Standardisation of methods and processes

Fabio Ricciato

European Commission, Eurostat
Unit A5 – Methodology; Innovation in Official Statistics

ISTAT Workshop on methodologies for official statistics
Rome, 5-6 December 2022

2 papers in this session

Representation in a structured and formalized way

- ***Standardization*** *of* ***methods*** *and* ***processes****: overview of the Istat activities and open problems*

- ***Metadata*** *for statistical processes on registers: how to organize facts with GSIM*

European Commission

[Excerpt from]
Standardization of methods and processes: overview of the Istat activities and open problems

3.1 Description of processes

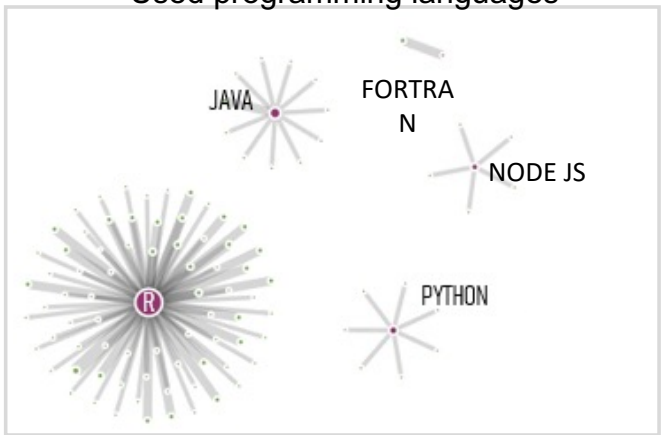
3.2 The 'Collection of methods' website

3.3 Process documentati

Obviously, one of the main difficulties with such a platform with a large and growing availability of information is its maintenance: this will be a challenging task from both a methodological and a technical perspective. Indeed, the website can be considered as a "photograph" of the state of the offering of methodological service for a specific period; but as methods are updated, discovered or made obsolete the need will arise to keep track of such changes. At the same time, the website will have to follow the changes of the

	Methodological tools		
Row number	Statistical methods	Procedures	Process contro
1		In R: read.table; read.csv read csv2 Relais: function dataset	Number of records; variables number, name and type
2		In R: read.table; read.csv, read csv2 Relais: function dataset	Number of records; var
3	Fraction of rows/units with no missing values	Relais: data profiling R: ad hoc function	The best variables are t
4	Fraction of rows/units with a correct value	Relais: data profiling R: ad hoc function	The best variables are t
5	Fraction of rows/units with values correctly linked between the variable and the correlated variable	Relais: data profiling R: ad hoc function	The best variables are t
6	Gini index computed on the frequencies of the values	Relais: data profiling R: ad hoc function	The best variables are t
7	Cramer's V	Relais: data profiling R: ad hoc function	The best variables are t

Used programming languages

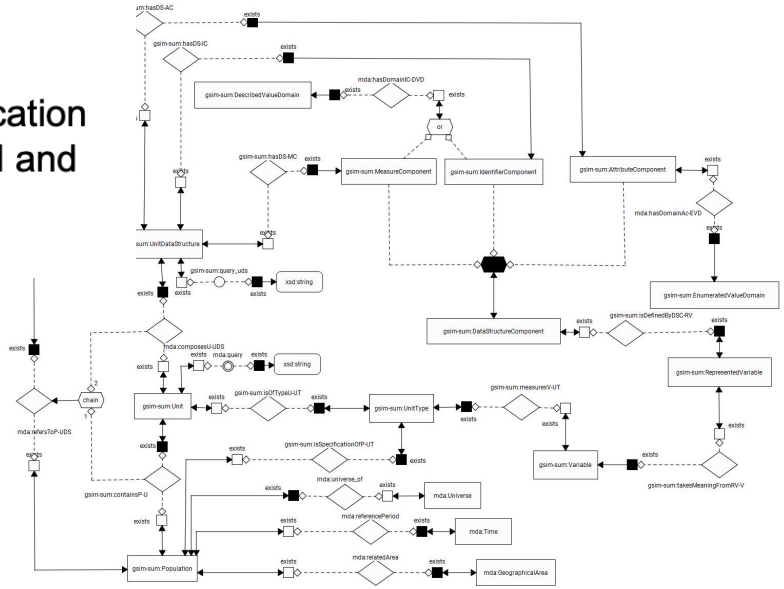


European Commission

[Excerpt from]
Metadata for statistical processes on registers: how to organize facts with GSIM

A standard conceptual framework: The Generic Statistical Information Model is a conceptual model whose objective is the standard definition of the information objects that are input or output of the different steps of a production process. We used GSIM extensively, although we had the need to add some new concepts

Ontologies: In computer science, an Ontology is a conceptual specification of the domain of interest expressed through a formal language, shared and unambiguous



Person(x,y,z,k,l,s,b,r)← select id, name,surname, age,adult,sex, birthdate,resident from Human
where alive=TRUE
Household(x,y,z) ← select * from Household
belong_to_household(x,y) ← select * from B_to_h
Income(x,y,z,k) ← select id, amount, ref_year, source from Income
has_income(x,y) ← select id_person, id_income from Income
Municipality(x,y) ← select code, description from Municipality
resident_in_municipality(x,y) ← select id_household, id_municipality from H_to_M
in_region(x,y) ← select code, code_region from Muncipality
Region (x,y) ← select code, description from Region

- *The proposed approach governs the whole production pipeline from registers' data to output datasets by means of a defined and coherent metadata asset.* This is maybe the most interesting and practical aspect to pursue. As an example, Istat has the objective to create a framework that allows users to create their own analyses from registers[1] and the work here presented is an important step in this direction.

European Commission

- *Formalization of processes, methods and (meta)data structures*

- *Why? → How?*

What happened to the car industry

- *Moderate **complexity** of product/process*
- ***Design** of product/process instructions: by few humans*
- ***Execution** by (many) humans*

- *High **complexity** of product & process*
- ***Design** of product/process instructions: by many humans*
- ***Execution** (mostly) by machines + human **supervision***

Official Statistics (OS) ≠ car industry

Some points of difference (where analogy fails) …

- *Data, information ≠ metal, plastics*
- *Statistics ≠ Industry*

- *100% automation perhaps impossible (nor desirable) in OS*
  - Some level of human supervision unavoidable

- *Distinction between 'process innovation' and 'product innovation' perhaps not so sharp in OS*
  - '***What*** you measure is defined by ***How*** you measure it'
- *…*
- *…*

European Commission

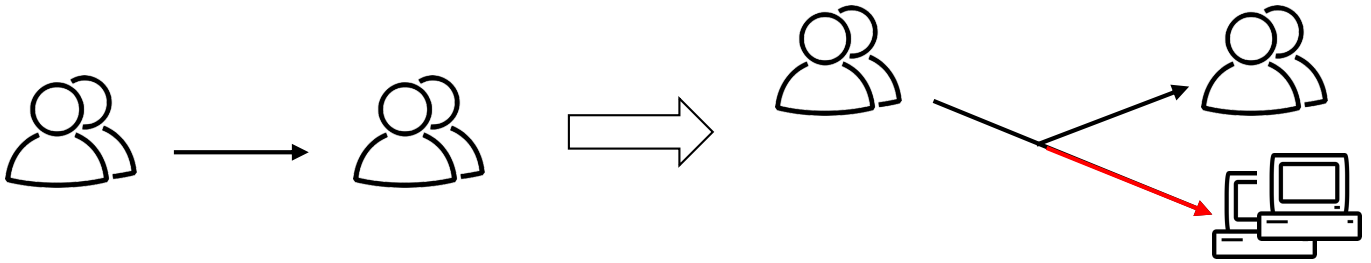
Complexity → Automation

… but still some elements may be inspiring

- ***Automation** necessary to deal with increased **complexity** of product/process*
  - Especially relevant for new statistics based on 'big data'
- *Human work shifts from 'executing instructions' to 'formulating instructions'*
  - More instructions and more complex, need more human designers
  - Skilled workers get 'upgraded' to engineers
  - Execution shifts from humans to machines
- *Paradigm change from:*

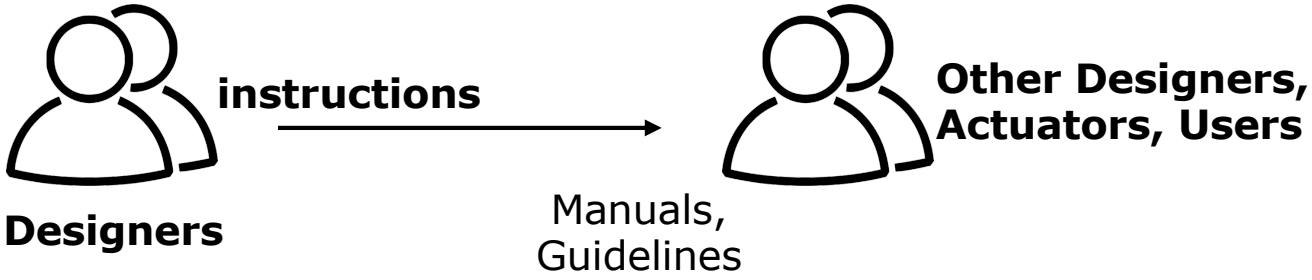
'instructions written by humans **for humans**' *to*

'instructions written by humans **for machines and humans**'

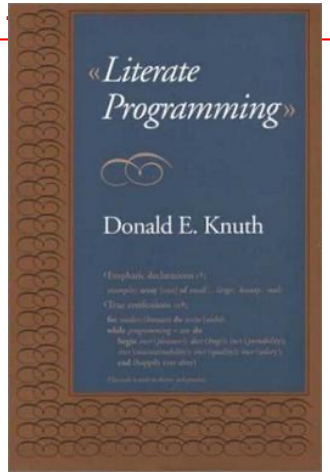


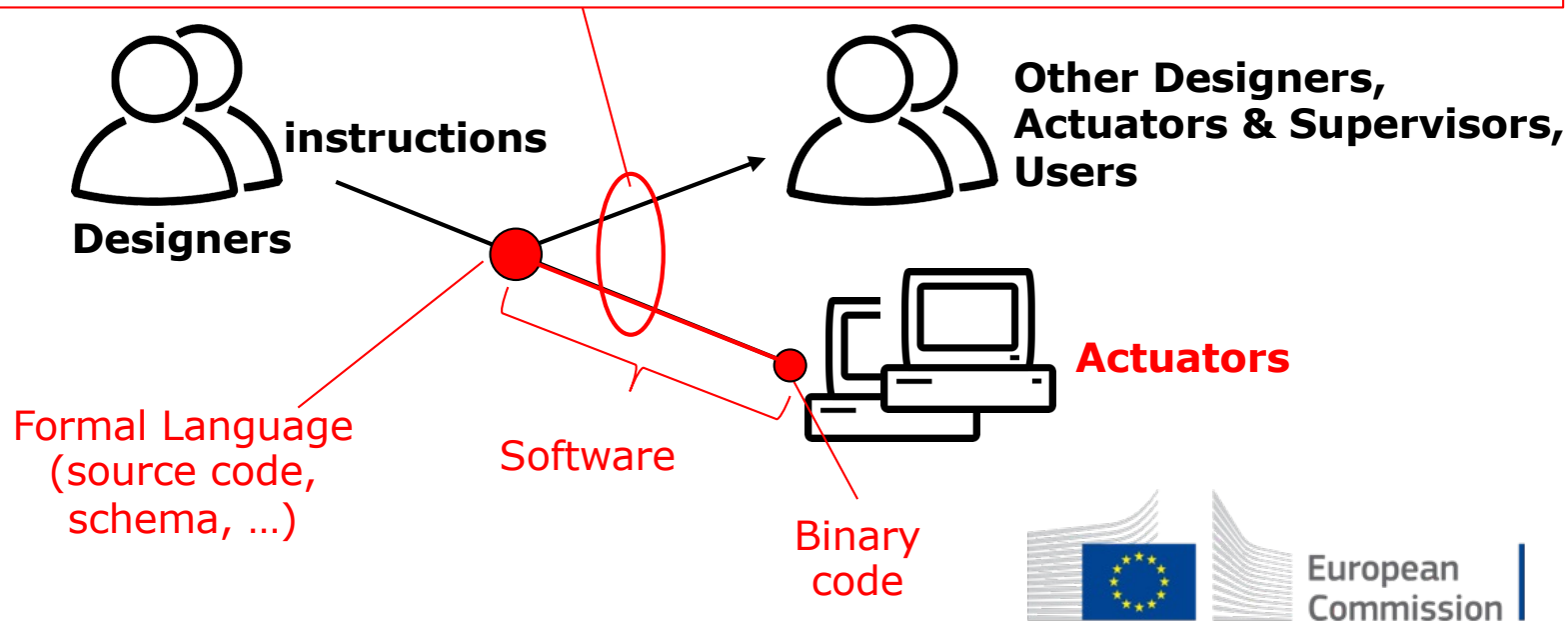
European Commission

Automation → Softwarization



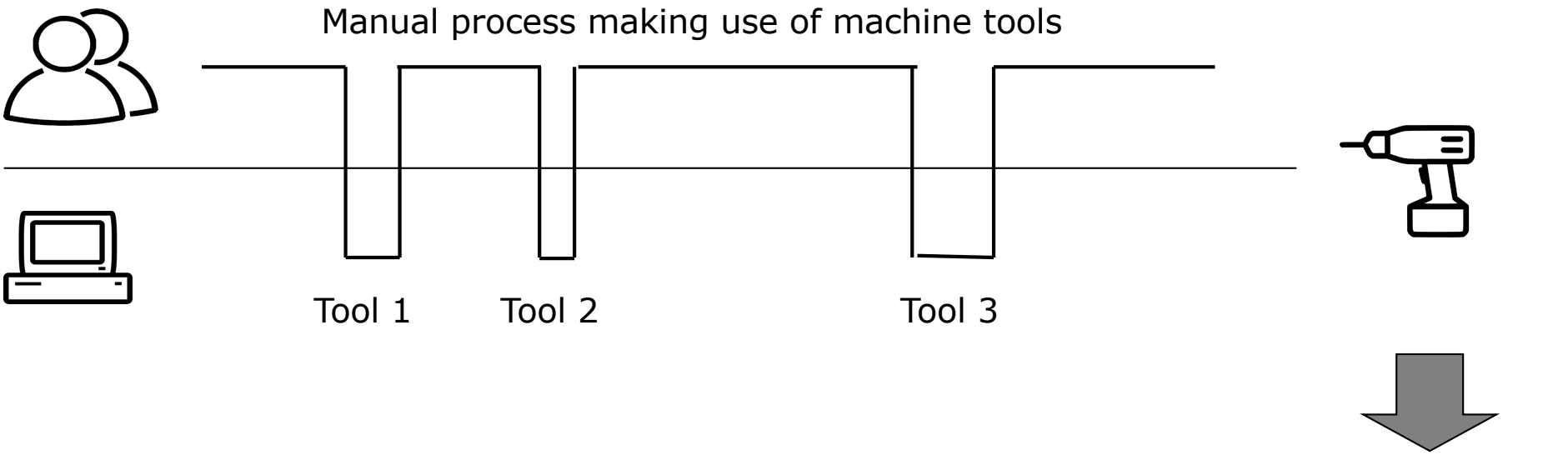
These two description levels are not independent and should be implemented together, as two sub-parts of the same methodological development process.
Code + Documentation; **Ontologies** + Documentation; Notebooks (Literate Programming)





European Commission

From tools to pipelines



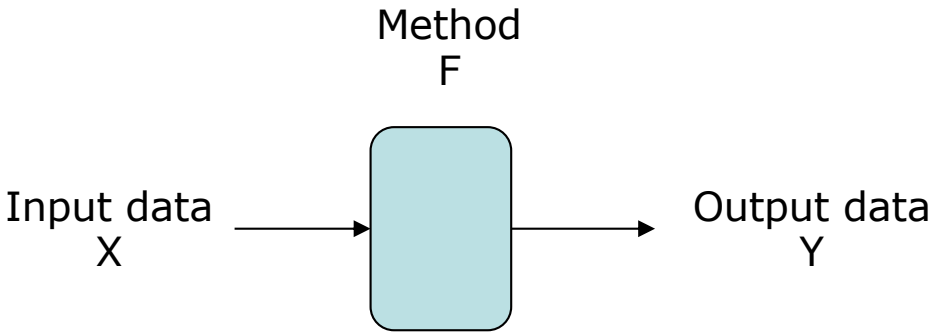
Machine pipeline with human checkpoints

Softwarization of statistical methods

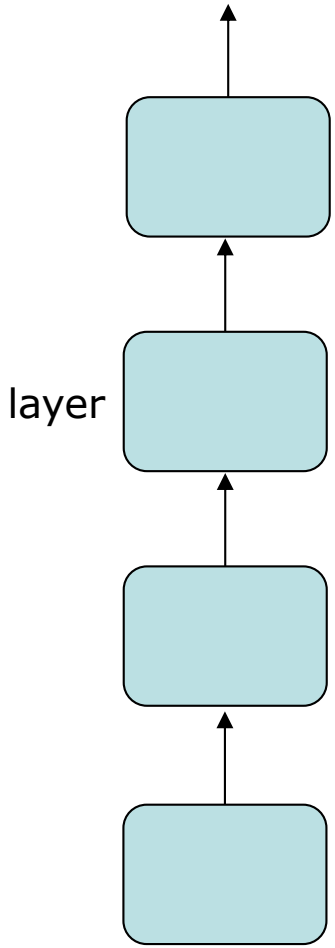
- *More process/product complexity → automatization*
- *Automatization → **Softwarization** of statistical methodologies**
- *Implications*
  - Software quality as part of methodological quality
  - Software development practices in methodological development (e.g. modularity, collaborative development, versioning)
  - Open-source code release as part of methodological transparency.
  - Benefit of open-source platforms and programming languages
  - Sharing code with other NSI, ease harmonisation, pool resources
  - New statistical services to users (custom on-demand table-building based on onthologies)
  - …

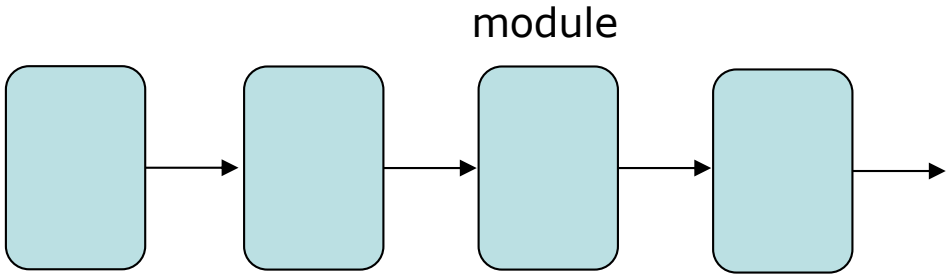
(*) A reflection on methodological sensitivity, quality and transparency in the processing of new 'big' data sources, Q2022 conference, Vilnius https://q2022.stat.gov.lt/scientific-information/papers-presentations/session-20

Code as “process metadata”

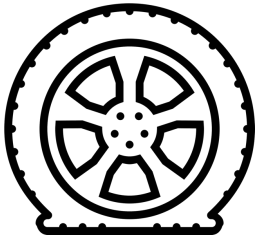


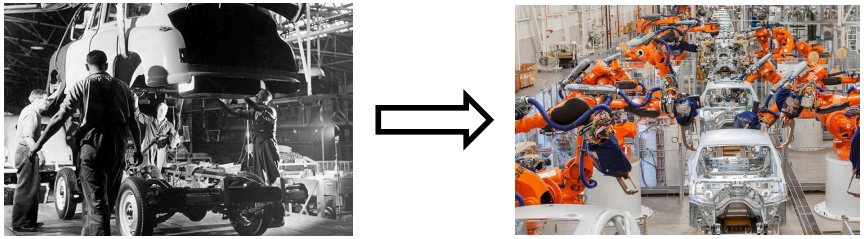
Description of Output Data (output meta-data)
= Description of Input Data (input meta-data)
+ description of Method (software code for F as process meta-data)



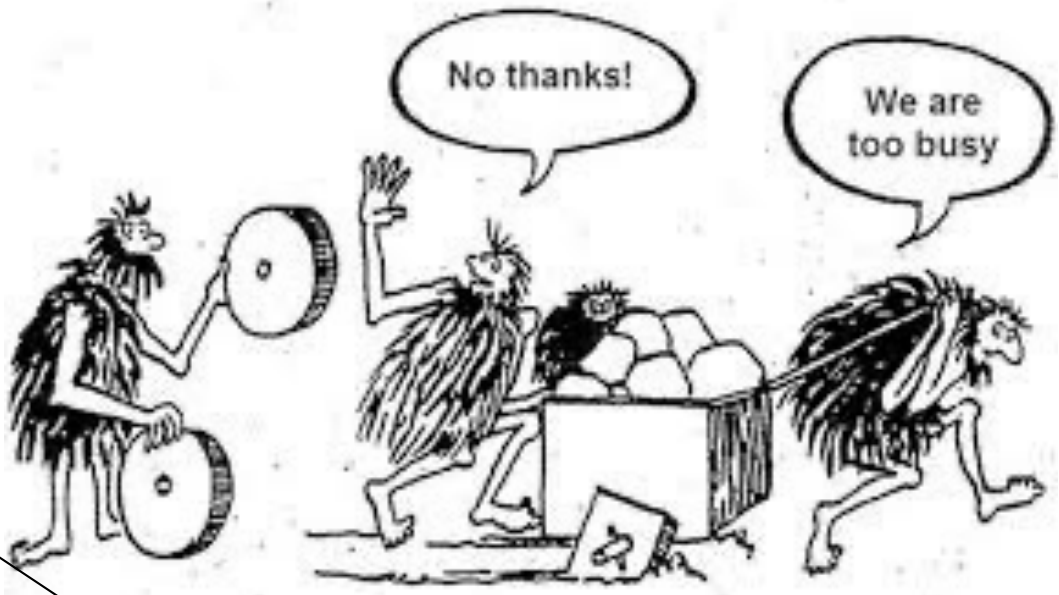


A good idea can be implemented in a bad way …



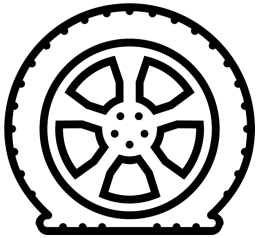


- *Many things can go wrong during the **transition** from human-only to human-cum-machine system, e.g. …*
  - The final desired state is designed, but not the migration path to there
  - The promise (that automated components will relief humans from repetitive and error-prone work and let'em move to more rewarding tasks) is not lived up in a reasonable time, causing lack of acceptance
  - A good general idea (e.g., replacing wood wheels with inflatable rubber tires) is implemented in a poor way (e.g., *that* particular tire was holed). → The failure of that specific (bad) implementation may lead to rejection of the whole (good) general idea!



European Commission

Examples



- *Introducing an onthology is an excellent idea, but is* that *particular proposed onthology fit for the purpose?*
  - Are domain experts happy about that? If not, is that because they do not understand the general concept of onthology, or because they have spotted a problem in that particular proposed onthology?

- *Standardisation based on international standard is an excellent idea, but are GSIM/GSBPM models fit also for new "big data" sources?*
  - They were developed for statistical data, maybe they can be "extended" (or "stretched") to deal with big data sources, or maybe new int'l models are needed …

- *Moving to open-source tools is an excellent idea, but is that specific package fit for the purpose? Do we have a good versioning system?*
  - Etc etc.

European Commission