

BNU 2018: Population and Housing Censuses using administrative sources. Part I-III

The past, present and future of population censuses:
Methodology and quality aspects when data sources are
being reused and combined to transform a census system.

21-24 August 2018



Statistisk sentralbyrå
Statistics Norway

anders.holmberg@ssb.no

Contents Part I - Part III

- Census history and background
- Motives for using administrative data
- The Nordic model
- Registers, data infrastructures and methods
- Other census models
- What is quality in a Census ? Assessment methods.

History

- American colonies, Quebec
- 1749 Sweden (Finland)
- Italian and German cities 1770
- Denmark 1787
- US 1790, England 1801
- France (tried) 1800, 1806
- China 1953



Role of the census

Benchmark in a National statistical system

1. For baseline estimates and small groups/domains
2. Assisting data collection; Organisation, Frames, Techniques, Routines.

Planning on local and national level

Use Internationally and for the National Accounts

Direct Topics in a Population Census

Geographic: place enumerated and/or place of usual residence;

Family circumstances: relation to head of household or family;

Demographic: sex, age, marital status, children ever born, birthplace;

Economic: type of activity, occupation, industry, employer-employee status;

Social and political: citizenship, language, ethnic or religious affiliation

Educational: literacy or level of education, school attendance.

Examples of Statistics from Population Census

Total population (Counts)

population of towns and local areas

urban-rural distribution

household or family composition

Plus Additional topics

internal migration, income, labour-force participation, duration of marriage, etc...

Statistics from the Housing part of the Census

Total population! (Counts of Dwellings)

Housing arrangements (used as residence, vacant or secondary use)

Occupancy status

Type of Ownership

Useful floor area

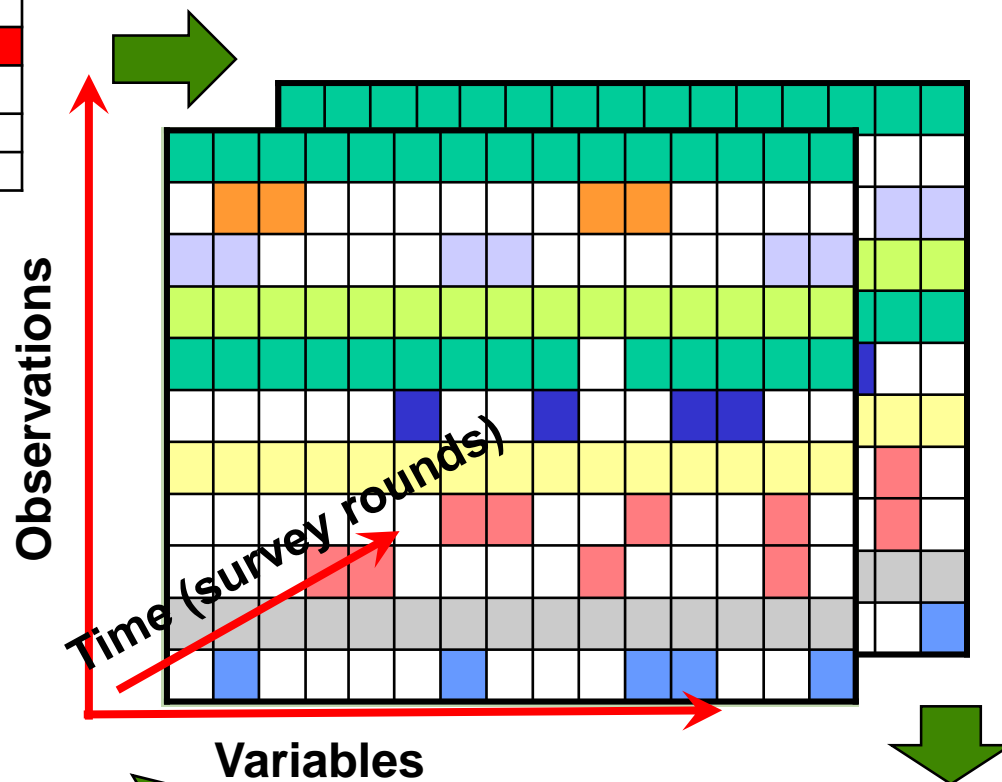
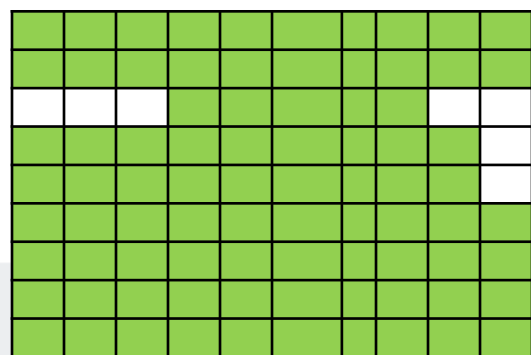
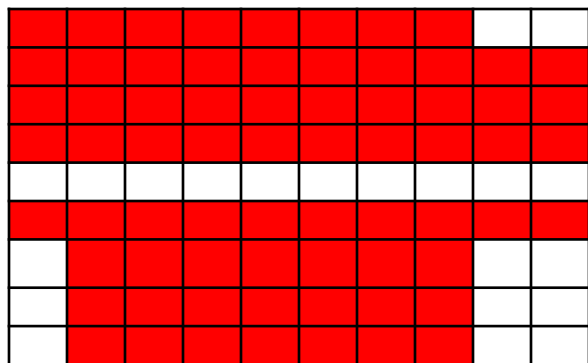
Standards of dwelling (Water, toilets, heating)

Dwellings by Type of Building, Size

Motives for a register-based census

- Cost-efficiency
- Reduced response burden on the population
- Provides the means for recurrent register-based production of (some) social statistics
 - Improves accuracy and timeliness
- Provides the means for new, better statistics (e.g. on living conditions)
 - Improves relevance

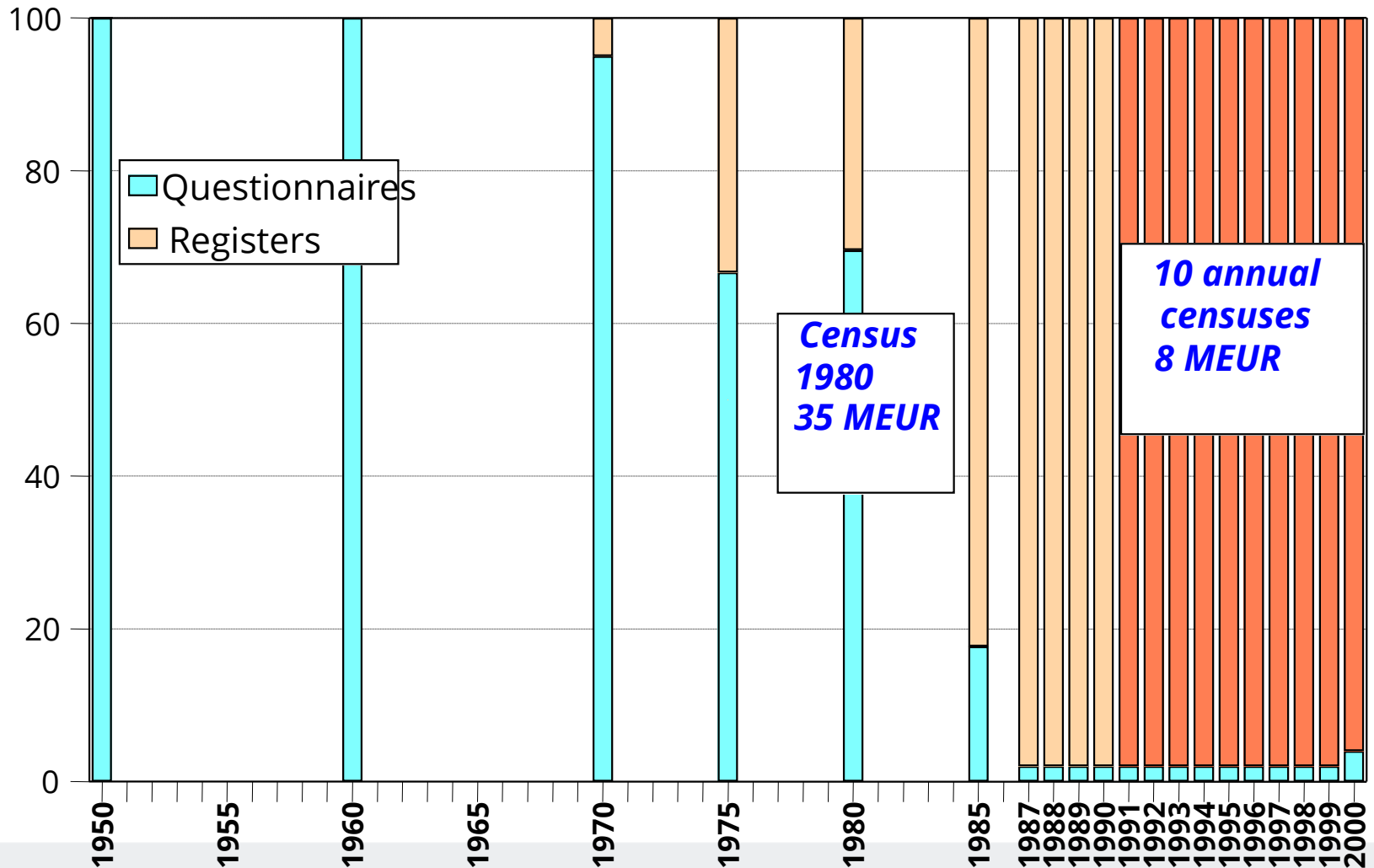
NSO-Target!? To create (new) knowledge from recycling existing data



Processes, Statistics och Analyses
**Decision, long term use in EU,
UN mm**
anders.holmberg@ssb.no

Cost of Censuses in Finland

Census data 1950-2000 according to the data source



More countries move from a traditional census

All countries try to modernize collection

UNECE keeps a good overview

<https://statswiki.unece.org/display/censuses/2010+Population+Census+Round>

Slovenia, Austria, The Netherlands (Reg)

Belgium, Iceland (reg plus a few surveys)

Estonia,, Germany, Latvia, Lithuania, Switzerland (Mix
Reg + surveys)

France (Rolling census system)

US, Canada, UK, New Zealand, Australia, Ireland, Italy
(Trad.)

Ukraine (no census)

Table. The year of establishing registers/introducing registers in census statistics by type of register and country

Type of register	Denmark		Finland		Norway		Sweden	
	<i>Established</i>	<i>First used in census</i>	<i>Established</i>	<i>First used in census</i>	<i>Established</i>	<i>First used in census</i>	<i>Established</i>	<i>First used in Census</i>
Central Population Register	1968	1981	1969	1970	1964	1970	1967	1975
Business Register	1975	1981	1975	1980	1965	1980	1963	1975
Dwellings	1977	1981	1980	1985	2001	2011	2008?	2011?
Housing conditions	1977	1981	1980	1985	2001	2011	2008?	2011?
Education	1971	1981	1970	1975	1970	1980	1985	1990
Employment	1979	1981	1987	1990	1978	2001	1985	1985
Family	1968	1981	1978	1980	1964	1980	1960	1975
Household ^a	1968	1981	1970	1975	2001	2011	2011?	2011?
Income	1970	1981	1969	1970	1967	1980	1968	1975
Totally register-based census		1981		1990		2011		2011?

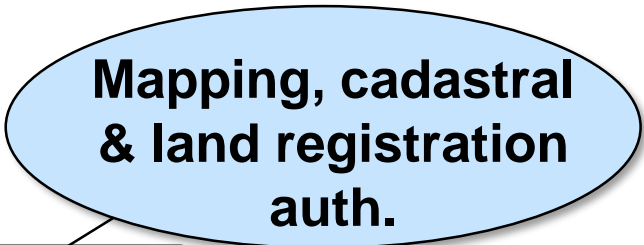
^a Household-dwelling unit, i.e. all the persons living in one dwelling

Swedish Censuses in the 20:th century

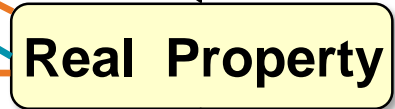
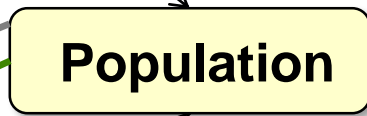
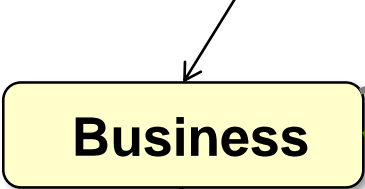
- Modern Population and housing censuses every 5th year between 1960 – 1990. Traditional censuses with increasing use of register data
- 1985 and 1990
 - High degree of computerization (main frame)
 - 1985 Public inquiry on increasing the use of registers.
 - Huge protest campaigns in media
- Cancelled 1995, 2000 (some data to Eurostat)

Input to the Swedish register based census

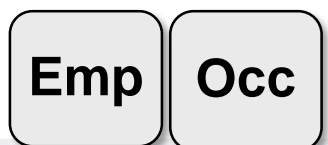
Important external register keepers



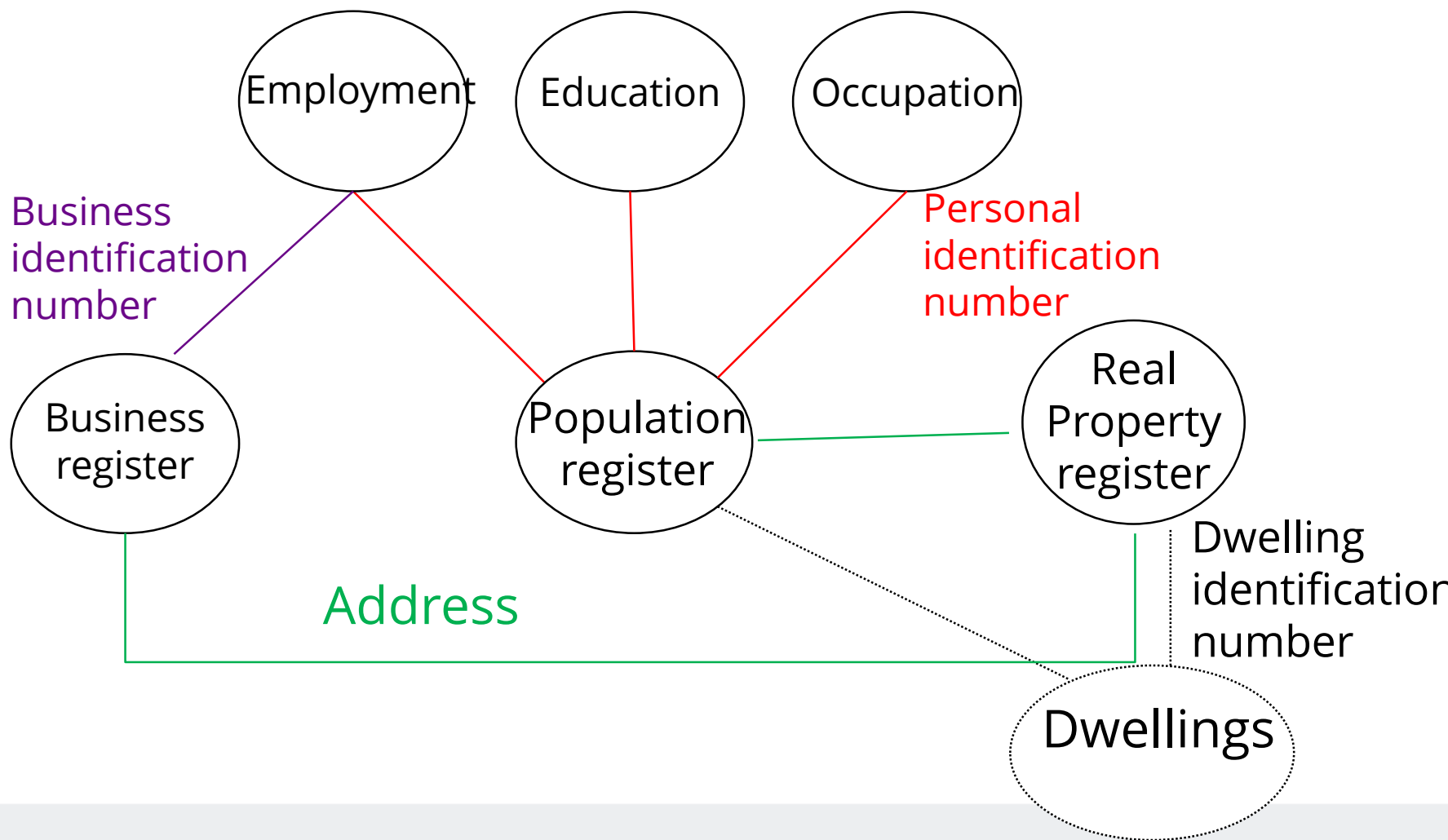
Registers at Statistics Sweden



Subject matter Reg



Links



The weakest link - the Dwelling ID-key

- Two important steps:
 1. Construction of Dwelling register
 - Unique address for real property (building)
 - Unique dwelling number within building
 - Unique dwelling ID-key
 2. Registration of (registered) population on dwelling
- Complex processes and tight schedule

Constructing the Dwelling register

- Complex interaction required between
 - Municipalities
 - Mapping, cadastral & land registration authority
 - Real property owners
- Residents living in multi-dwelling buildings were informed about their dwelling numbers by the property owner

Populations, units and definitions

With a well functioning population register, individual person is a basic unit.

Households and Families are trickier.

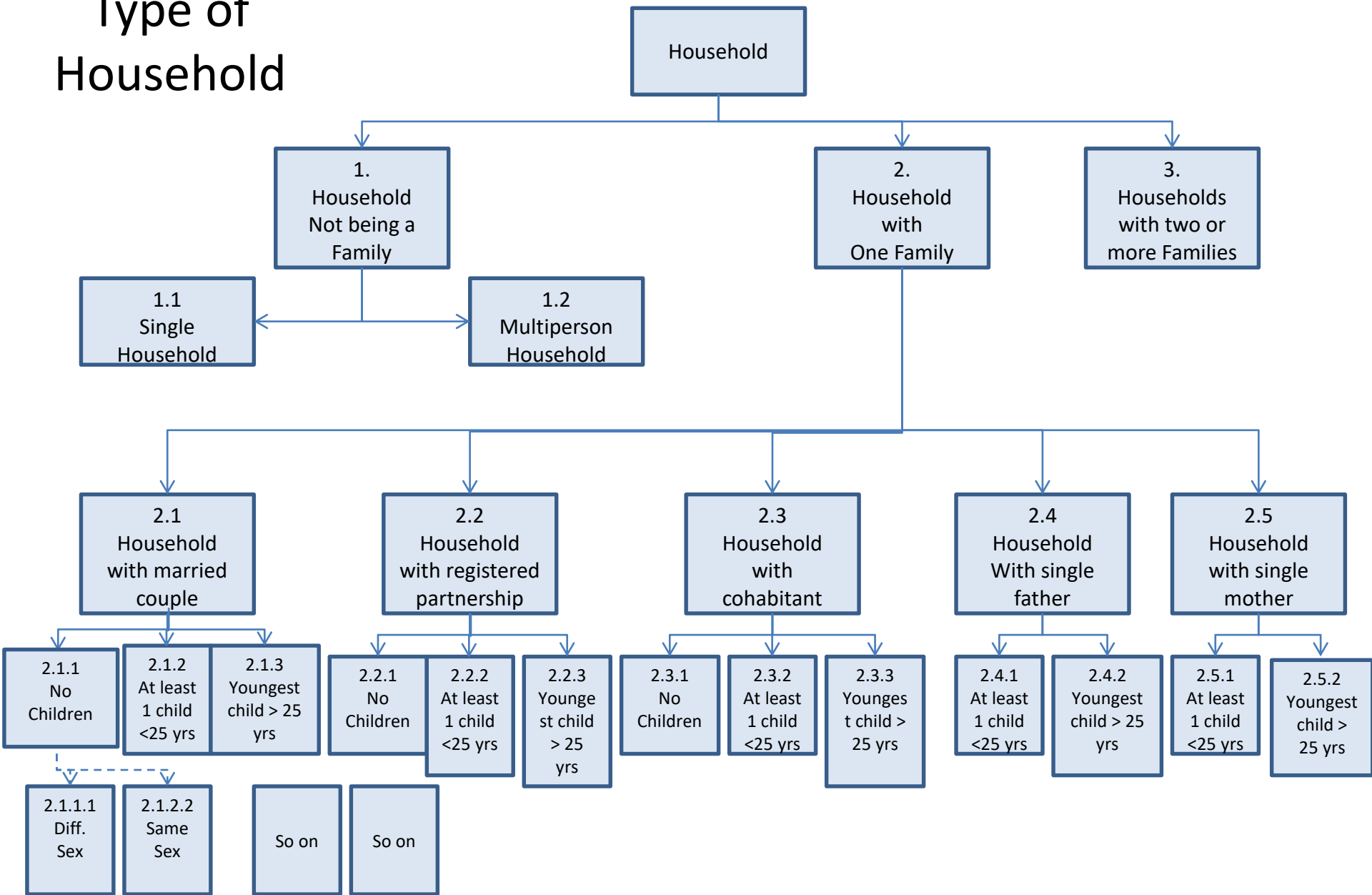
On the housing side Properties is the basic (legally regulated) unit. Buildings and Dwellings are units assigned to them where people live businesses are active etc. Addresses, Geocodes are attributes.

Populations, units and definitions

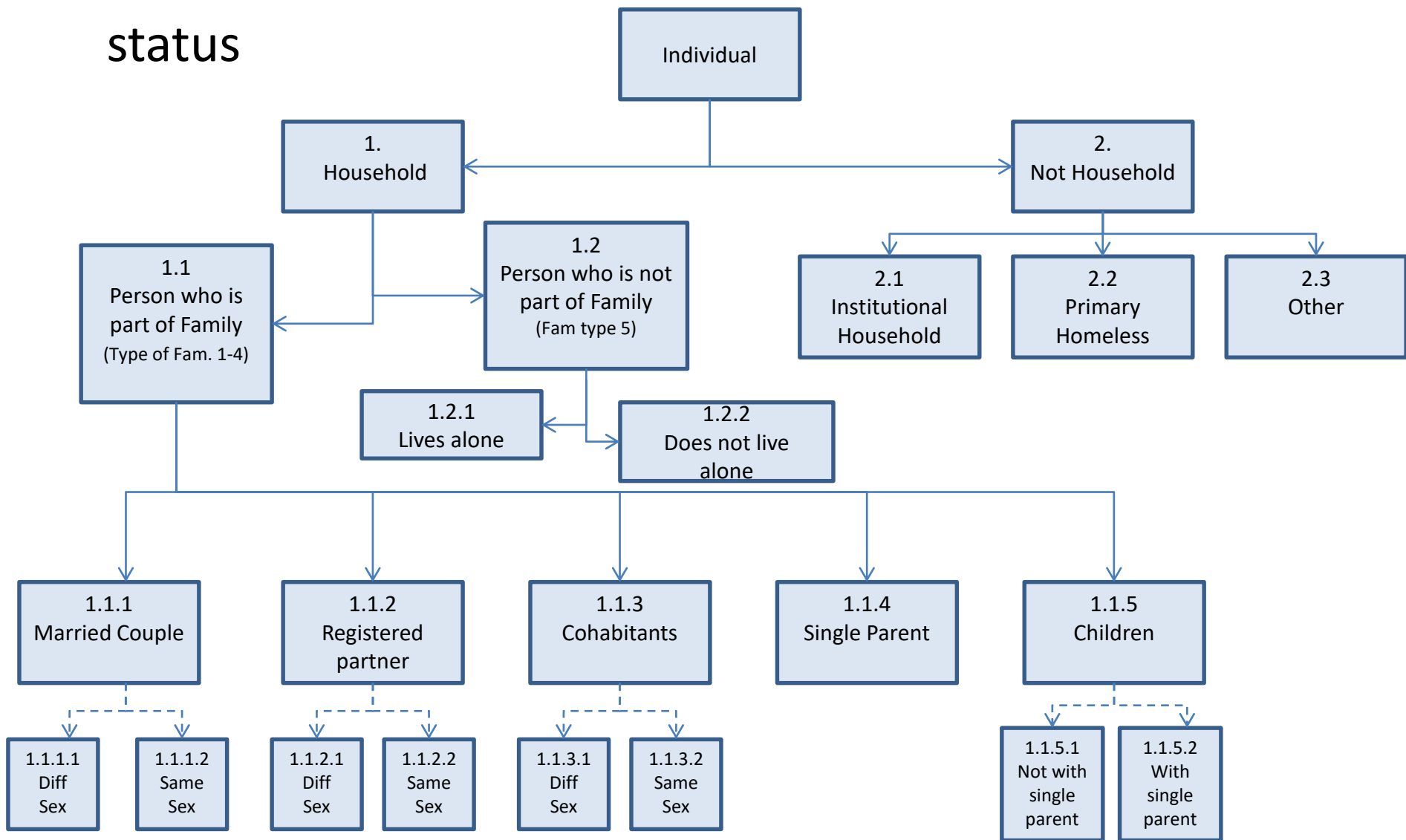
De jure or de facto population?

Institutions, dormitories, Age limits?

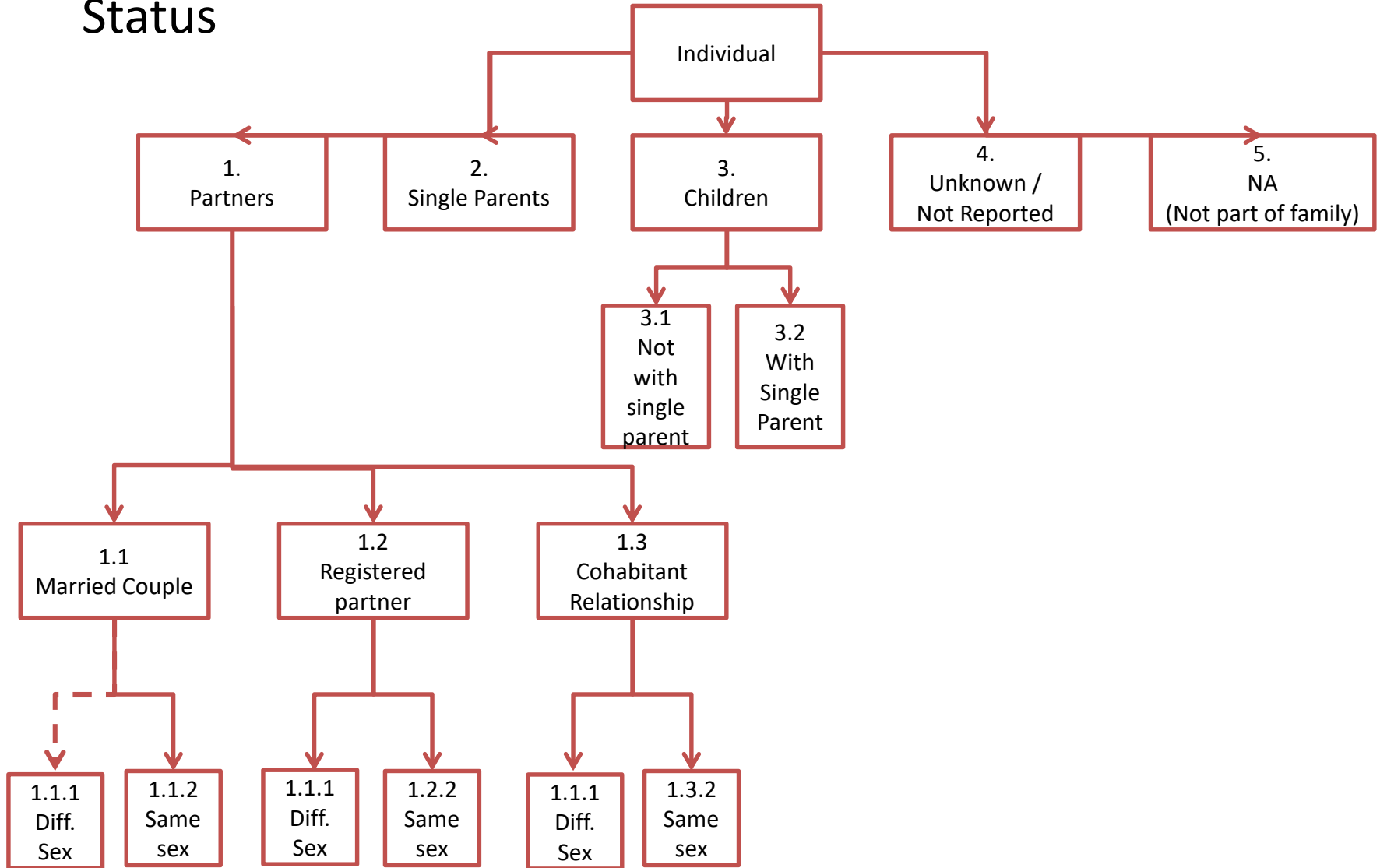
Type of Household



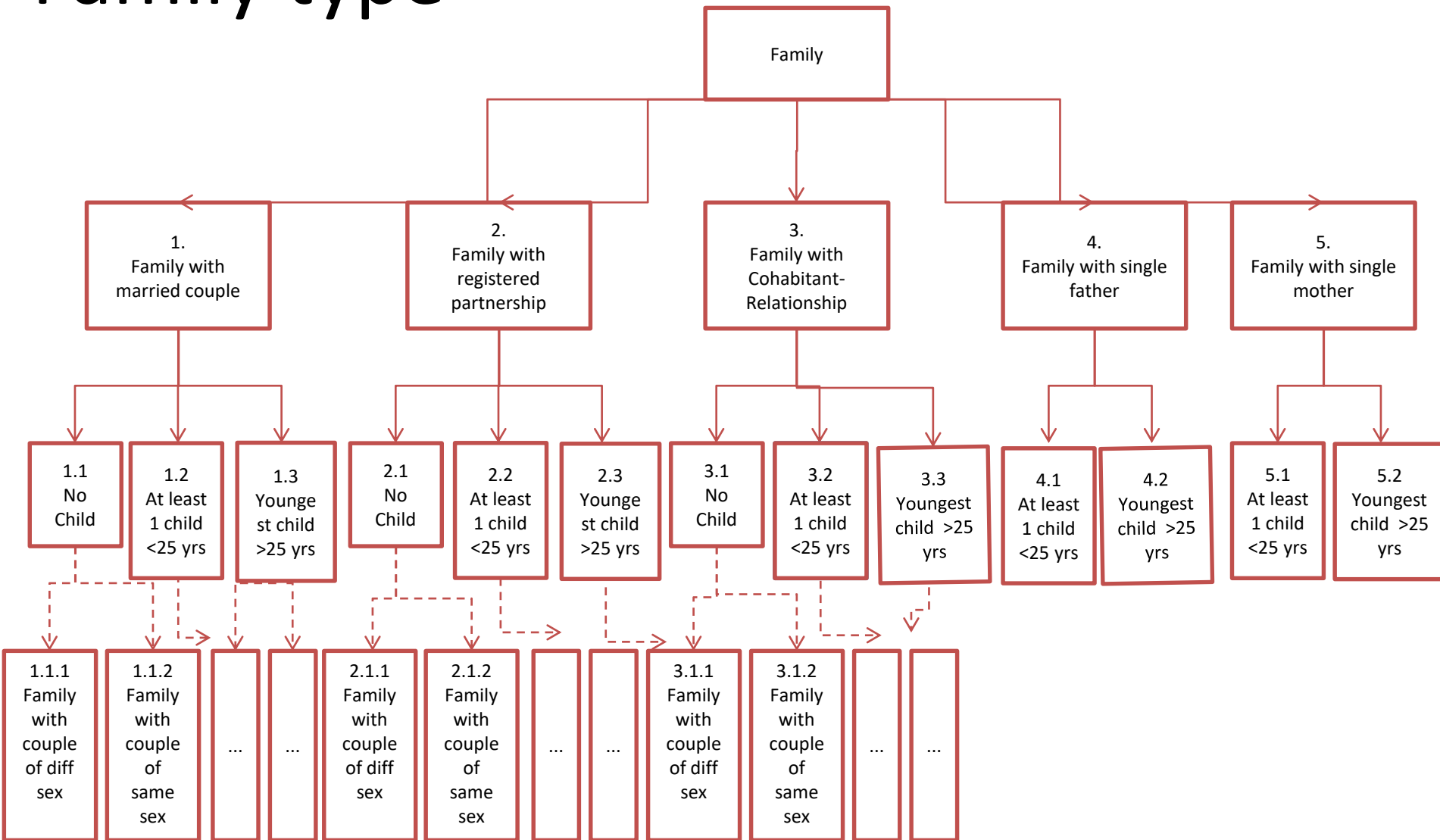
Household status



Family Status



Family type



Notifications to the population register



Definition of the census population

- Registered population (12 month criteria)
- Place of birth – A newborn children born in Sweden, is registered in the parish in which the mother was registered at the time of birth. If the mother was not registered at the time of birth, the child is registered only if the father was registered and guardian. A newborn baby, born abroad, are registered only if the mother was registered at the time of birth.

Families

- *The family nucleus is defined in the narrow sense, that is as two or more persons who belong to the same household and who are related as husband and wife, as partners in a registered partnership, as partners in a consensual union, or as parent and child.*
 - Cohabitation
 - One person is not a family;
 - how many families does a household of a married couple and a lodger consist of. 1+0?
 - The lodgers household status is "none" but is he included in a type of private household – "Married couple households"?

Definitions of cohabitation

- *Two persons are considered to be partners in a 'consensual union' when they*
 - *belong to the same household*
 - *have a marriage-like relationship with each other*
 - *are not married to or in a registered partnership with each other.*
- Sex and age criteria ?

(Assumptions set to minimize misclassification errors)

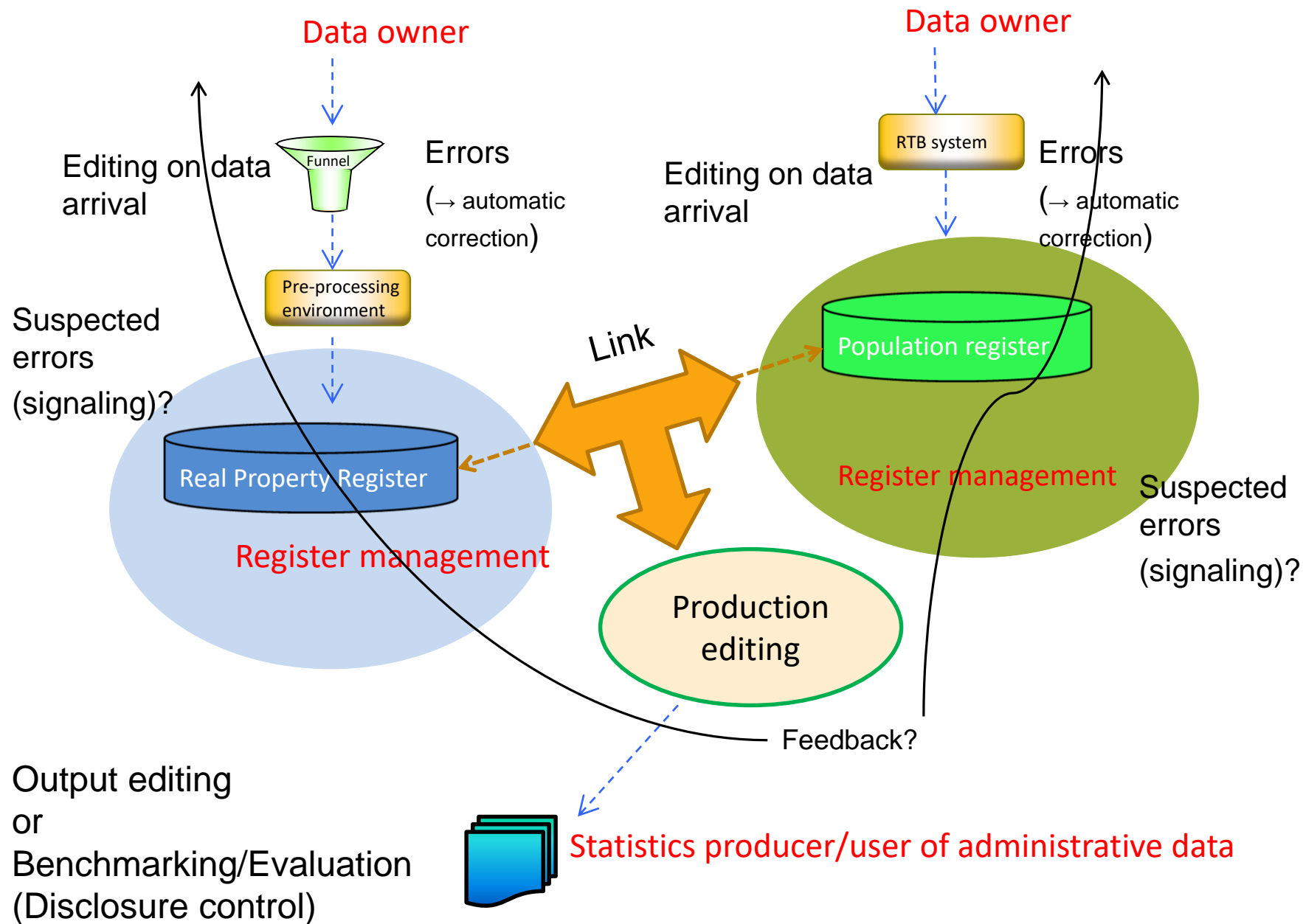
A selection of concerns

- ALL of the traditional error sources in surveys are present!?
- Many different authorities are involved
- In Sweden: 290 different municipalities and about 2 million real estate owners has to be informed and willing to provide information.
- Different actualities in the registers.

A selection of concerns

- SWE: Different input (processes) for one- and two-dwelling buildings and multi-dwelling buildings.
- Both real property owners, residents and municipalities are respondents, what are their incentives?
- Relevance of some of the variables
- Large amounts of data
- The quality of the other registers ? What happens when they are joined?

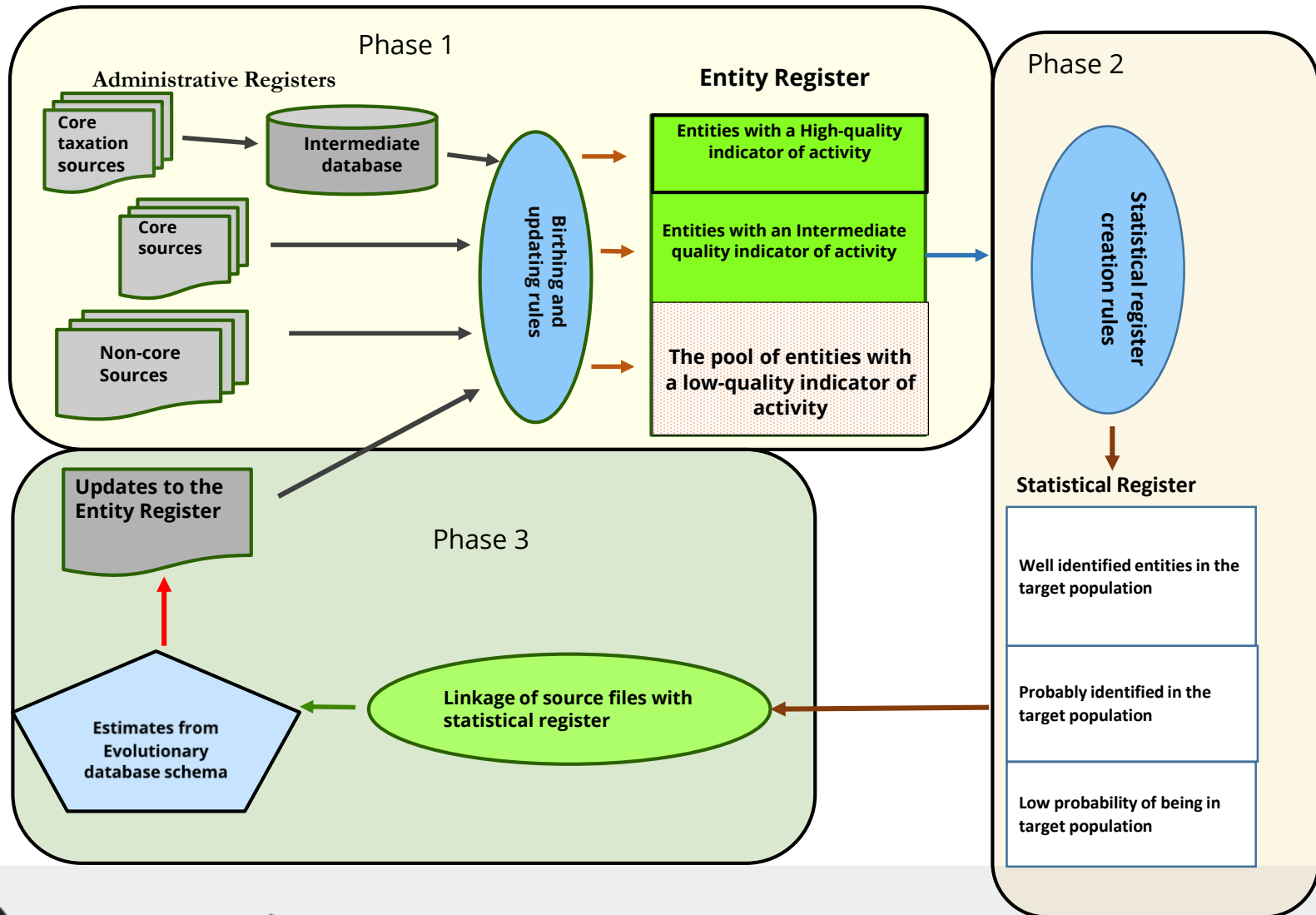
Processes and Editing of the Census



What if the administrative registers do not exist?

- Can the NSO create statistical versions for census or other purposes?
- Hybrid censuses (Combining registers and 'direct' survey type data collections)
- Investments in statistical systems

A evolutionary model for administrative data in statistics

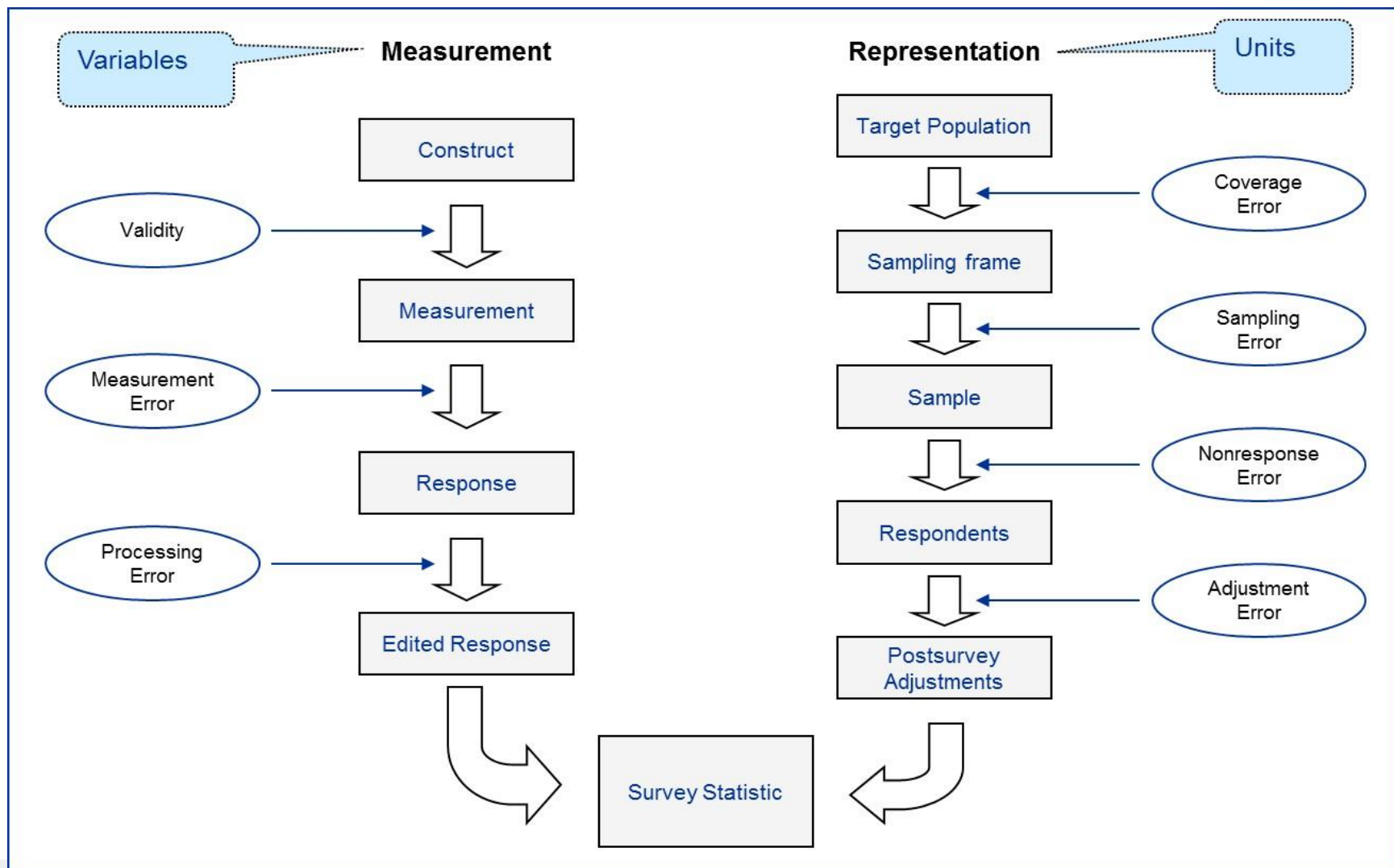


Quality aspects of a census based on administrative data

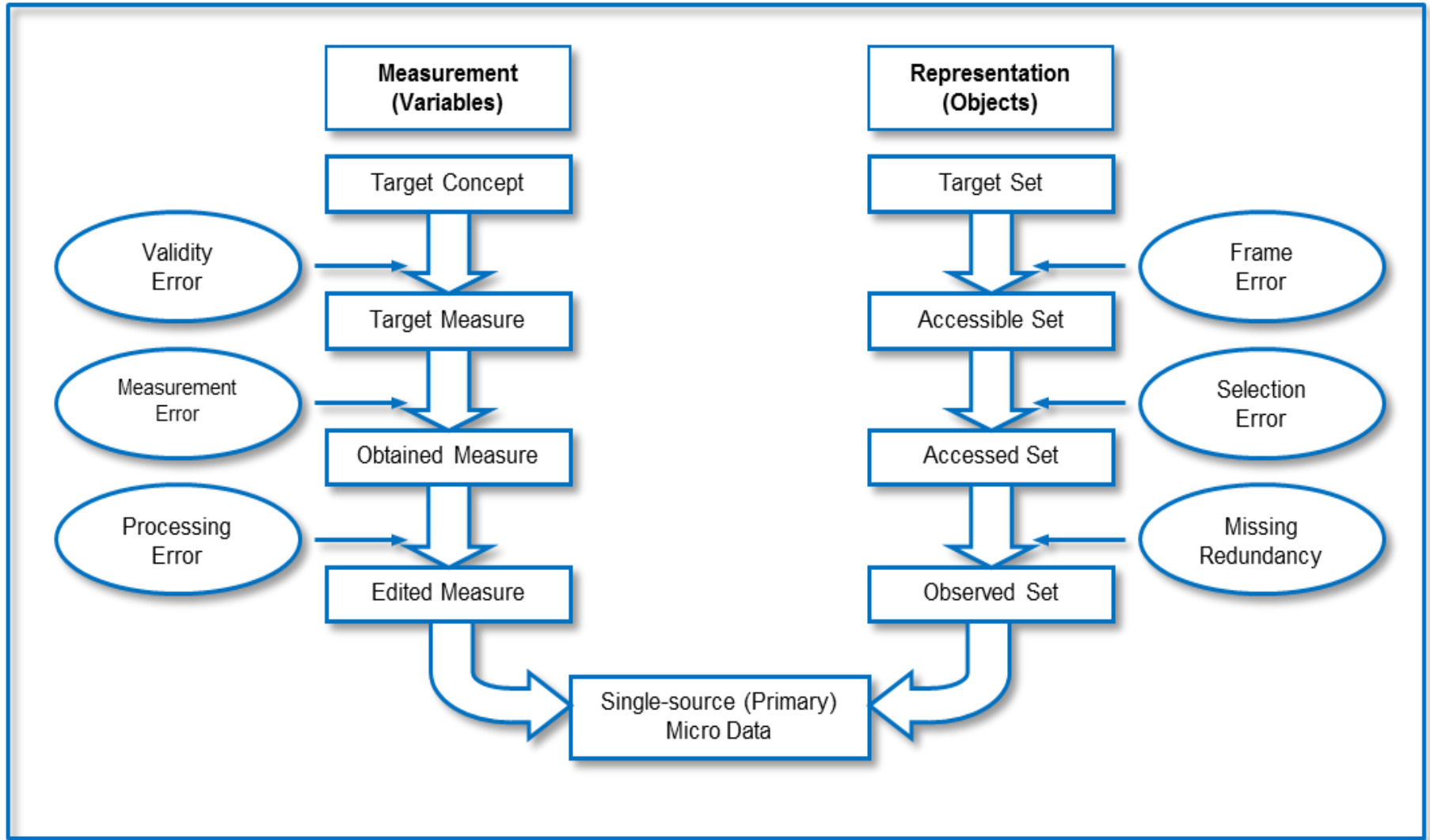
- Relevance, Accuracy, Timeliness, Punctuality, Accessibility, Clarity, Comparability and Coherence
- How to assure quality?
- How to measure quality ?
- How to enable continuous improvement?

Total survey error framework

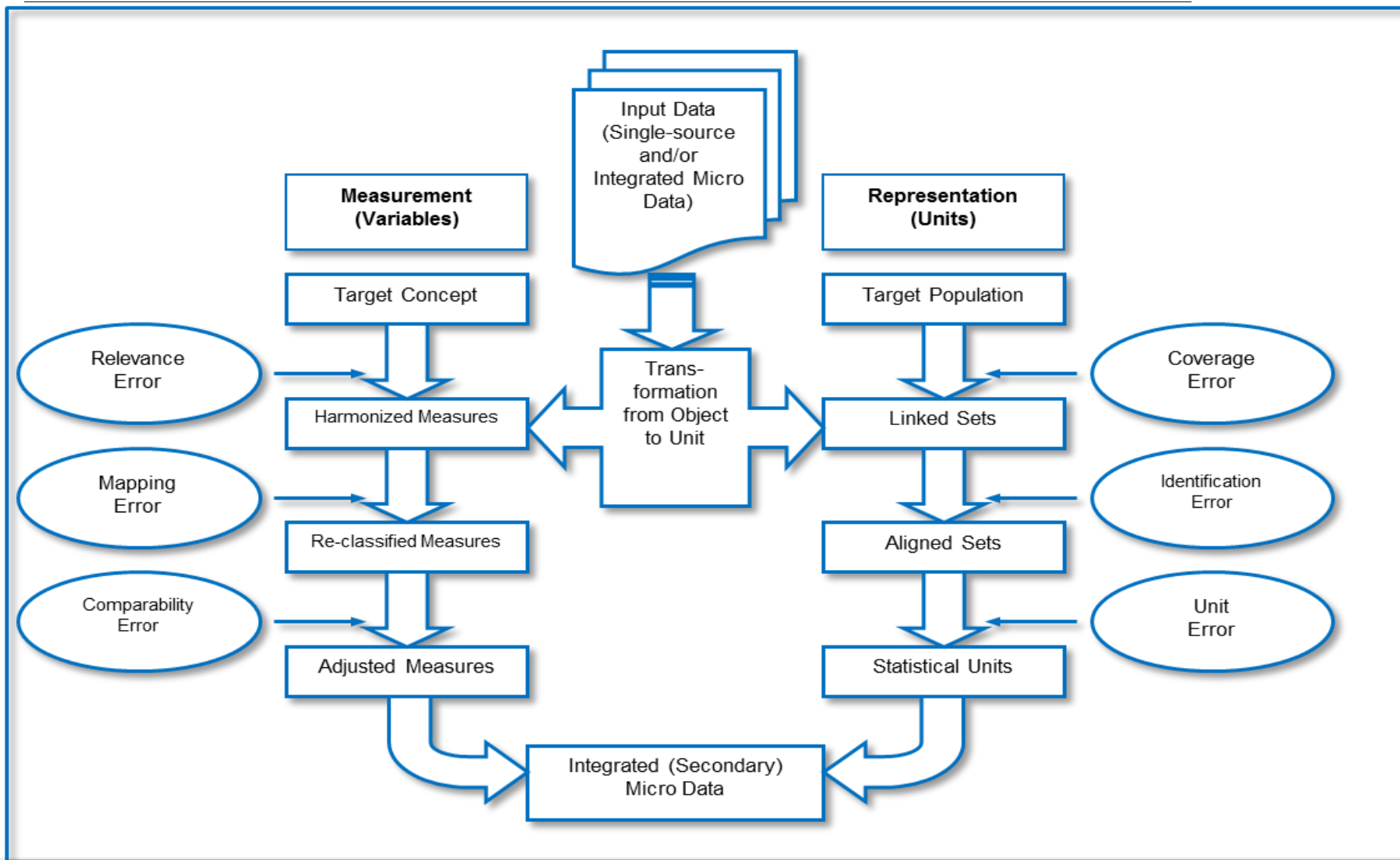
(Groves et al. 2004, Fig. 2.5)



Two-phase life-cycle (I): primary source data (Zhang, 2012)



Two-phase life-cycle (II): secondary integrated data



Quality reporting in Eurocensus: Accuracy

- Accuracy: Composition of data sources

The data sources used for census purposes often have more than one origin. Data from two registers might be merged or data obtained by means of a field enumeration might be linked to data available in a register.

'Record linkage' means the process of merging information from different data sources by comparing the records for the individual statistical units and merging the information for each statistical unit where the unit to which the records refer is the same.

1. How well do the published data reflect the 'true' size of the population?
2. How much are the published data based on real observations?
3. How many of the records in the data source have been used for the census?

Quality reporting: an assessment of coverage and of the imputed and deleted data records.

1. The match between the estimated target population (the best estimate for the 'true' population) and the census population (mostly affected by coverage errors),
2. The effect of record imputations and deletions
3. The effect of any method or technique that adjusts the weights of data records to obtain a better fit of the census population to the target population and/or to increase the plausibility and consistency of the data,
4. The effect of frames that are samples.

Quality reporting Accuracy: an assessment of coverage

Coverage assessment means a study of the difference between a specified target population and its census population.

'Post-enumeration survey' means a survey conducted shortly after the enumeration for coverage and content assessment purposes.

'Under-coverage' means the set of all statistical units that belong to a specified target population, but are not included in the corresponding census population.

'Over-coverage' means the set of all statistical units that are included in a census population used to report on a specified target population without belonging to that target population.

Quality reporting Accuracy:

an assessment of the imputed and deleted data records. Attributing weights $\neq 1$ to data records

Weights different from 1 can be set for the following reasons.

Records are imputed not physically, but only virtually (e.g. a record is weighted by 2 to substitute for an invalid record, which is weighted 0)

NSO can scale some data up or down by systematically increasing or decreasing the weights of parts of the census population

Methods that adjust the weights of data records to improve the plausibility and/or consistency of the data are used.

Quality reporting Accuracy : QUANTITATIVE INFORMATION ON THE TOPICS

Reasons for data records where information on a topics is missing

The data source is a composed data source. It is the result of record linkage from original data sources, some of which might not cover the complete census population.

Unit no-information might have lead to data not being available

There might have been *item no-information* for that topic

Quality assessment

- Aim is to assess quality of both input to and output from a register based sources
- Questions that need to be answered: How to
 - derive data on variables not available in registers?
 - distinguish between no data and missing data in registers?
 - resolve discrepancies between sample data and (derived) register data?

A key issue: reconciliation

- Reconciliation through renewed contact via telephone with respondents will be used, but ...
- ...other methods, e.g., models and rules, will have to be used
- Nonresponse in the evaluation survey causes additional problems

Errors in registers categorical variables

	Register		
Evaluation source	Category 1	Category 2	Sum
Category 1			$N_{1\bullet}$
Category 2			$N_{2\bullet}$
Sum	$N_{\bullet 1}$	$N_{\bullet 2}$	

The proportion of **gross error** in a category is given by $GE_j = (N_{\bullet j} + N_{j\bullet} - 2N_{jj}) / N_{\bullet j}$ where N_{jj} is the number of objects of category j in both the evaluation and register. $N_{\bullet j}$ and $N_{j\bullet}$ indicate summation over rows and columns, respectively.

Evaluating base register quality: an example from 2011 deciding on a derived a variable.

- Register variable based on operational definition of Cohabitant (person living in consensual union)
- Sample data from the LFS

	According to register ,model based on <15 years difference	
According to LFS	Cohabitant	Not Cohabitant
Cohabitant	19.7 %	3.3 %
Not Cohabitant	3.5 %	73.4 %

Net error $3.5 - 3.3 = 0.2 \%$

Gross error $3.3 + 3.5 = 6.8 \%$

Evaluating base register quality: an example from 2011 deciding on a derived a variable.

- The same background data, but only for people
 - Resident in large cities (according to register)
 - 15 – 25 years old

	According to register ,model based on <15 years difference	
According to LFS	Cohabitant	Not Cohabitant
Cohabitant	11.0 %	5.0 %
Not Cohabitant	9.5 %	74.5 %

Net error $9.5 - 5.0 = 4.5 \%$

Gross error $9.5 + 5.0 = 14.5 \%$

Evaluating base register quality: an example from 2011 deciding on a derived a variable.

- The same data, but only for people
 - Resident in large cities (according to register)
 - 26 – 35 years old

	According to register ,model based on <15 years difference	
According to LFS	Cohabitant	Not Cohabitant
Cohabitant	29.8 %	7.1 %
Not Cohabitant	8.0 %	55.1 %

Net error $8.0 - 7.1 = \mathbf{0.9 \%}$

Gross error $8.0 + 7.1 = \mathbf{15.1 \%}$

Base register quality: An operational definition of cohabitants

Age	Column 1	2	3	4
15-25	11 %	5 %	9,5 %	74,5 %
26-35	29,8 %	7,1 %	8 %	55,1 %
36-45	21 %	4 %	3,5 %	71,5 %
46-55	14,3 %	3,9 %	2,8 %	79 %
56-65	9,8 %	2,2 %	2,7 %	85,3 %
66-74	7,9 %	1,8 %	2,1 %	88,2 %
75+	5,6 %	0	2,2 %	92,2 %
Total	17,7 %	4,3 %	5 %	73 %

Larger cities Register definition < 15 years age difference

Column 1: Cohabitant in both sources, 2: Cohabitant in LFS Not according to Base register, 3: Not Cohabitant in LFS Cohabitant in Base register, 4: Not Cohabitant in both sources.

Statistics' Quality and Evaluations

- Missing values: their effects and how to treat it.
 - Imputation techniques or weighting techniques
- Quality Evaluation
 - EU quality requirements on the Census data
 - Data Quality (Micro and/or macro) ?
 - Production of recurrent social statistics with the register system
 - An integrated register and sample survey QC-scheme?

Swedish Census

Household statistics - an evaluation study

- To evaluate the quality of
 - the Dwelling register
 - the registration of people on dwelling
 - their combined effect on household statistics
- If possible, provide information that can be used to improve the quality of census statistics
- Sample survey in order to measure to what extent answers were consistent with the register information at the reference date.

Some errors: causes & consequences

Property register/ Dwelling register	Population register	Error type
Address missing	No dwelling key	Missing value
All dwellings missing/All except one dwelling missing	No dwelling key Wrong type of housing code, no questionnaire sent out	Missing value Errors giving large HH-sizes
A few dwellings missing on an address	Various cases of mismatches between registers	Missing value & HH-size errors
Correct information	Classical data collection problems (measurement/response process, data processing)	In most cases, Missing values

The setup of the evaluation study

- Select sample of individuals
- Collect vital household data for sampled individuals
- Compare sample and register data at the individual level and try to resolve true circumstances
- Estimation of parameters of interest

Questionnaire and data collection

- Limited amount of questions regarding
 - Address and no. of dwellings at the address
 - The dwelling in which respondent lives
 - Other people living in the same dwelling
- Questionnaire distributed via mail
- Data capture via web or self-administered questionnaire
- Follow-up of non-respondents via mail and telephone

Issues in design of an evaluation

- Collecting data on household status with a self administered questionnaire (Web/paper) is not straightforward!
- Nonresponse in the evaluation survey is problematic!
 - Many possible causes, which call for different actions
- Reconciling observed discrepancies requires large resources
 - The amount of possible combinations in error is large and requires manual handling

Sampling design

- Stratified SRS, $n=15,000$, 108 strata, defined by register-data on
 - Existence of Dwelling ID-key
 - Residence Municipality (grouped)
 - Age class
 - Type of dwelling
 - Number of families in dwelling

Sample size allocation

- Oversampling of persons from strata for which dwelling ID-key is missing (1/3), as they
 - will cause errors in register-based production
 - are more likely to be nonrespondents in the evaluation study
- Proportional allocation, with upper and lower boundaries, for remaining strata

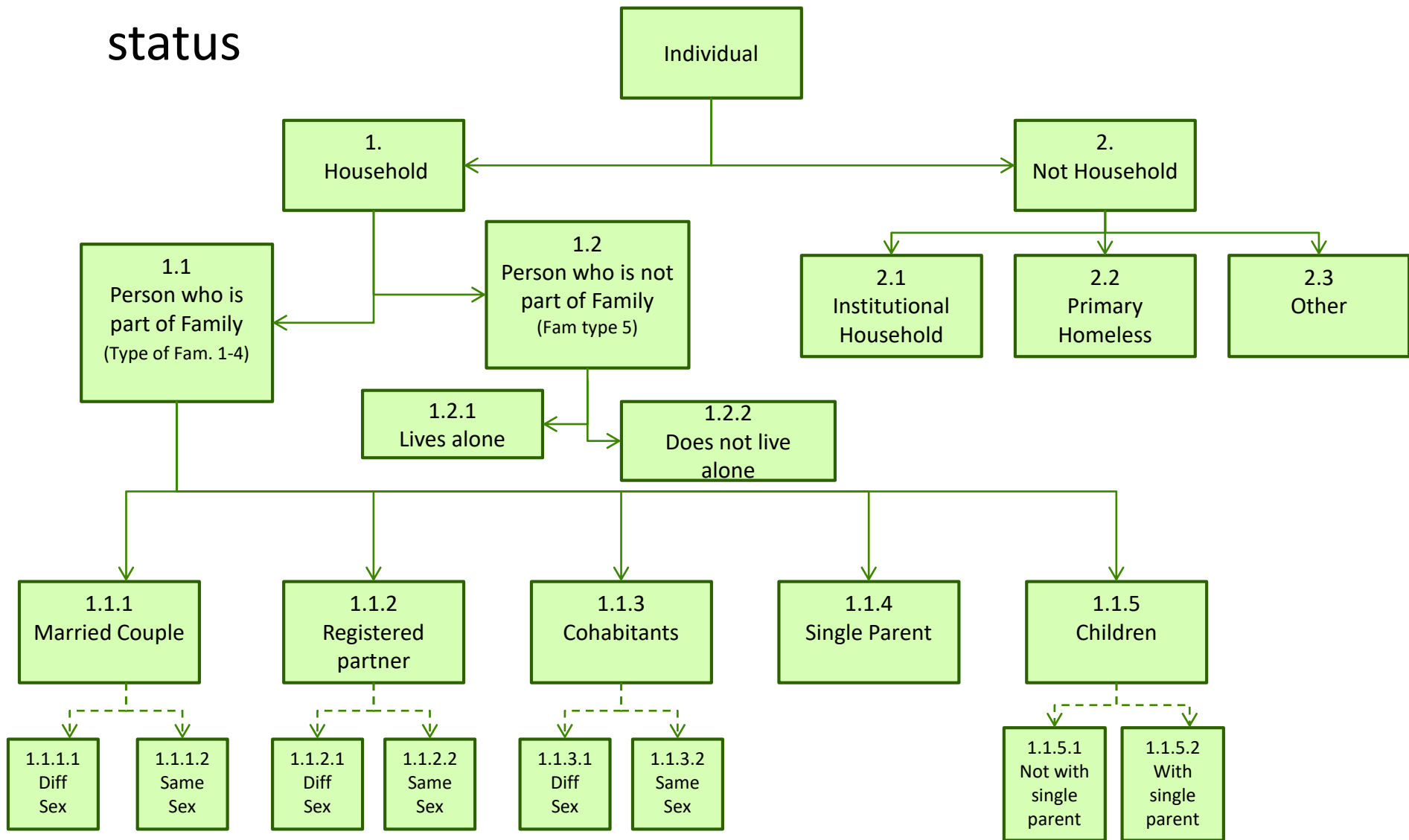
Data collection

- First invitation by mail to respond by web only. Second invitation by mail to those not yet responding, including a printed copy of the questionnaire. Finally a third possibility to respond, by a telephone interview.
- Data collected between January and May 2012.
- Response rate of 62 percent (65 percent weighted). More than 40 percent response rate in all strata.
- If household size or household type differed between survey and register: re-contact by telephone in order to establish the “true” value. Almost 3000 individuals contacted, 85 percent of the 3000 agreed to confirm.

Some important parameters

- Proportions
 - Persons linked to (partly) incorrect address
 - Persons/households living in a rented dwelling
- Gross and net errors (using sample and register data)
 - Persons after household status
 - Private households after type of household

Household status



Some results based on the evaluation study

Variables; household size, household type, type of ownership

- Gross error: data quality for different types of analysis
- Net error: data quality for aggregates

Some results based on the evaluation study (Werner 2014)

Estimated no. of hhs in r and c :

$$\hat{Y}_{rc} = \frac{X_{.c}}{\hat{X}_{.c}} \sum_{h=1}^H \frac{N_h}{n_h} \sum_{k \in rc} y_k$$

$X_{.c}$ 'known' Hhs in class c from census

$\hat{X}_{.c}$ estimated no. Hhs in c from census

$y_k = 1/a_k$, a_k # persons ≥ 18 yrs in hh from *ev Surv*

$x_k = 1/b_k$, b_k # persons ≥ 18 yrs in hh from *Census*

Some results based on the evaluation study

92,1 % ($\pm 0,8$ %) of hhs have correct size (# inhabitants).

Errors higher for class 6-10 and 11+
(5-10 46% net error 36" +/- 8").

Net error estimates indicate show 1 and 2 person hhs are underestimated by 10.3 and 9.3 % respectively.

Some results based on the evaluation study

The results indicate that the number of smaller households are underestimated and the number of larger households are overestimated by the register.

	1	2	3	4	5	6-10	11+	Total
Total	-	-	0	0	+	+	+	-
Stockholm	-	-	0	0	0	+	+	-
Gothenburg	-	-	0	0	+	+	+	-
Malmo	-	-	0	0	0	+	+	-
Municip. > 70"	-	-	0	0	0	+	+	-
Smaller Municip	-	-	+	0	0	+	+	-

Evaluation: Which source is most reliable?

Variable	# Miss-matches	# Where correct information was collected	Register information correct	Questionnaire information correct	Other information correct
# inhabitants	2 410	2 034	239	1 653	142
HH-type	2 251	1 881	469	1 205	207
Total	2 817	2 394	-	-	-

Other evaluations done in the Swedish census

Evaluation of dwelling ownerships

Comparison of Household counts depending on definitions.

Comparison of the dwelling register and a survey of available rental apartments

Quality treatments in the census

- Imputation?
 - No useful auxiliary information to aid imputation, since no reliable household information existed from another source
- Matching persons without dwelling keys to apartments?
 - Assuming that the relationship between size of apartment and size of household is similar for individuals with and without registered dwelling keys would have yielded doubtful quality of the result: 48 percent of the persons with missing dwelling keys had addresses in buildings with no apartments according to the Real Property Register.
- Adjustment for missing data was not required by Eurostat

The number of persons with missing identification key has decreased (2,3% in 2017).

To use the LFS to assess Quality in a Census system?

The Census:

Straightforward goal and delivery

Once done the results are what they are

The population of individuals fairly well known

Only some variables not known

Requirements fairly modest

System of social statistics

More detailed statistics

Flexible output, a core not decided,

User Requirements

Household population, definitions, uses, models

Maintenance over time

Microdata Quality

Sketching an Integrated Register and Sample Survey QC-scheme

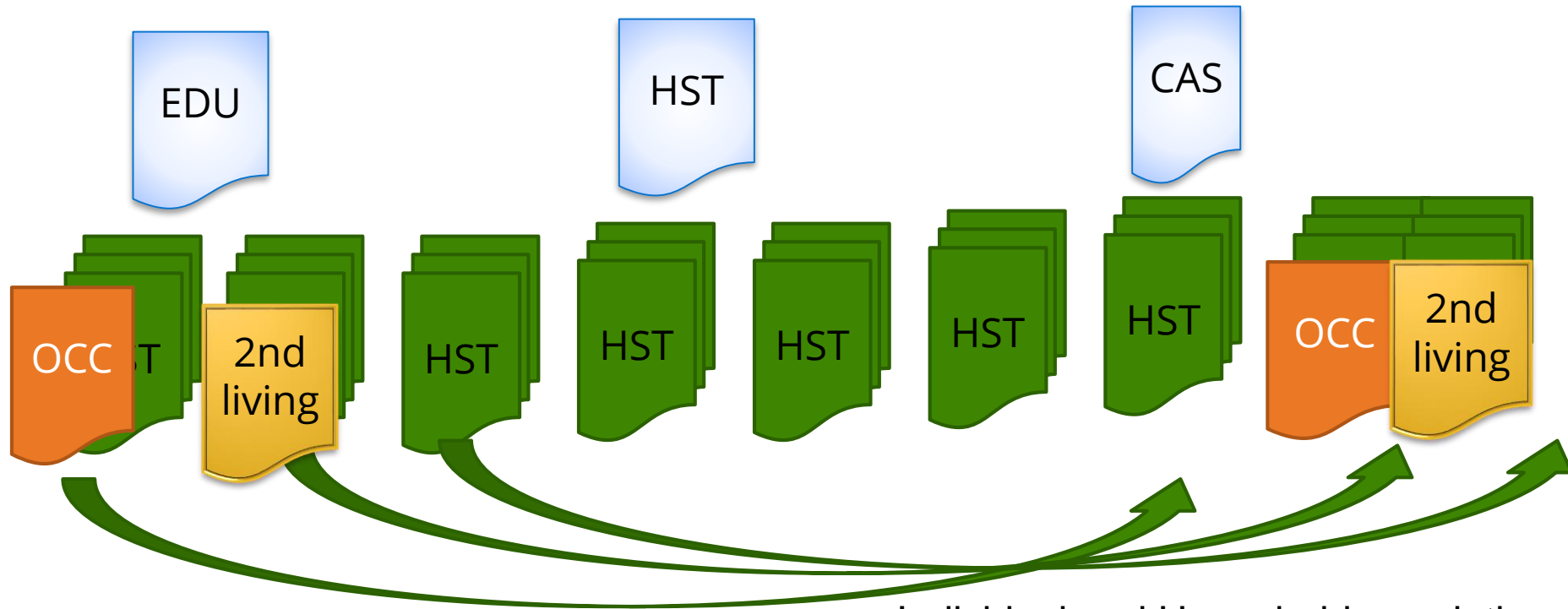
Recurrent and inbuilt system to maintain populations and quality assure standardised core variables in a national statistics production system.

How to cope with the Housing statistics and Property and Building population?

Other sources: E.g. Smart meters,

Cross checks with other registers?

Sketching an Integrated Register and Sample Survey QC-scheme



Individual and Household population

The LFS-panels, A tailored complement survey, geographic rotation? Subpopulations? Use of other surveys and sources. Test of assumptions as well , frequency is a function NSOs quality dedication.

References

Baffour, B., King, T. and Valente P., (2013). *The Modern Census: Evolution, Examples and Evaluation* . International Statistical Review (2013), 81, 3, 407–425 doi:10.1111/insr.12036

Berka, C., S. Humer, M. Lenk, M. Moser, H. Rechta, and E. Schwerer. (2012). *Combination of Evidence from Multiple Administrative Data Sources: Quality Assessment of the Austrian Register-Based Census 2011*. Statistica Neerlandica Vol 66 nr. 1 18–33.

Axelsson, M., Holmberg, A., Jansson I. and Westling S., (2018) *A Register-Based Census: The Swedish Experience*, Chapter 17 in *Administrative records for Survey Methodology*, Wiley New York. (forthcoming)

Dunne, J. and MacFeely, S., (2014) Laying the foundations for a new approach to Census taking in Ireland Statistical Journal of the IAOS 30 (2014) 359–365 DOI 10.3233/SJI-140848

Eurostat (2011) EU Legislation on the 2011 Population and Housing Censuses — Explanatory Notes. <http://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-RA-11-006>

Groves, R.M., Fowler Jr., F.J., Couper, M., Lepkowski, J.M., Singer, E. and Tourangeau, R. (2004). *Survey Methodology*. New York: Wiley.

References

Holmberg, A., (2012). *Discussion on assessing quality of administrative data*. Statistica Neerlandica, Vol 66, nr. 1 pp 34-40. doi: 10.1111/j.1467-9574.2011.00507.x

Holmberg, A., and Bycroft, C. (2017). Statistics New Zealand's Approach to Making Use of Alternative Data Sources in a New Era of Integrated Data. Pp 474-478 In Total Survey Error in Practice, Wiley, New York.

Holmberg, A., Blomqvist, K., Engdahl, J., Irebäck, H., Lundell L.-G. and Svensson, J. (2011). A strategy to improve the register system to store, share and access data and its connections to a generic statistical information model (GSIM). Invited paper presented at Work Session on Statistical Data Editing, UNECE, Ljubljana, Slovenia, 9-11 May.

Nordbotten, S. (2010) The Use of Administrative Data in Official Statistics – Past, Present, and Future – With Special Reference to the Nordic Countries, Official Statistics – Methodology and Applications in Honour of Daniel Thorburn, 2010, pp. 205–223

Laitila, T., Wallgren, A. and Wallgren B., (2011). Quality Assessment of Administrative Data. Research and Development–Methodology Reports from Statistics Sweden, 2011:2, Statistics Sweden.

Lindgren, K (2014). *Evalvering av Census 2011 – Utvärdering av variabel ägarkategori*, Internal report, Statistiska Centralbyrån, 2014-02-07 (in Swedish)

References

Lothian, J., Holmberg, A., & Seyb A., (2018). *An Evolutionary Schema for Using "it-is-what-it-is" Data in Official Statistics*. Journal of official statistics (in press).

Reid, G., Zabala, F. and Holmberg A., (2016). *Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ*. *Journal of Official Statistics*, Vol. 33, No. 2, 2017, pp. 1–35, <http://dx.doi.org/10.1515/JOS-2017-00xx>

Schnetzer, M., Astleithner, F., Cetkovic, P., Humer, S., Lenk, M., and Moser, M., (2015). *Quality Assessment of Imputations in Administrative Data*. *Journal of Official Statistics*. Vol. 31, No. 2, 2015, pp. 231–247.

Schulte Nordholt, E., M. Hartgers and R. Gircour (eds.), 2004. *The Dutch Virtual Census of 2001, Analysis and methodology*. (Voorburg/Heerlen, 2004). (<https://www.cbs.nl/-/media/imported/documents/2005/17/b-57-2001.pdf>)

Skinner, C., Hollis, J., and Murphy, M. (2013). *Beyond 2011: Independent review of Methodology*. ONS report.

Statistics Finland (2004): *Use of Registers and Administrative Data Sources for Statistical Purposes – Best practices in Statistics Finland*.

References

Statistics New Zealand (2012). Transforming the New Zealand Census of Population and Dwellings: issues, options and strategy. Wellington: Statistics New Zealand.

Thygesen, L. (2010) The Importance of the Archive Statistical Idea for the development of Social Statistics and Population and Housing Censuses in Denmark

UNECE (2007) *Register-based statistics in the Nordic countries : Review of best practices with focus on population and social statistics*, ISSN 0069-8458

UNECE (2012), 60:th Conference of the European Statisticians (CES) , *Paris, 6-8 June 2012*, Papers submitted by several countries f.ex. Austria, Switzerland, Slovenia, Italy and Finland.
<https://www.unece.org/index.php?id=28580>

Zhang, L.-C. (2012). *Topics of statistical theory for register-based statistics and data integration*, Statistica Neerlandica, vol 66, nr 1, 41-63.

Zhang, L.-C. (2014). Data integration. The Survey Statistician. No. 70, July 2014. The International Association of Survey Statisticians.

Werner, P., (2014). *Evaluering av Census 2011*, Slutrapport, Statistiska Centralbyrån, (in Swedish)