TECHNICAL REPORT

ESPON

Coding scheme to label indicators: Guiding principles to TPGs involved in ongoing and future ESPON projects

MAIN RESULTS

- The rationale defined by each TPG to label indicators and indices increases the degree of ambiguity
- Develop an harmonised scheme to code ESPON indicators and indices
- Illustrative examples demonstrate the usefulness of such approach

ESPON 2013 DATABASE

MARCH 2011

EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

12 pages

LIST OF AUTHORS

Nuno Madeira, University of Luxembourg
Geoffrey Caruso, University of Luxembourg

Contact

E-mail: nuno.madeira@uni.lu; Tel. +352 46 66 44 9691
E-mail: geoffrey.caruso@uni.lu; Tel. +352 46 66 44 6625

Tables of contents

1. Introduction

This report provides background information and guidelines to those responsible for the delivery of ESPON data. One of the issues that emerged during the implementation of the ESPON 2013 DB concerns the rationale defined by Transnational Project Groups (TPGs) to label variables. Despite the usefulness of the different methods applied by TPGs, the fact is that these methods increase the degree of ambiguity within the ESPON 2013 DB. The implications are relevant, because if no harmonisation is applied it will invariably lead to incoherence.

Therefore, the aim of the present document is to provide guidelines on how to create codes for datasets that will be integrated in the database. Taking into consideration that different TPGs may have to deliver the same variable, it is then essential to apply a harmonised and identical code for those that share the same designation. In a way, this would bring some coherence to the database.

2. Coding scheme to label indicators

Prior considerations

Naming indicators is an important component of indicator development. Therefore, TPGs should strive to be objective and consistent. Taking into consideration the ultimate list of ESPON indicators we noticed a wide variation of naming conventions that differ according to the criterion defined by each research team. Indeed, some applied research projects have chosen to use more simplistic names that capture the essence of what is being measured while others tend to be very descriptive. This same list discloses, however, some similar labels that essentially reveal different methodologies, data providers and, most likely, implications for the variable that is being classified as such.

In this section we introduce a coding scheme to label ESPON indicators and indices. Its purpose is to assemble relevant information about data using a minimum number of characters. We believe that it is impossible to fully harmonize a coding scheme given the variety of indicators delivered up to date. However, some harmonisation should be adopted.

With this regard, those responsible for the delivery of datasets within a TPG should first browse the list of acronyms and abbreviations proposed to identify variables (see Annex 1). In the event that a specific variable already integrates the above mentioned list, TPGs are encouraged to use the existing code. If not, TPGs should simply create a new code according to the scheme described in these guidelines.

Therefore, we invite TPGs to apply a coding scheme that comprises three fields or divisions. The information to be included is each division is based on the current state of the database and, essentially, tries to focus on the indicator itself, i.e. subject, nature of data, and calculation method(s). Other elements that be might be used to label data will not be considered as they are already mentioned on metadata (i.e. time, space, data provider).

Procedure

Each field or division that we consider for the coding scheme is accompanied by a list of abbreviations and/or acronyms that best identify those above-mentioned categories. The first two lists will certainly be adapted and enriched over time as the database integrates new data. For this purpose, we provide some abbreviations widely used to label indicators (see Annex 2). The third list should preferably remain fixed since it corresponds to measurement scales as recognised in the geographical and/or statistical literature (see

Annex 3). No maximum number of characters will be applied in whatever division. However, we recommend metadata developers to keep the code as short as possible.

The process for structuring the coding scheme is the following:

1) Start with three lower case letters best identifying the subject. If the main subject is described as an acronym use upper case letters. If possible, choose those three letters from the list of acronyms and abbreviation provided as an annex to this document. Try to use only three letters. However, some flexibility is allowed, especially for composed indicators.
2) Refine the subject using at least one character to identify derivations or restrictions when measuring the indicator. The goal here is to add items that could give more detailed information about the variable that is being measured;
3) Conclude the coding scheme by proving some information on the levels of measurement. For this purpose, we invite metadata developers of each TPG to use the list of abbreviations that relate to the nature of data, preferably the list of abbreviations for measurement scales.

Guiding principles

The procedure is not constrained to a limit of characters. This means that those who have the responsibility to deliver data will have flexibility enough to code indicators. However, it is important to stick to the above-mentioned structure and, more importantly, use the list provided as an annex to these guidelines. Obviously, this list is not exhaustive but rather based on the current state of the database. Moreover, we should be well aware that in some cases adaptations will be necessary, especially to obtain more degree of freedom when facing rather complex, but similar, indicators.

The list of acronyms and abbreviations we recommend in Annex 1 for potential subjects is rather straightforward, and usually tries to capture the first three consonants of a word, or even other letters representing at best the core subject of a variable.

The second part of the list refers to widely used abbreviations that impose restrictions and/or use derivations. A typical example in demographic data is the distinction of gender by "f" for female and "m" for male. Other abbreviations refer to commonly used terms to describe a variable such as *index*, *rate*, or *change*. Sometimes it directly relates to the nature of data (particularly terms such as *volume*, *absolute*, *relative*), though it is only loosely related to the measure itself.

To what regards the third part of the coding scheme, we believe that it should contain an unambiguous description of the level of measurement (expression coined by Stevens, 1946) since data needs to be described as accurately as possible. Levels of measurement are particularly important because it allows the user to draw a direct relationship between data classification and the cartographic representation of data, as well as to capture its use within mathematical operations. Together with metadata, it is an important feature that could lead in the future to an automatic, clever data management system. In literature, four levels of measurement are commonly distinguished following the proposals made by Stevens (1946) on the theory of scales and measurements. In ascending order of precision these are: nominal, ordinal, interval, and ratio data. Statistical analysis and most of the spatial analysis handbooks refer to those four scales (Chrisman, 2002; Haining, 2003; Slocum et al., 2005, Haining, 2010).

At **nominal** level of measurement, the numbers are used to classify data. On the contrary, the **ordinal** scale illustrates some ordered or ranked relationship between categories. Despite the fact that both levels correspond to categorical data the major difference between them lies on the hierarchical, non-sequential relationship. For instance, land use types are classified as nominal data while income categories are explicitly ordinal.

The **interval** and **ratio** scales are quantitative data (i.e. numerical measures). The interval level has equal units of measurement, thus making it possible to interpret not only the order of scale but also the distance between them. Nevertheless, the zero point of an interval scale is arbitrary and is not a true zero. The ratio scale of measurement has fixed origin or zero point. This is the most advanced scale. Most of common statistical methods of analysis however require only interval level of measurement.

Geographers, particularly those involved in GIS and cartography argue that scales should be refined when dealing with geographic data (Forrest, 1999; Chrisman, 2002; Slocum et al., 2005). We choose to follow the naming and definition of measurement scales proposed by Forrest (1999). In addition, Chrisman (2002, 31-33) and Slocum et al. (2005, 60-61) also discuss the rationale for these subdivisions.

Against this background, the above-mentioned authors state that nominal data are divided up into four categories: *unique* (i.e. no duplicated value), *dichotomous* (i.e. binary data), *categorical* and *categories with graded membership*; ordinal data are subdivided into *complete* and *classed ordinal* depending on whether all values are unique or not and, finally, ratio data are subdivided into six subtypes. The first two are often referred to as *volumes* of *absolute numbers* in cartographic literature and mapped with proportional symbols: *extensive ratio* (where additive properties apply) and *count* (number of something). Then follows those ratios that are mapped using choropleth maps and often referred to as *relative* data: *derived ratio* (resulting from the division of any quantity by another), *density ratio* (the denominator is a geographical surface), and *constrained ratio* (values bounded between 0 and 1, representing proportions or probabilities). The last subtype, less in use within the EU territorial development research field, is the *cyclic* ratio (e.g. angles).

The corresponding list acronyms and abbreviations are displayed in both Annex 1 and 2 and a short description with examples for each level of measurement is provided in Annex 3. Whenever possible, TPGs should use the most descriptive abbreviation.

3. Illustrative examples

We now apply the coding scheme on several indicators to demonstrate the usefulness of this approach. For this purpose, we have randomly selected a set of indicators that show distinct features when are subject of conversion by the coding scheme. Table 1.1 illustrates the result of this method that wishes to capture, as much as possible, the content of each indicator. First we add the subject(s) attributed to each indicator that is being measured. Secondly, we determine the most characteristic derivations or restrictions induced by those indicators and, lastly, we determine the level of measurement as accurately as possible. In order to differentiate each category we suggest metadata developers to use underscore or low line symbol. If needed, TPGs should use dots to separate items that belong to the same category or division[1]. For age intervals, we recommend dashes.

(a)

Subject(s)								Derivations / Restrictions							Level of measurement					
m	**i**	**g**	**.**	**p**	**o**	**p**	**_**	**c**	**h**	**.**	**t**				**_**	**r**	**t**	**c**			

(b)

[1] The option for using dots as a separator is to reduce potential misunderstandings when importing variables to files with the CSV extension.

Subject(s)								Derivations / Restrictions								Level of measurement						
a	**c**	**c**	**.**	**a**	**i**	**r**	**_**	**a**	**b**	**s**					**_**	**r**	**t**	**e**				

(c)

Subject(s)								Derivations / Restrictions								Level of measurement						
e	**d**	**u**	**.**	**s**	**c**	**d**	**_**	**t**							**_**	**r**	**t**	**c**				

(d)

Subject(s)								Derivations / Restrictions								Level of measurement						
p	**o**	**p**					**_**	**2**	**0**	**-**	**3**	**9**	**.**	**t**	**_**	**r**	**t**	**c**				

(e)

Subject(s)								Derivations / Restrictions								Level of measurement						
C	**O**	**2**	**.**	**r**	**o**	**d**	**_**	**v**	**o**	**l**					**_**	**r**	**t**	**e**				

Table 1.1 Some illustrative examples of ESPON indicators that have applied the coding scheme, where (a) reflects "Migratory population change", (b) "Potential accessibility by air [absolute level]", (c) "Number of persons with secondary education degree", (d) "Population aged 20-39 years", and (e) "CO2 emissions by road traffic".

4. Conclusions

The result is rather helpful and easy to comprehend. However, these guidelines should be seen as an attempt to harmonise coding schemes. Most likely, additional improvements will be needed to further increase the quality of this proposal. At this point, is not possible to foresee or describe many of the indicators that will come out from the current and future applied research projects. This will require the involvement of the ESPON research community through a continuous, dynamic process.

References

Chrisman, N. (2002) Exploring Geographic Information Systems. New York: Wiley.

Forrest, D. (1999) Geographic information. Its nature, classification, and cartographic representation. *Cartographica*, 36(2): 31-53.

Haining, R. (2003) Spatial Data Analysis. Theory and Practice. Cambridge: Cambridge University Press.

Haining, R. (2010) The Nature of Georeferenced Data. In Fisher M. M. & A. Getis (eds.) *Handbook of Applied Spatial Analysis. Software Tools, Methods and Applications*. Springer-Verlag: Berlin, pp. 197-217.

Slocum et al. (2005) Thematic Cartography and Geographic Visualisation. New Jersey: Prentice Hall Series in Geographic Information Science.

Stevens, s. (1946) On the theory of scales of measurement. *Science*, 103(2684): 677-680.

Annex 1: Non-exhaustive list of acronyms and abbreviations

The list of acronyms and abbreviations compiles some of the subject-matter terms that have been identified in ESPON indicators. As stated elsewhere, the goal is to have some consistency on metadata structures and therefore facilitate information-sharing. For this purpose, we recommend TPGs to apply this list whenever feasible on the coding scheme in order to facilitate communication and understanding. The level of coverage is limited to data delivered up to June 2010. However, regular updates will inform TPGs about codes that have been used to harmonise terminology and coding schemes.

A

Access(ibility)	acc
Administration	adm
Airports	air
Arable (land)	arb

B

Birth	bth
Border(s)	brd
Broadband access (to Internet)	bro

C

Carbon dioxide	CO2
Cinema(s)	cnm
City	cty
Clean	cln
Coastal	cst
Community	com
Component(s)	cmp
Construction	con
Core city	cc
Cultural	clt

D

Death	dth
Density	den
Dependency	dep
Desiderability	dsb
Diversity	div
Doctor(s)	drs

E

Easy-to-find-a-job	job
Easy-to-find-good-housing	hou
Economic old age dependency	eod
Economic	eco
Education	edu
Efficiency	eff
Electricity (Gas and Water Supply)	gas
Emigration	emi
Employment	emp
Energy	ene
Environment(al)	env
European Union	EU
Evaporation	eva

F

Facilities	fcl
Farm(ers)	frm
Fertility	fer
Firm(s)	fir
Force	frc
Foreigner(s)	fgn
Frost	frt
Freight	frg

G

Greenspace	grs
Gross Domestic Product	GDP
Growth	gwh

H

Heritage	hrt
High and medium tech manufacturing activities	htc
Holdings	hld
Hospitals	hst
Higher Education Institutions	hei
Hotels and Restaurants	hot
Household(s)	hsd

I

Identity	idn
Immigration	imi
Industry and or/industrial	ind
Internet	itn
Island	isl
Infrastructure	inf

L

Labour	lab
Landscape	lds
Large Urban Zone	LUZ
Life Expectancy	lif
Life-long learning	lll

M

Manufacturing	man
Market	mrk
Metropolitan	met
Migration	mig
Mining (and Quarrying)	min
Mortality	mor
Mortality	mrt
Mountain(ous)	mtn

N

Natural grassland, heathland	grl
Natural	nat
Neighbourhood	nbh
Net extra-Europe migration	nee
Net inter-country migration	nie
Net internal Migration	nim
Noise	nse
Maritime	mar
Nomenclature des unités territoriales statistiques	NUTS

O

Old Age Dependency	oad

P

Pastures and mosaic farmland	ptr
Pollution	pol
Population	pop
Potential (impact)	pim
Precipitation	prc
Primary	prm
Public	pub
Purchasing power parity	ppp
Private	pvt

Purchasing power standards	pps

Q

Quality	qua

R

Rail	ral
Rainfall	rfl
Real State (renting and business activities)	res
Replacement	rpl
Residential (areas)	rsd
Resources	res
Retail trade (and wholesale)	ret
Road	rod
Rural	rur

S

Safe(ty)	sfe
Schools	sch
Service(s)	ser
Secondary	sec
(Post) Secondary	pos
(Upper) Secondary	ups
Student(s)	stu
Sensitivity	sns
Sparsely	spr
Sport	spt
Summer	smm

T

Temperature	tmp
Territorial (Impact)	tim
Tertiary	trt
Tourism	tou
Transport (and Storage)	tra

U

Unemployment	ump
Urban	urb

V

Vegetation	veg
Very old age dependency	vod
Vulnerability	vln

W

Website	web
Working age population	wap

Y

Young age dependency	yng

Annex 2: Non-exhaustive list of derivations and/or restrictions

Term	*Recommended code*	*Comments*
Absolute	abs	The absolute value of an indicator implies a real number. This means that its value is closely related to the notions of magnitude, distance or norm and therefore is always greater than or equal to 0. An illustrative example for absolute values is the number of births.
Average	av	The value represents a central tendency. An easy way to obtain this value is to calculate the sum and then divide by the number of occurrences. There are many different descriptive statistics that can be chosen as measurement of the central tendency, including arithmetic mean, median distribution, and mode.
Change	ch	This value explains a temporal change over a period of time. In a way, it can be interpreted as a pattern that allows us to understand a particular spatial phenomenon. For example, population growth is the change of population over time, and can be quantified as the change in the number of individuals.
Index	ix	One of the most important features in the construction of an index number are its coverage, base period, weighting system, and method of averaging statistical results.
Share	sh	This value corresponds to a distribution, a portion that is calculated as a percentage of the derived total quantity. For instance, the prevalence of unemployment is usually measured using the 'unemployment rate' which is defined as the percentage of those in the labour force who are unemployed.
Weight	wgt	This concept is relevant for index numbers and transformations, such as 'GDP as constant prices'.
Female	f	The code for this term is applied if data needs to be categorised by gender.
Male	m	The code for this term is applied if data needs to be categorised by gender.
Total	t	This code corresponds to the total membership of a defined class of people, events or objects.
20-39	20-39	Age class between 20 and 39 years old of a certain phenomenon. For instance, 'sex ratio at age 20-39'. The same applies to other classes.
65+	65+	In this case, the membership is just focused on the age class over 65 years old.

Non-exhaustive list of derivations and/or restrictions

Annex 3: Levels and scales of measurement

	Stevens' scales (1946)	*Forrest's extended levels (1999)*	*Required information and nature of data*	*Illustrative examples*	*Recommended abbreviation*
Qualitative	**Nominal**				**nom**
		Unique	All different (no duplication)	Country names, NUTS identificators	nou
		Dichotomous	Membership (presence or absence)	Coastal areas, regions benefitting from EU funds	nod
		Categorical	Categories	Land use type, main religion	noc
		Graded membership	Categories plus degree of membership	Soil type with percentage conformance	nog
Ranking (no quantity)	**Ordinal**				**ord**
		Complete ordinal	Unique ordering	Any ranking of regions or cities (without ex-aequo)	oru
		Classed ordinal	Categories plus ordering	Densely/intermediate/weakly populated regions	orc
Quantitative	**Interval**		Measure plus arbitrary zero point	Temperatures in degrees	**int**
	Ratio				**rto**
		Extensive ratio	Measure (additive rules apply)	GDP, CO2 emissions, transported tons	rte
		Count	Measure (with unit =1)	Population, number of households, firms, births	rtc
		Derived ratio	Measure (quantity divided by)	GDP per inhabitant, labour productivity, cars per household	rtd
		Density ratio	Measure (quantity divided by area)	Population density	rde
		Cyclic ratio	Measure plus length of cycle	Angles, slope orientation	rty
		Constrained ratio	Probability or proportion, range [0,1]	Unemployment rate, share of young people	rtp

Source: Levels and scales of measurement proposed by Stevens (1946) and Forrest (1999).