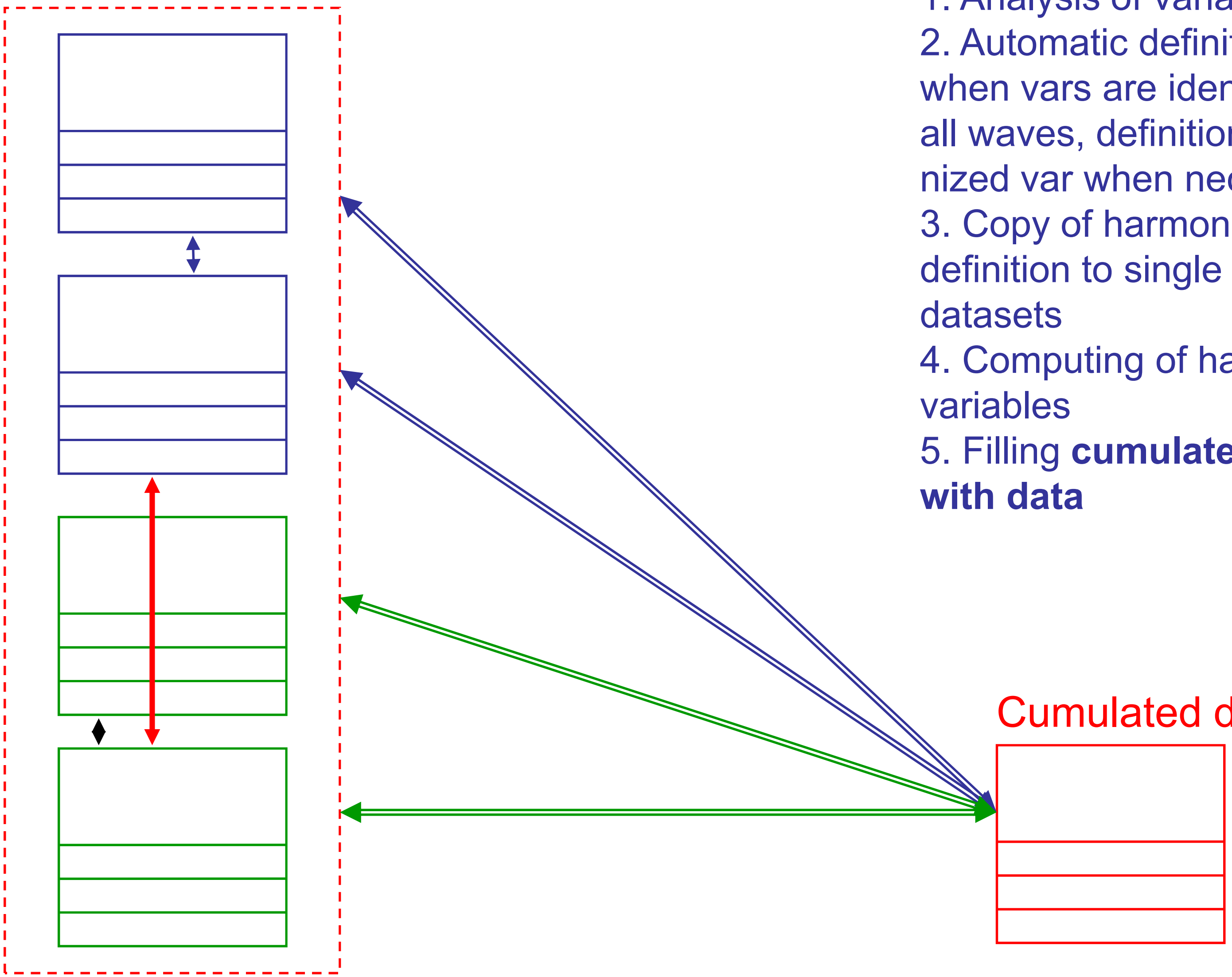
…and be presented as a whole with metadata in a hyperlinked format



Four datasets

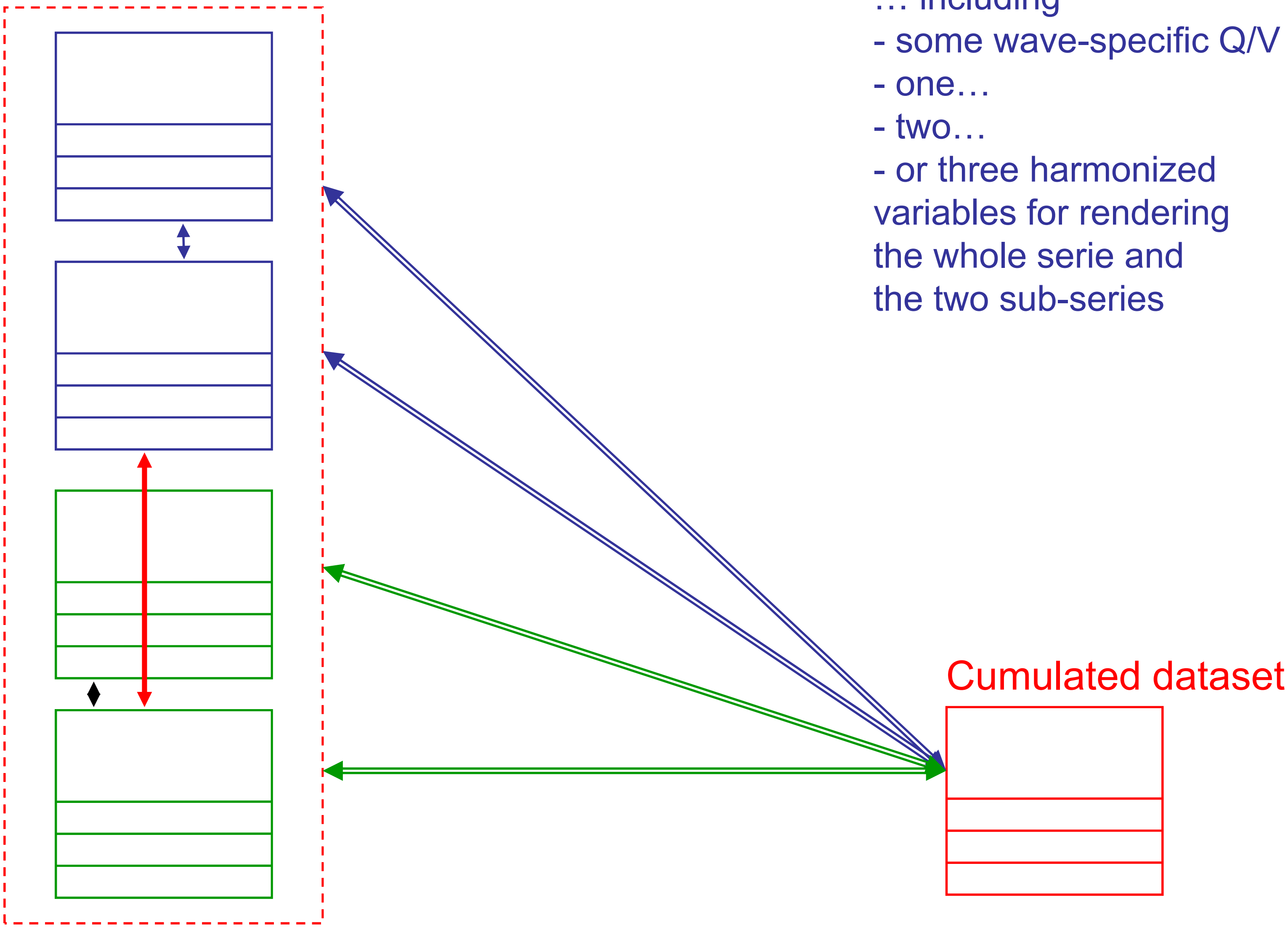
One hyperlinked metadata product

And we can readily define a cumulated dataset



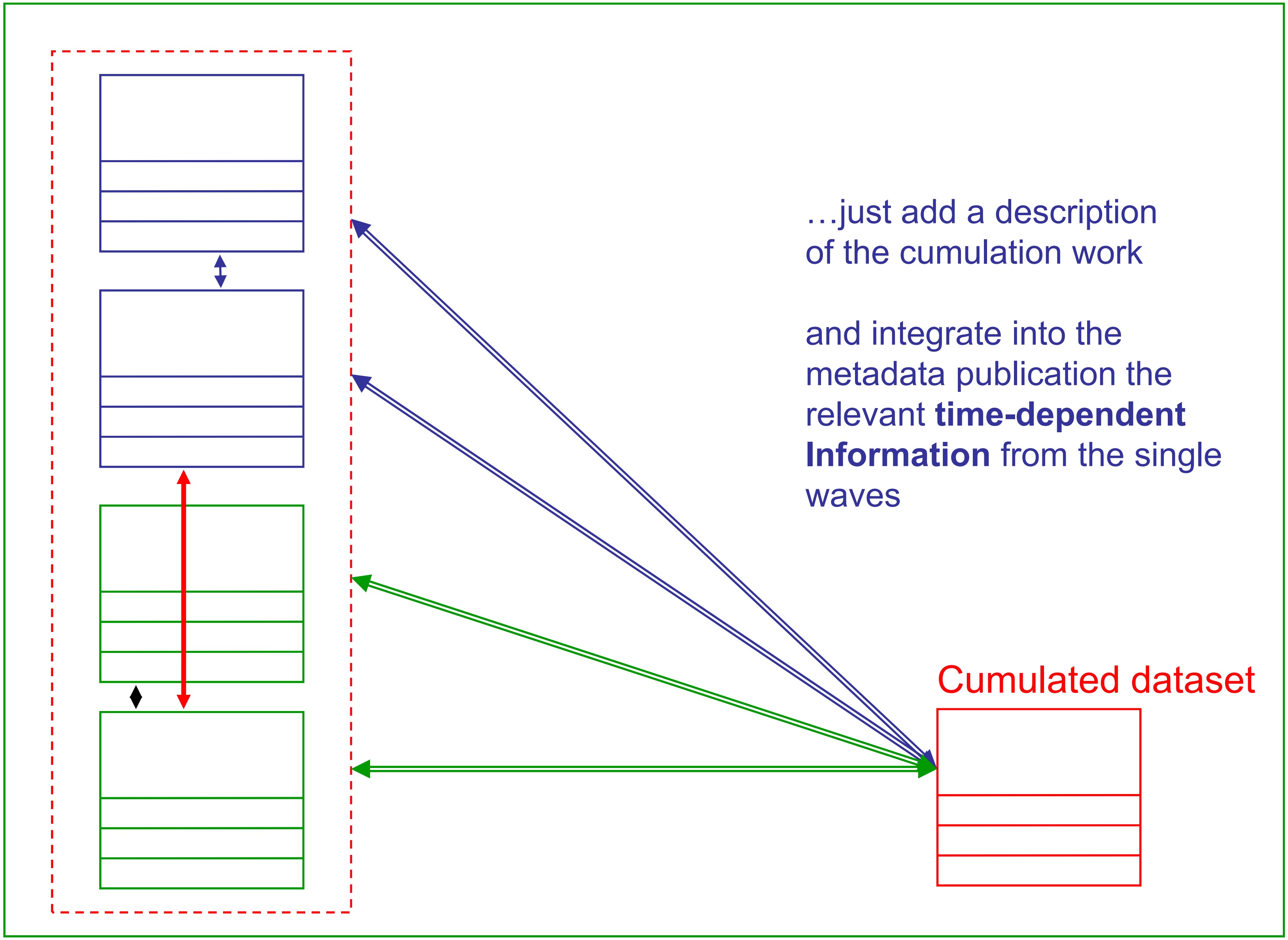
1. Analysis of variations
2. Automatic definition of vars when vars are identical over all waves, definition of harmonized var when necessary
3. Copy of harmonized definition to single wave datasets
4. Computing of harmonized variables
5. Filling **cumulated dataset with data**

And we can readily define a cumulated dataset



… including
- some wave-specific Q/V
- one…
- two…
- or three harmonized variables for rendering the whole serie and the two sub-series

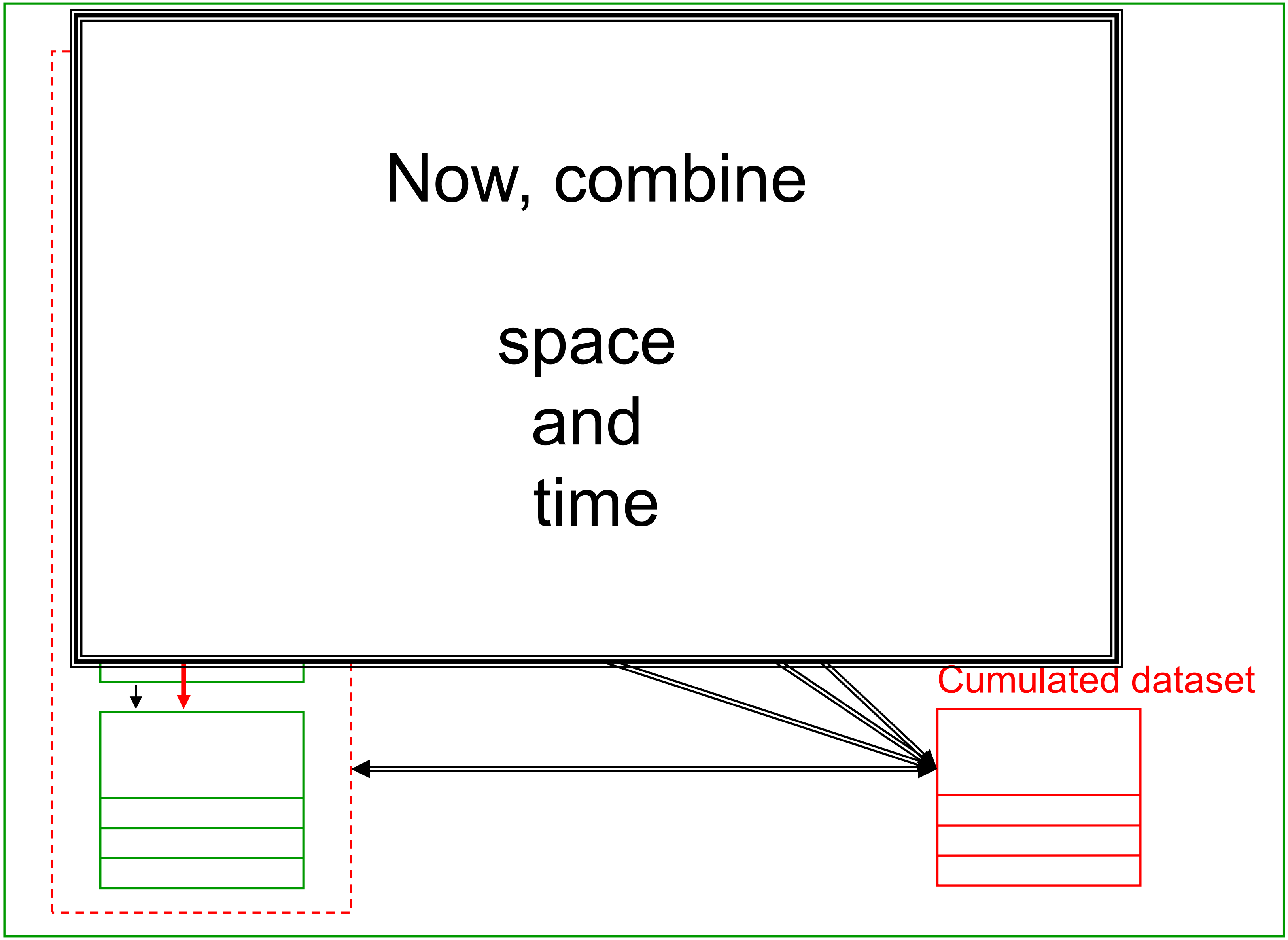
The overall study is here also the reference for the cumulated dataset



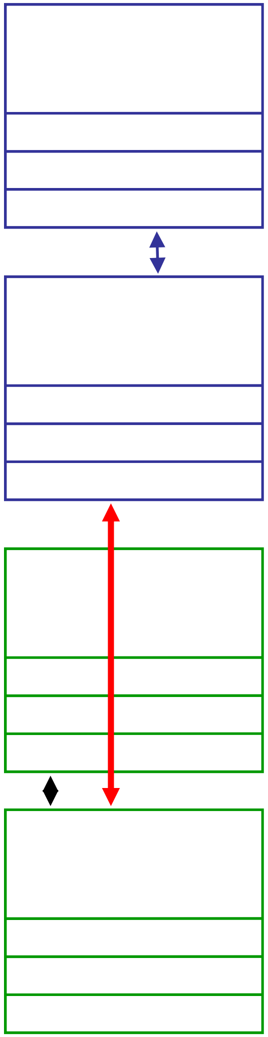
…just add a description of the cumulation work

and integrate into the metadata publication the relevant **time-dependent Information** from the single waves

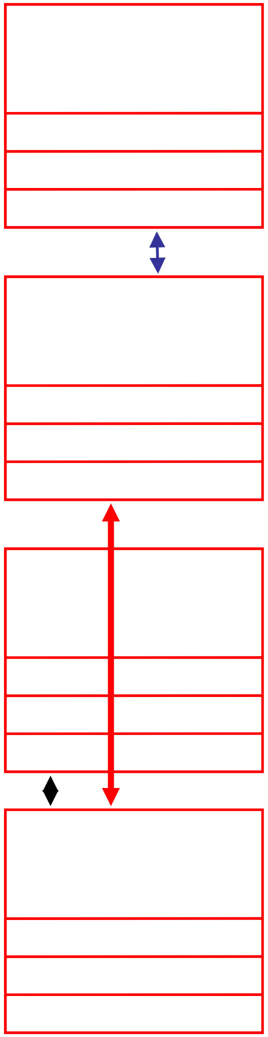
And we can readily define a cumulated dataset for integration



Imagine the standard definition evolves in time as a longitudinal study

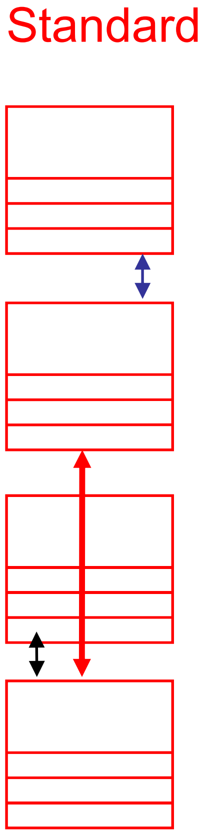


Imagine the standard definition evolves in time as a longitudinal study



Series of successive standard definitions

Imagine the standard definition evolves in time as a longitudinal study

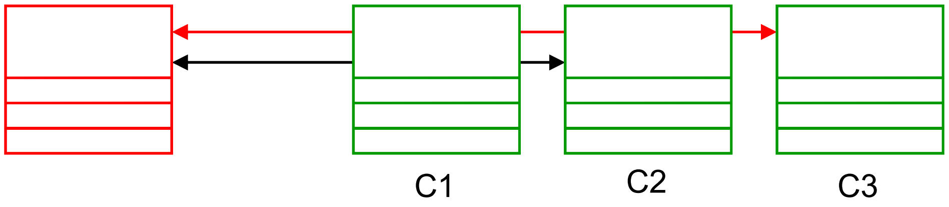


Now imagine the countries are distributed on the horizontal axis

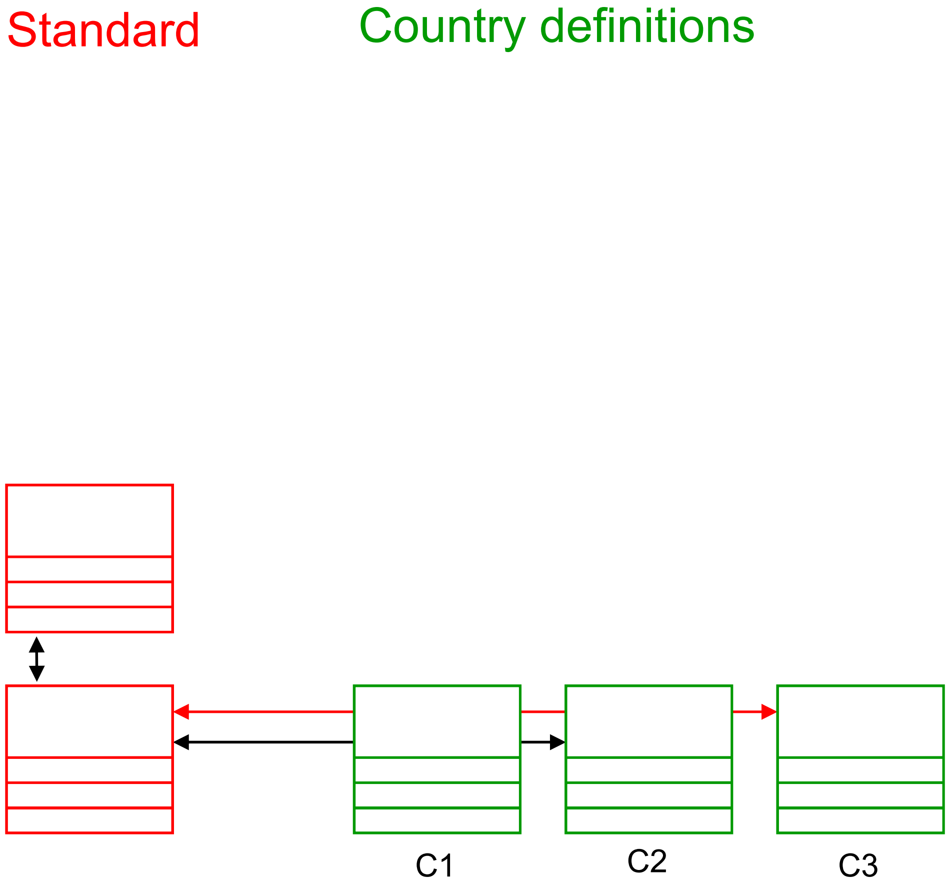
Standard

Country definitions

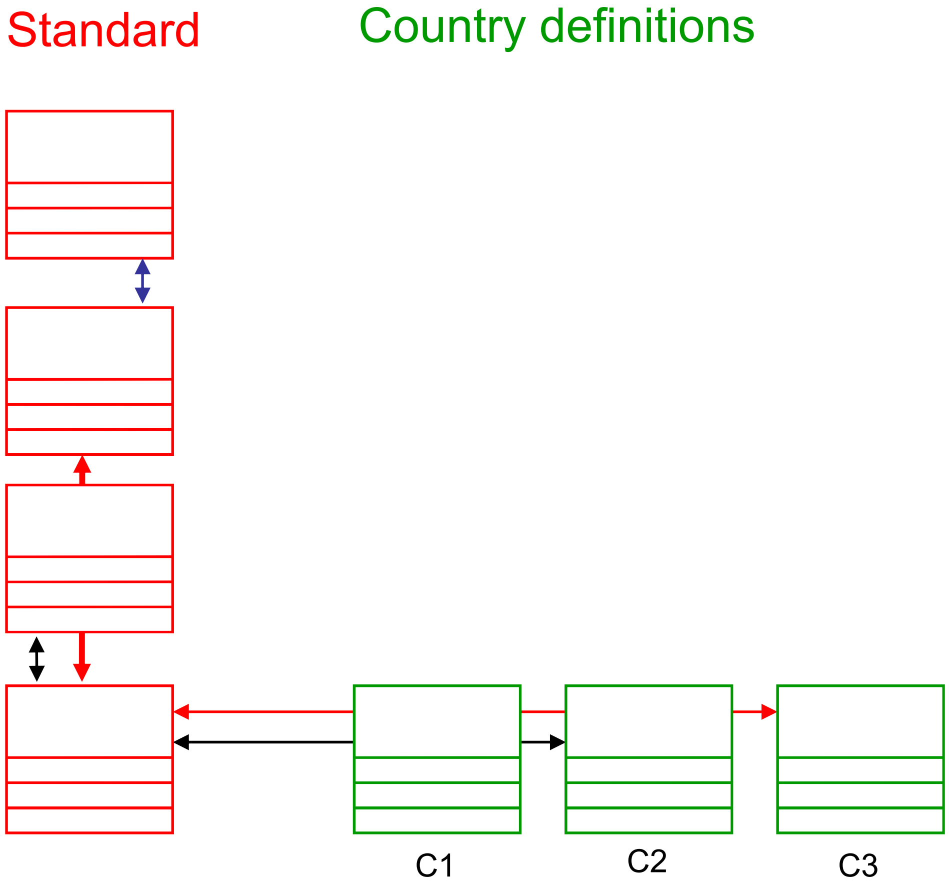
… still symbolizing a country-specific variable in C1,
an identical variable in C2 and
a variant in C3



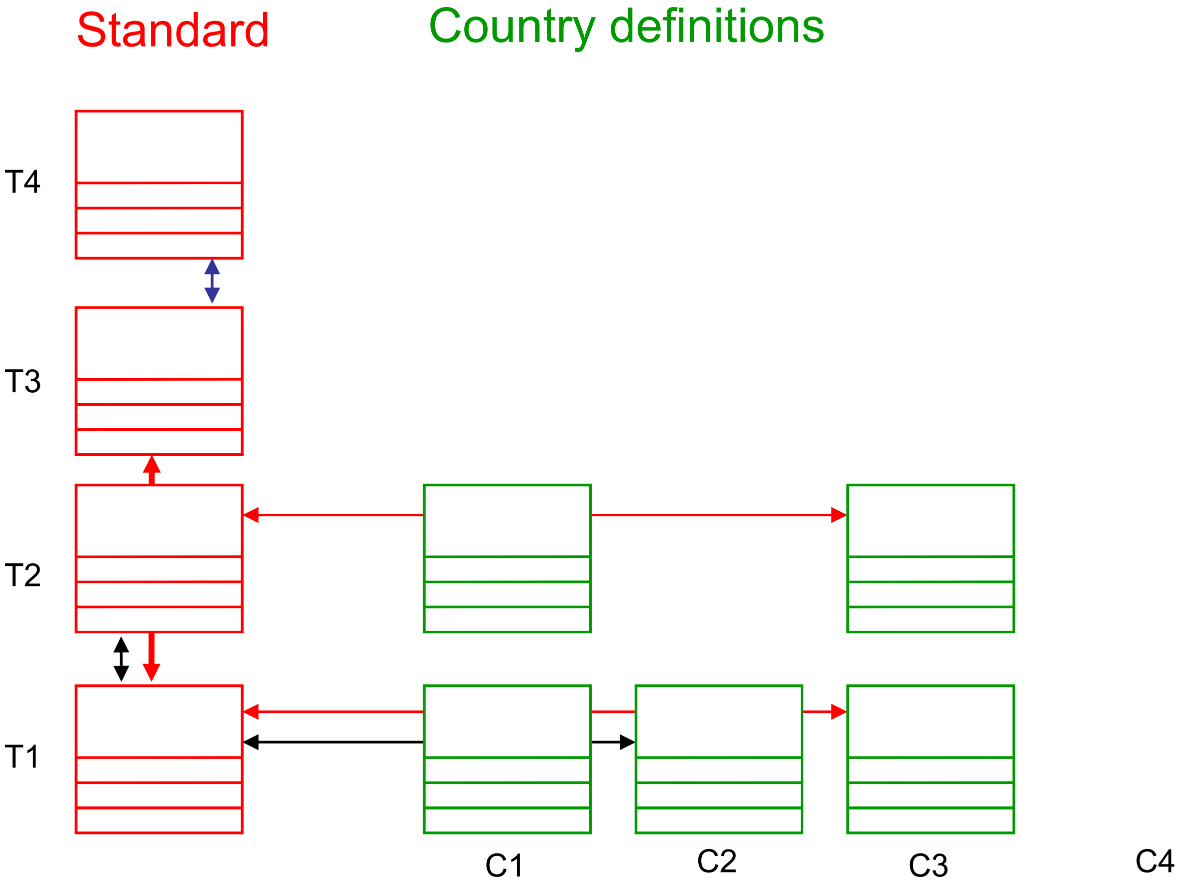
… the standard for the second cross-national wave can be defined using the relationships for the longitudinal study



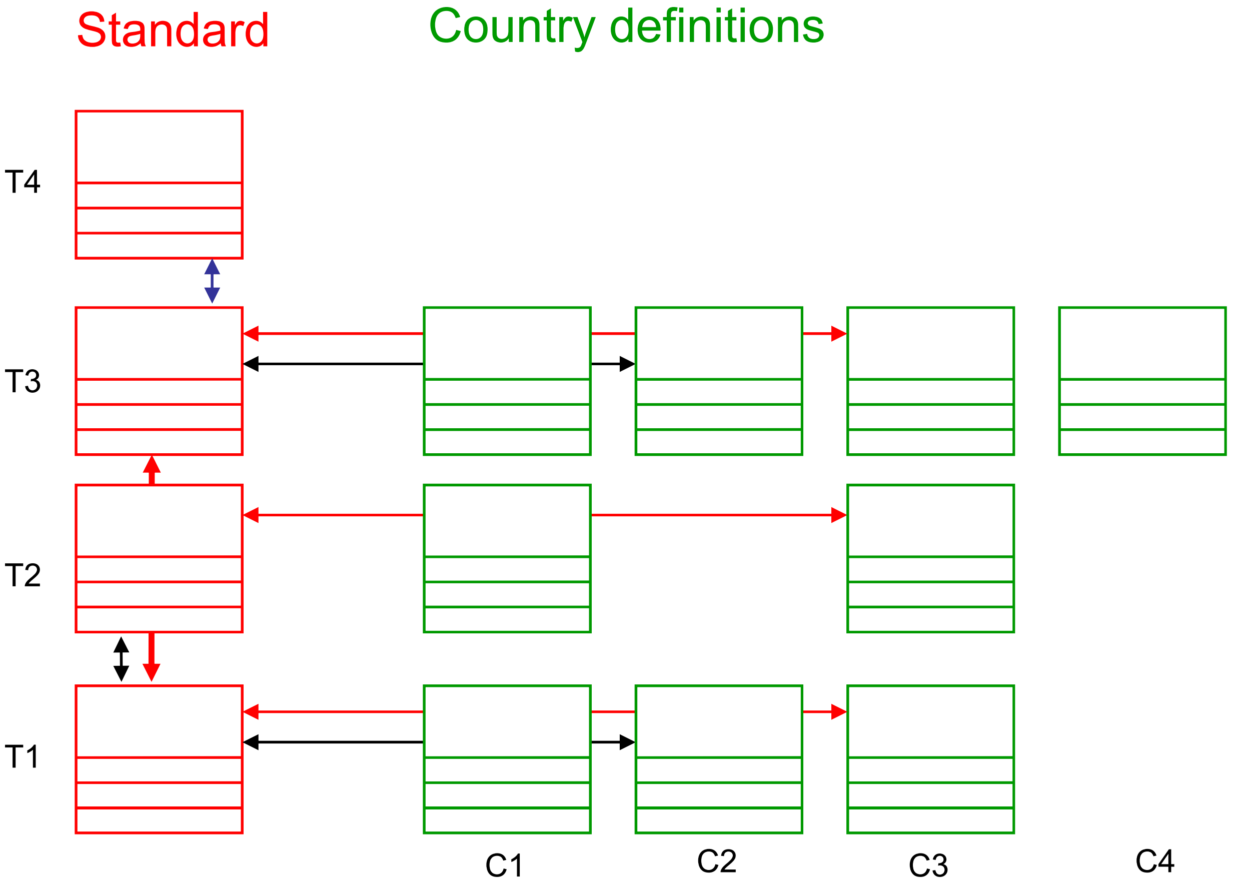
… and so on:



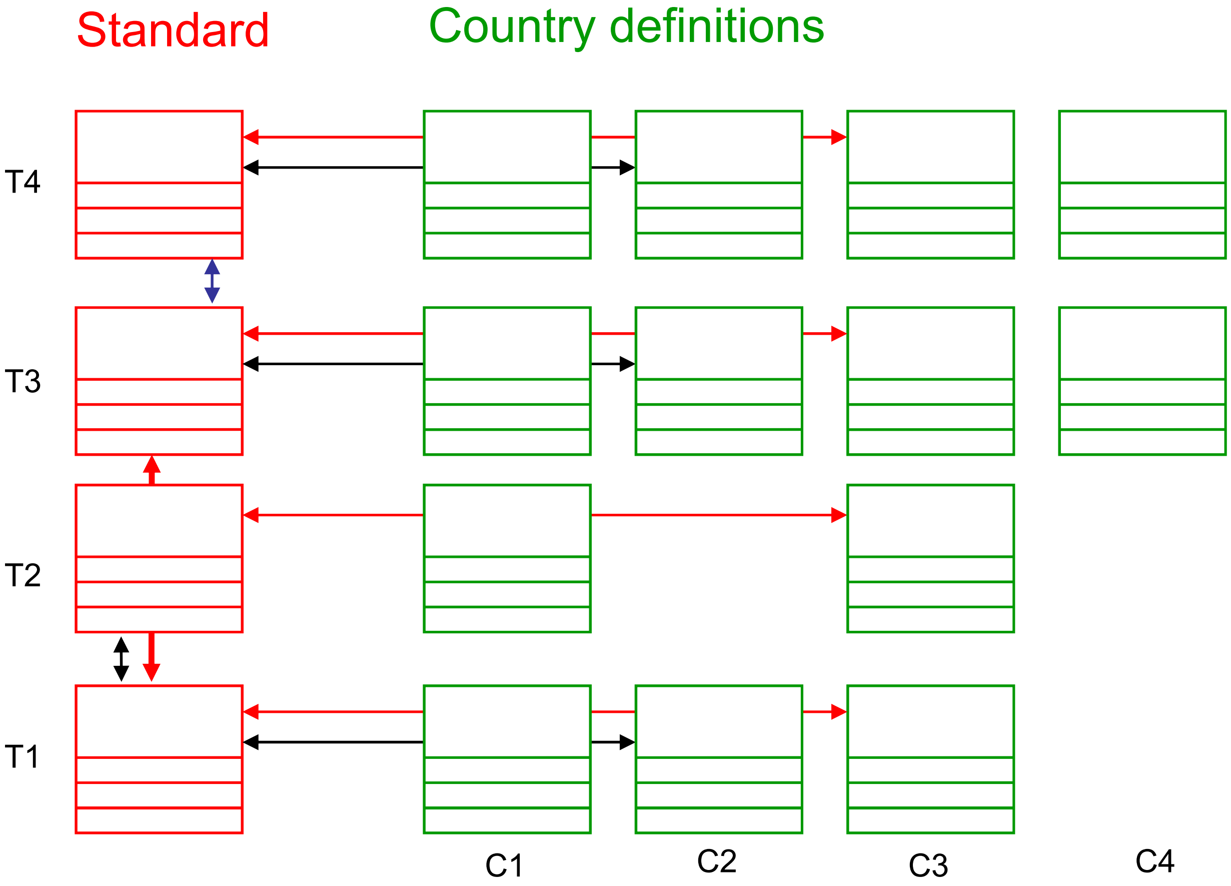
The country datasets appear to grow like the branches of a christmas tree



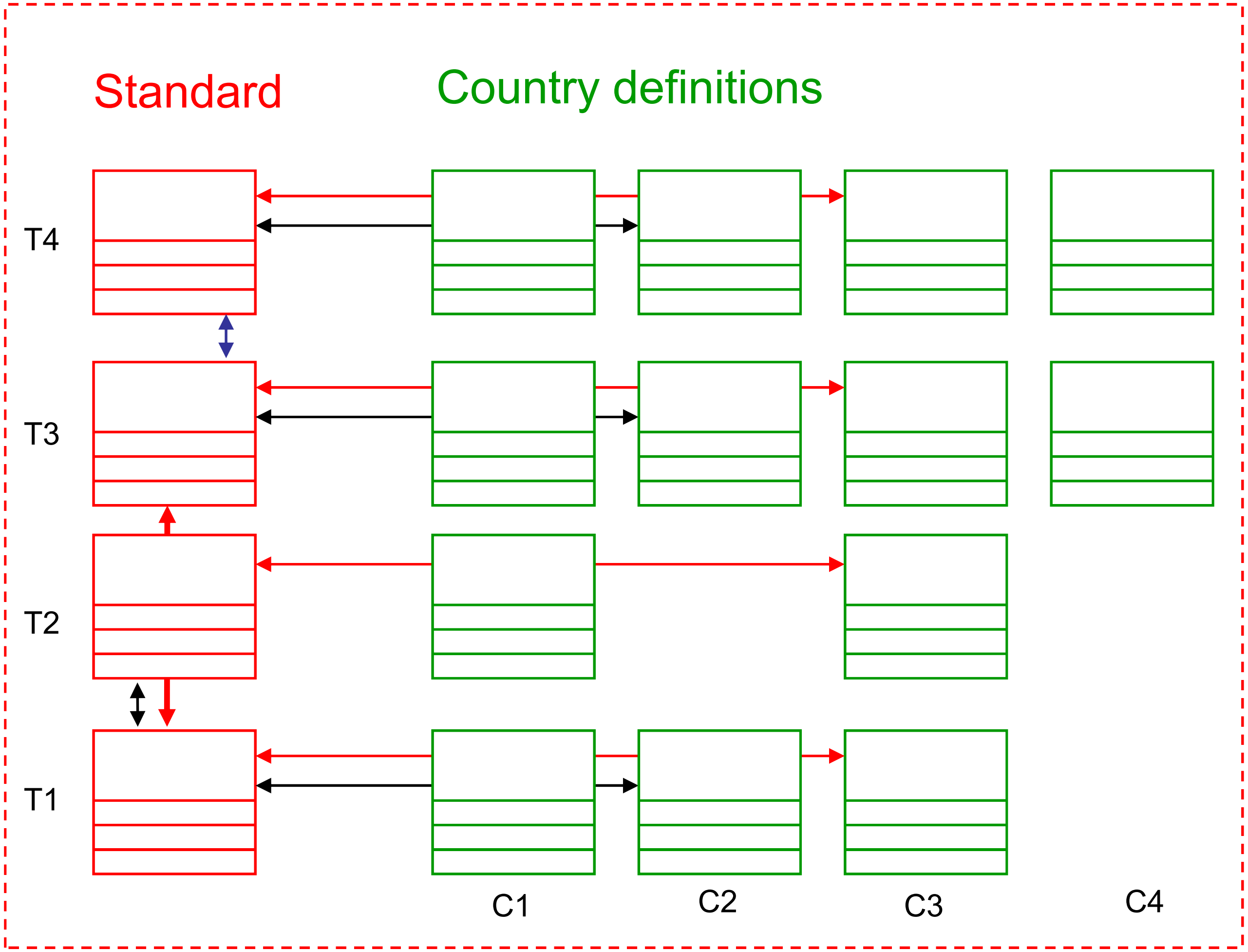
The country datasets appear to grow like the branches of a christmas tree



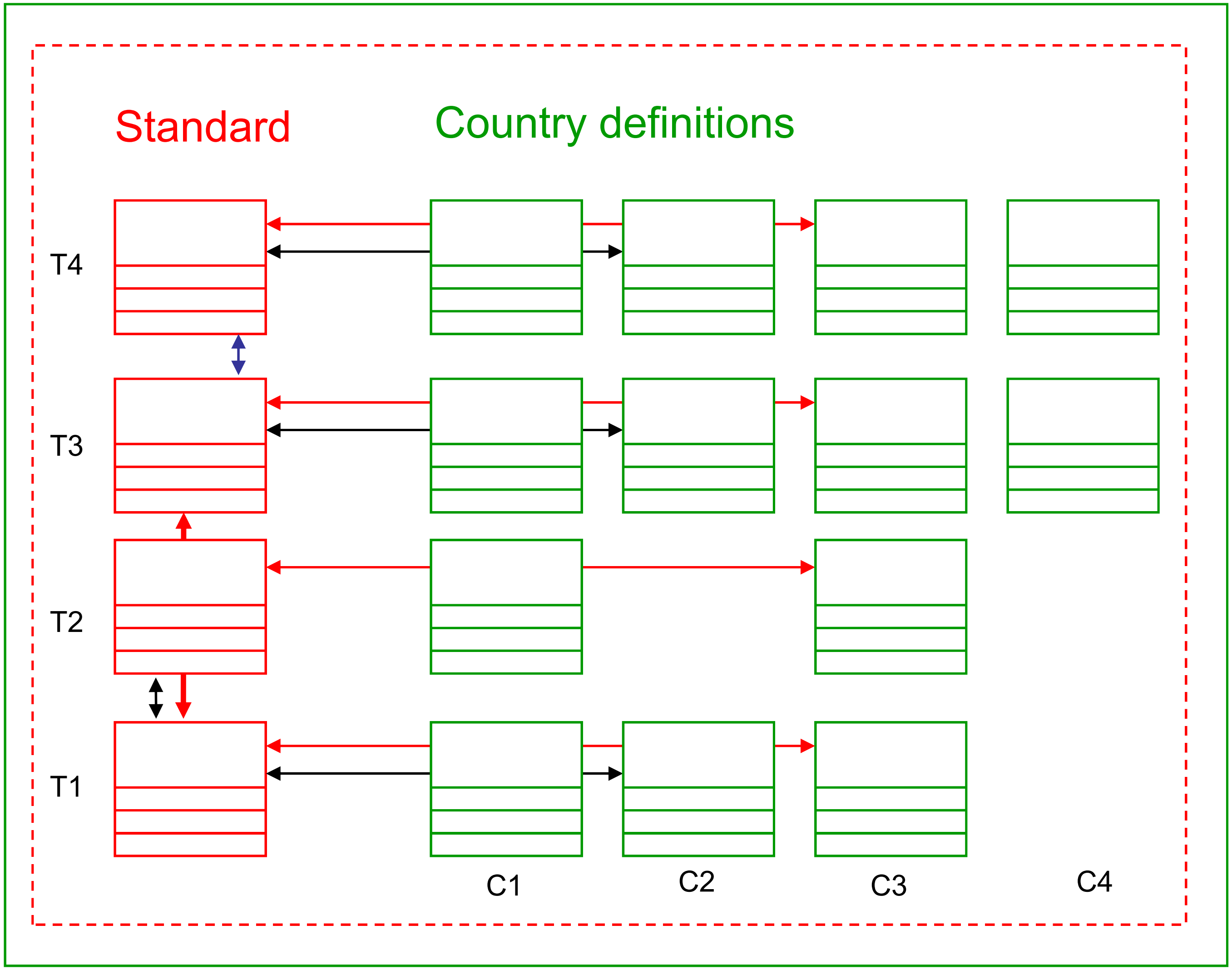
The country datasets appear to grow like the branches of a christmas tree



We get a space and time hyper-compound dataset
(a higher level dataset of higher degree)

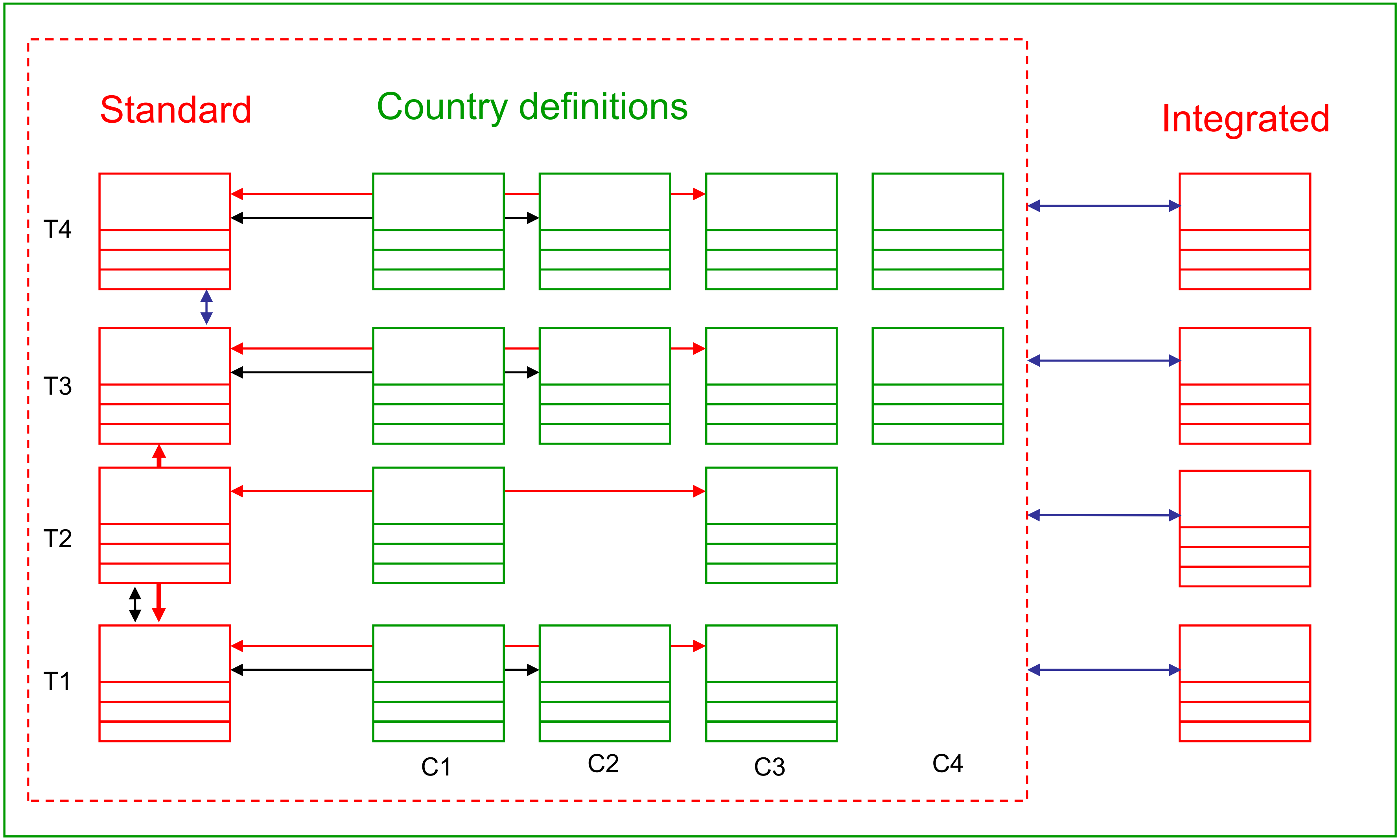


…which fairly well corresponds to the overall cross-national program

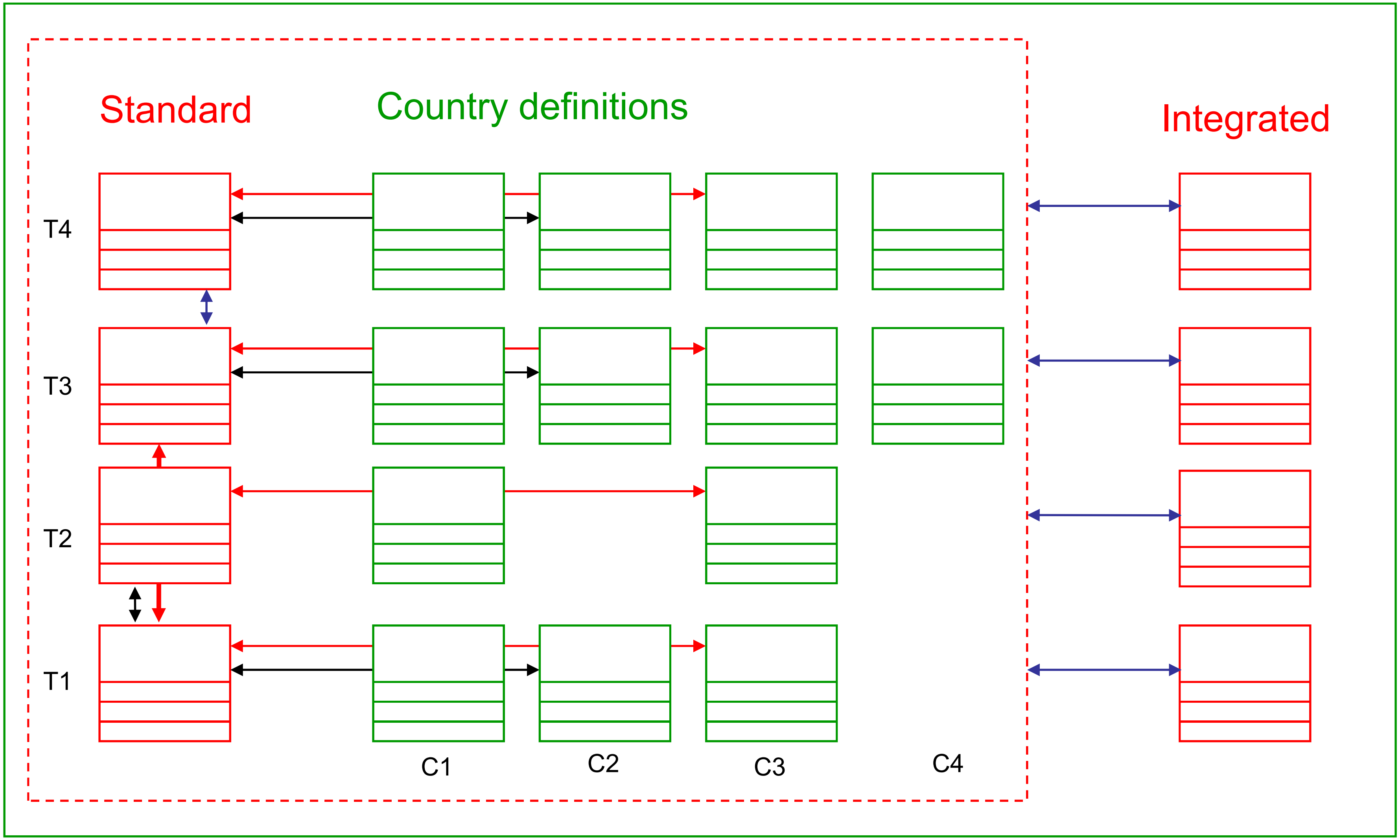


Metadata can
be published in
hyperlinked format

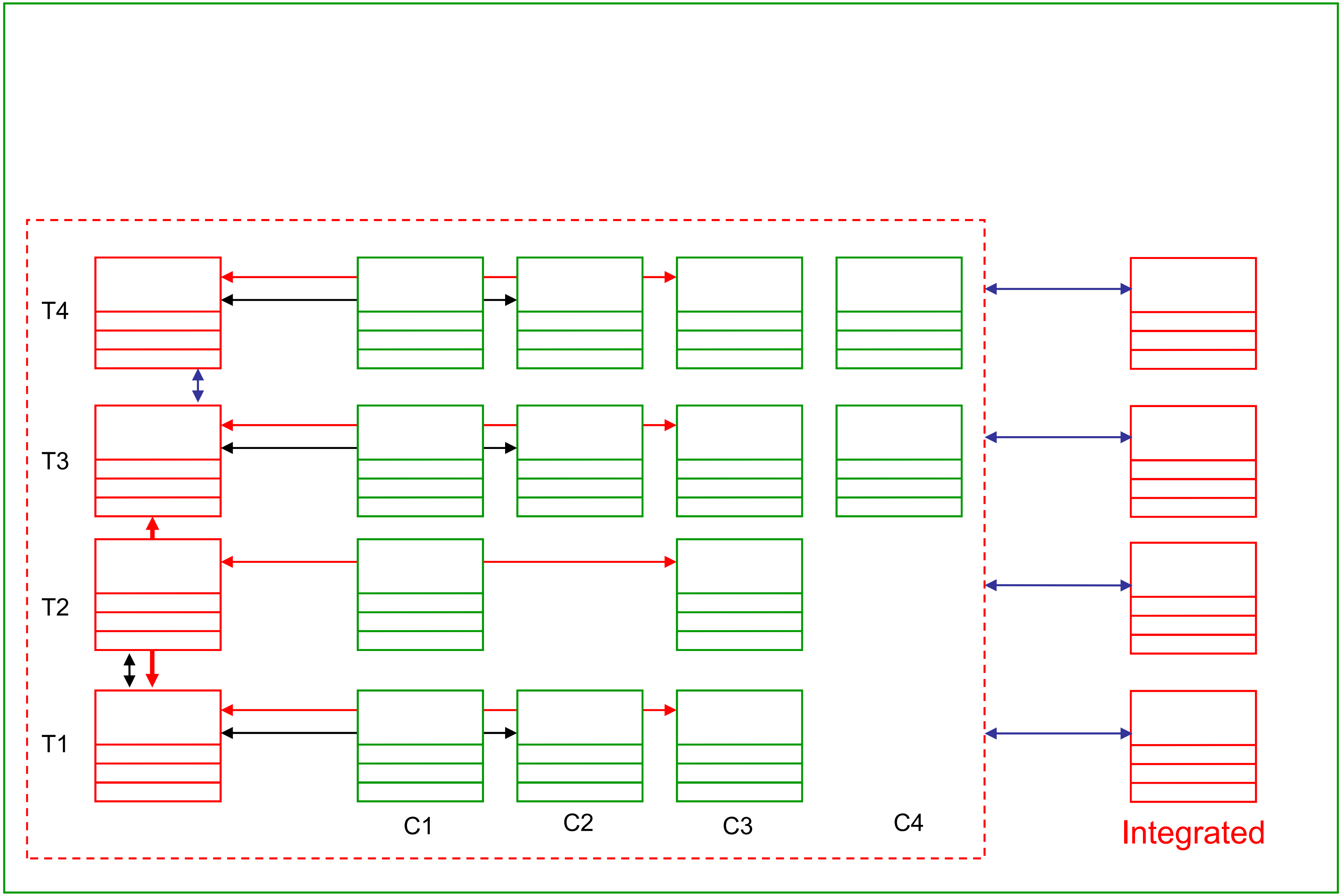
Integrate


Standard
Country definitions
Integrated
T4
T3
T2
T1
C1
C2
C3
C4

The study description will still work as an overall wrapper but it must include **time-dependent information** to account for changes in the program



Let’s simplify again…



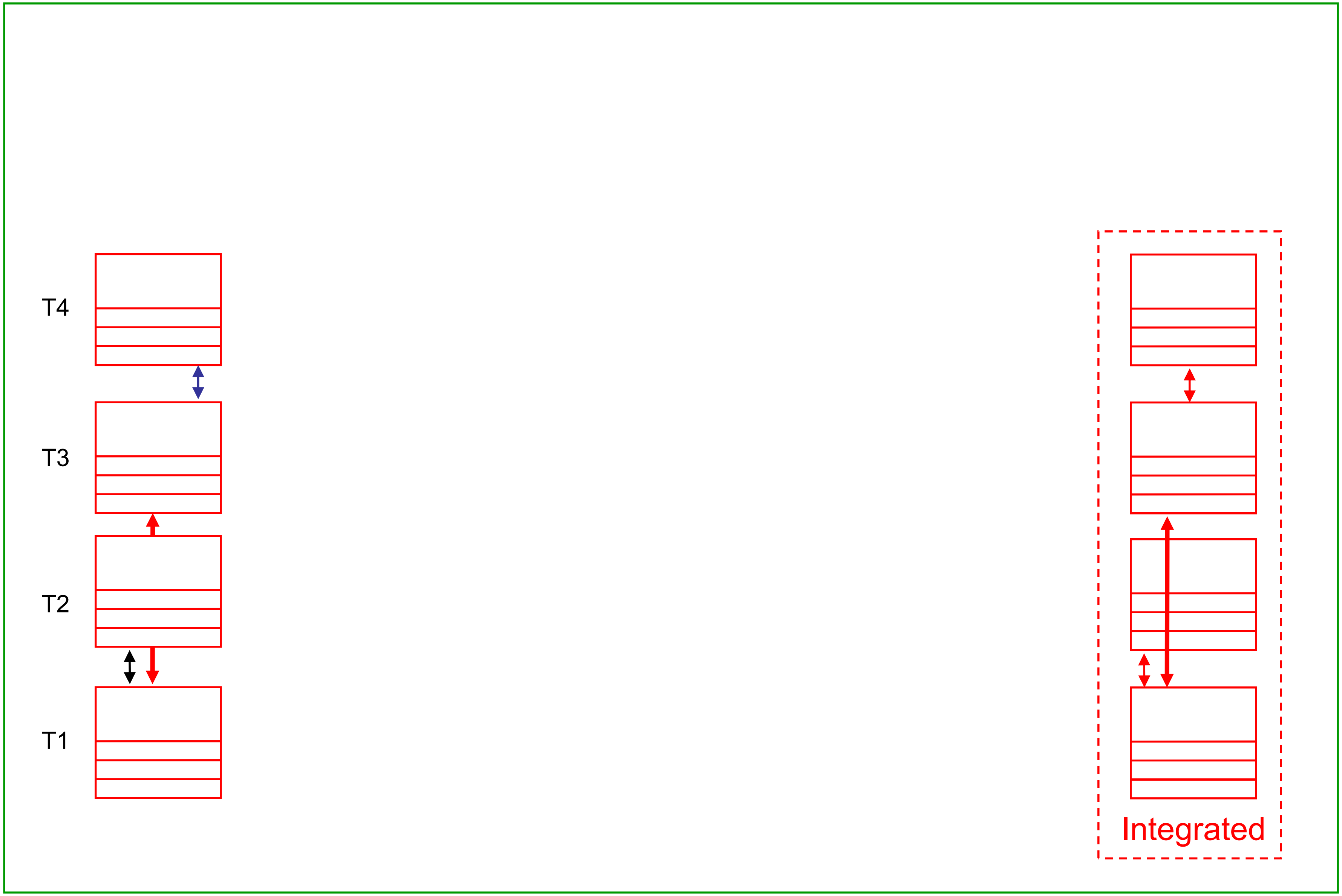
And keep only the integrated datasets:

Integrated

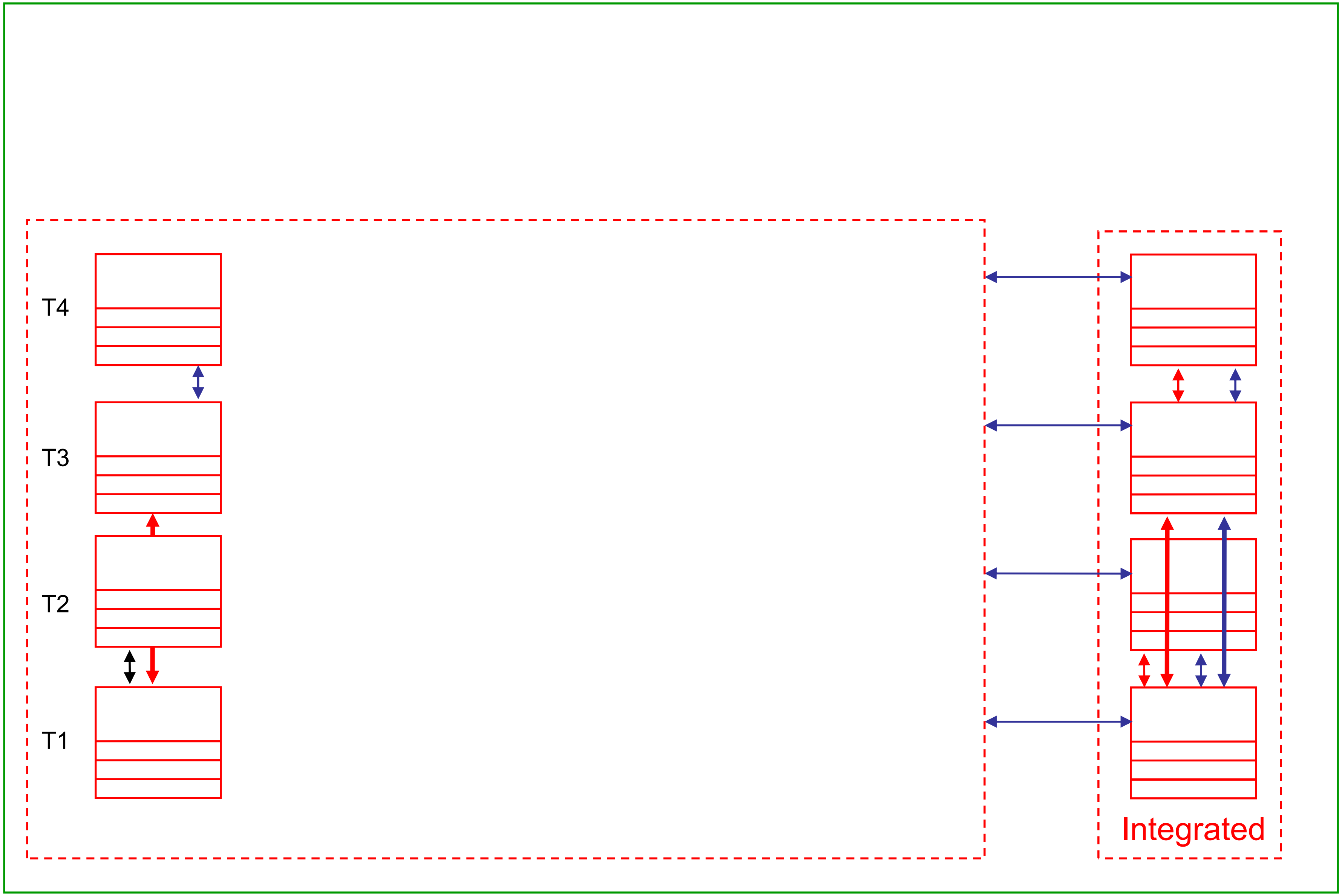
We get a nice time-compound dataset:

Integrated

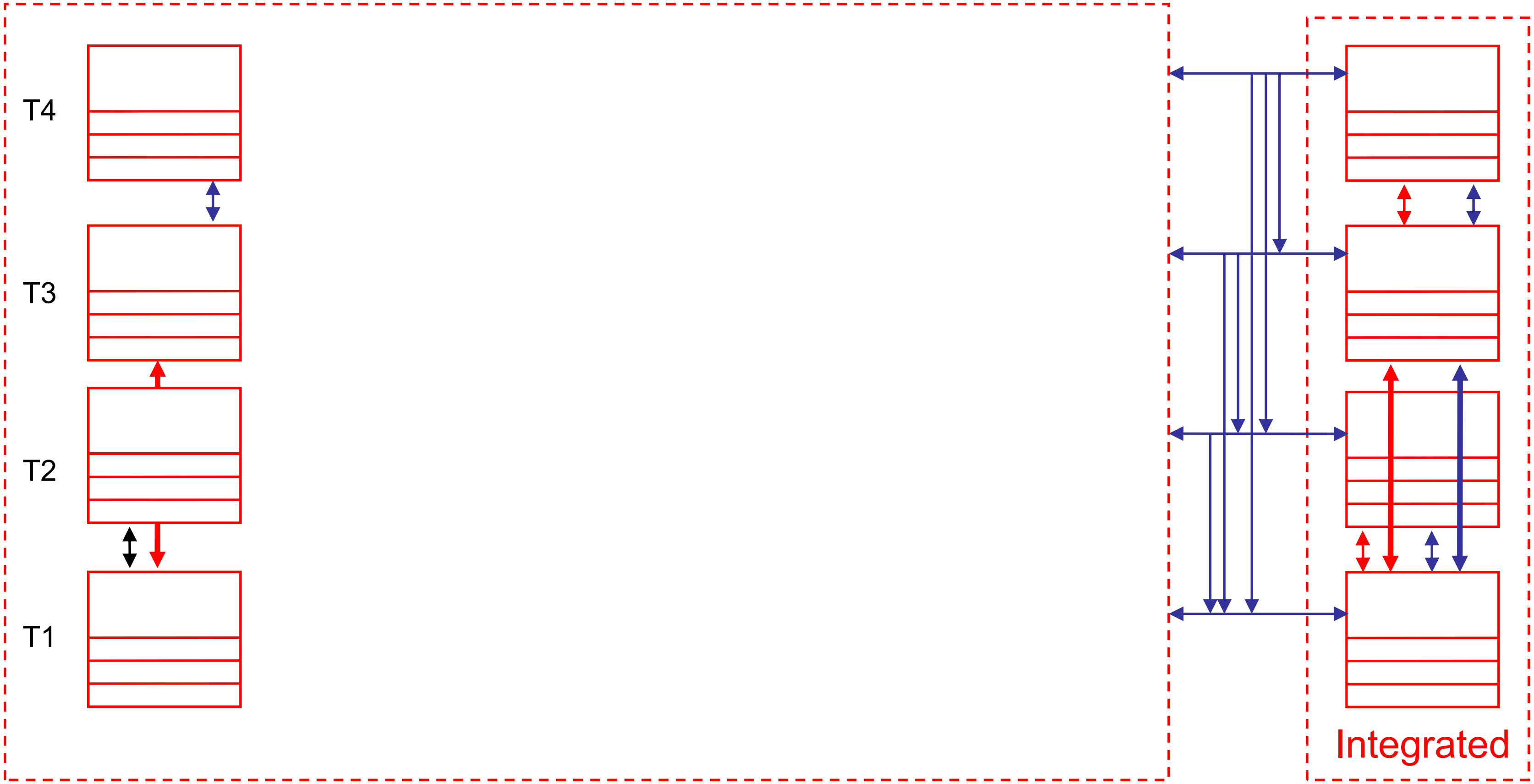
Some of the references over time are inherited from the series of standards:



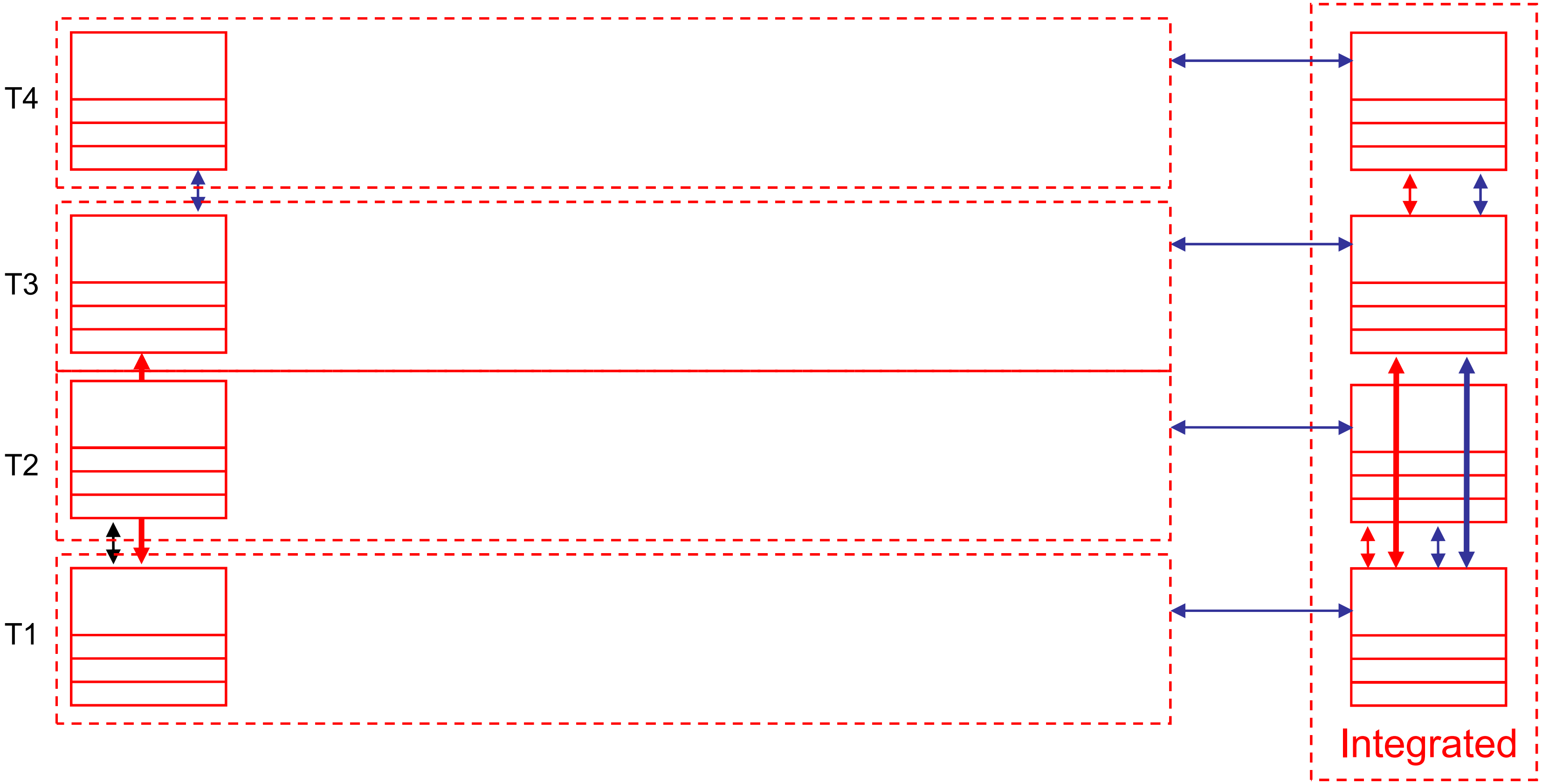
Additional references must be built for the harmonized variables



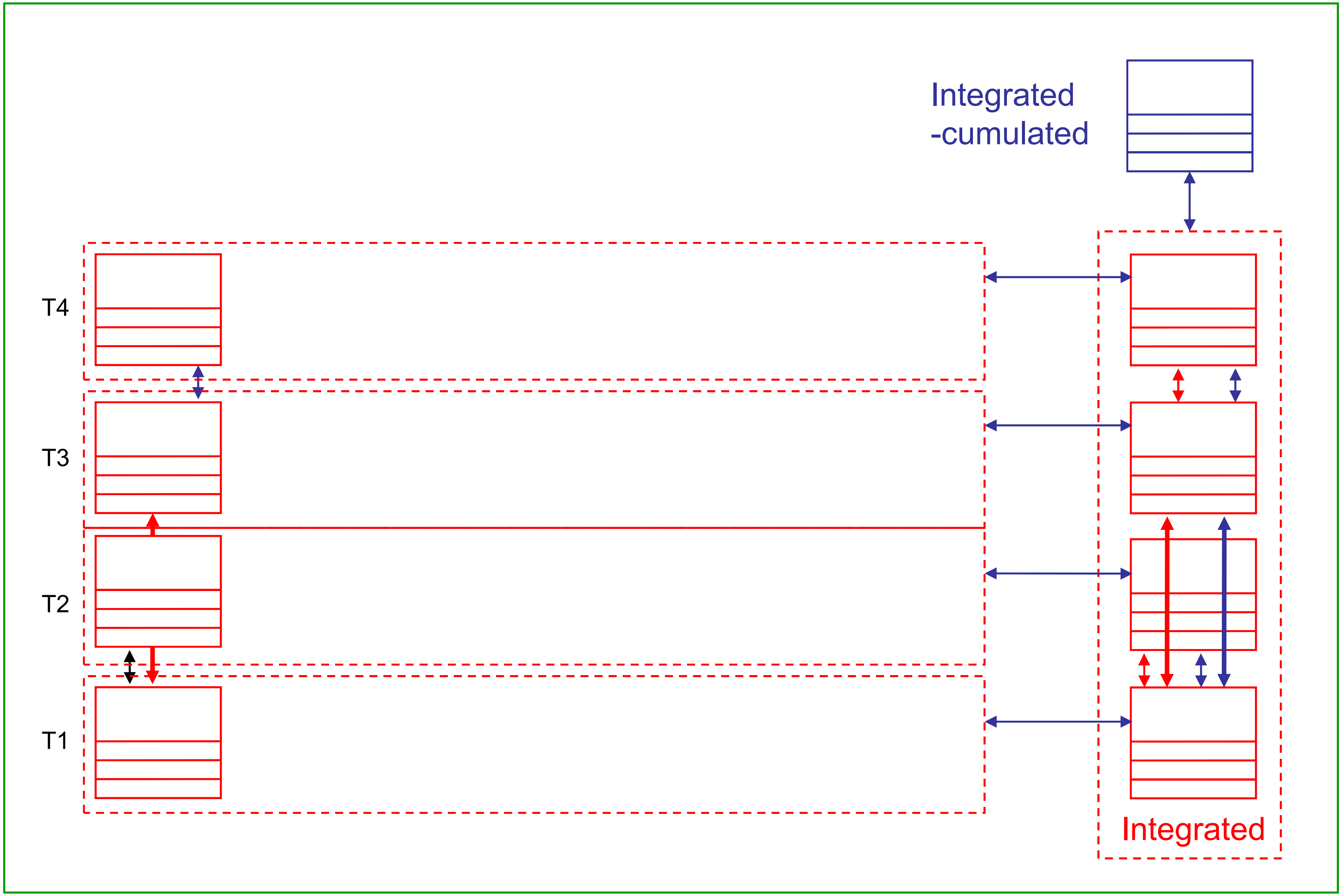
Most of the time, the computation of an integrated dataset for a new wave will take into account the choices made for the precedent ones



The variables in the integrated datasets will be referenced in both the single space-compound datasets and the time-compound of integrated datasets

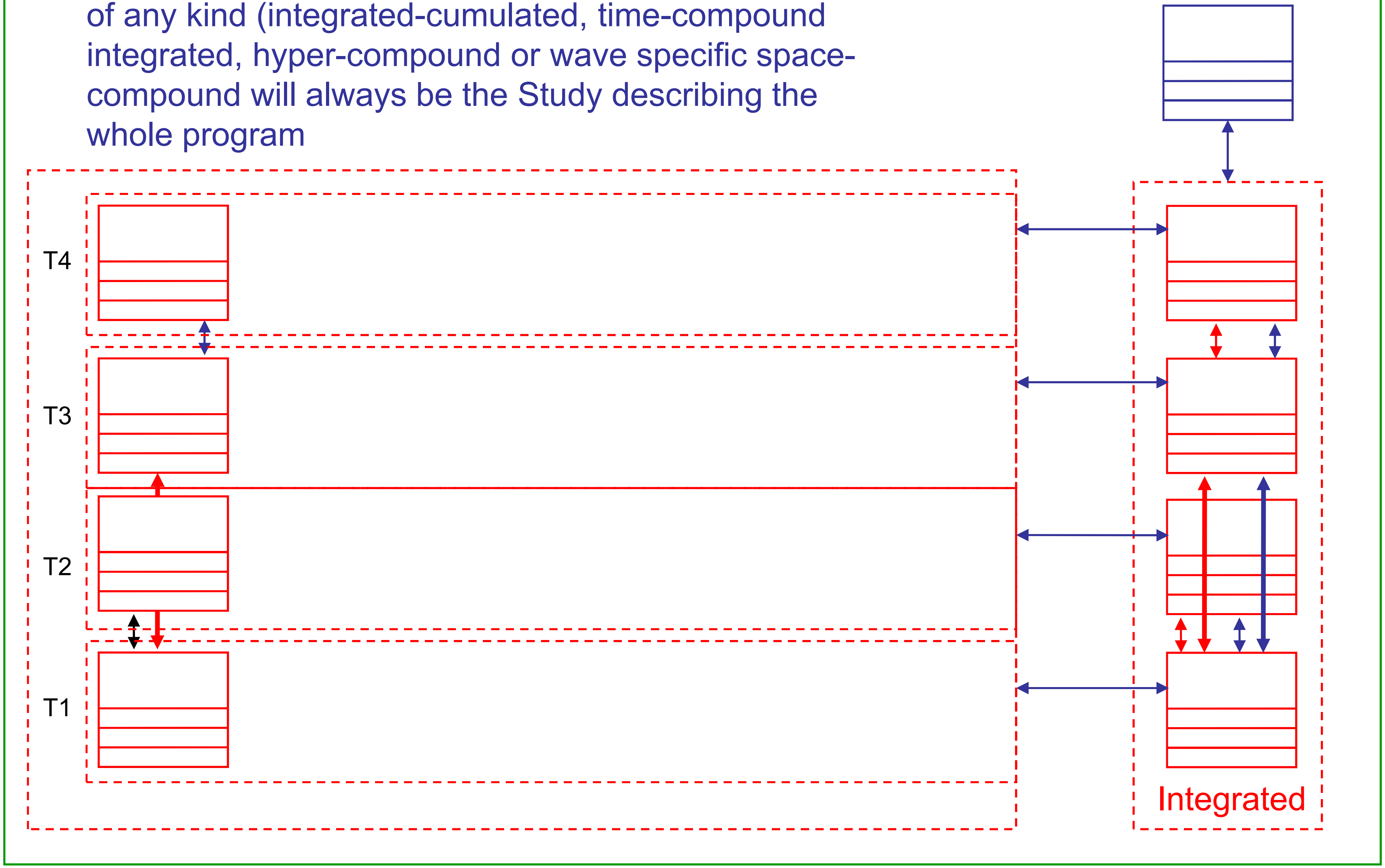


The cumulation of the single integrated datasets is now straightforward:

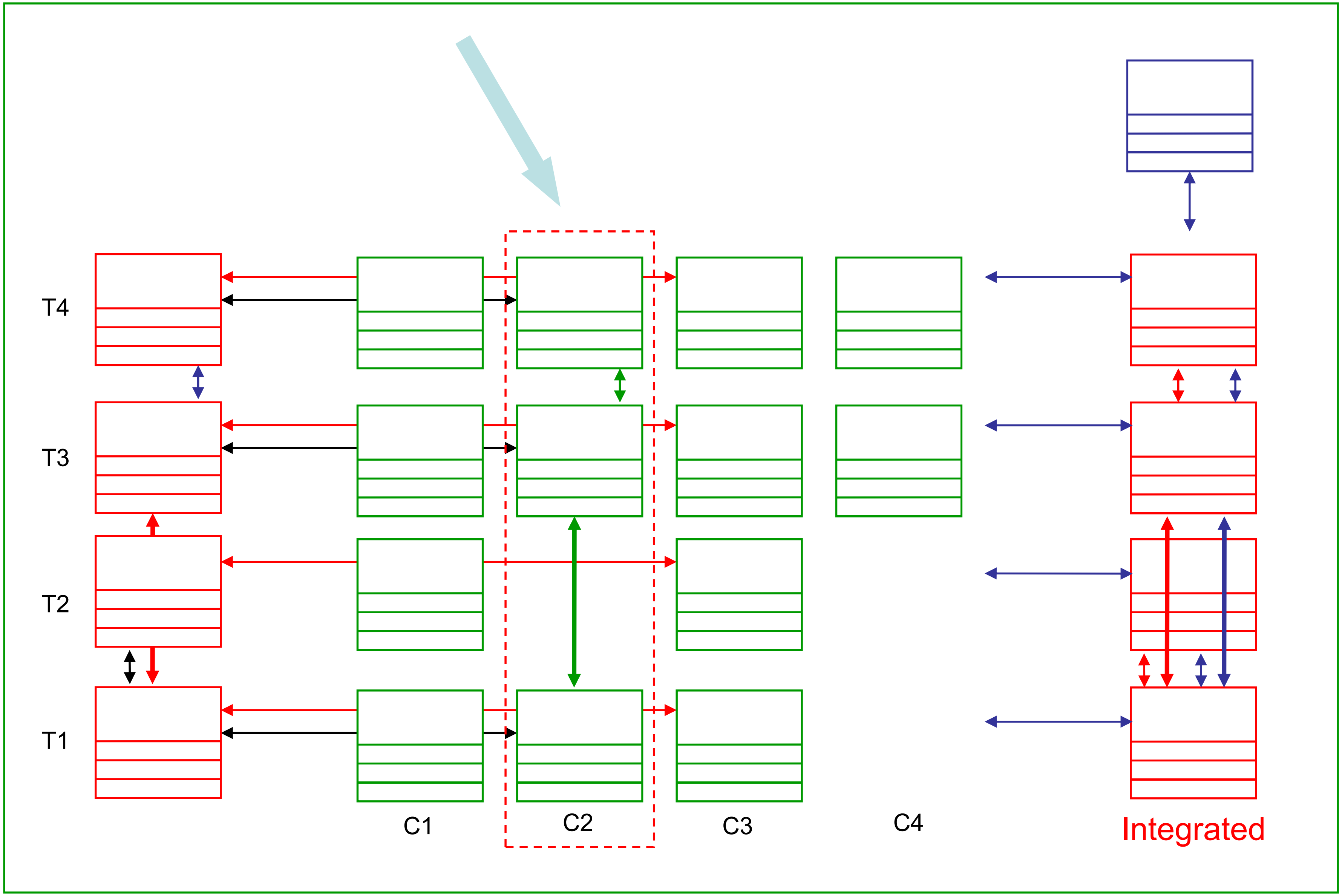


Metadata publication:

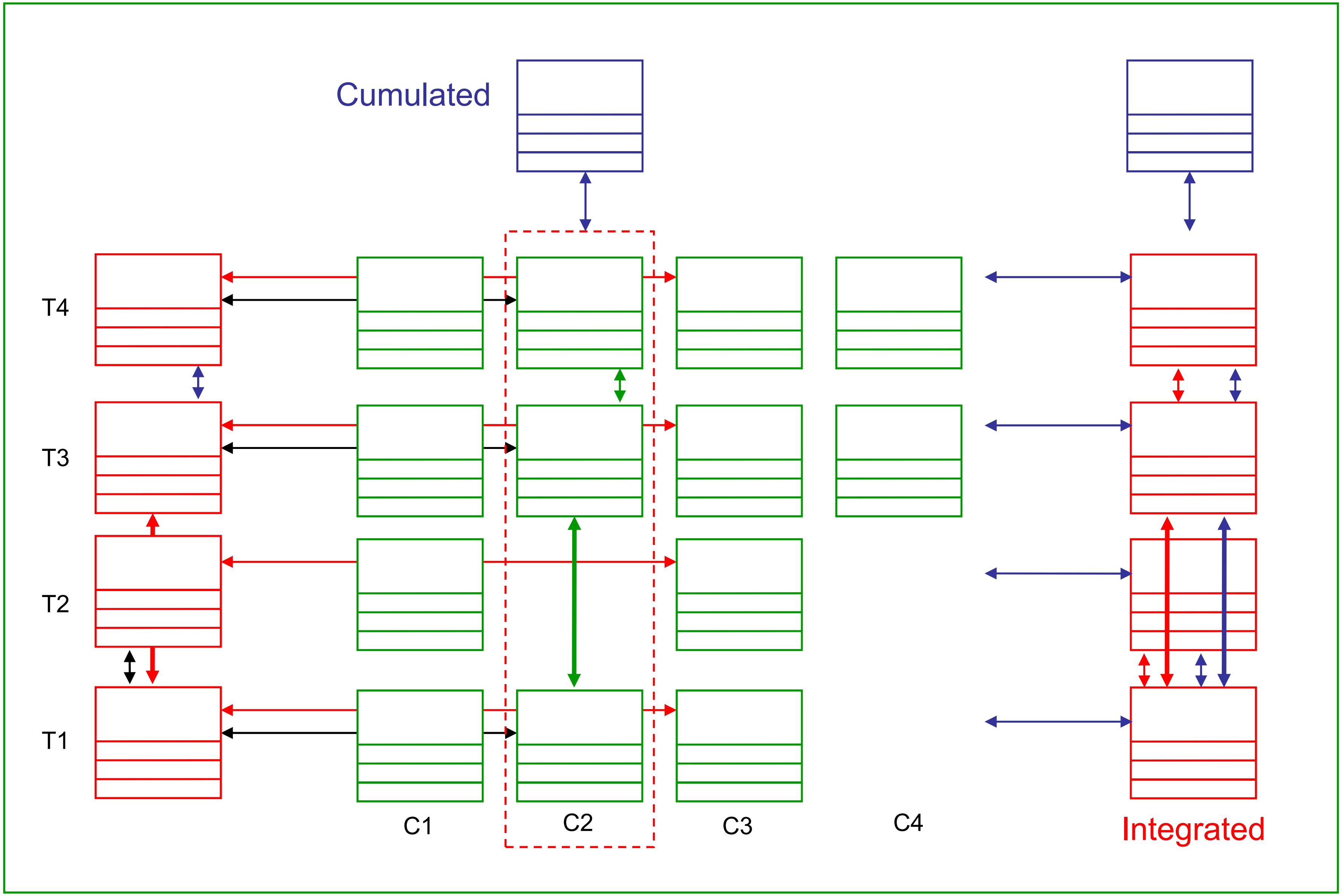
The ultimate wrapper for the publication of metadata of any kind (integrated-cumulated, time-compound integrated, hyper-compound or wave specific space-compound will always be the Study describing the whole program



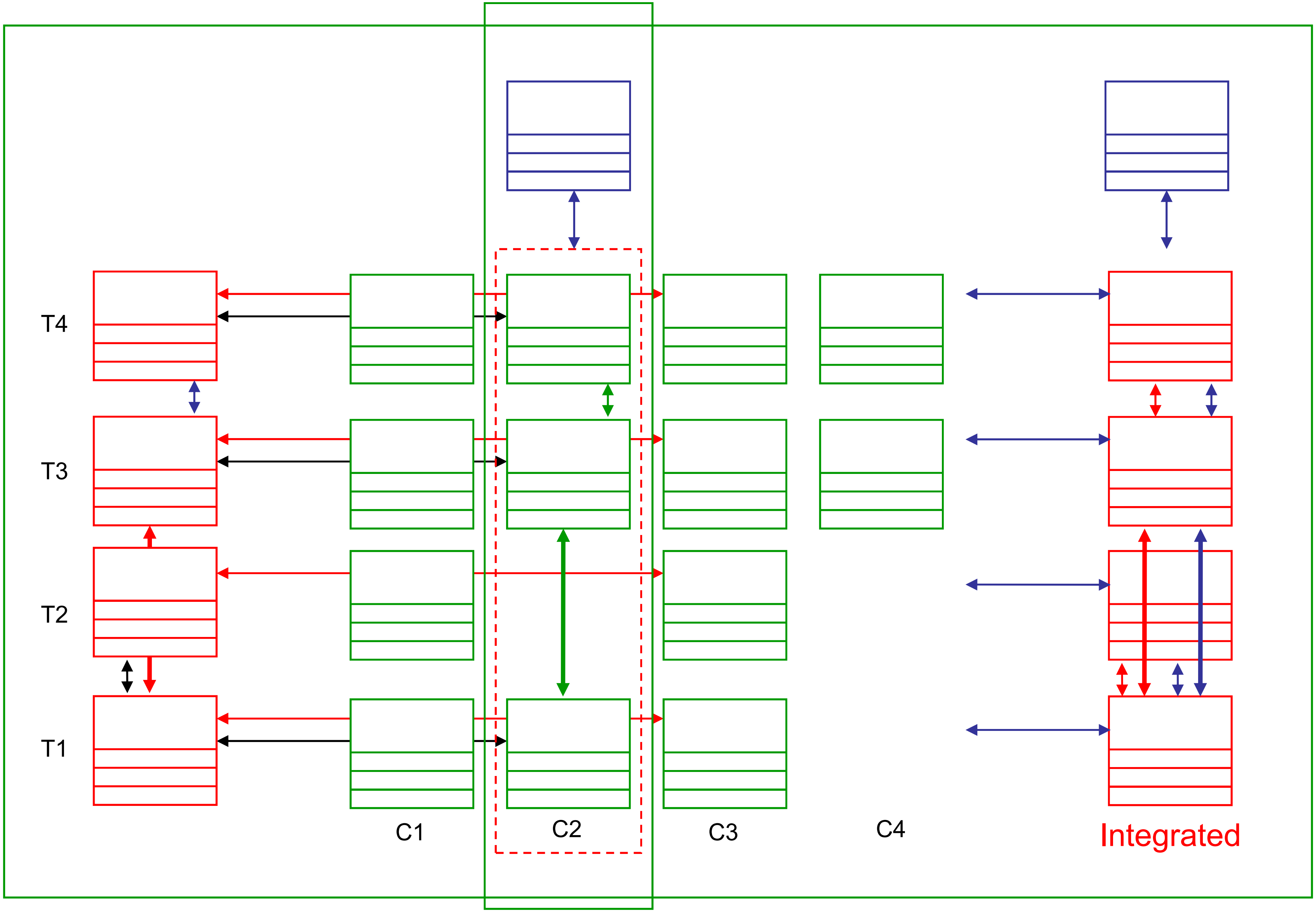
Some countries will care for their datasets over time…



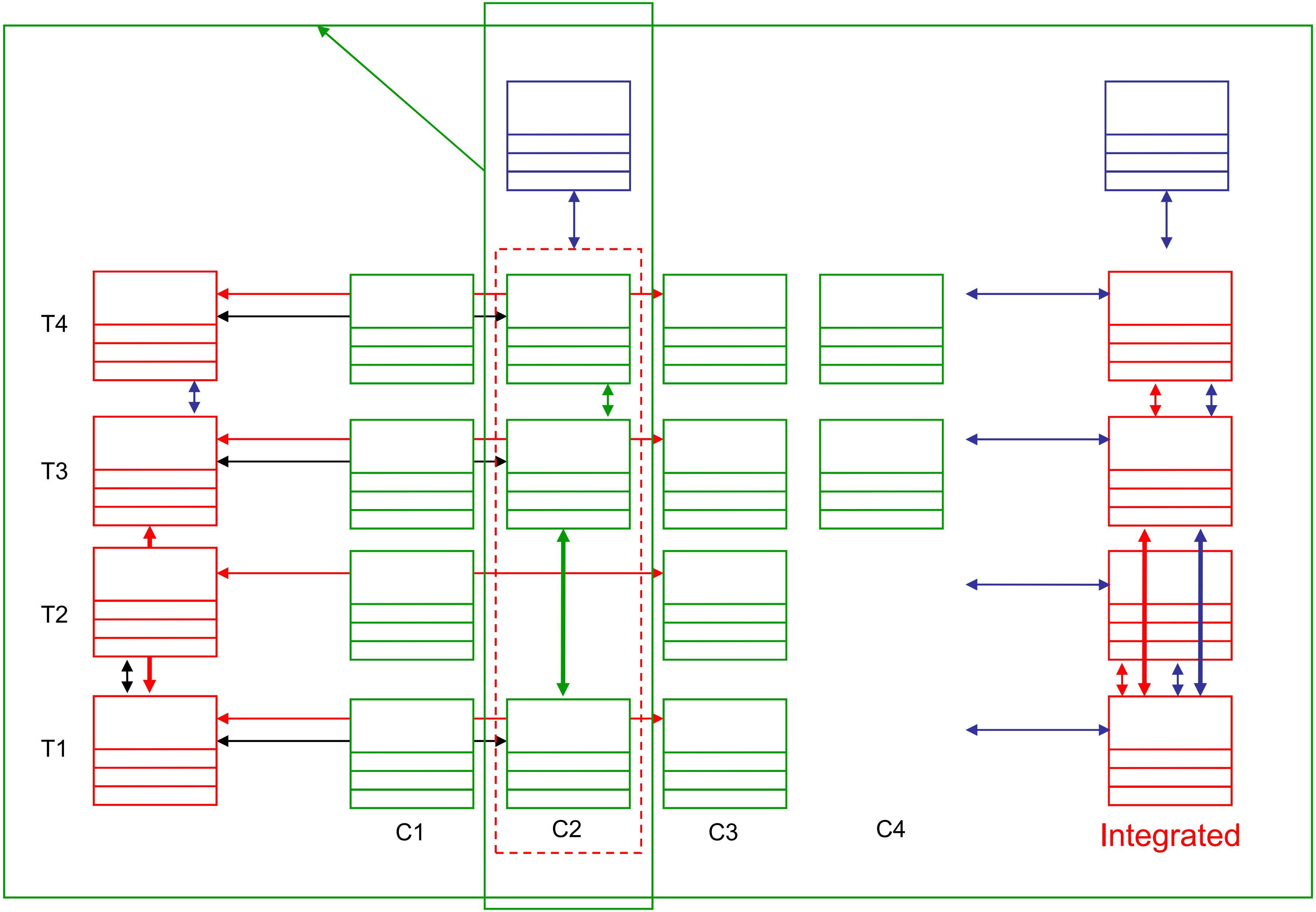
… cumulate them…



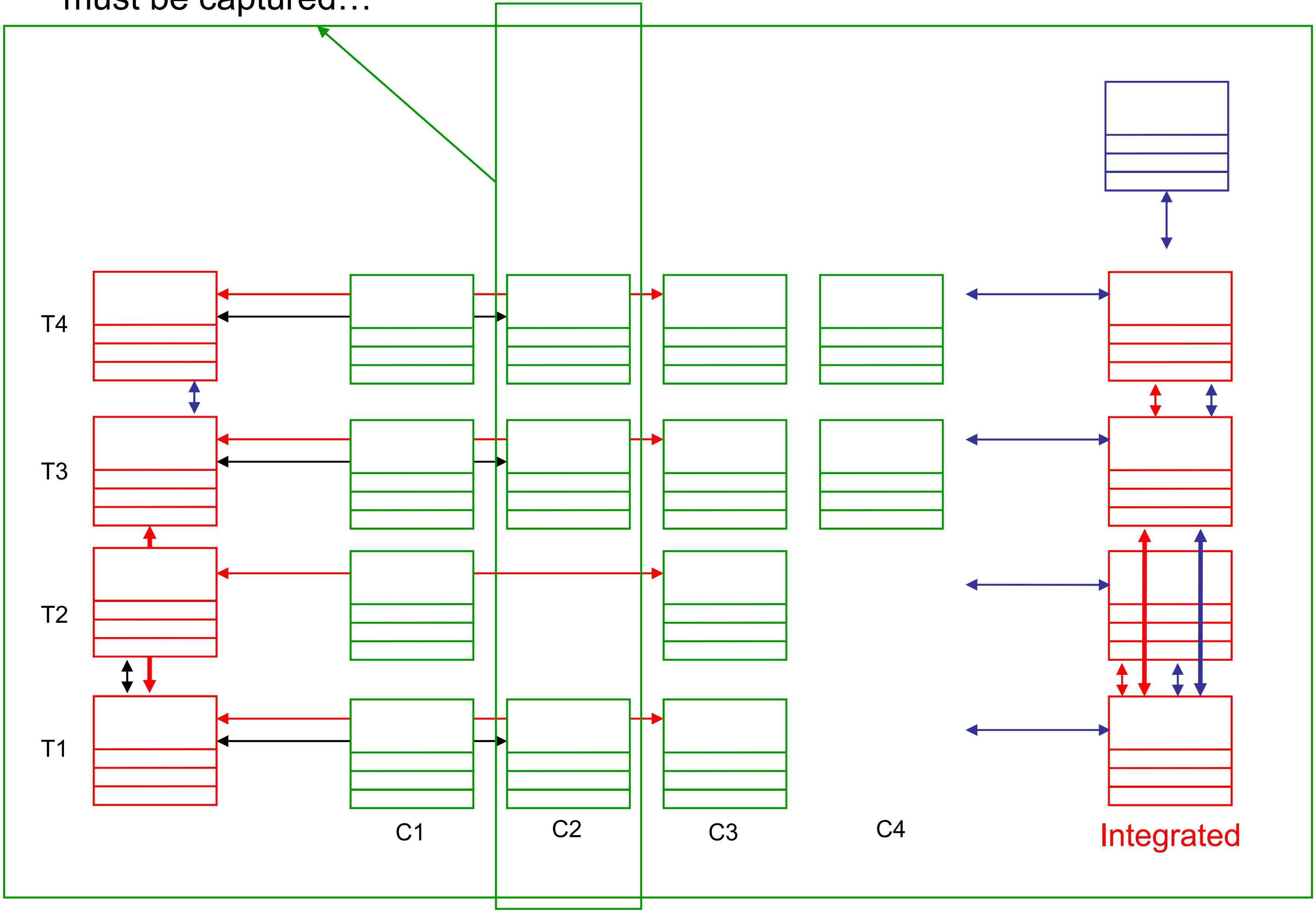
… and care for their own study description:



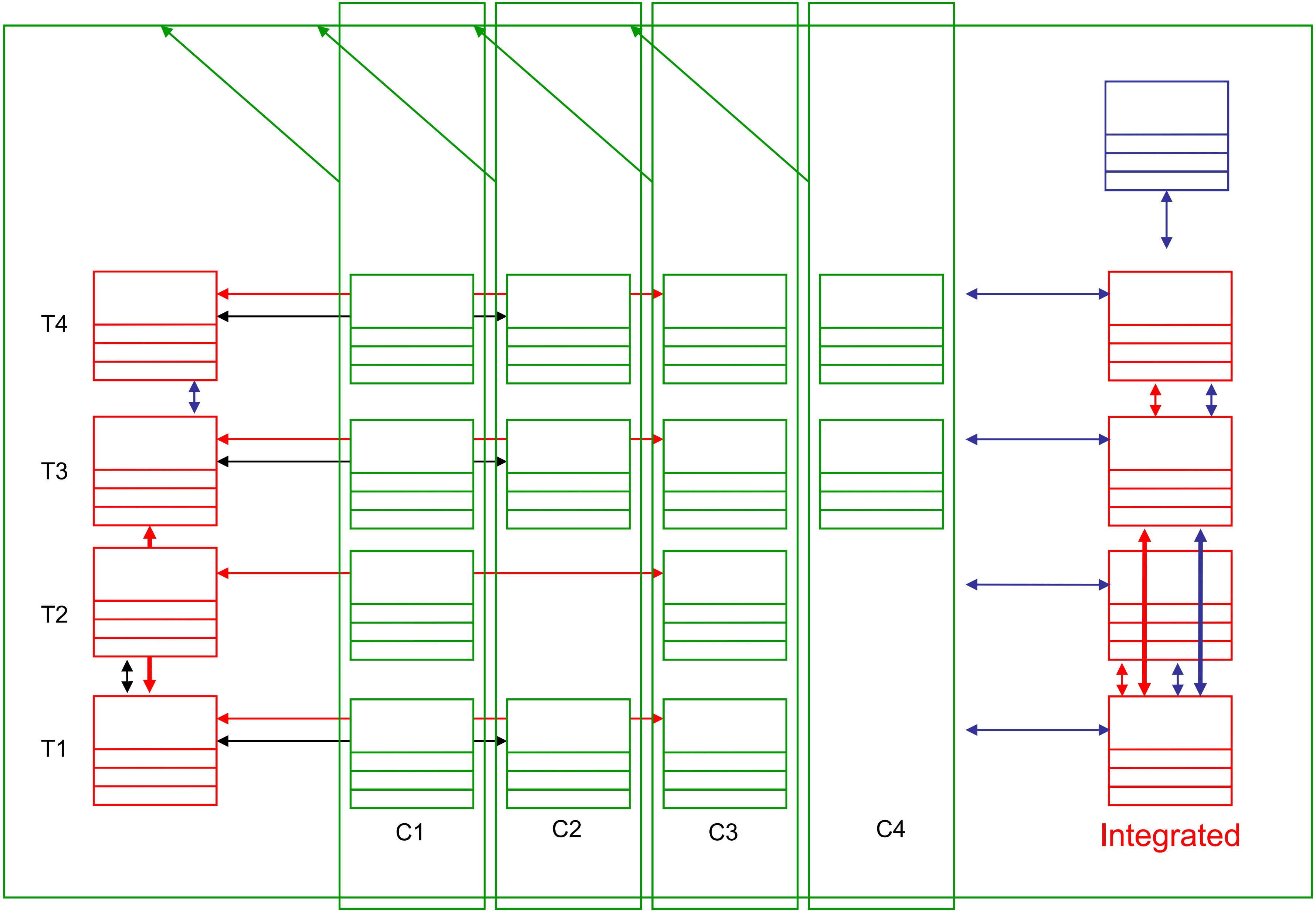
For sure, that study description will refer to the program



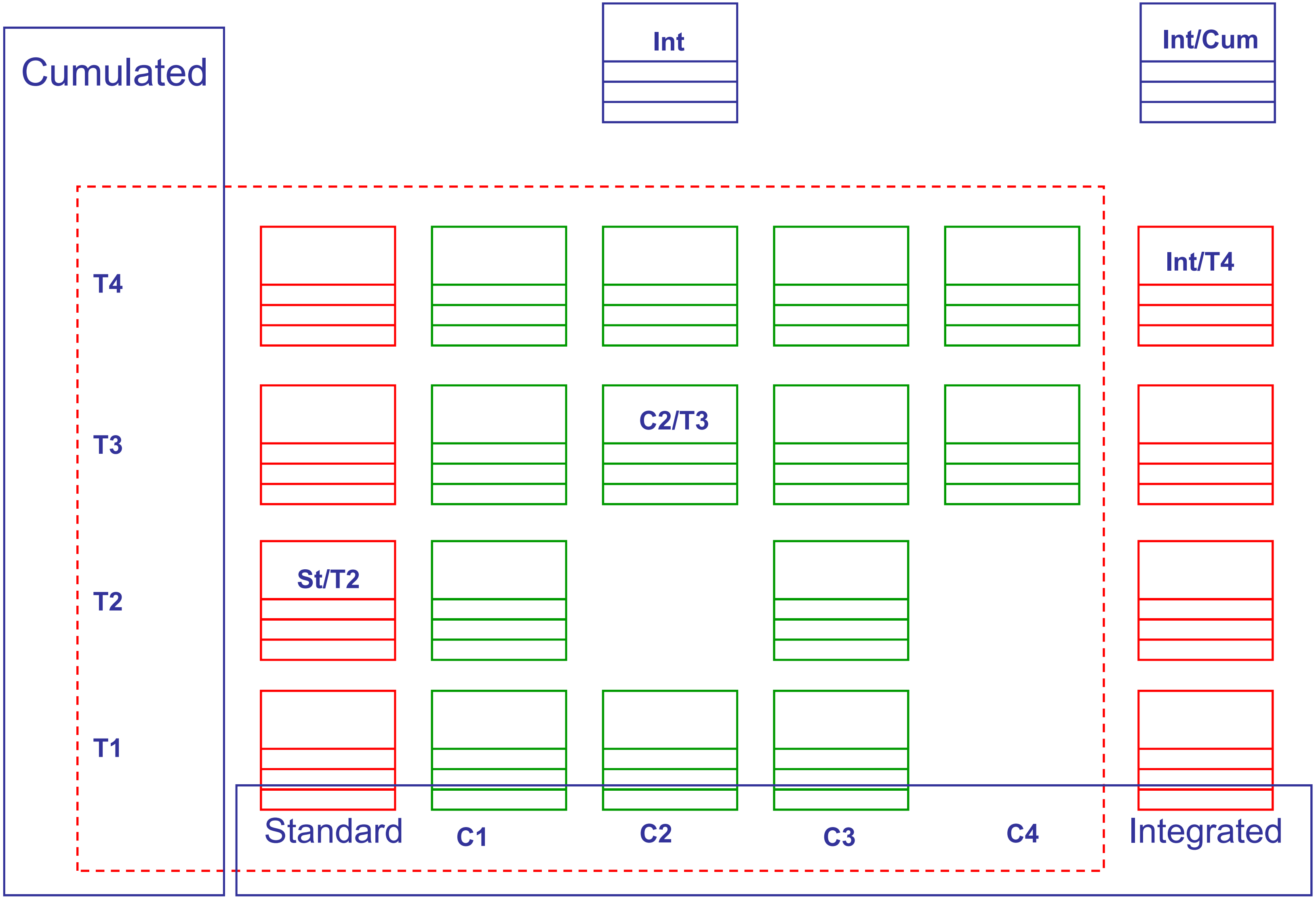
Even where no cumulation takes place, country specific study level information must be captured…



… for all countries, to be included in all relevant metadata products



Countries and waves serve as coordinates to navigate the dataset space and select the one to work on



Int

Int/Cum

Cumulated

T4

Int/T4

T3

C2/T3

T2

St/T2

T1

Standard

C1

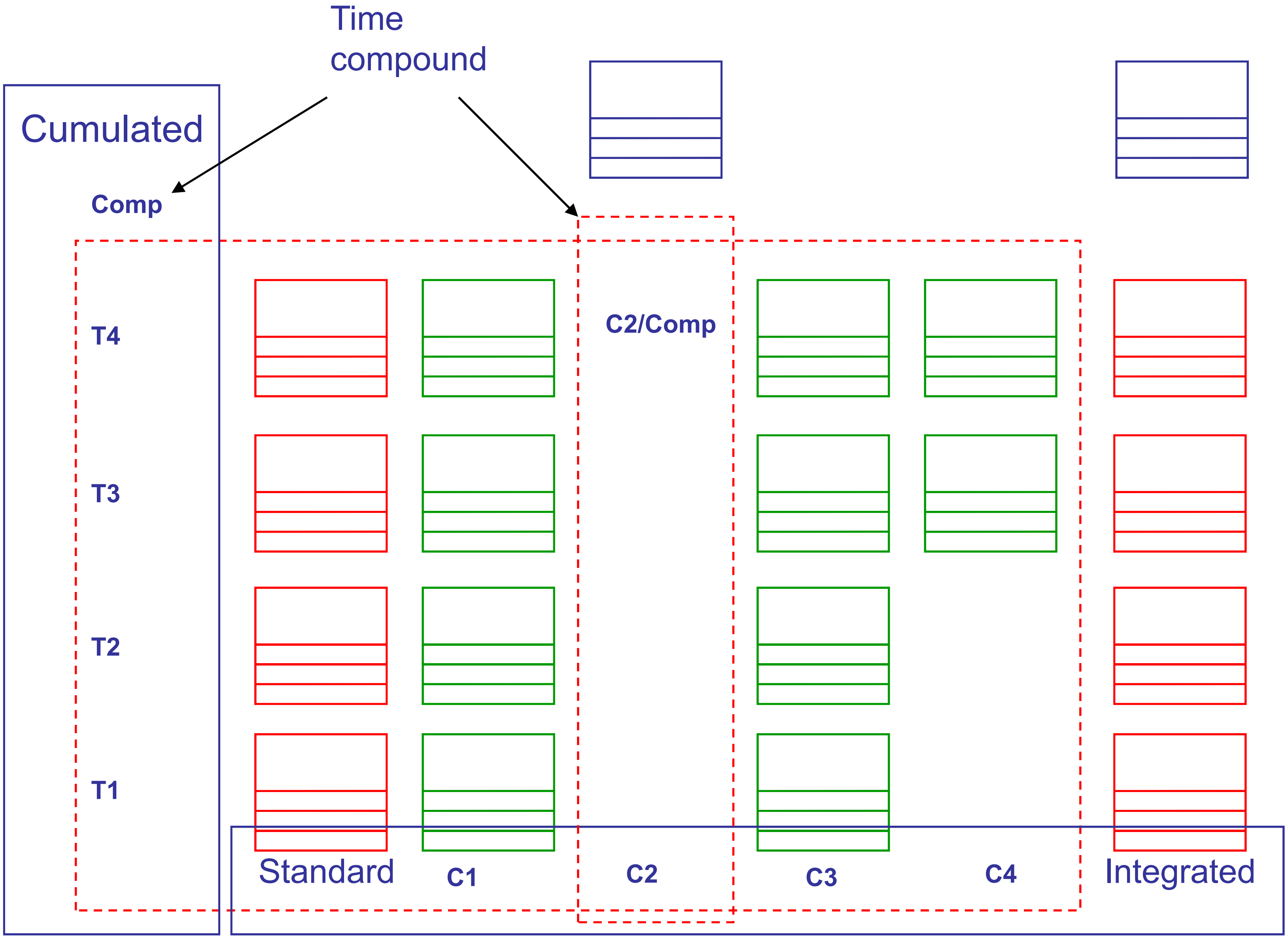
C2

C3

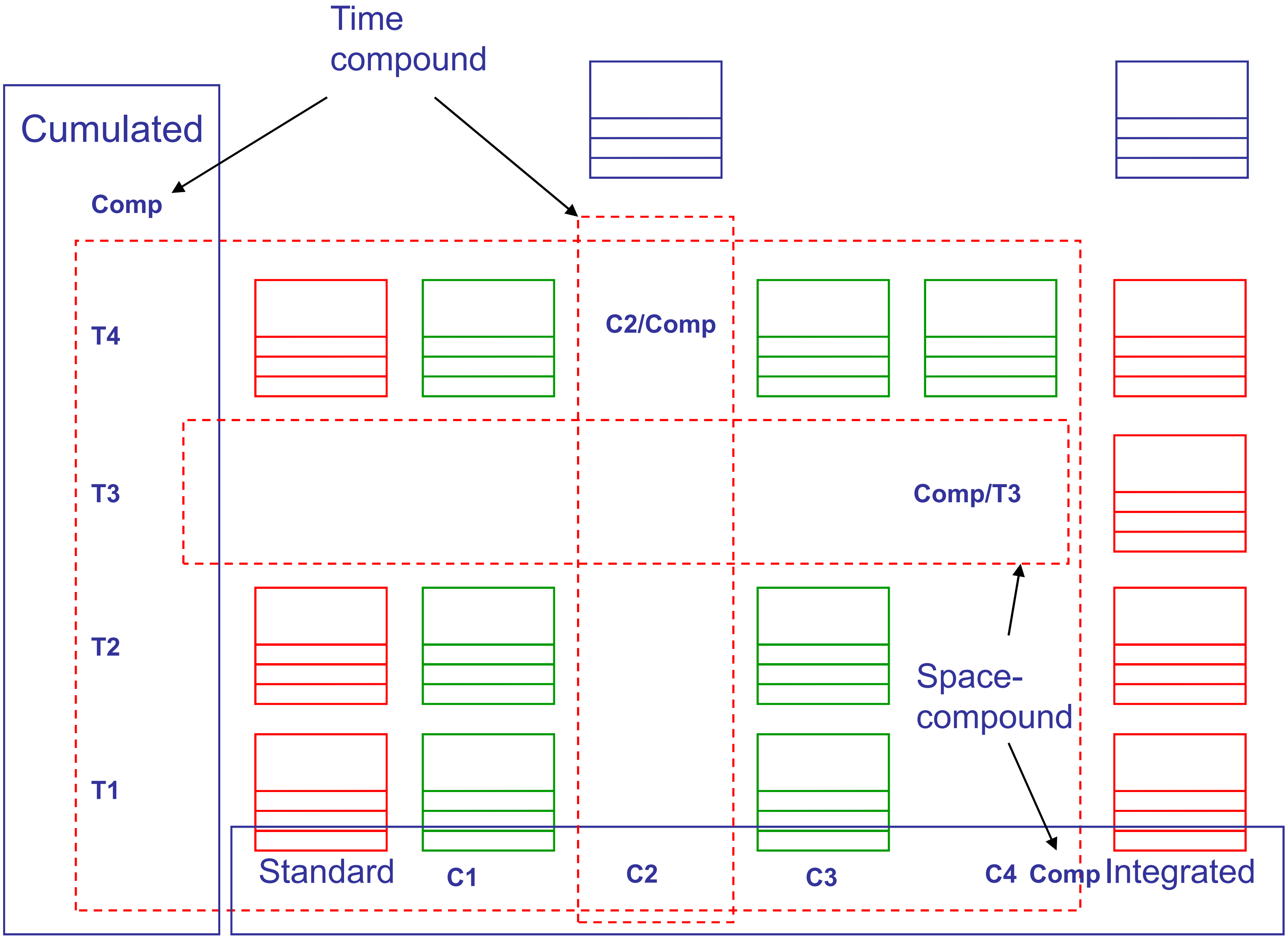
C4

Integrated

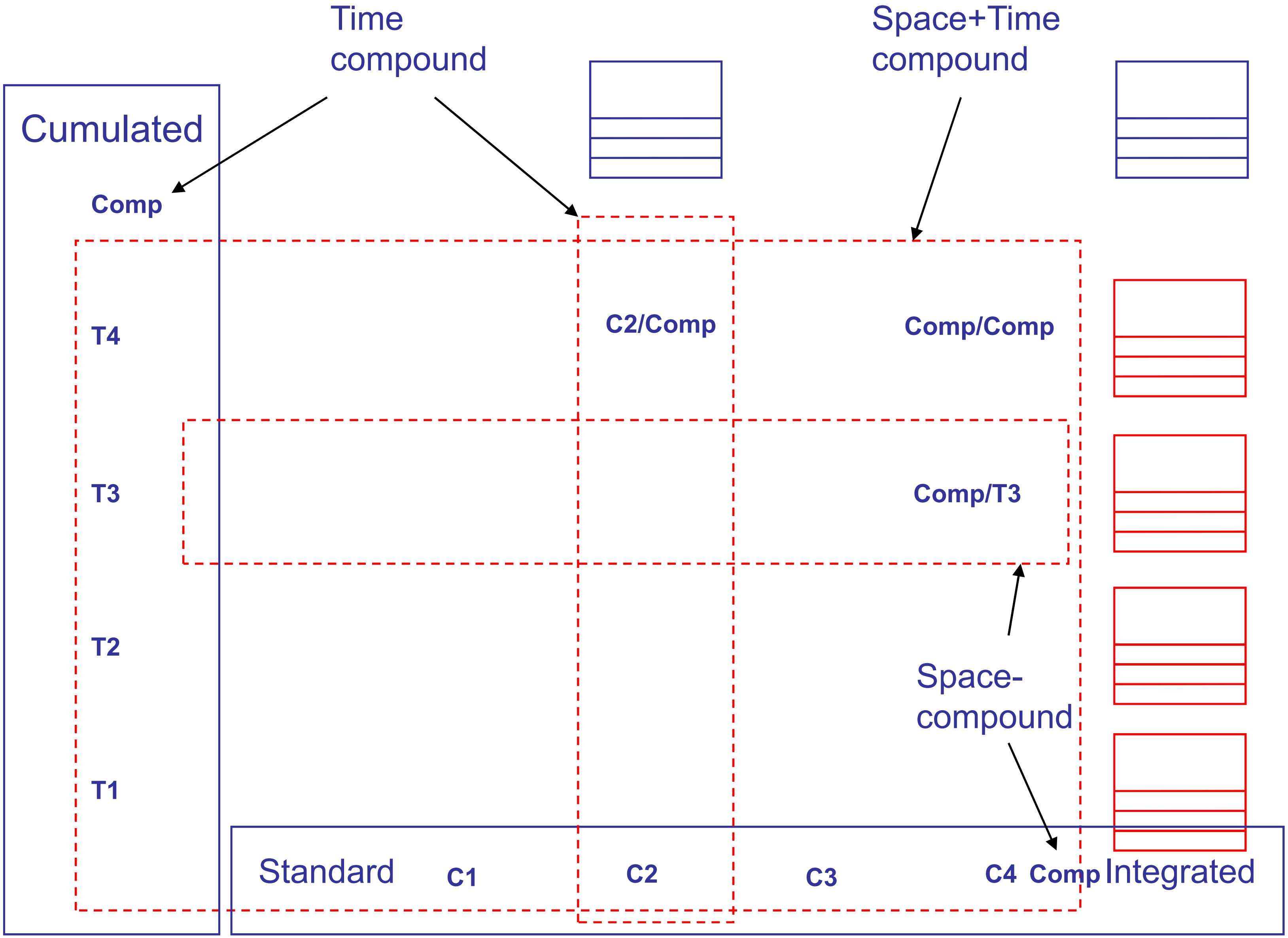
Values have to be added on the two scales to reach the compound datasets



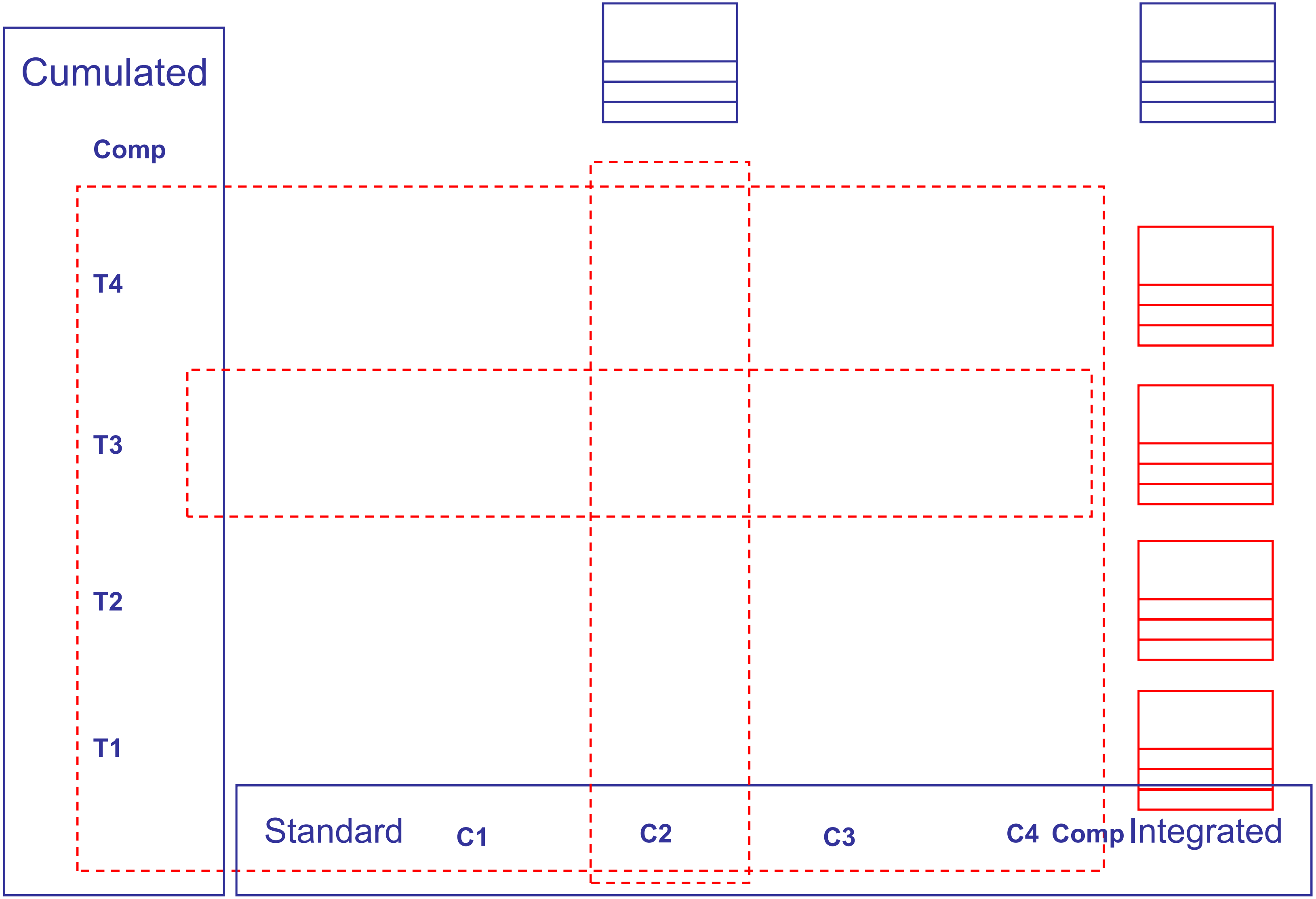
Values have to be added on the two dimensions to reach the compound datasets



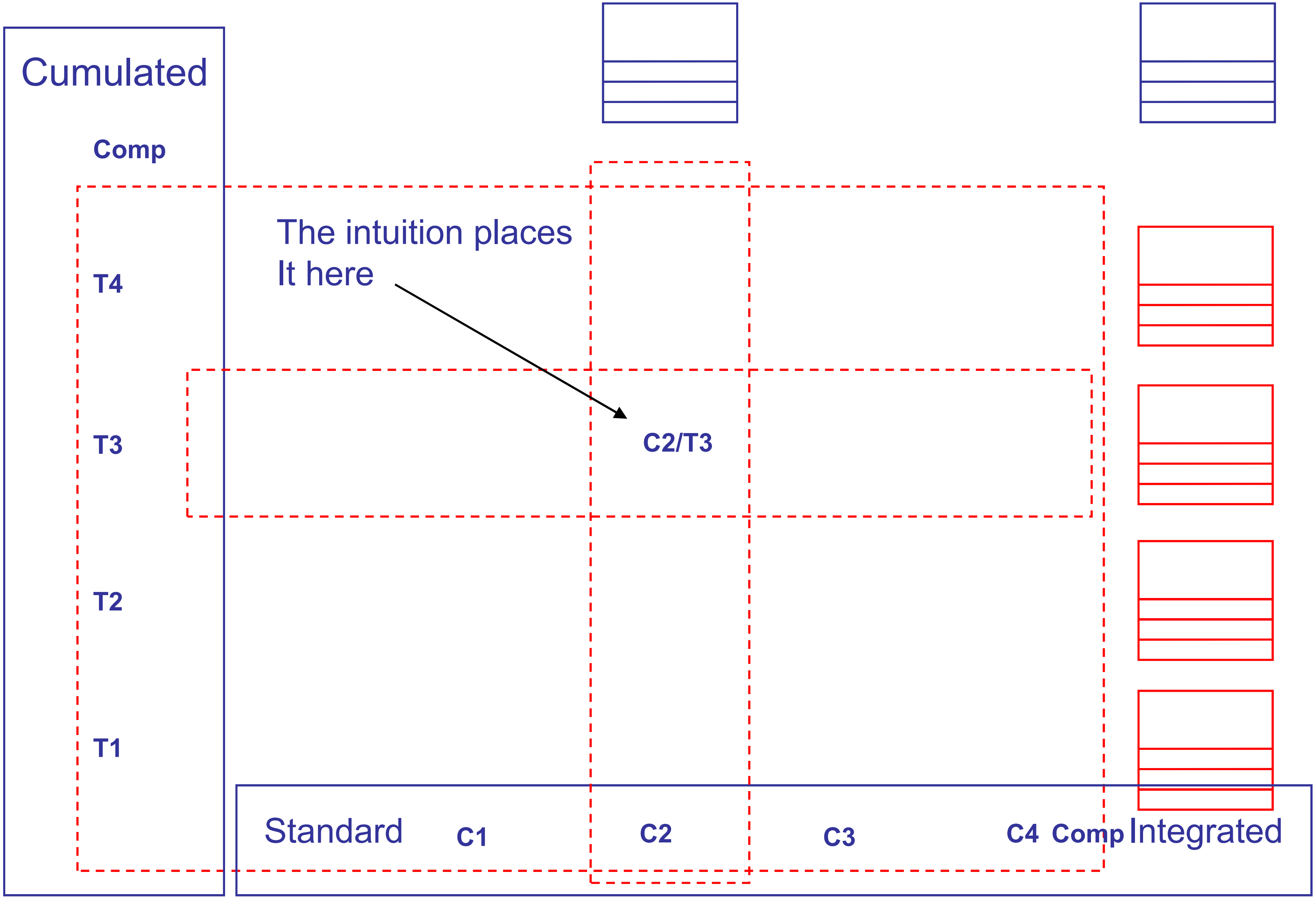
Summarizing the compound datasets:



… And the one time cross-section?!!!



… And the one time cross-section?!!!



Cumulated

Comp

T4

T3

T2

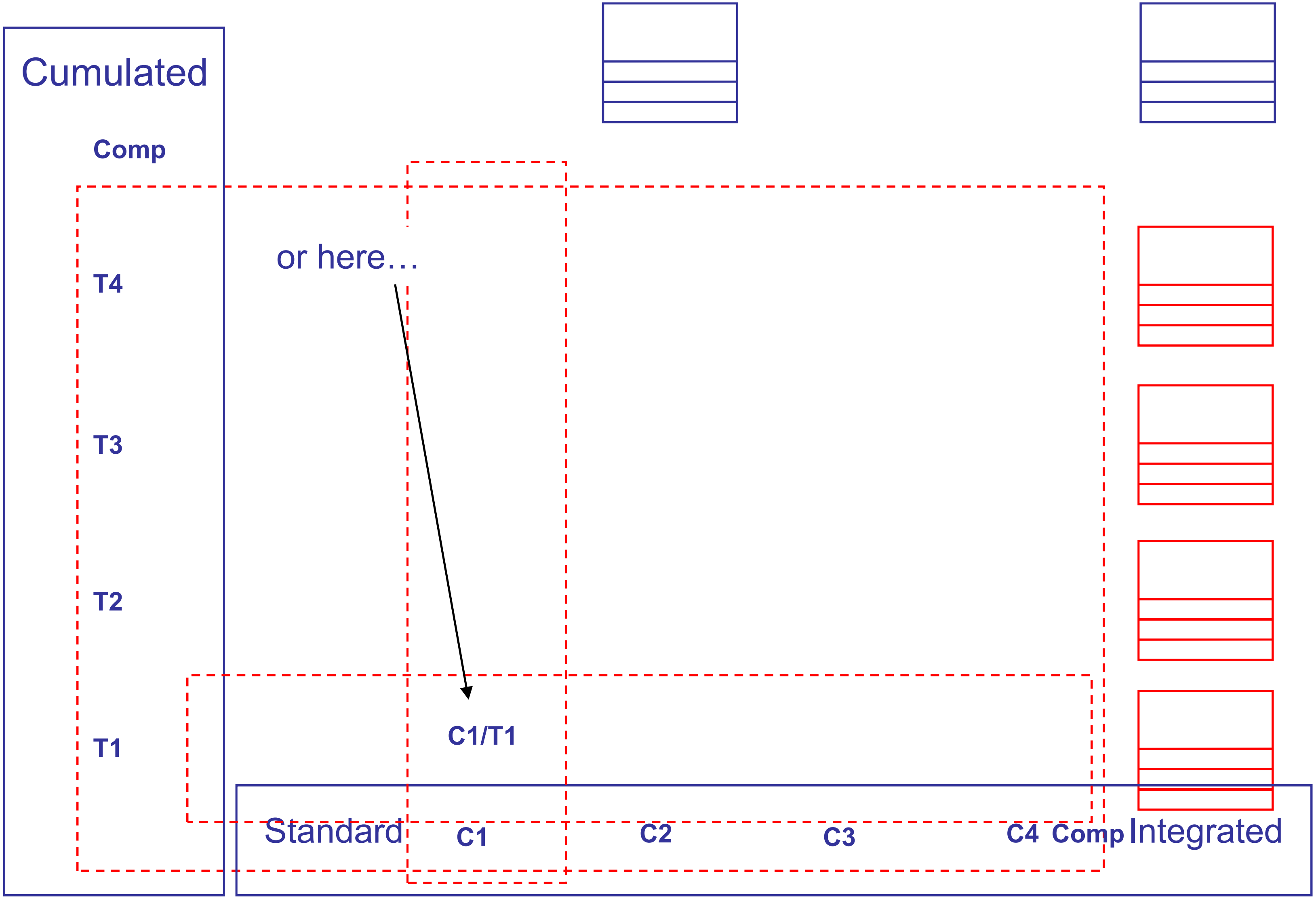
T1

The intuition places
It here

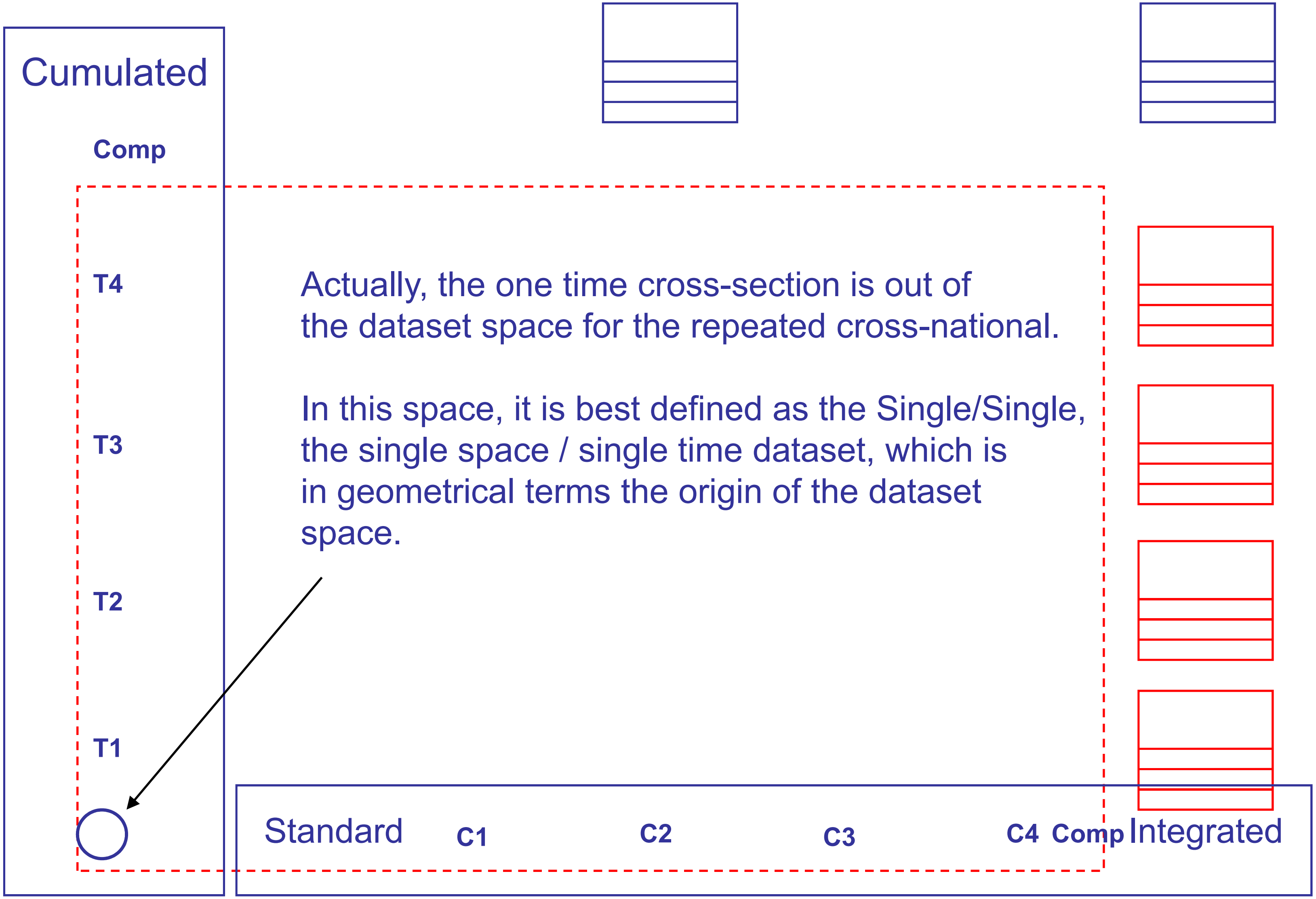
C2/T3

Standard C1 C2 C3 C4 Comp Integrated

… And the one time cross-section?!!!



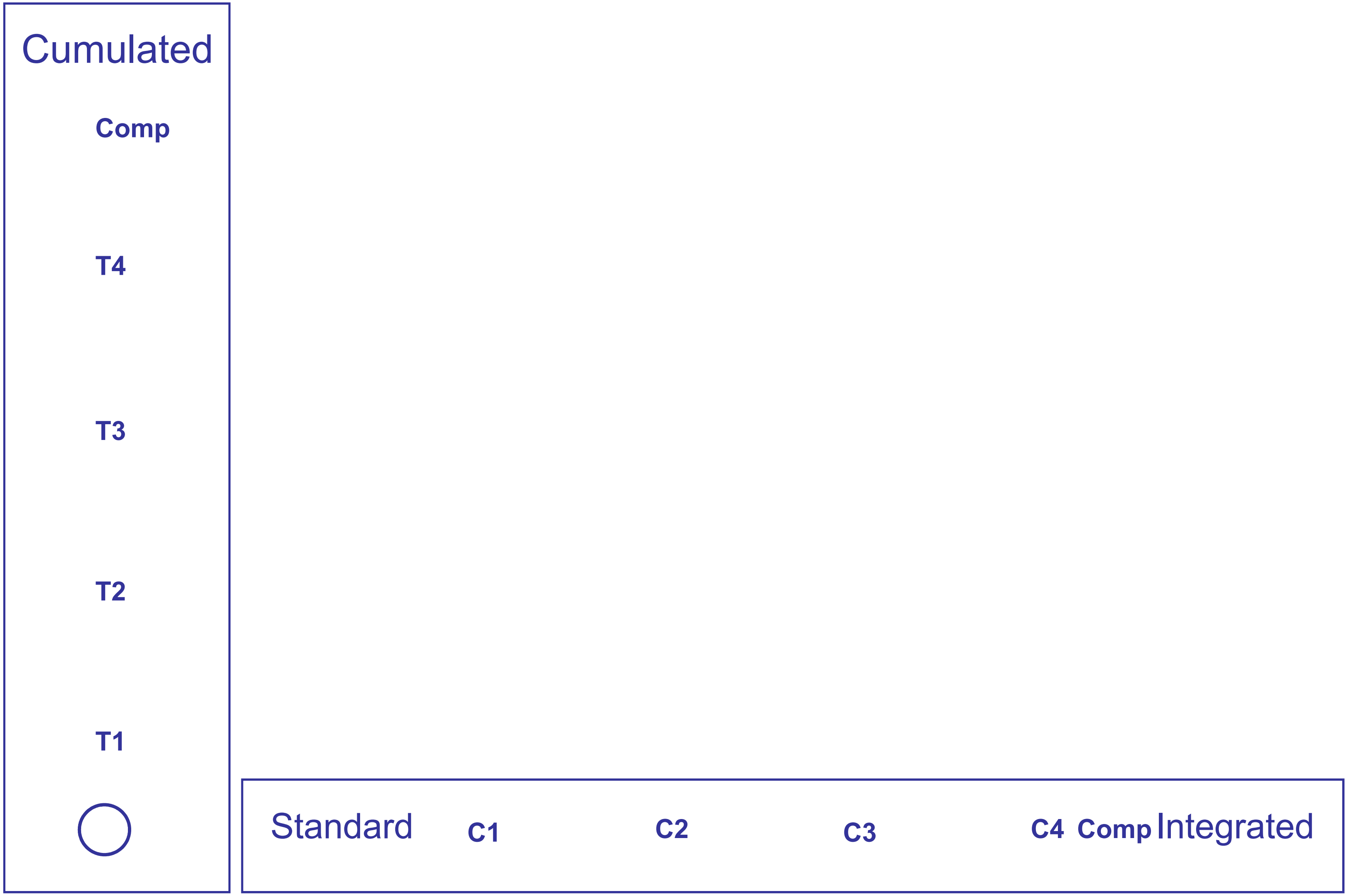
… And the one time cross-section?!



Actually, the one time cross-section is out of the dataset space for the repeated cross-national.

In this space, it is best defined as the Single/Single, the single space / single time dataset, which is in geometrical terms the origin of the dataset space.

Now, we can define in a systematic manner the types of datasets, which are defined in the dataset space to be used for a repeated cross-national program



Cumulated

Comp

T4

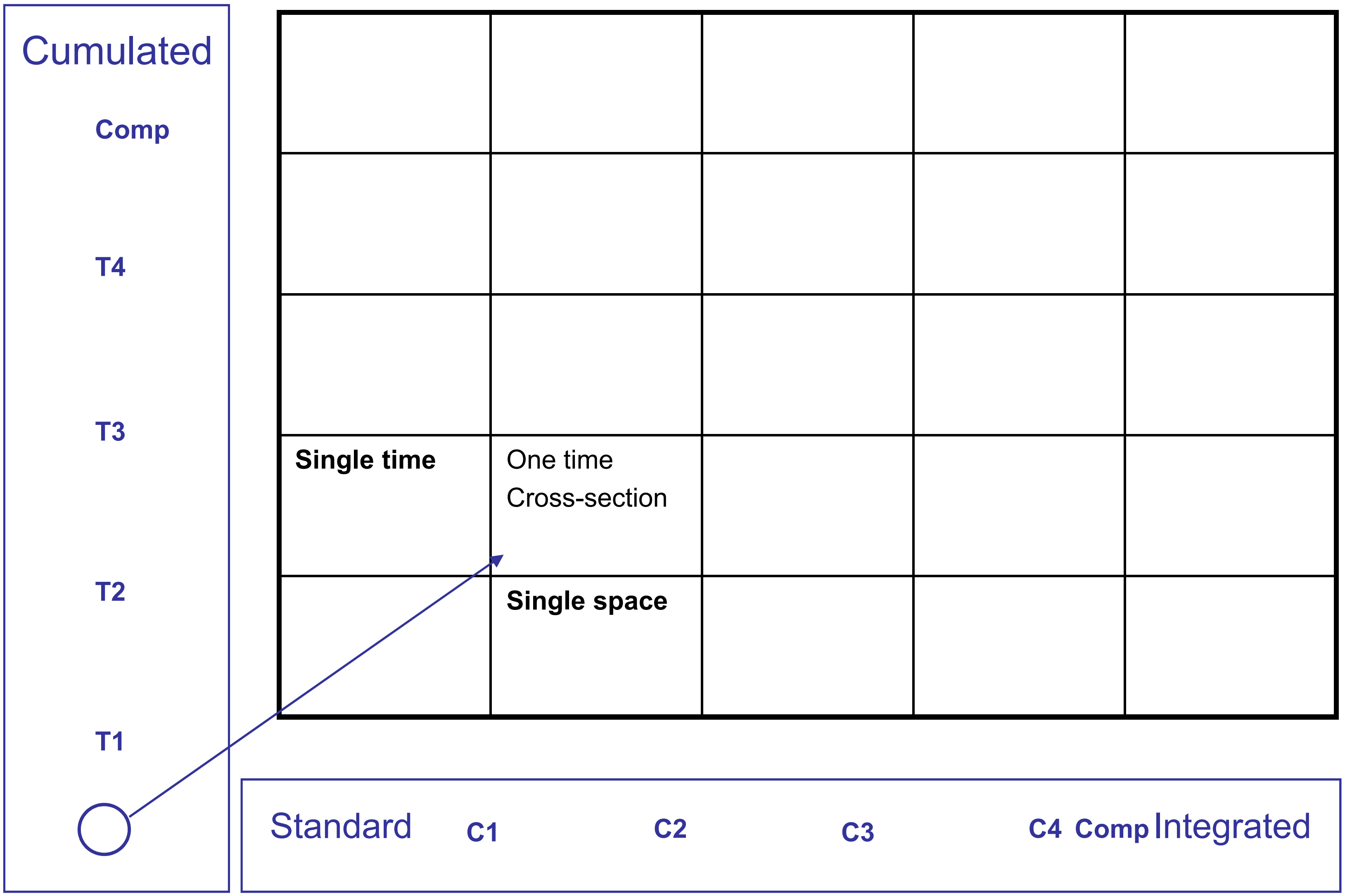
T3

T2

T1

Standard C1 C2 C3 C4 Comp Integrated

Start… with the origin:



Cumulated

Comp

T4

T3

T2

T1

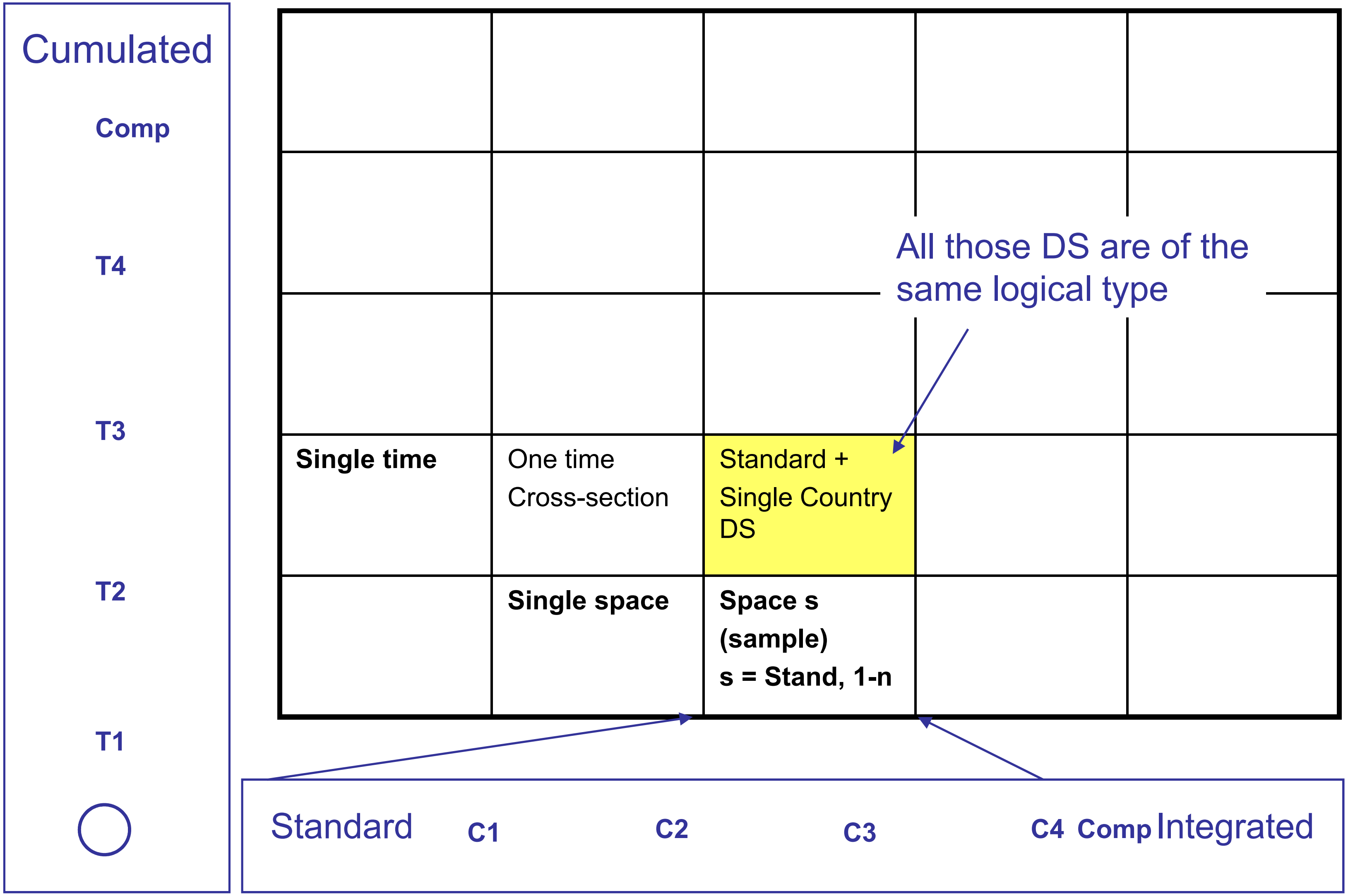
Single time

One time
Cross-section

Single space

Standard C1 C2 C3 C4 Comp Integrated

If we start a cross-national program, we need a standard definition and several country datasets



Cumulated

Comp

T4

T3

T2

T1

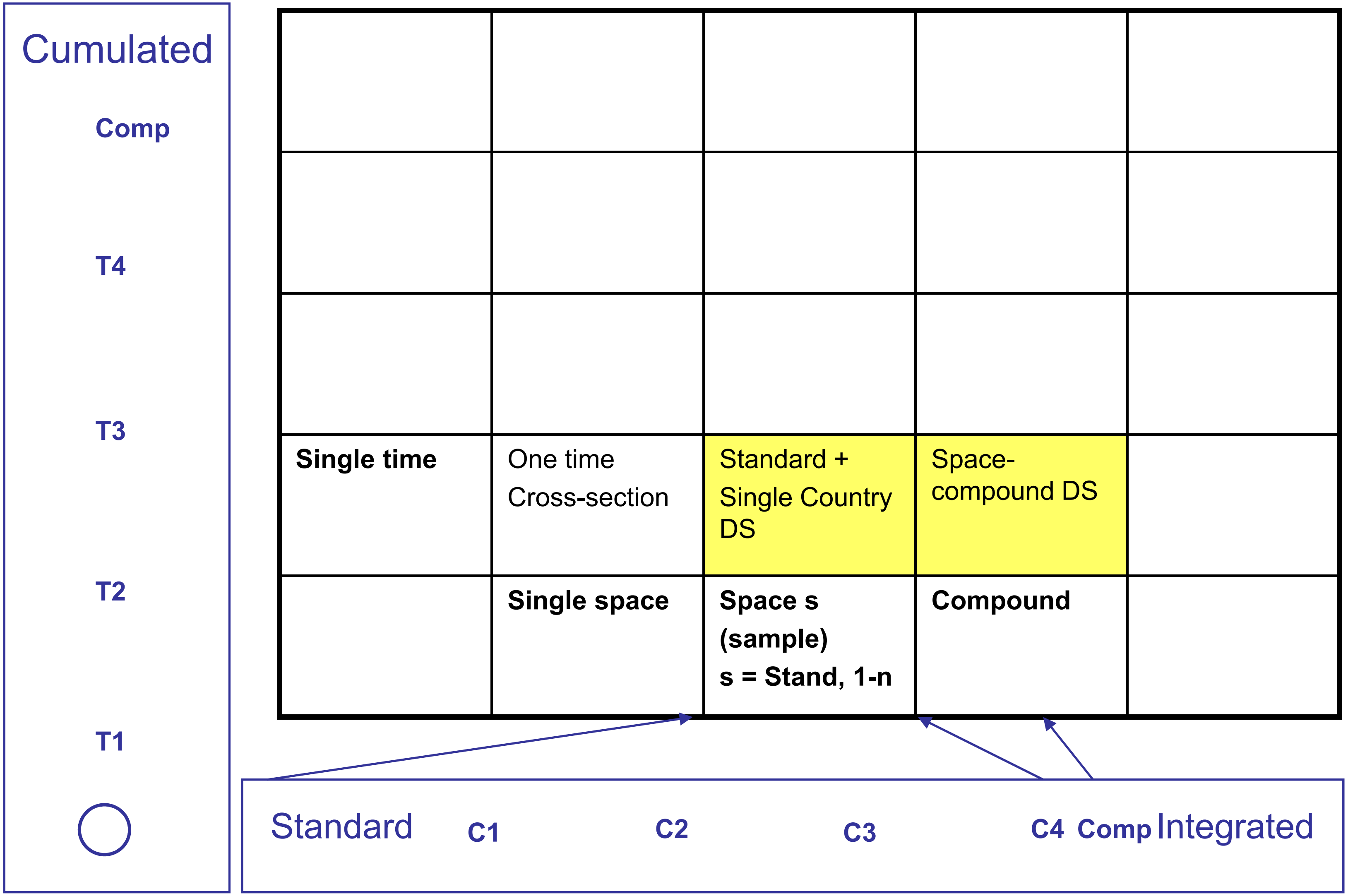
○

Single time	One time Cross-section	Standard + Single Country DS		
	Single space	**Space s (sample)** **s = Stand, 1-n**		

All those DS are of the same logical type

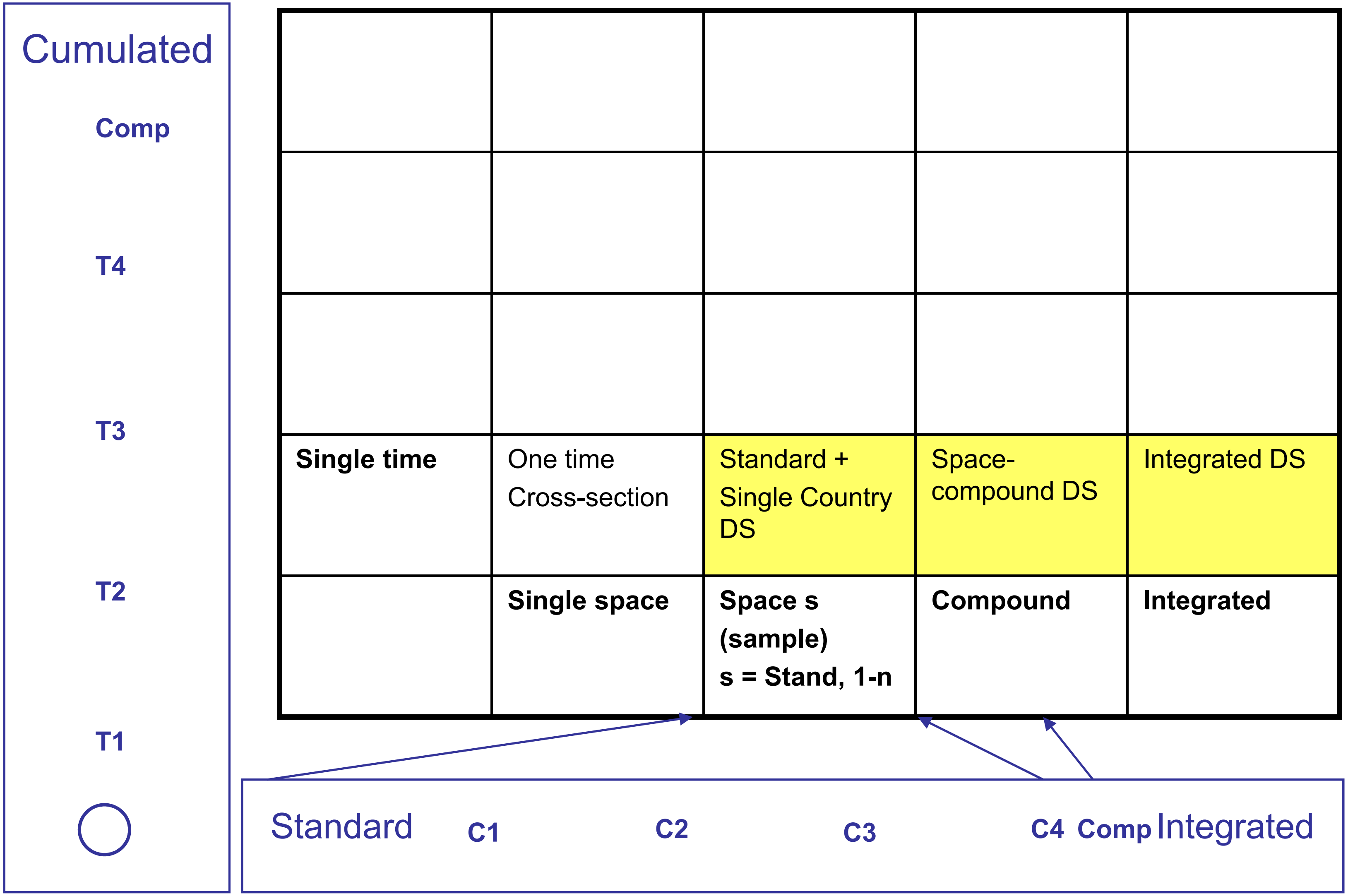
Standard **C1** **C2** **C3** **C4** **Comp** Integrated

By building anetwork of references from the country datasets to the standard we make this be a space-compound dataset, a higher level dataset



Single time	One time Cross-section	Standard + Single Country DS	Space-compound DS	
	Single space	**Space s (sample)** **s = Stand, 1-n**	**Compound**	

Integration…



Single time	One time Cross-section	Standard + Single Country DS	Space-compound DS	Integrated DS
	Single space	**Space s (sample)** **s = Stand, 1-n**	**Compound**	**Integrated**

Cumulated

Comp

T4

T3

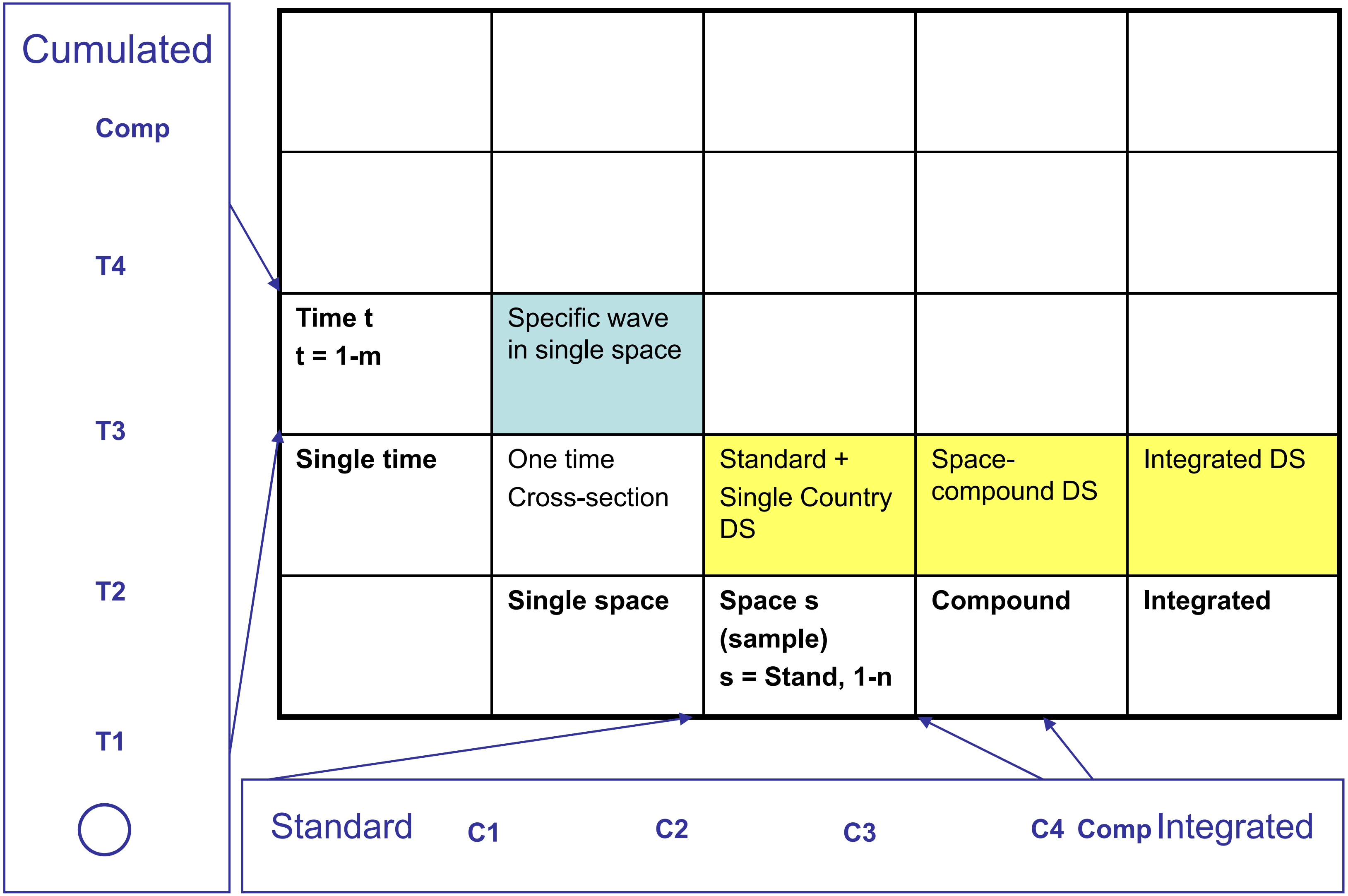
T2

T1

○

Standard **C1** **C2** **C3** **C4** **Comp** Integrated

Repeating the one-time cross-section, we get a series of datasets for a longitudinal study:



Time t **t = 1-m**	Specific wave in single space			
Single time	One time Cross-section	Standard + Single Country DS	Space-compound DS	Integrated DS
	Single space	**Space s** **(sample)** **s = Stand, 1-n**	**Compound**	**Integrated**

Cumulated

Comp

T4

T3

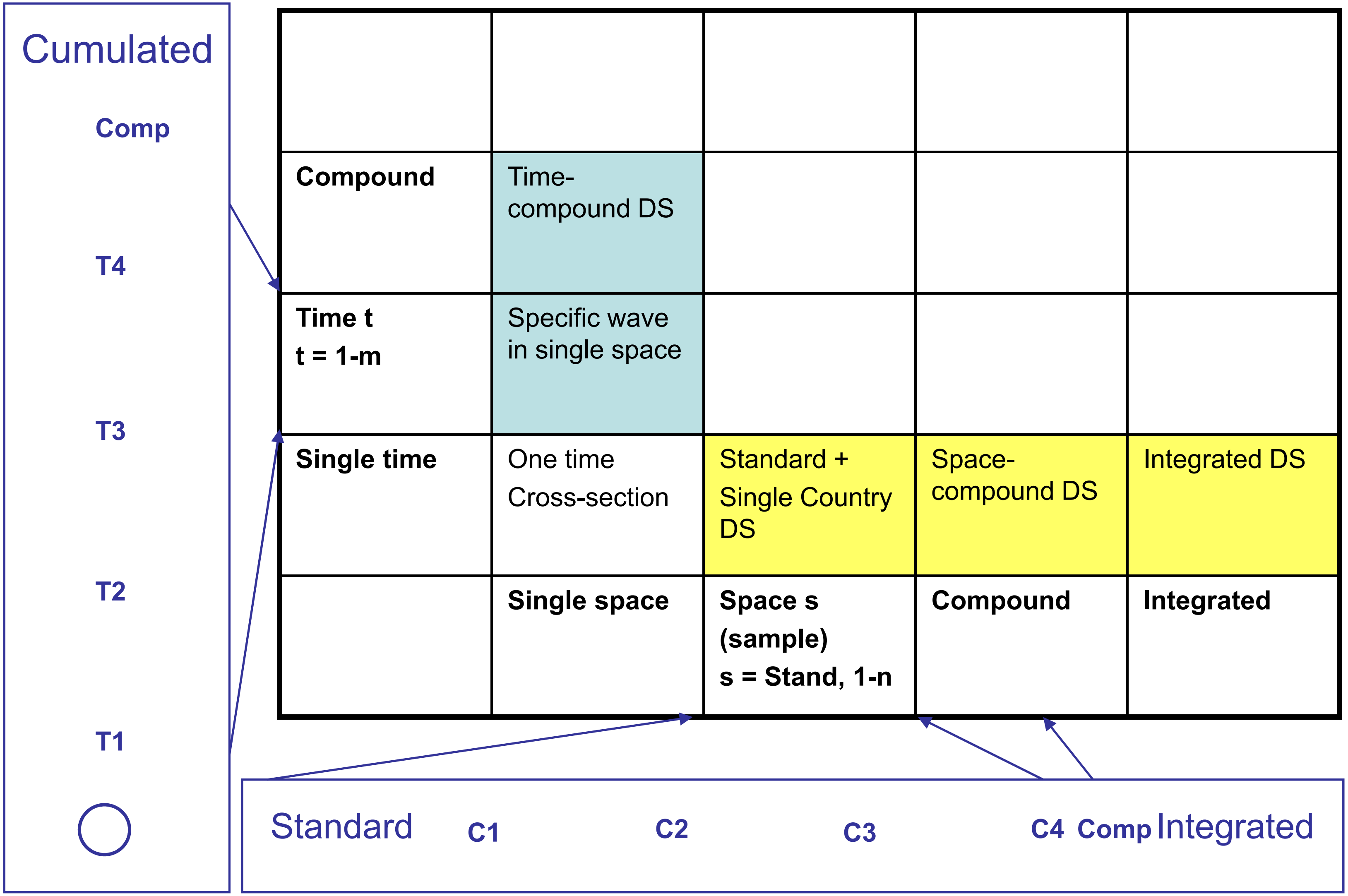
T2

T1

○

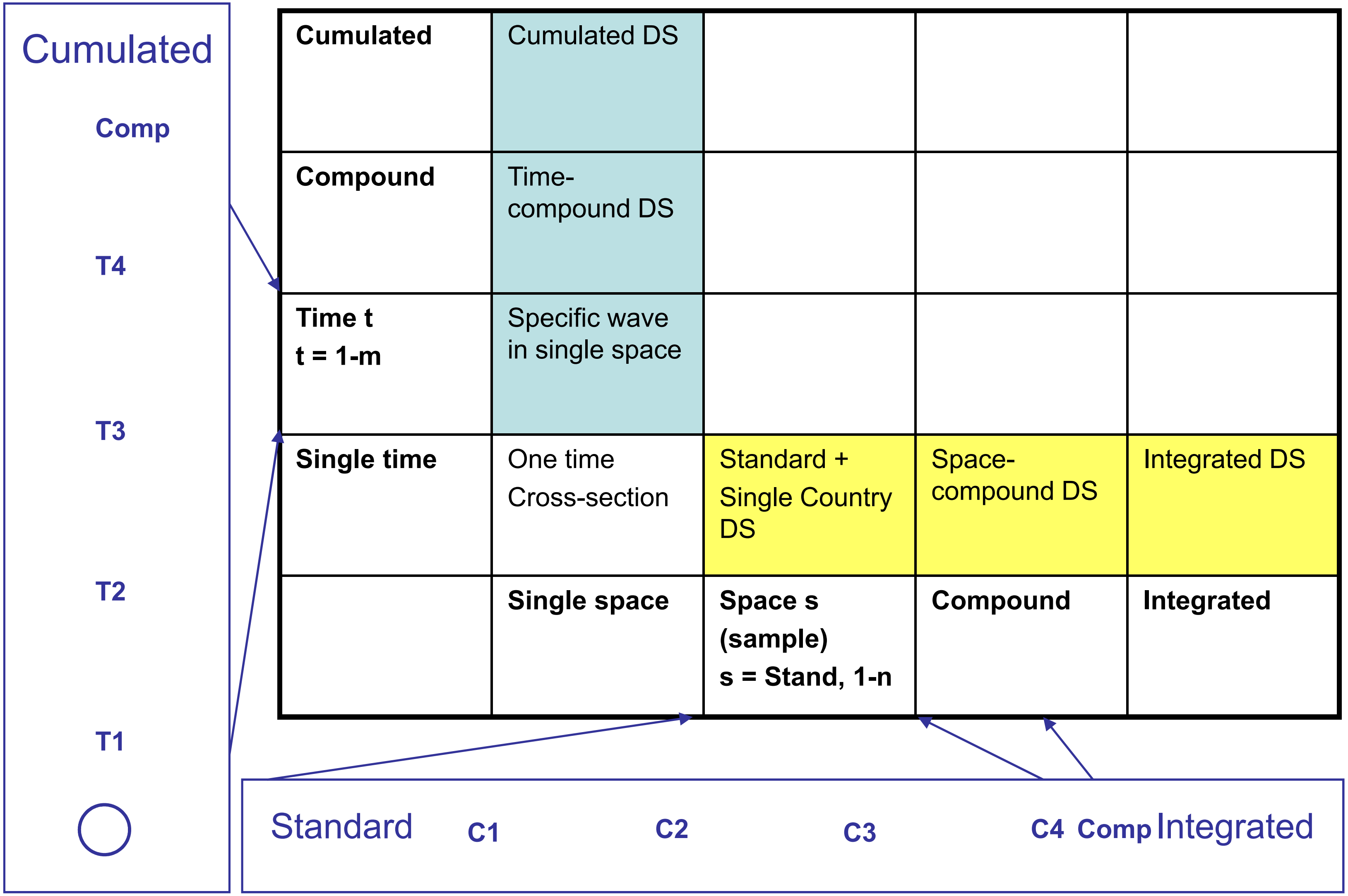
Standard **C1** **C2** **C3** **C4** **Comp** Integrated

Constructing the references from posterior waves to anterior waves, we get the time-compound dataset, a higher-level dataset



Compound	Time-compound DS			
Time t **t = 1-m**	Specific wave in single space			
Single time	One time Cross-section	Standard + Single Country DS	Space-compound DS	Integrated DS
	Single space	**Space s (sample)** **s = Stand, 1-n**	**Compound**	**Integrated**

Cumulate…



Cumulated	Cumulated DS			
Compound	Time-compound DS			
Time t **t = 1-m**	Specific wave in single space			
Single time	One time Cross-section	Standard + Single Country DS	Space-compound DS	Integrated DS
	Single space	**Space s (sample)** **s = Stand, 1-n**	**Compound**	**Integrated**

Cumulated

Comp

T4

T3

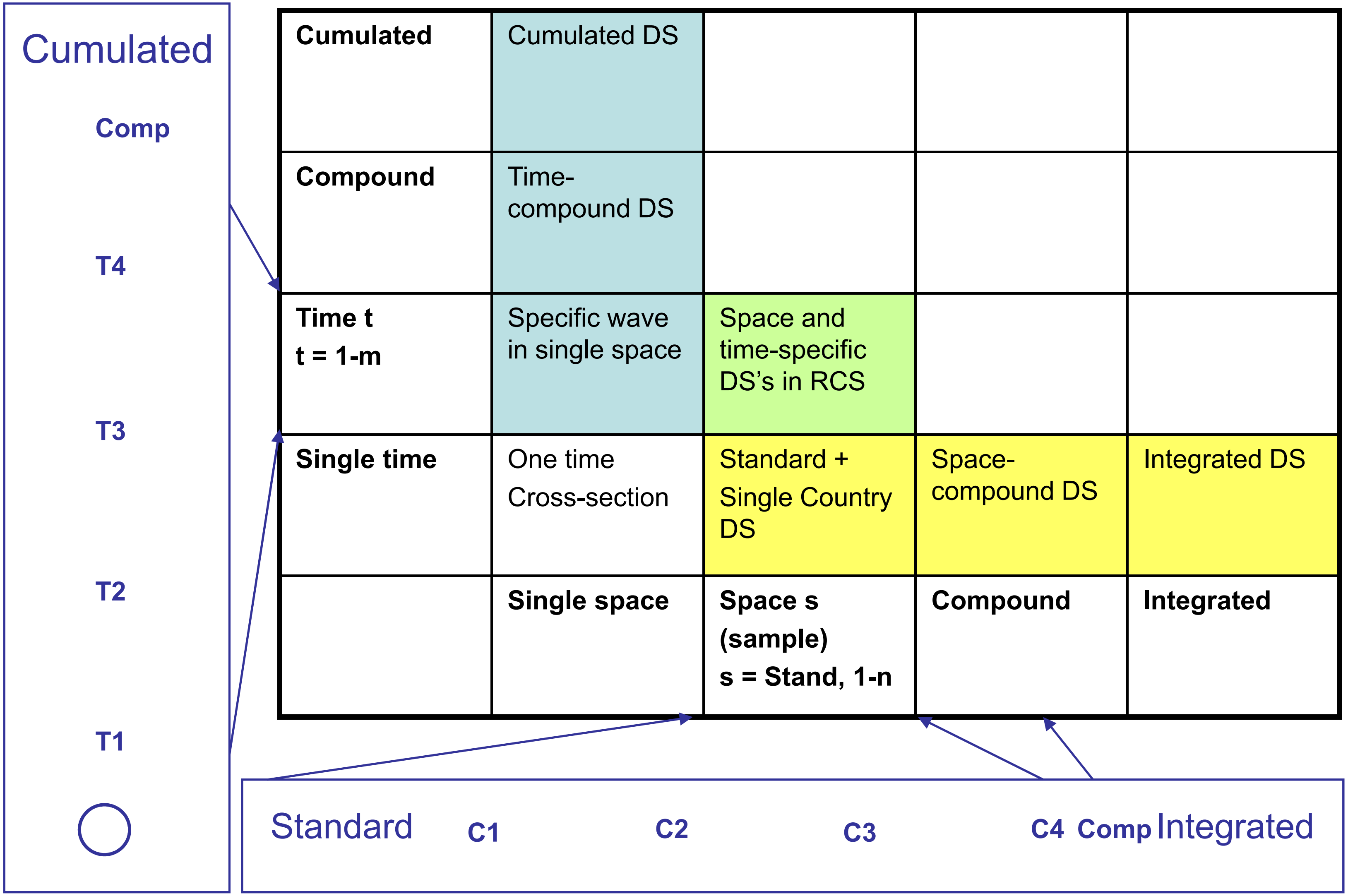
T2

T1

○

Standard **C1** **C2** **C3** **C4** **Comp** Integrated

You can also repeat a cross-national study program, so the space-specific DS's must also be defined in time



Cumulated

Comp

T4

T3

T2

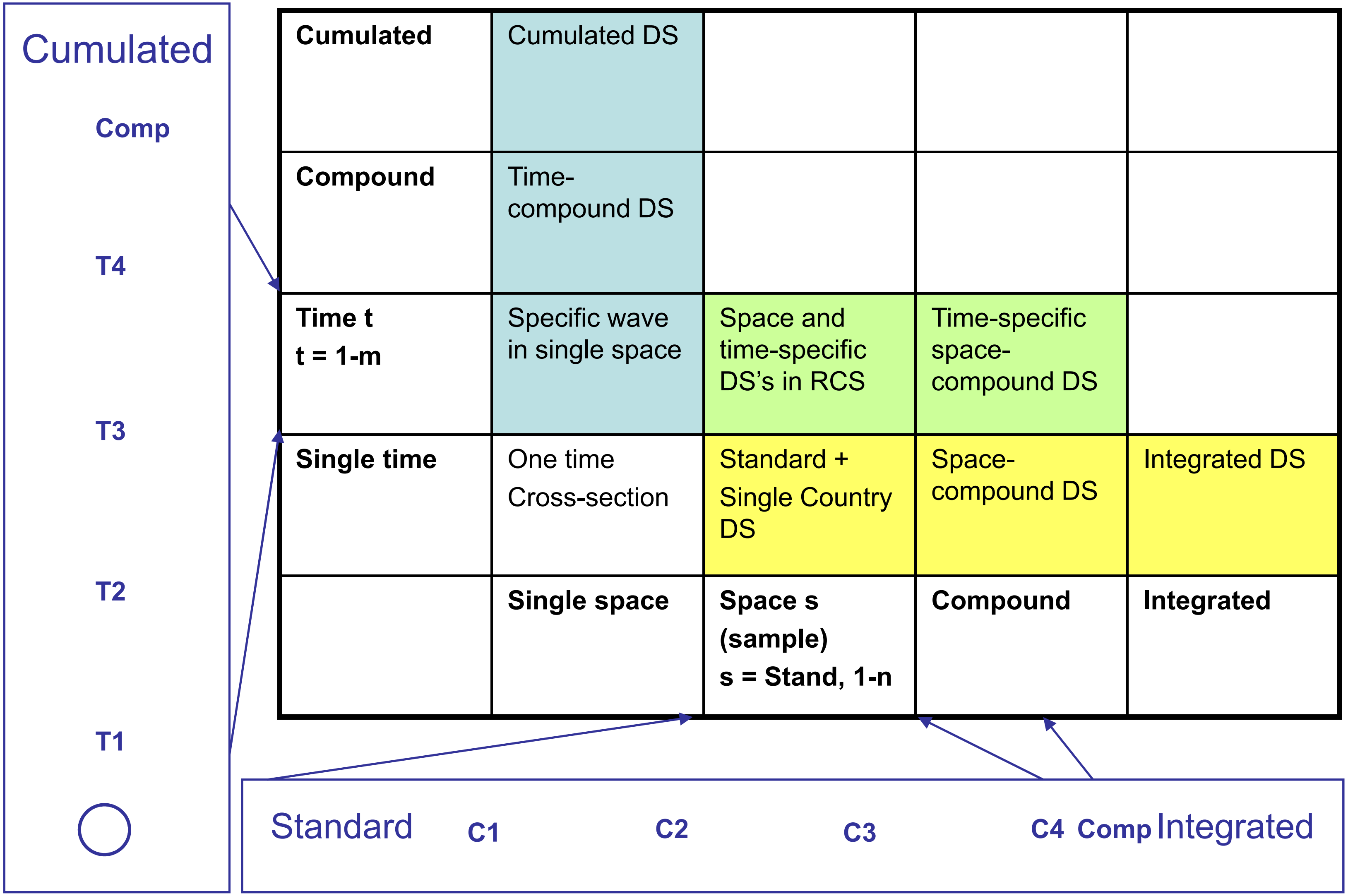
T1

○

Cumulated	Cumulated DS			
Compound	Time-compound DS			
Time t **t = 1-m**	Specific wave in single space	Space and time-specific DS's in RCS		
Single time	One time Cross-section	Standard + Single Country DS	Space-compound DS	Integrated DS
	Single space	**Space s** **(sample)** **s = Stand, 1-n**	**Compound**	**Integrated**

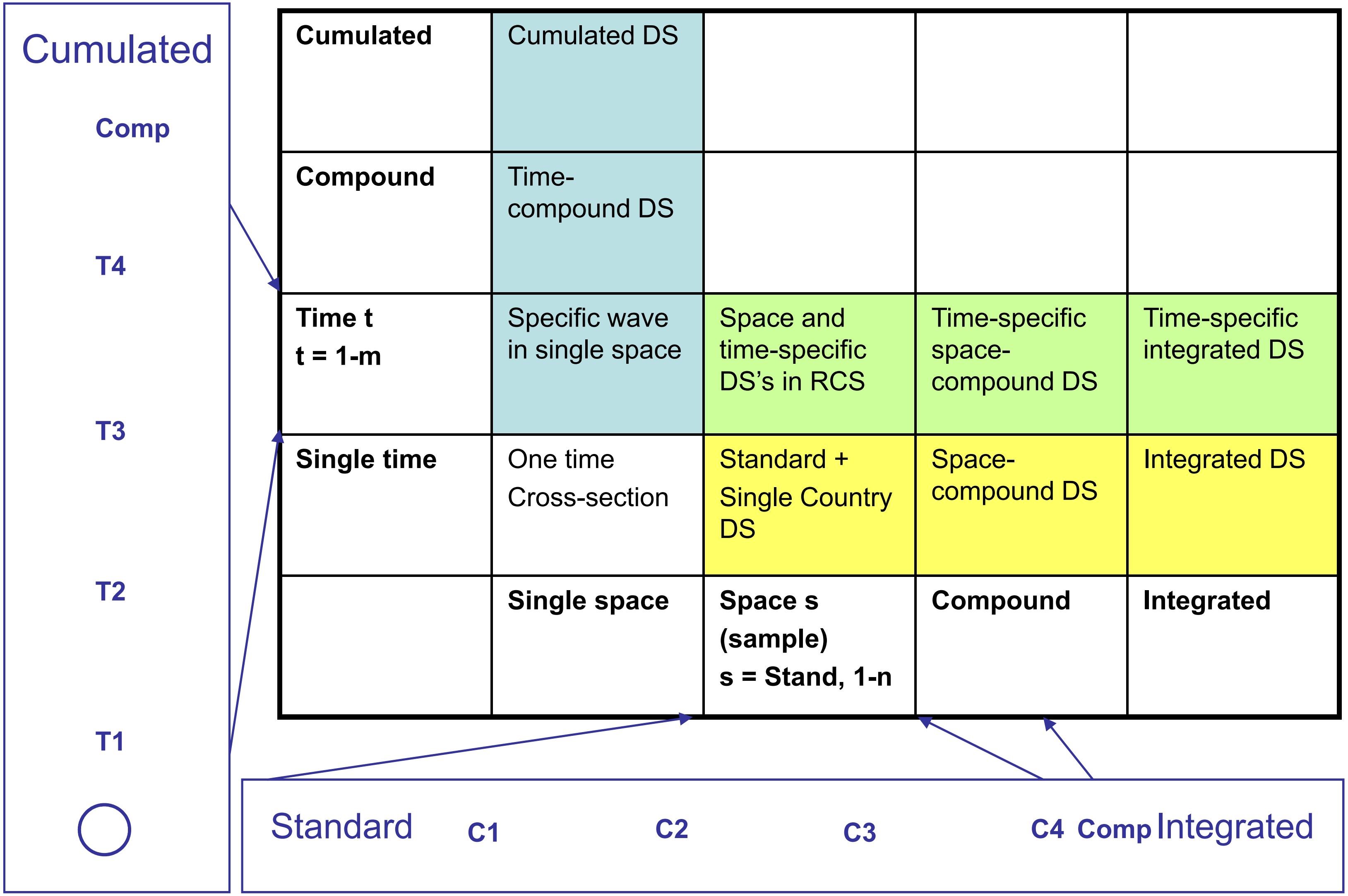
Standard C1 C2 C3 C4 Comp Integrated

Wave after wave, a time-specific compound dataset is made from the single-space datasets by building the network of references to the standard



Cumulated	Cumulated DS			
Compound	Time-compound DS			
Time t **t = 1-m**	Specific wave in single space	Space and time-specific DS's in RCS	Time-specific space-compound DS	
Single time	One time Cross-section	Standard + Single Country DS	Space-compound DS	Integrated DS
	Single space	**Space s** **(sample)** **s = Stand, 1-n**	**Compound**	**Integrated**

... and integrate wave after wave the compound dataset into a time-specific integrated dataset



Cumulated	Cumulated DS			
Compound	Time-compound DS			
Time t **t = 1-m**	Specific wave in single space	Space and time-specific DS's in RCS	Time-specific space-compound DS	Time-specific integrated DS
Single time	One time Cross-section	Standard + Single Country DS	Space-compound DS	Integrated DS
	Single space	**Space s** **(sample)** **s = Stand, 1-n**	**Compound**	**Integrated**

Waves following the one another, the standard will grow as a space-specific time-compound dataset, and some country datasets as well

Cumulated	Cumulated DS			
Compound	Time-compound DS	Space-specific time-compound DS		
Time t **t = 1-m**	Specific wave in single space	Space and time-specific DS's in RCS	Time-specific space-compound DS	Time-specific integrated DS
Single time	One time Cross-section	Standard + Single Country DS	Space-compound DS	Integrated DS
	Single space	**Space s** **(sample)** **s = Stand, 1-n**	**Compound**	**Integrated**

Cumulated

Comp

T4

T3

T2

T1

○

Standard **C1** **C2** **C3** **C4** **Comp** Integrated

The resulting christmas tree can be seen as a hyper-compound dataset, since
It is composed on two successive logical levels

Cumulated

Comp

T4

T3

T2

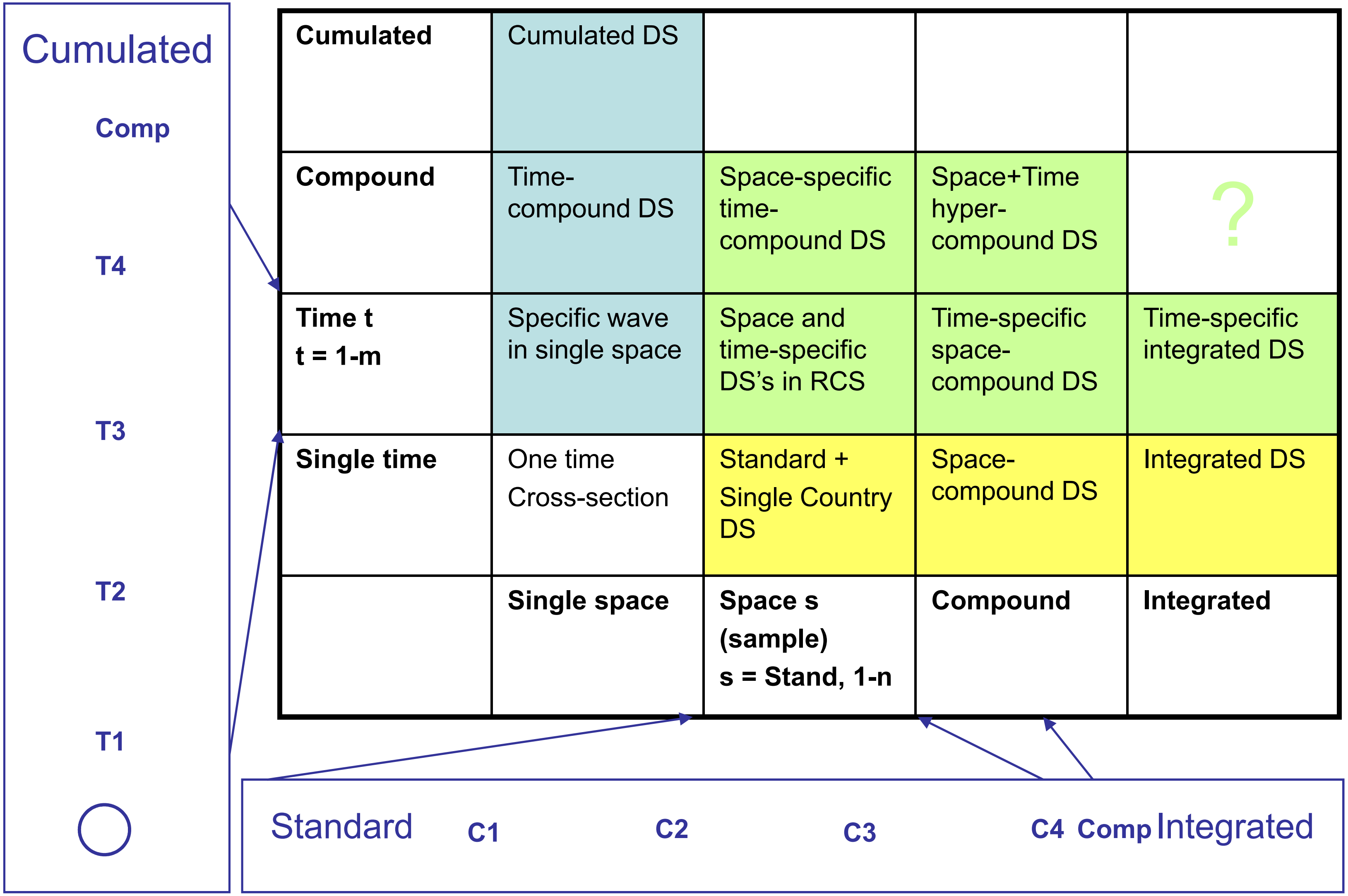
T1

○

Cumulated	Cumulated DS			
Compound	Time-compound DS	Space-specific time-compound DS	Space+Time hyper-compound DS	
Time t **t = 1-m**	Specific wave in single space	Space and time-specific DS's in RCS	Time-specific space-compound DS	Time-specific integrated DS
Single time	One time Cross-section	Standard + Single Country DS	Space-compound DS	Integrated DS
	Single space	**Space s** **(sample)** **s = Stand, 1-n**	**Compound**	**Integrated**

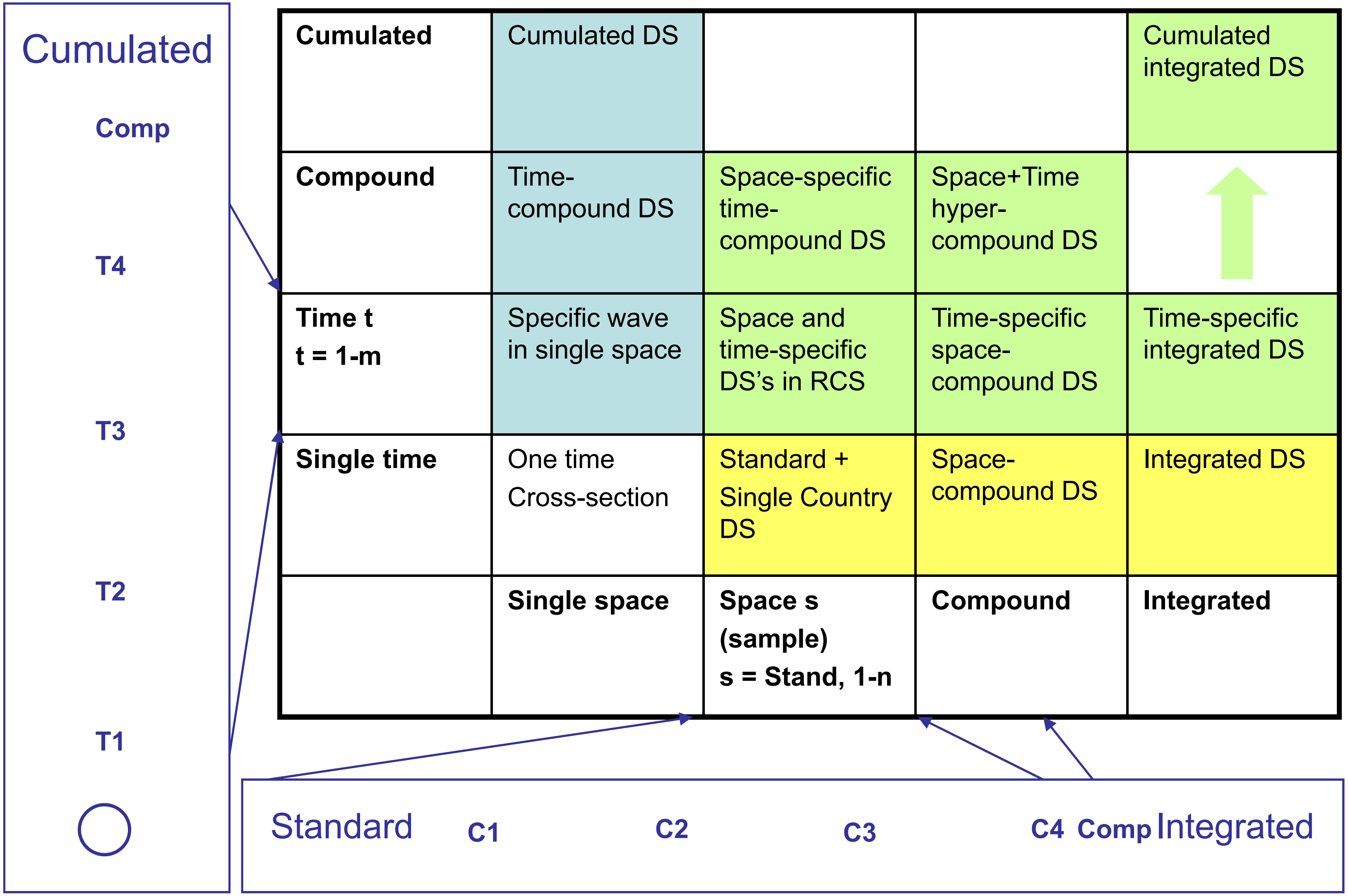
Standard **C1** **C2** **C3** **C4** **Comp** Integrated

… but you would probably not integrate the hyper-compound dataset



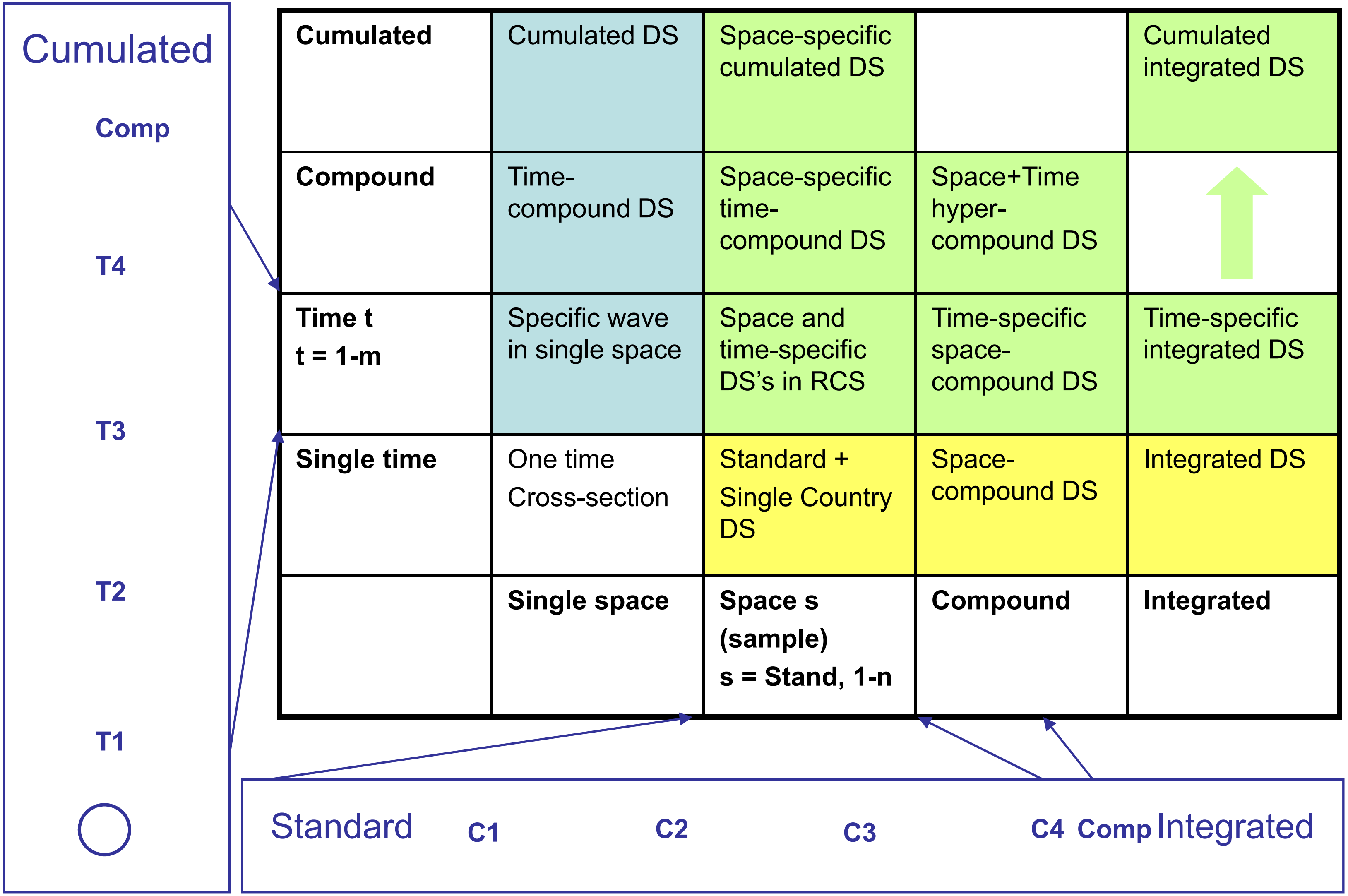
Cumulated	Cumulated DS			
Compound	Time-compound DS	Space-specific time-compound DS	Space+Time hyper-compound DS	?
Time t **t = 1-m**	Specific wave in single space	Space and time-specific DS's in RCS	Time-specific space-compound DS	Time-specific integrated DS
Single time	One time Cross-section	Standard + Single Country DS	Space-compound DS	Integrated DS
	Single space	**Space s (sample)** **s = Stand, 1-n**	**Compound**	**Integrated**

Instead, you will cumulate the time-specific datasets:



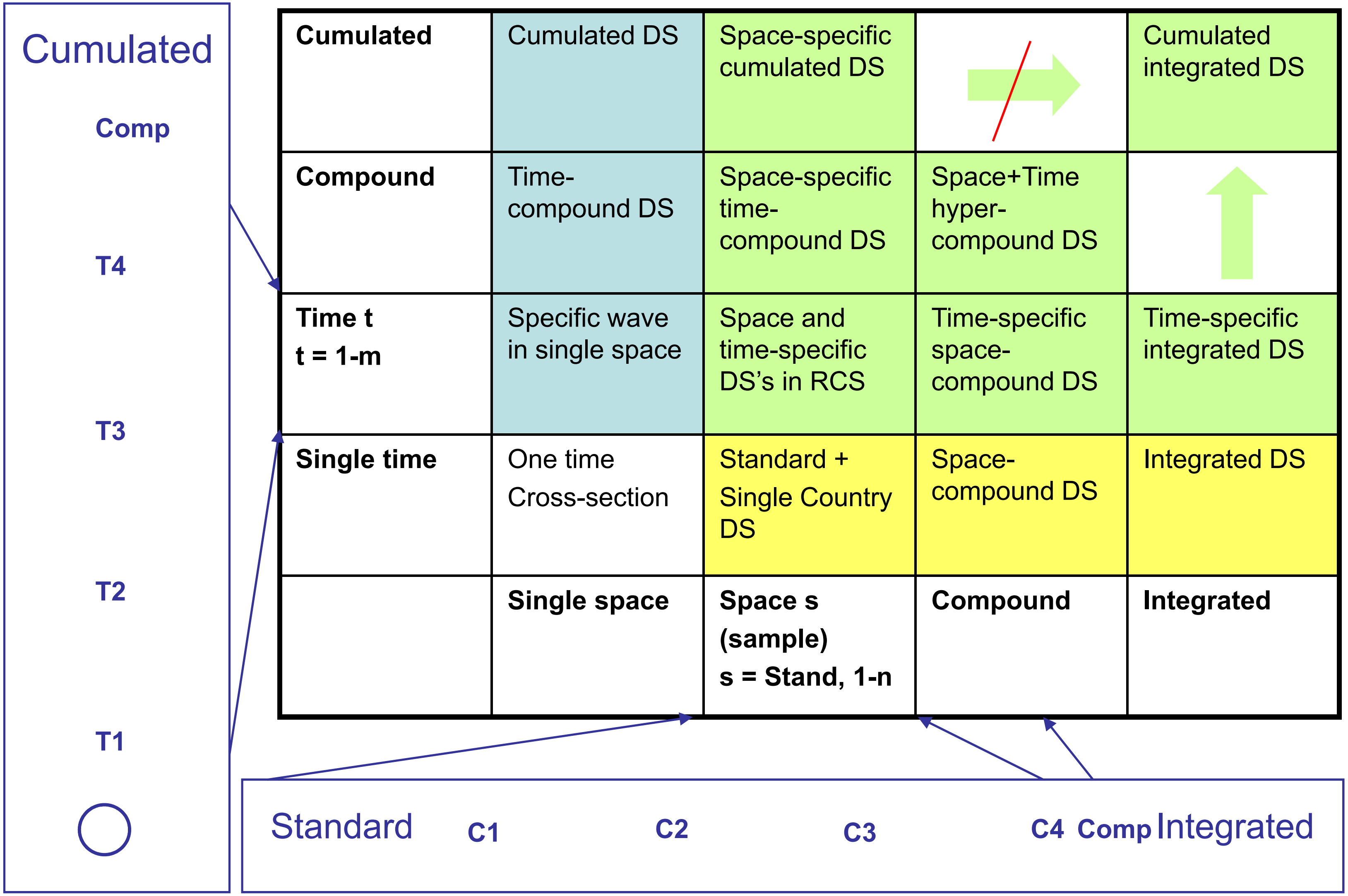
Cumulated	Cumulated DS			Cumulated integrated DS
Compound	Time-compound DS	Space-specific time-compound DS	Space+Time hyper-compound DS	↑
Time t **t = 1-m**	Specific wave in single space	Space and time-specific DS's in RCS	Time-specific space-compound DS	Time-specific integrated DS
Single time	One time Cross-section	Standard + Single Country DS	Space-compound DS	Integrated DS
	Single space	**Space s (sample)** **s = Stand, 1-n**	**Compound**	**Integrated**

Some countries will cumulate the datasets they cared for as local longitudinal studies



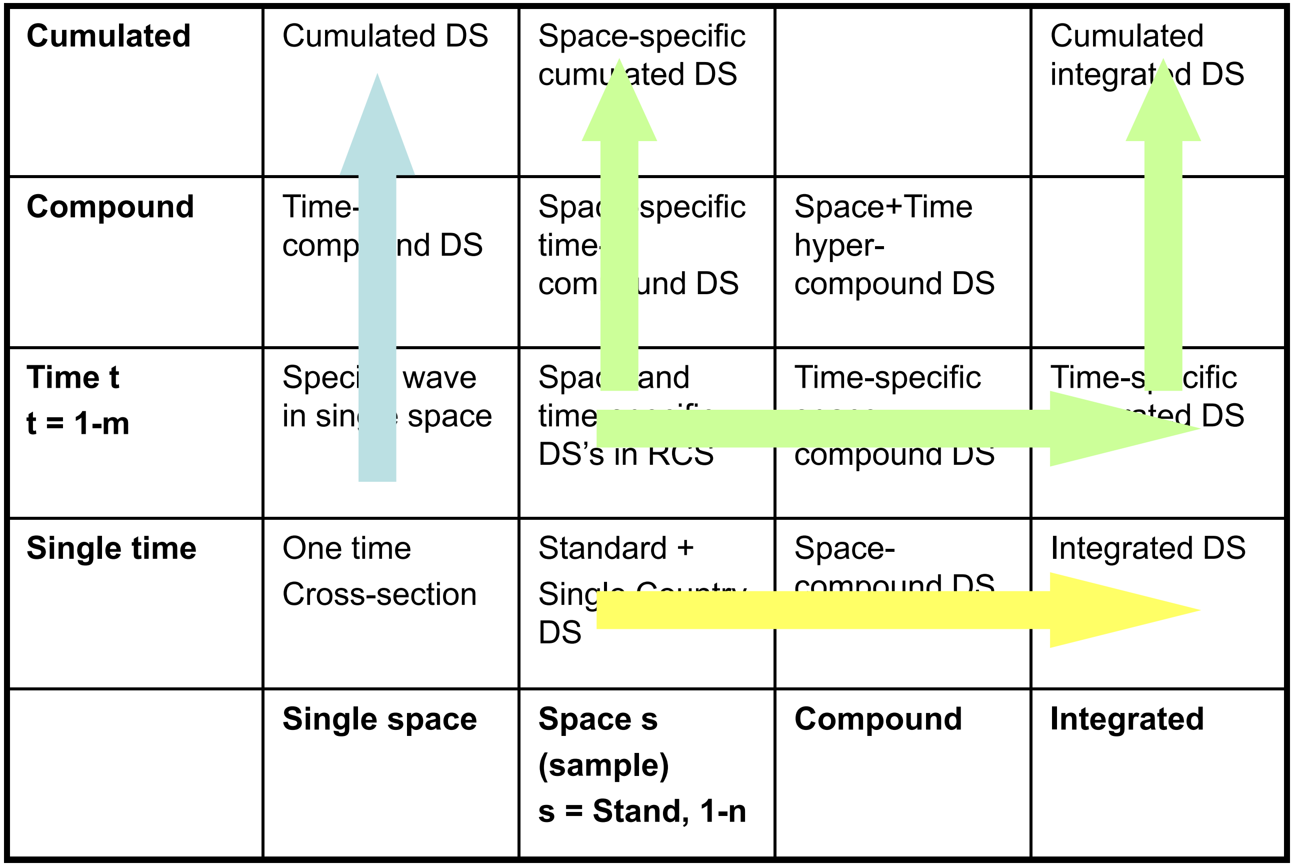
Cumulated	Cumulated DS	Space-specific cumulated DS		Cumulated integrated DS
Compound	Time-compound DS	Space-specific time-compound DS	Space+Time hyper-compound DS	
Time t **t = 1-m**	Specific wave in single space	Space and time-specific DS’s in RCS	Time-specific space-compound DS	Time-specific integrated DS
Single time	One time Cross-section	Standard + Single Country DS	Space-compound DS	Integrated DS
	Single space	**Space s** **(sample)** **s = Stand, 1-n**	**Compound**	**Integrated**

…but there will probably be no attempt at composing them nor at integrating Integrating them.



Cumulated	Cumulated DS	Space-specific cumulated DS	↛	Cumulated integrated DS
Compound	Time-compound DS	Space-specific time-compound DS	Space+Time hyper-compound DS	↑
Time t **t = 1-m**	Specific wave in single space	Space and time-specific DS's in RCS	Time-specific space-compound DS	Time-specific integrated DS
Single time	One time Cross-section	Standard + Single Country DS	Space-compound DS	Integrated DS
	Single space	**Space s** **(sample)** **s = Stand, 1-n**	**Compound**	**Integrated**

So we end up with the following typology of datasets:

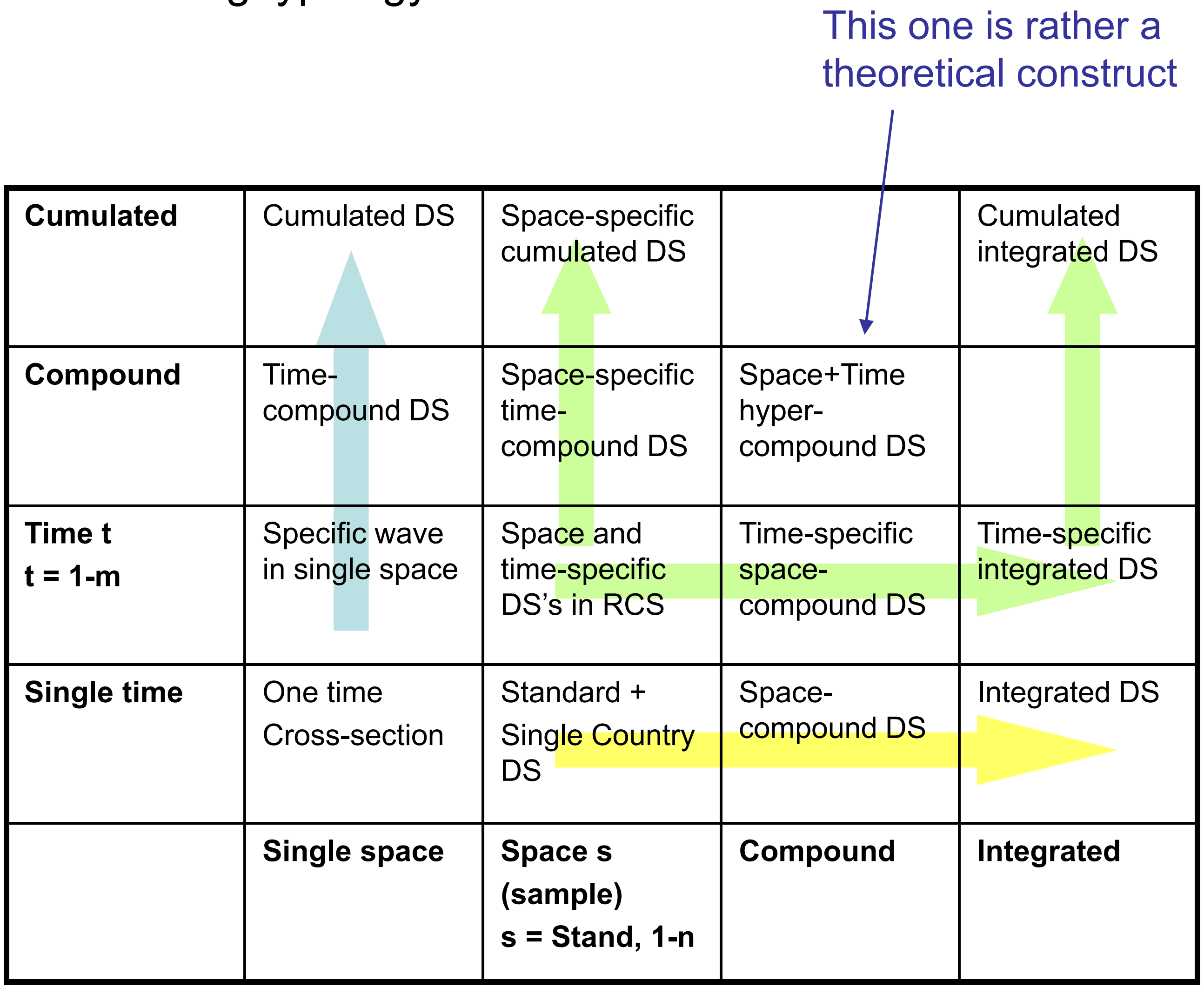


Cumulated	Cumulated DS	Space-specific cumulated DS		Cumulated integrated DS
Compound	Time-compound DS	Space-specific time-compound DS	Space+Time hyper-compound DS	
Time t **t = 1-m**	Specific wave in single space	Space and time-specific DS’s in RCS	Time-specific space-compound DS	Time-specific integrated DS
Single time	One time Cross-section	Standard + Single Country DS	Space-compound DS	Integrated DS
	Single space	**Space s** **(sample)** **s = Stand, 1-n**	**Compound**	**Integrated**

So we end up with the following typology of datasets:

Cumulated	Cumulated DS	Space-specific cumulated DS		Cumulated integrated DS
Compound	Time-compound DS	Space-specific time-compound DS	Space+Time hyper-compound DS	
Time t **t = 1-m**	Specific wave in single space	Space and time-specific DS's in RCS	Time-specific space-compound DS	Time-specific integrated DS
Single time	One time Cross-section	Standard + Single Country DS	Space-compound DS	Integrated DS
	Single space	**Space s** **(sample)** **s = Stand, 1-n**	**Compound**	**Integrated**

So we end up with the following typology of datasets:



This one is rather a theoretical construct

Cumulated	Cumulated DS	Space-specific cumulated DS		Cumulated integrated DS
Compound	Time-compound DS	Space-specific time-compound DS	Space+Time hyper-compound DS	
Time t **t = 1-m**	Specific wave in single space	Space and time-specific DS’s in RCS	Time-specific space-compound DS	Time-specific integrated DS
Single time	One time Cross-section	Standard + Single Country DS	Space-compound DS	Integrated DS
	Single space	**Space s** **(sample)** **s = Stand, 1-n**	**Compound**	**Integrated**

Based on this typology, we can define a system of coordinates for the identification of the dataset of reference for work or publication

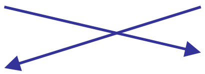
Space	Time
Single	Single
Germany	2001
France	2003
Italy	2005
UK	2007
etc.	etc.
Compound	Compound
Integrated	Cumulated

… with standard values and program specific values

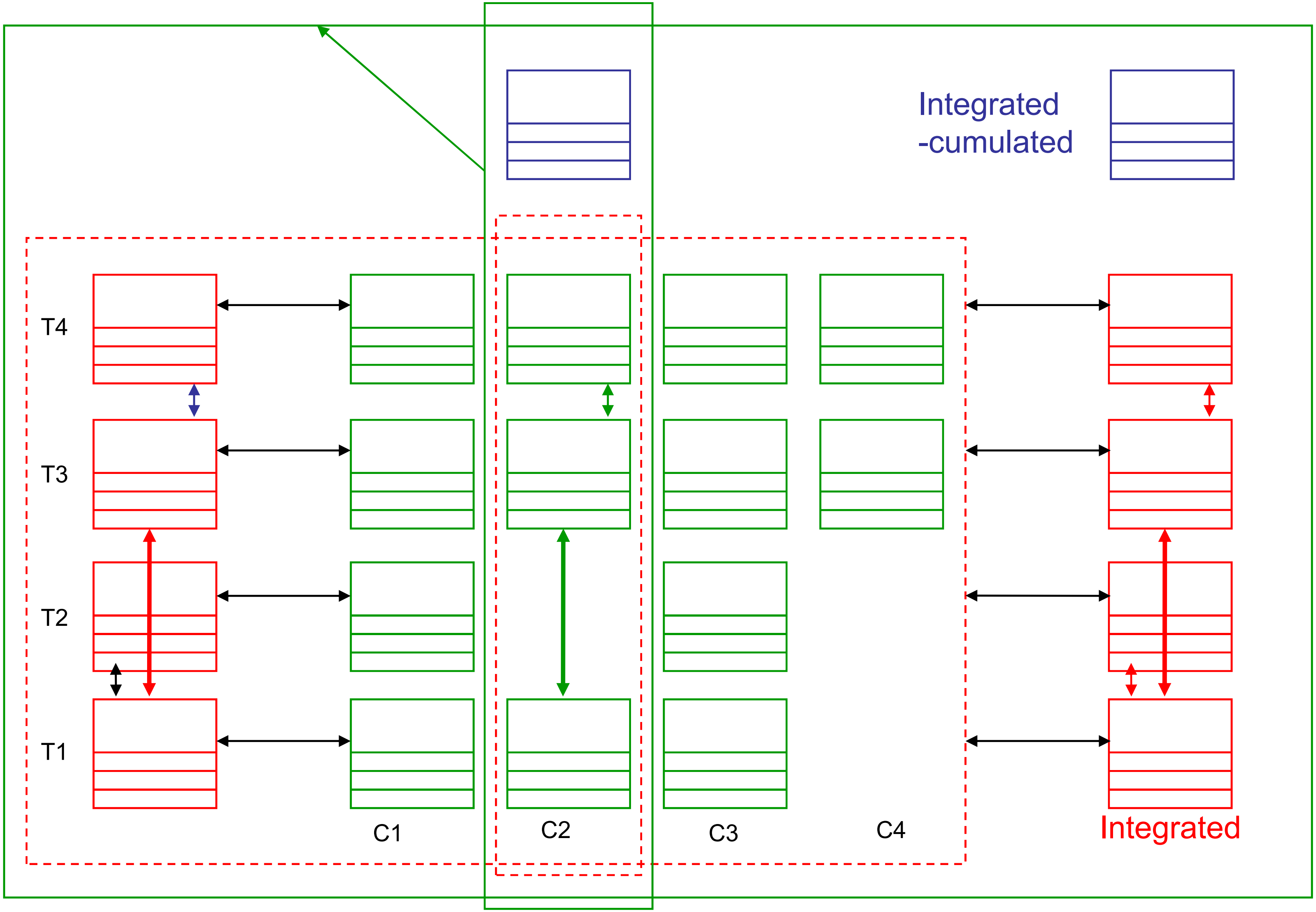
Space	Time	
Single	Single	(Standard)
Germany	2001	(Variable)
France	2003	
Italy	2005	
UK	2007	
etc.	etc.	
Compound	Compound	(Standard)
Integrated	Cumulated	

… and some forbidden combinations

Space	Time	
Single	Single	(Standard)
Germany France Italy UK etc.	2001 2003 2005 2007 etc.	(Variable)
Compound Integrated	Compound Cumulated	(Standard)



Let’s turn back to a previoius view to screen the involved studies



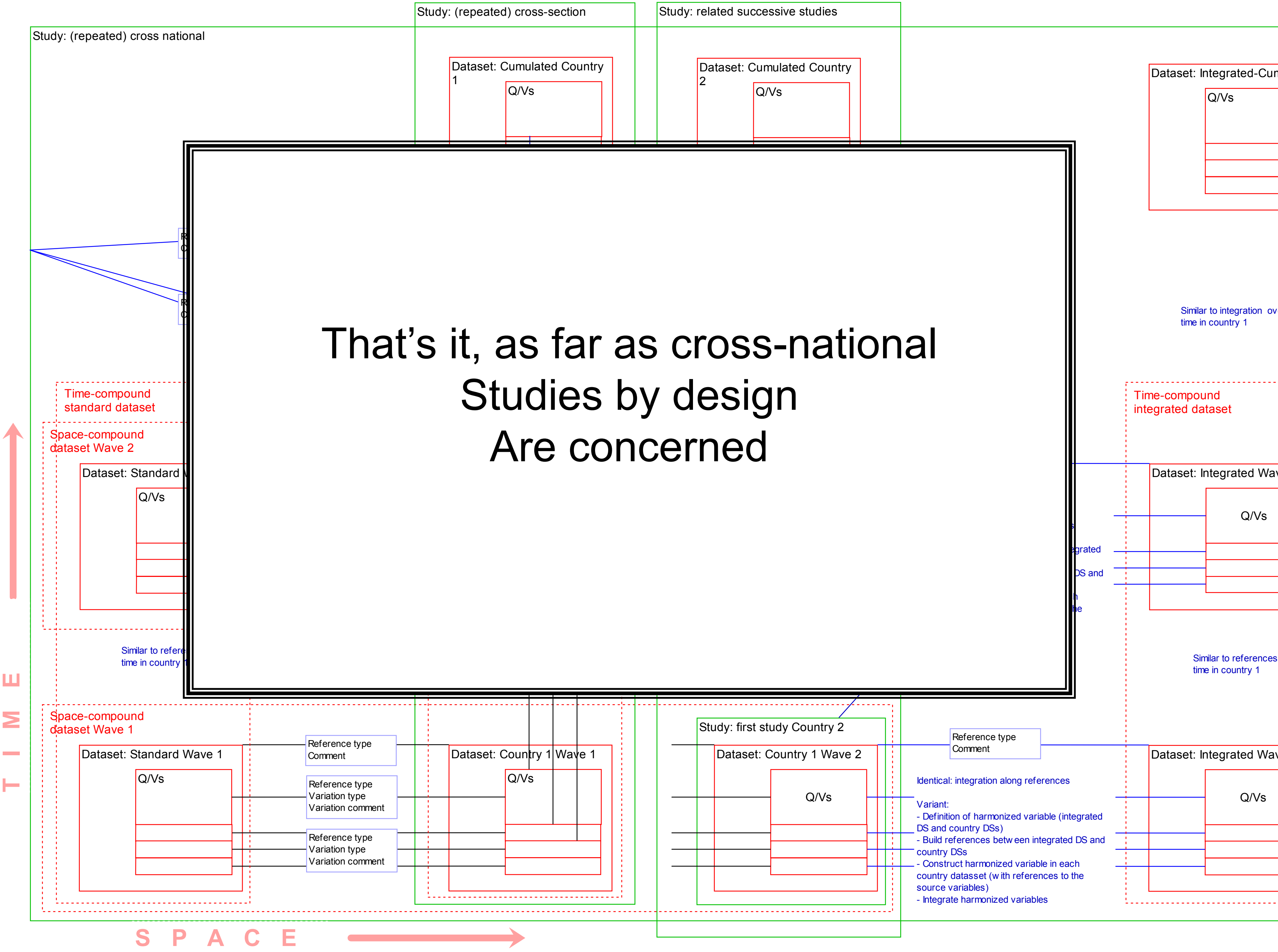
Let’s turn back to a previoius view to screen the involved studies

Only two types of studies are needed:

the **coordination study** and
the **country study**

Provided that there is a possibility
for describing different activities
within the same study
  coordination
  implementation
  integration (space)
  cumulation (time)
  (and a few others)
together with the respective authority
Information, notes and citation statements

the same information structure can be used for both types of studies

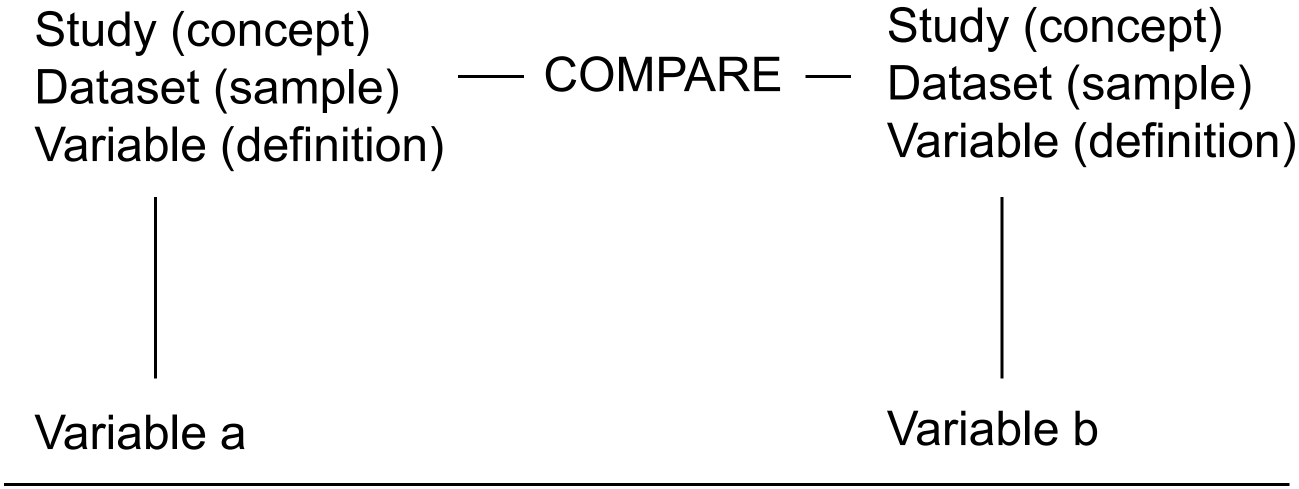

Study: (repeated) cross-section
Study: related successive studies
Study: (repeated) cross national
Dataset: Cumulated Country 1
Q/Vs
Dataset: Cumulated Country 2
Q/Vs
Dataset: Integrated-Cum
Q/Vs
R
C
R
C
Similar to integration ov
time in country 1
Time-compound standard dataset
Space-compound dataset Wave 2
Dataset: Standard W
Q/Vs
Time-compound integrated dataset
Dataset: Integrated Wav
Q/Vs
s
egrated
DS and
n
he
Similar to refere
time in country 1
Similar to references
time in country 1
Space-compound dataset Wave 1
Dataset: Standard Wave 1
Q/Vs
Reference type
Comment
Reference type
Variation type
Variation comment
Reference type
Variation type
Variation comment
Dataset: Country 1 Wave 1
Q/Vs
Study: first study Country 2
Dataset: Country 1 Wave 2
Q/Vs
Reference type
Comment
Dataset: Integrated Wav
Q/Vs
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 datasset (with references to the source variables)
- Integrate harmonized variables
T I M E
S P A C E

That’s it, as far as cross-national
Studies by design
Are concerned

Now…
just pick two variables…

(the harmonization study)

… and compare the respective metadata hierarchies:



Study (concept)
Dataset (sample)
Variable (definition)

— COMPARE —

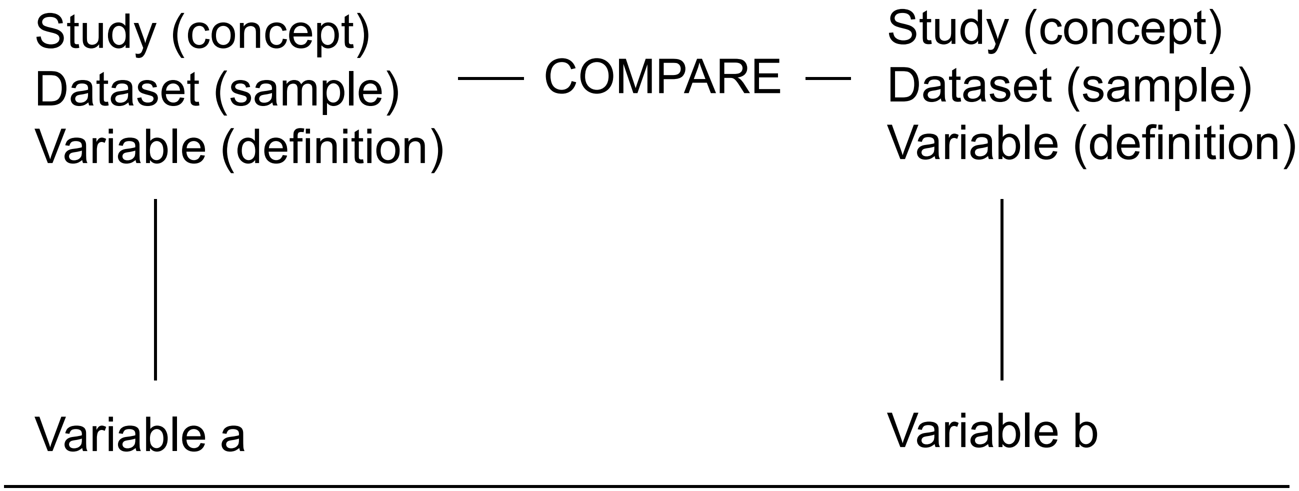
Study (concept)
Dataset (sample)
Variable (definition)

Variable a

Variable b

If the metadata are close enough on all levels, compare the data.
If not, take hands off.

Choosing the two variables in your favorite metadata handling application, you should be offered the information elements to compare



Study (concept)
Dataset (sample)
Variable (definition)

— COMPARE —

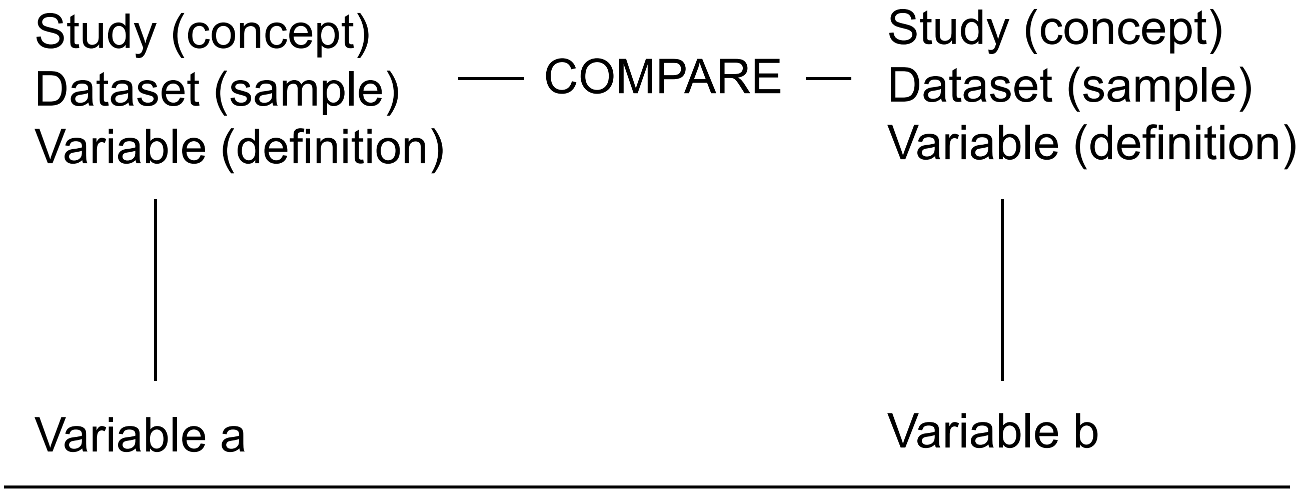
Study (concept)
Dataset (sample)
Variable (definition)

Variable a

Variable b

This is not a matter of metadata structure. At this stage, it is a matter of the application

But you may wish to store the comparison as a re-usable relationship in the metadata and even to compute a harmonized variable…



Study (concept)
Dataset (sample)
Variable (definition)

COMPARE

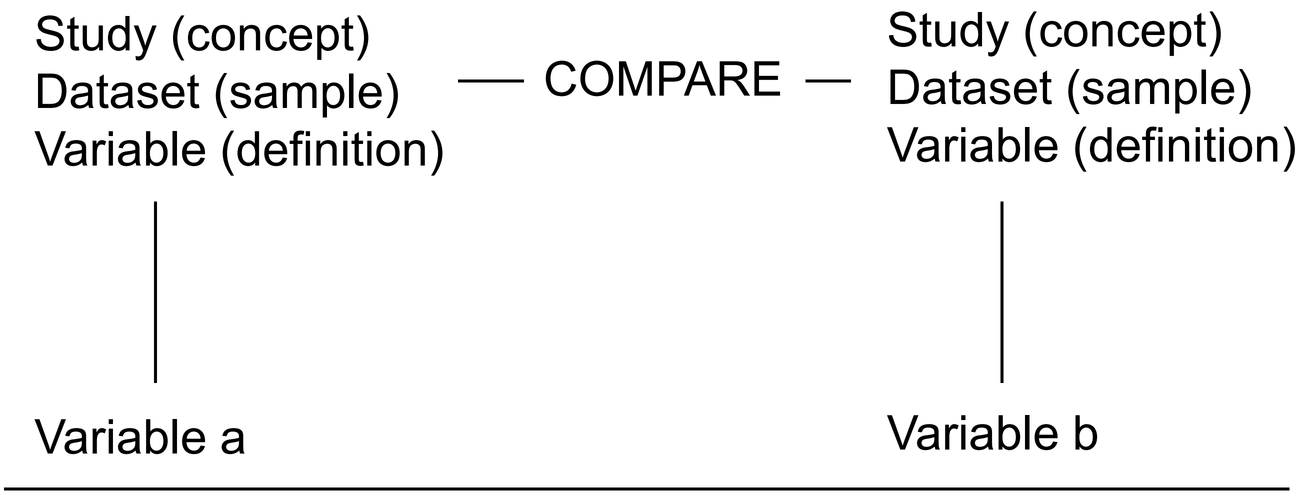
Study (concept)
Dataset (sample)
Variable (definition)

Variable a

Variable b

Then you will create a study for its own, describe in it your comparison project and let the variables refer to the variables in the original datasets.

You should also be able to choose studies, which you know to be kin in methods and contents, create a virtual compound dataset, and check the degree of comparability on all levels



Study (concept)
Dataset (sample)
Variable (definition)

COMPARE

Study (concept)
Dataset (sample)
Variable (definition)

Variable a

Variable b

…before defining the sets of variables, which can be harmonized over space or over time

The end

Issues

Abstract conceptual model

- Reference case:
  - All operations handled with one instance of the application under one single authority
- Real life cases:
  - Coordination group and local implementers
  - Changes in organization over time
- Extension: multiple authors…
  - working on a single system
  - wording on communicating systems

Levels

- Some of the mechanisms have been tested in a SIDOS owned prototype (series of questions and variables).
- Studies and datasets of various types, depending on their level in the structure (higher level objects composed of lower level objects)

Reference or inheritance?

- DDI V 3.0 introduces a grouping structure
  - Association?
  - Composition?
  - Inheritance (with local override)? (o-o thinking)
- Relational data model
  - References
- Combination of both ways of thinking?
  - More detailed analysis necessary

To be continued…