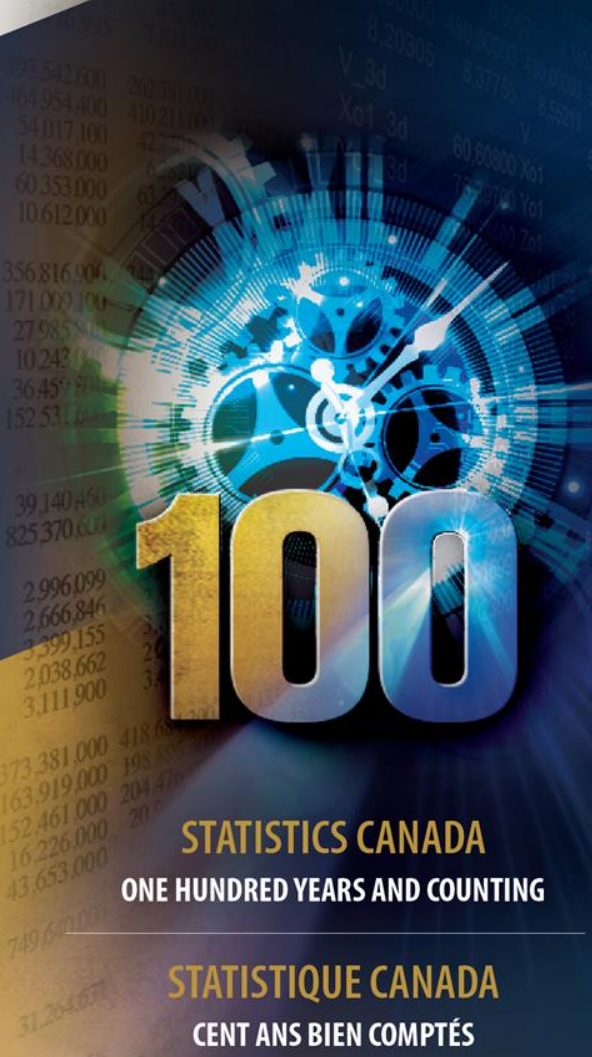
(R)evolution of generalized systems and statistical tools at Statistics Canada

Susie Fortier and Steven Thomas
International Cooperation and Corporate
Statistical Methods Division



6th Conference on the Use of R in Official Statistics
The Hague, September 12-14 2018

Content

- Context and background
  - Modernization initiative
  - StatCan's Generalized Systems
- Recent and on-going exploration
- Enablers and challenges
- Next steps

Large modernization initiative at Statistics Canada

The vision statement includes the following:

- identifying new methods of generating and collecting data that move beyond a survey-first approach
- finding new ways to integrate data from a variety of sources.

➔ *A modern, responsive statistical agency with a strong culture of **innovation** and desire to continuously improve and develop our products and services*

Modernization initiative

One *of the many* challenge*s*:

How to make the most use of all relevant sources of data in the most statistically/scientifically rigorous, informed and efficient manner?

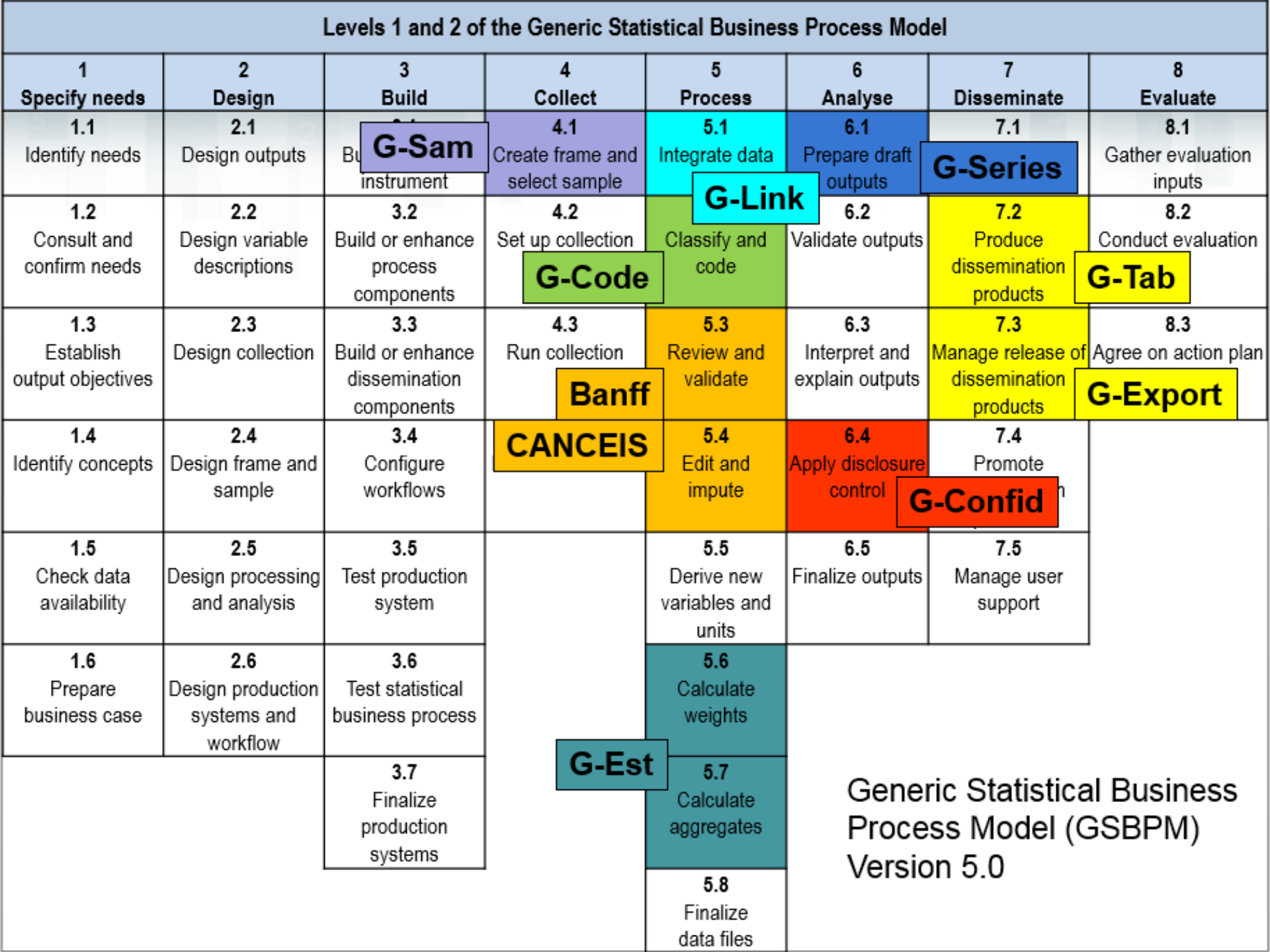
(but this challenge is not *that* new…)

One *of our many historical* answer*s*:

The use of corporate and generalized systems

The Generalized Systems

- A suite of closed-source, (mostly) SAS based, statistical systems developed, supported and maintained by the Methodology Branch and the Informatics Branch
- Available for free upon request
- Covering most statistical steps in the Generic Statistical Business Process Model


Levels 1 and 2 of the Generic Statistical Business Process Model
1 Specify needs
2 Design
3 Build
4 Collect
5 Process
6 Analyse
7 Disseminate
8 Evaluate
1.1 Identify needs
2.1 Design outputs
Bu instrument
4.1 Create frame and select sample
5.1 Integrate data
6.1 Prepare draft outputs
7.1
8.1 Gather evaluation inputs
1.2 Consult and confirm needs
2.2 Design variable descriptions
3.2 Build or enhance process components
4.2 Set up collection
Classify and code
6.2 Validate outputs
7.2 Produce dissemination products
8.2 Conduct evaluation
1.3 Establish output objectives
2.3 Design collection
3.3 Build or enhance dissemination components
4.3 Run collection
5.3 Review and validate
6.3 Interpret and explain outputs
7.3 Manage release of dissemination products
8.3 Agree on action plan
1.4 Identify concepts
2.4 Design frame and sample
3.4 Configure workflows
5.4 Edit and impute
6.4 Apply disclosure control
7.4 Promote
1.5 Check data availability
2.5 Design processing and analysis
3.5 Test production system
5.5 Derive new variables and units
6.5 Finalize outputs
7.5 Manage user support
1.6 Prepare business case
2.6 Design production systems and workflow
3.6 Test statistical business process
5.6 Calculate weights
3.7 Finalize production systems
5.7 Calculate aggregates
5.8 Finalize data files
G-Sam
G-Link
G-Series
G-Code
G-Tab
Banff
G-Export
CANCEIS
G-Confid
G-Est
Generic Statistical Business Process Model (GSBPM) Version 5.0

The Generalized Systems pros & cons

✓ Cost-saving

- for dev and maintenance
- easier staff transition (less training)
- ***savings materialized only with strong governance and mandatory use…*

✓ Increased quality

- fully vetted methods
- robust systems with fewer bugs
- dev by expert group

✓ Limitations

- available functionalities
- timely innovations

✓ Closed-source approach

- limited collaboration opportunities (both internal and external)
- potential for Black Box pitfall

The Generalized Systems

Traditional Approach to Software Development

1. Initial research
   - Supported by literature and technical/advisory committees
2. Development
   - Proof of concept/prototype; generalization of approaches ; business requirement and software specification
3. Execution
   - Investment proposal with full project mgnt (2-3 years); certification and user acceptance; doc & training; full support and maintenance.

Recent and on-going exploration

What has changed?

- Innovation and modernization
  - New methods, new data sources, new software solutions, new workforce
- Open and accessible
  - New opportunities, new culture, data-driven approach
- Agility
  - Fast-changing environment, evolving data needs, evolving data sources, evolving statistical methods

Recent and on-going exploration

(R)evolution @ StatCan

- Data Science / Machine Learning projects using R spread throughout the department
- Internal R user group and CRAN committee
  - Methodology, subject matter and IT staff
  - Create an R environment in StatCan (accessible to all staff)
  - *(Most) R packages can be made available upon request for R&D work and kept on an internal 'CRAN'*
- Exploratory research phase for GenSys and methodology

Recent and on-going exploration

Exploring R for Generalized Systems

- `sae` - small area estimation (Molina and Marhuenda (2015) The R Journal)

- `Stratification` - generalization of the Lavallée-Hidiroglou method (Baillargeon and Rivest (2011) Survey Methodology)

- `SamplingStrata` – Optimal Statification of Sampling Frames for Multipurpose Sampling Surveys (Ballin and Barcaroli (2014) Survey Methodology)

Recent and on-going exploration

Key findings

- Most packages are backed with a degree of methodological research
- They are actively supported by an individual (usually the researcher / programmer)
- Packages are documented with examples
- Packages can be ‘hidden’ in SAS environment (PROC IML)

Issues

- Individual for support may not be available
- Some packages are less solid then others
- Connection between R and SAS is a challenge (version control)

Recent and on-going exploration

Many other exploratory uses *(non-exhaustive list)*

- Machine learning techniques to enhance nearest neighbour imputation in the Census (`ReliefF` feature selection algorithm from the `CORElearn` package*)
- Recent extensive use in a large scale imputation Hackathon held over the summer (results under further evaluation ; scaling up challenge quickly identified)
- Creation of synthetic files for an external hackathon (with the `SynthPop` package)
- Shiny package
- Data visualisation
- …

Enablers and challenges

Aside from our modernization initiative…

- ✓ Strong engagement and initiative from internal users
- ✓ Bimodal IT development and environment
- ✓ Exploring repository manager (JFrog Artifactory)
- ✓ Philosophical change in IT security *(blacklisting instead of whitelisting *)*
- ✓ Sharing experience with other NSOs

Enablers and challenges

Our progress is still modest…

- ✓ Mode 2 to Mode 1 bridge is work in progress
- ✓ Some technical challenges :
  - ✓ Packages with code needing a compiler remain an issue
  - ✓ Scaling up for use with big data
- ✓ Communication gap in the definition of “vetted” package (security vs accuracy/utility)
- ✓ Philosophical review of roles / responsibilities / accountabilities for quality assurance

Next Steps

- Review and act on recommendations from our Advisory Committee on Statistical Method on the use of R and open source for our generalized systems
  - Idea generally well received but strong caution advised
- Prototype of R module (function) in one of our Generalized Systems
- Review our current licensing strategy
- Continue to build capacity
  - Successful experience with active learning for machine learning but formal training for R will likely be required

THANK YOU

For more information, please contact

MERCI

Pour de plus amples renseignements, veuillez contacter

Susie.Fortier@canada.ca

www.statcan.gc.ca

#StatCan100