Q 2016
European Conference on Quality in Official Statistics

Hagstofa Íslands

Big data and the integrated production of official statistics

Session 34

1. June 2016

Anton Örn Karlsson

Statistics Iceland

anton.karlsson@statice.is

INE Instituto Nacional de Estadística eurostat

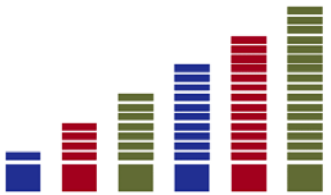
Madrid, May 31 - June 3

Q 2016
European Conference on Quality in Official Statistics

What is big data?

Huge datasets

- Organic data
- Found data
- n = all
- 3 V's
  - Volume, Velocity, Variety
    - 4th V: Veracity
    - 5th V: Value
- Not suitable for traditional processing or analysis
  - Alternative data processing methods may need to be applied
  - Big data analytical methods differ from traditional methods



Q 2016
European Conference on Quality in Official Statistics

Big data and stovepipes

The stovepipe approach

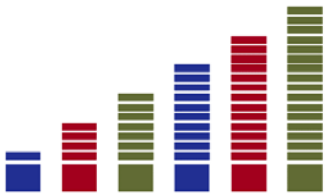
- A specific production process for each domain/statistic

Integrated approach

- A coordinated approach to provide a harmonious, interrelated production processes which can be reused across different domain

Big data can sometime lead to a more stovepipe related production processes

- Big data sources often seem to be rather linked to the stovepipe approach to statistical production
- Two Icelandic examples



Q 2016
European Conference on Quality in Official Statistics

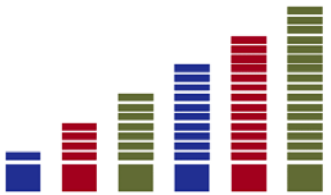
Two examples

Scanner data

- Contracts with the biggest supermarket chains in Iceland
  - They provide Statistics Iceland with information on every transaction made
  - Automatic transmission of data
  - Used for the Icelandic Price Index
    - And only for that!

The debts of households and enterprises

- Data gathered from every credit institute in Iceland
  - Commercial banks
  - Pension funds
  - The Housing Financing Fund
- Automatic transmission of data – high requirements of privacy and data security
- The data collection is based on Icelandic law passed by Alþingi in the summer of 2014.
  - A change to the statistical act was needed where it is explicitly stipulated that the data can only be used for this specific domain.

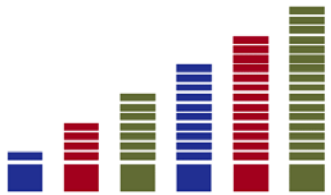


The challenge

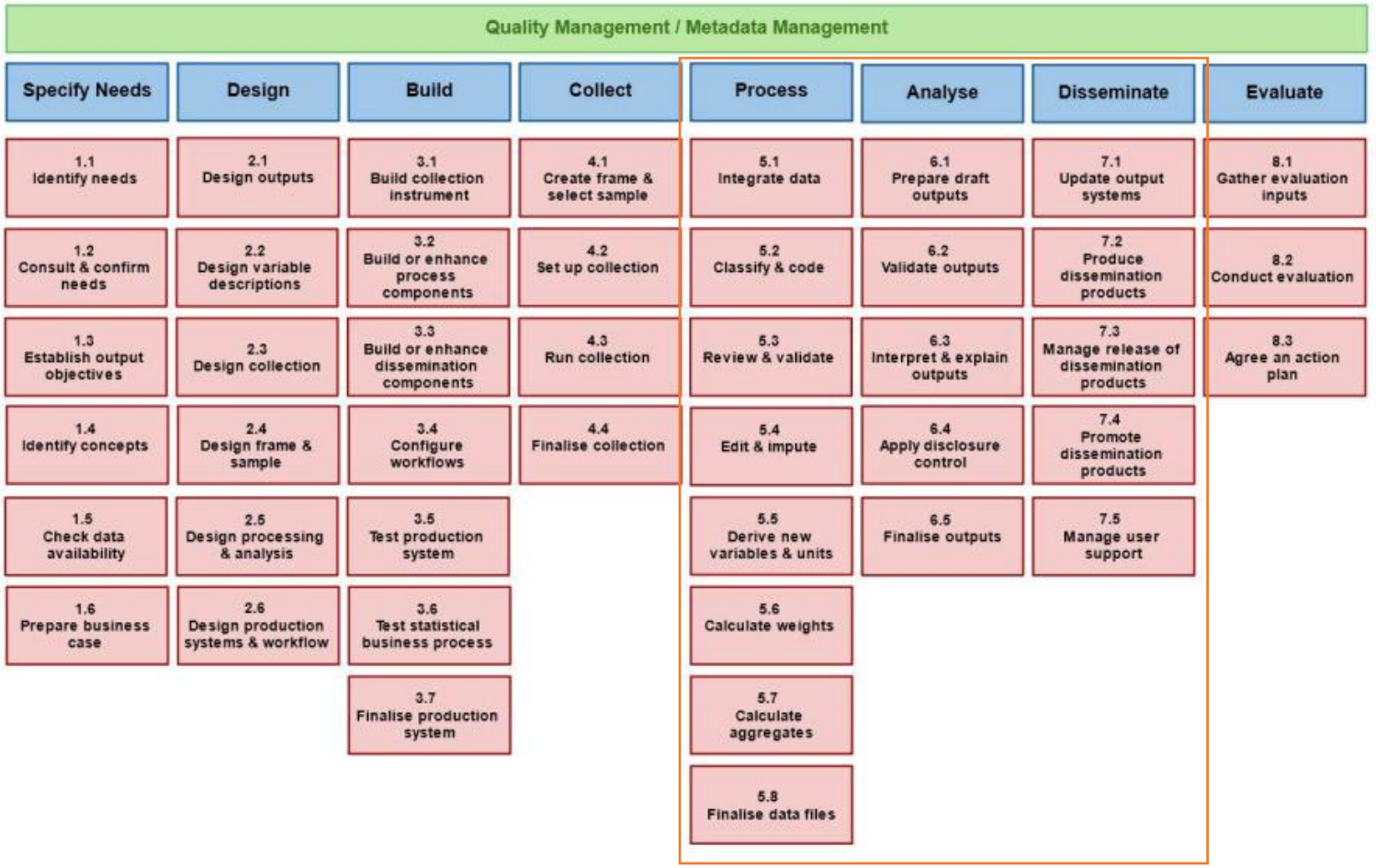
In both of these cases big data sources seem to increase reliance on the stovepipe approach

However, big data is more than just sources of data

- It also relates to processing and analyzing data
  - E.g. machine learning, unsupervised learning (data mining)
- And perhaps to the dissemination of data
  - E.g. visual presentation



GSBPM


Quality Management / Metadata Management
Specify Needs
1.1 Identify needs
1.2 Consult & confirm needs
1.3 Establish output objectives
1.4 Identify concepts
1.5 Check data availability
1.6 Prepare business case
Design
2.1 Design outputs
2.2 Design variable descriptions
2.3 Design collection
2.4 Design frame & sample
2.5 Design processing & analysis
2.6 Design production systems & workflow
Build
3.1 Build collection instrument
3.2 Build or enhance process components
3.3 Build or enhance dissemination components
3.4 Configure workflows
3.5 Test production system
3.6 Test statistical business process
3.7 Finalise production system
Collect
4.1 Create frame & select sample
4.2 Set up collection
4.3 Run collection
4.4 Finalise collection
Process
5.1 Integrate data
5.2 Classify & code
5.3 Review & validate
5.4 Edit & impute
5.5 Derive new variables & units
5.6 Calculate weights
5.7 Calculate aggregates
5.8 Finalise data files
Analyse
6.1 Prepare draft outputs
6.2 Validate outputs
6.3 Interpret & explain outputs
6.4 Apply disclosure control
6.5 Finalise outputs
Disseminate
7.1 Update output systems
7.2 Produce dissemination products
7.3 Manage release of dissemination products
7.4 Promote dissemination products
7.5 Manage user support
Evaluate
8.1 Gather evaluation inputs
8.2 Conduct evaluation
8.3 Agree an action plan

Big data methods, examples(1)

Calculating weights for social surveys

- Only used for SILC at the moment
  - The plan is to implement it for every social survey
- Regression forest used for finding the optimal pattern of auxiliary variables
  - Minimizes non-response bias in the data
  - Always using the same database for auxiliary variables
- Especially good for finding complex interactions in the data (which would take a very long time to uncover using traditional regression methods)

Big data methods, examples(2)

Imputations using random forest

- R-package missForest
- Already been used on census data, educational data and survey data (both social and enterprise) with promising results
- Provides a much needed quality measure of the imputation
- While multiple imputation results in multiple datasets (which can be challenging for users to understand) – random forest imputation only provides a single complete dataset.

Re-classification

- Used when classification systems have changed between two sets of data
  - A concrete example: Working with data in two different occupation coding schemes for the register based Icelandic Census
- Comparisons made between random forest, C5 and naïve bayes approaches.
  - Random forest had by far the least amount of categorization errors

Big data methods, examples(3)

Analysis using regression trees

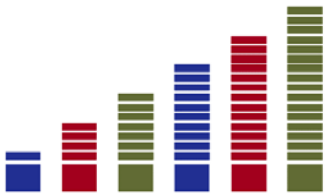
- Used for a cooperation project between Statistics Iceland and the Icelandic welfare monitor
- Two part project
  - Small area estimation
  - Regression trees
- Gives an opportunity to search for predictors of poverty using data from the Icelandic SILC.
- The results
  - Will have a noticeable effect on the work of the Welfare Monitor, e.g. because of the importance of self-defined health when predicting severe material deprivation
  - Were very well received by the Welfare Monitor – especially the regression trees.
  - An interesting way of presenting results
    - Very graphical, intuitive and easy to understand.

Q 2016
European Conference on Quality in Official Statistics

In summary

Big data offers both opportunities and challenges in integrated statistical production

- In some cases, big data sources can re-introduce the stovepipe approach
  - This is a challenge which has to be kept in mind when new data sources are introduced for statistical production
- The analytical methods of big data are an opportunity for further integrating the production processes for different domains of statistics



Q 2016
European Conference on
Quality in Official Statistics

Takk fyrir!