GSIM, CSPA, and Related Activities of the High-Level Group

Arofan Gregory
Metadata Technology NA
April 1. 2014
NADDI, Vancouver

Introduction

- It's all about acronyms:
  - UNECE
  - CES
  - GSBPM
  - GSIM
  - CSPA
  - DDI

Official Statistics

- UN/ECE – The United Nations Economic Commission for Europe
  - Based in Geneva (Palais de Nations)
  - *Not* strictly European in Scope
- CES – The Committee of European Statisticians
  - Not strictly European in Scope
  - Coordinates statistical activity among national and supra-national statistical agencies
- HLG – The High Level Group for the Modernization of Statistical Production and Services
  - What do you think?

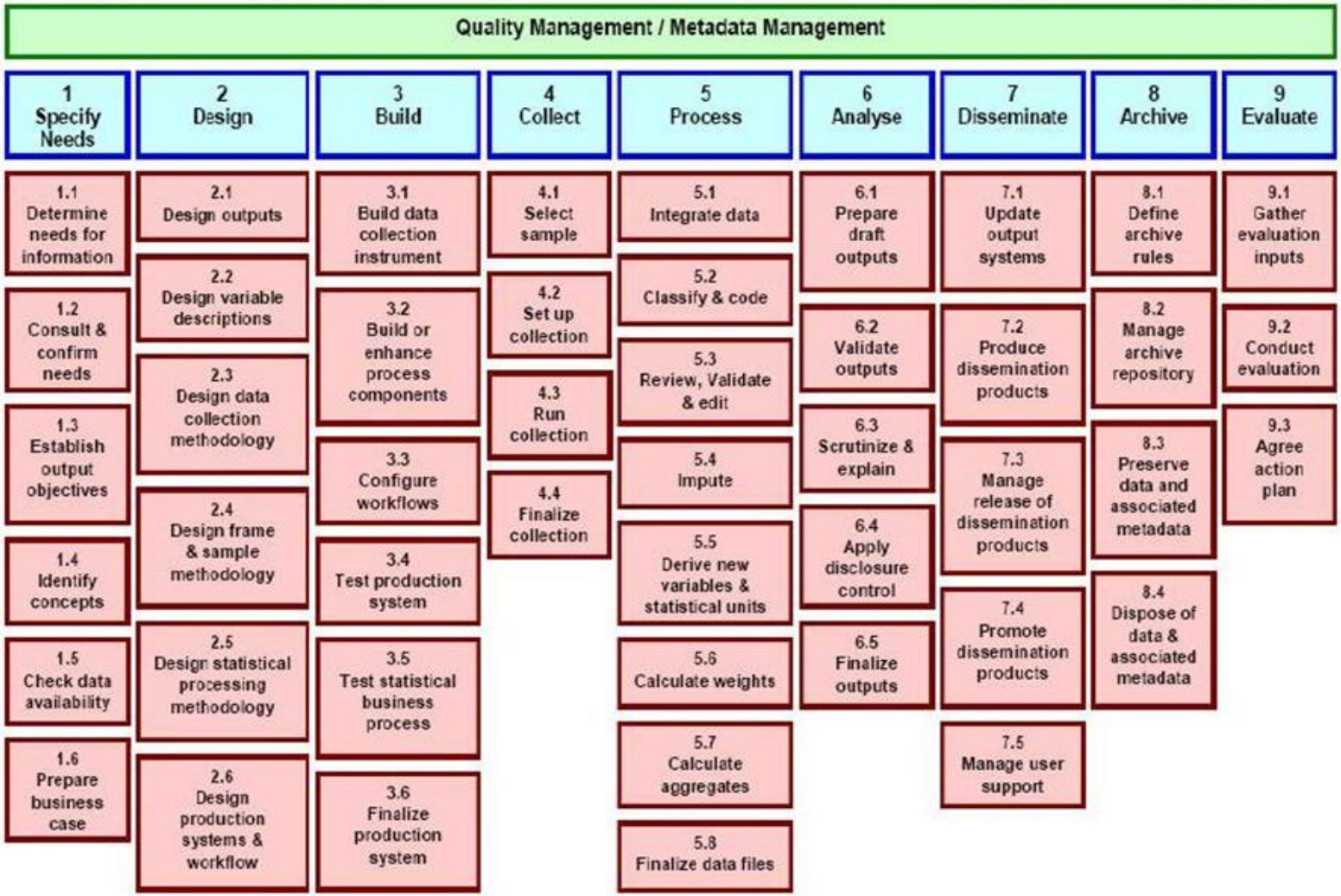
The Perceived Problem

- In the age of Google, “Big Data”, and Open Data, statistical agencies fear they may become irrelevant
- The demand for data is huge, but traditional statistical production is relatively slow, and very expensive
- Budgets are shrinking, but the legal requirements for official data are not changing

A Reference Model for Process

- Over the past few years, the CES community produced a “reference model” describing the statistical production process
- GSBPM – the Generic Statistical Business Process Model
  - Very effective in allowing statistical agencies to describe their processes, and to communicate among themselves
  - Defines a non-linear process model, with clear terms and definitions

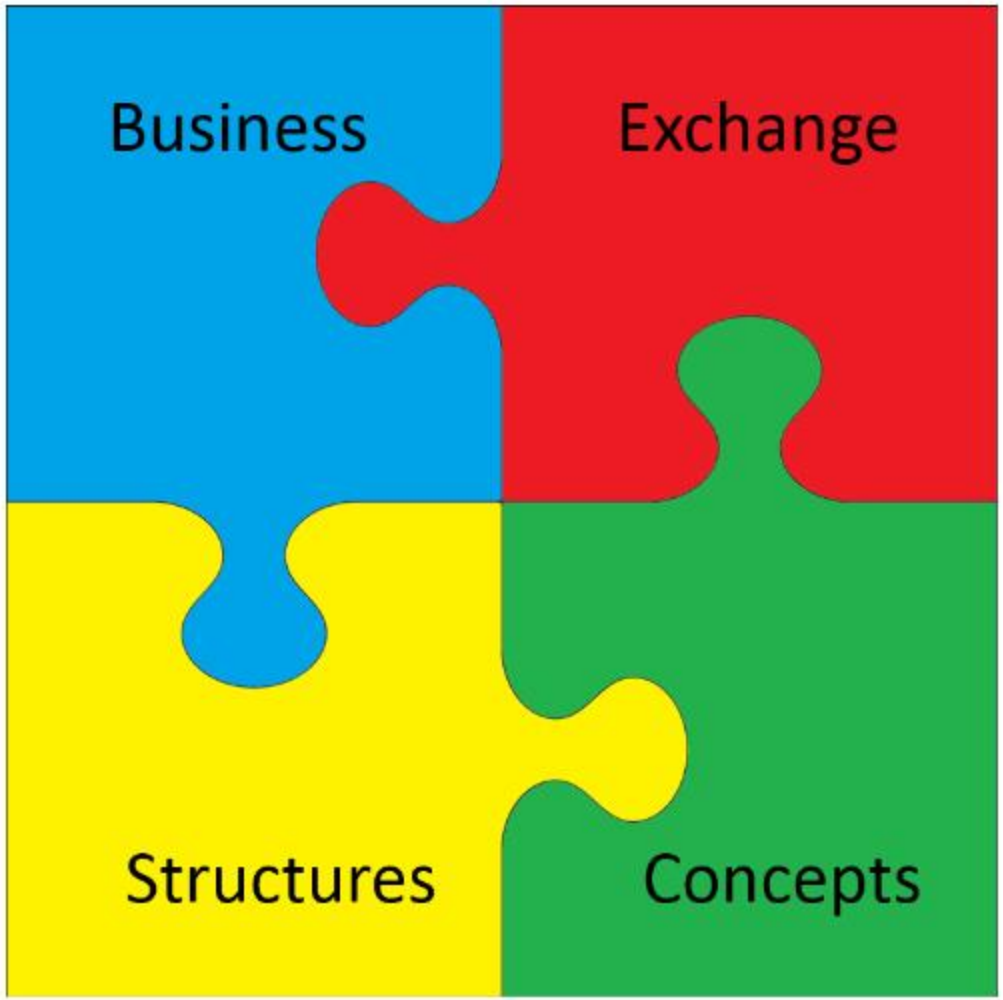
GSBPM


Quality Management / Metadata Management
1 Specify Needs
1.1 Determine needs for information
1.2 Consult & confirm needs
1.3 Establish output objectives
1.4 Identify concepts
1.5 Check data availability
1.6 Prepare business case
2 Design
2.1 Design outputs
2.2 Design variable descriptions
2.3 Design data collection methodology
2.4 Design frame & sample methodology
2.5 Design statistical processing methodology
2.6 Design production systems & workflow
3 Build
3.1 Build data collection instrument
3.2 Build or enhance process components
3.3 Configure workflows
3.4 Test production system
3.5 Test statistical business process
3.6 Finalize production system
4 Collect
4.1 Select sample
4.2 Set up collection
4.3 Run collection
4.4 Finalize collection
5 Process
5.1 Integrate data
5.2 Classify & code
5.3 Review, Validate & edit
5.4 Impute
5.5 Derive new variables & statistical units
5.6 Calculate weights
5.7 Calculate aggregates
5.8 Finalize data files
6 Analyse
6.1 Prepare draft outputs
6.2 Validate outputs
6.3 Scrutinize & explain
6.4 Apply disclosure control
6.5 Finalize outputs
7 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
8 Archive
8.1 Define archive rules
8.2 Manage archive repository
8.3 Preserve data and associated metadata
8.4 Dispose of data & associated metadata
9 Evaluate
9.1 Gather evaluation inputs
9.2 Conduct evaluation
9.3 Agree action plan

Success Leads to Success…

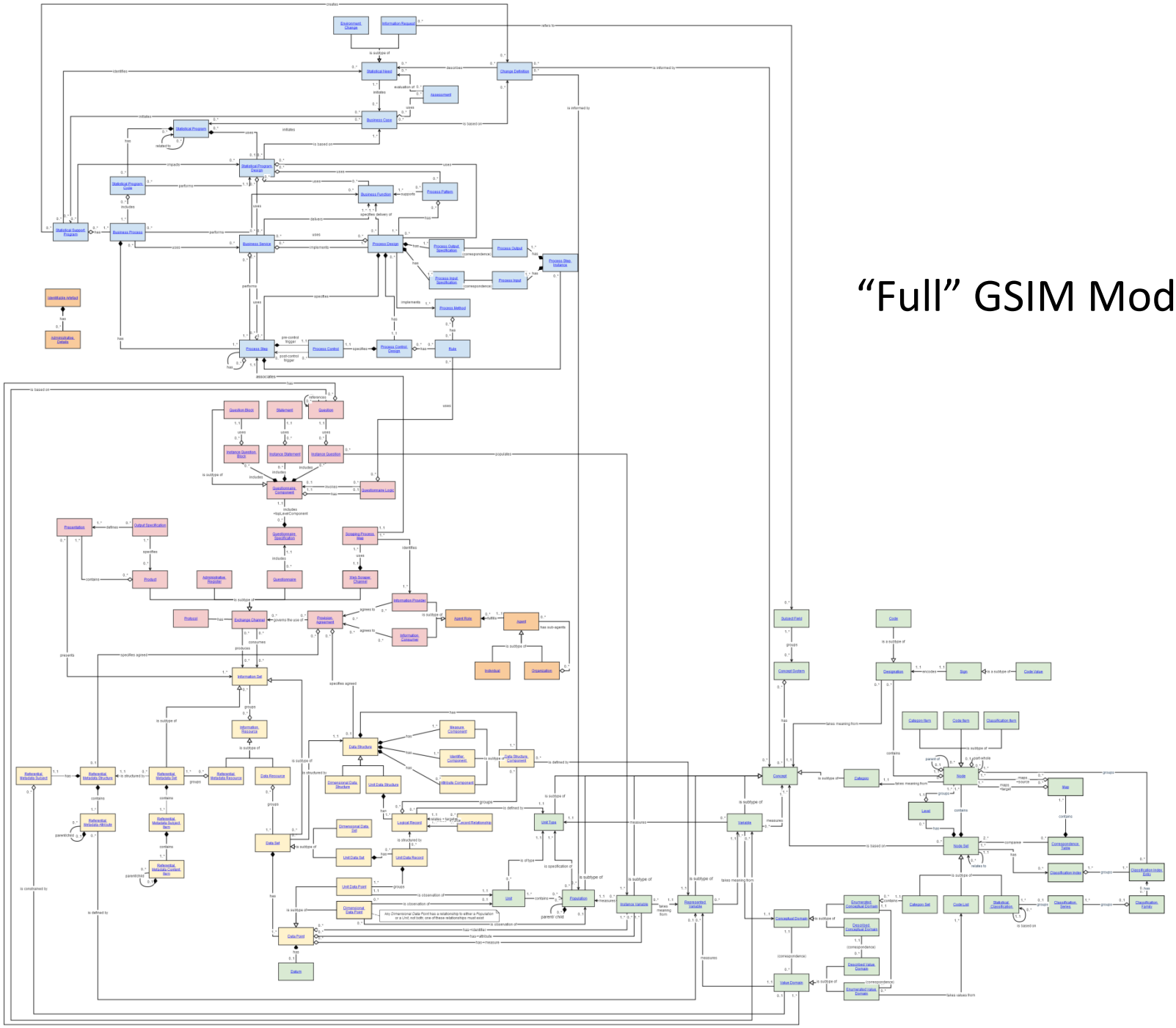
- Following the GSBPM model. The HLG decided to create an *information model…*
- GSIM – the Generic Statistical Information Model
  - One year to produce version 1.0 in a series of "sprints"
  - Version 1.1 delivered December 2013
  - A reference model for data and metadata

What is GSIM?

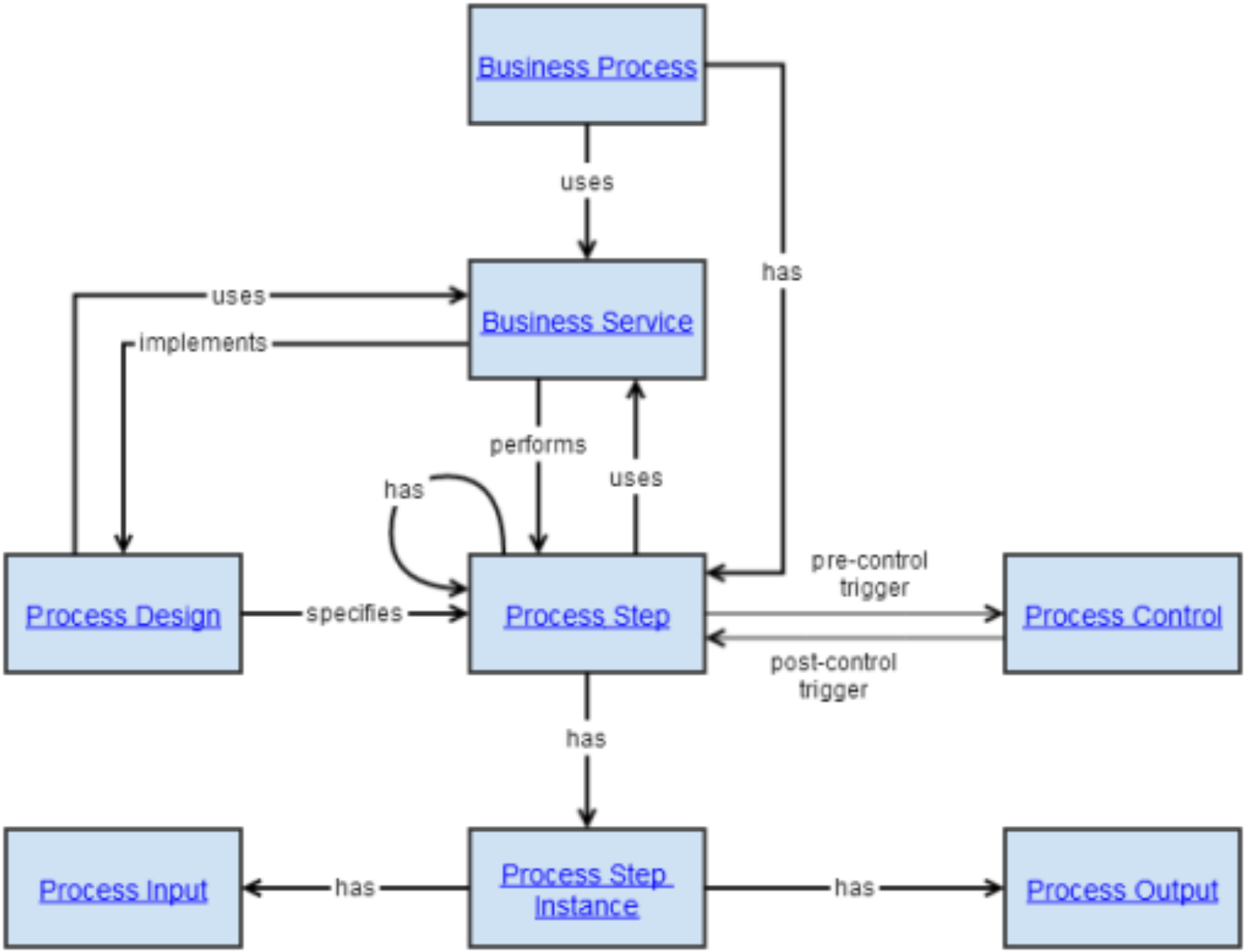


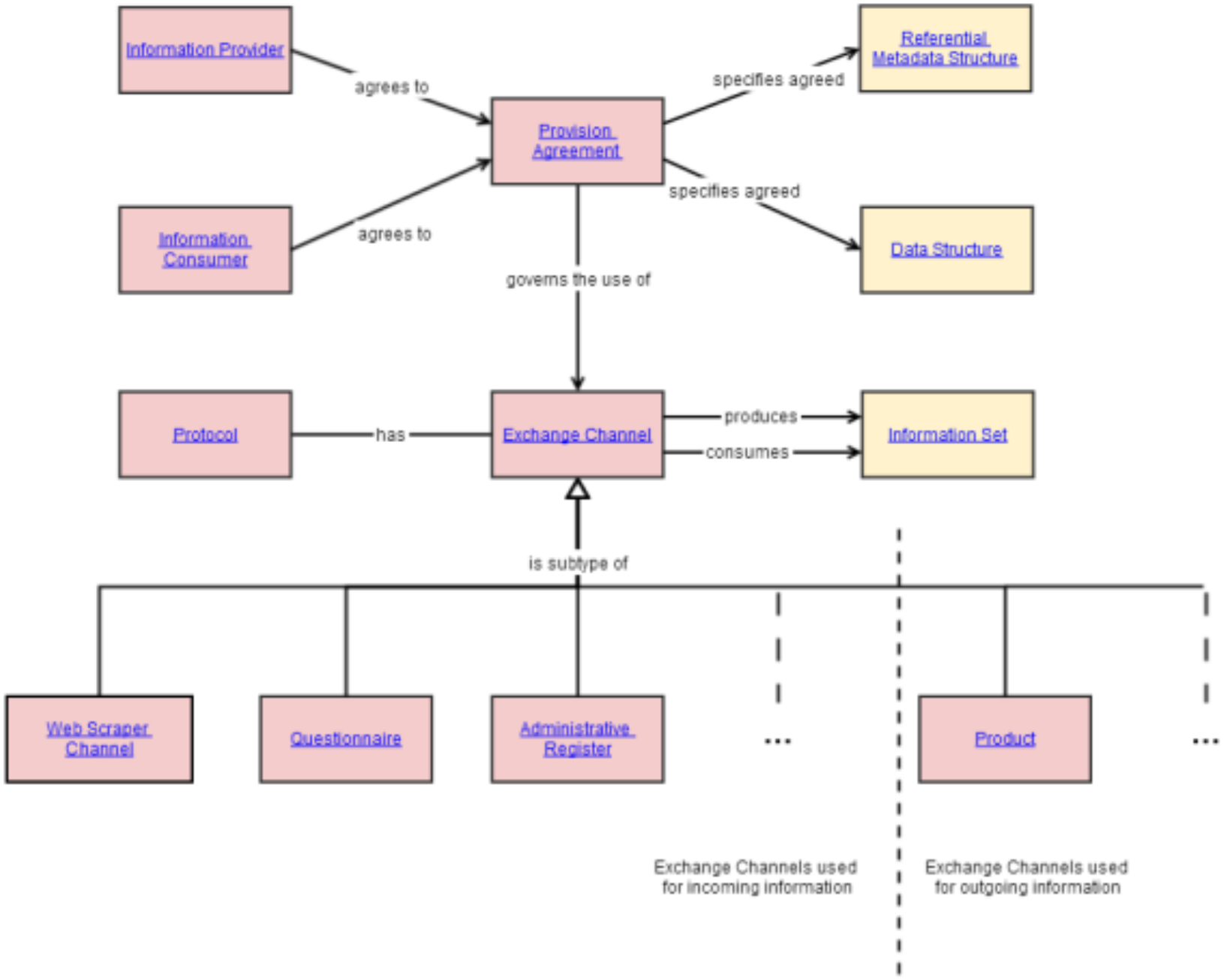
Version 1.1

- Version 1.1 re-organized and simplified the model:
  - Task-based models
    - Identifying statistical needs
    - Managing statistical programs
    - Designing and running statistical processes
    - Collecting, processing, and disseminating statistical information
  - "Foundational" information
    - Concepts, populations, codelists and classifications, variables, data sets, quality metadata
    - Neuchatel Classification model is now part of GSIM
- DDI 4 is aligning with the GSIM model
  - Some parts of DDI 3.2 were influenced by GSIM



“Full” GSIM Model


Business Process
uses
has
uses
Business Service
implements
performs
uses
has
Process Design
specifies
Process Step
pre-control trigger
Process Control
post-control trigger
has
Process Input
has
Process Step Instance
has
Process Output


Information Provider
agrees to
Provision Agreement
specifies agreed
Referential Metadata Structure
Information Consumer
agrees to
specifies agreed
Data Structure
governs the use of
Protocol
has
Exchange Channel
produces
consumes
Information Set
is subtype of
Web Scraper Channel
Questionnaire
Administrative Register
...
Product
...
Exchange Channels used for incoming information
Exchange Channels used for outgoing information

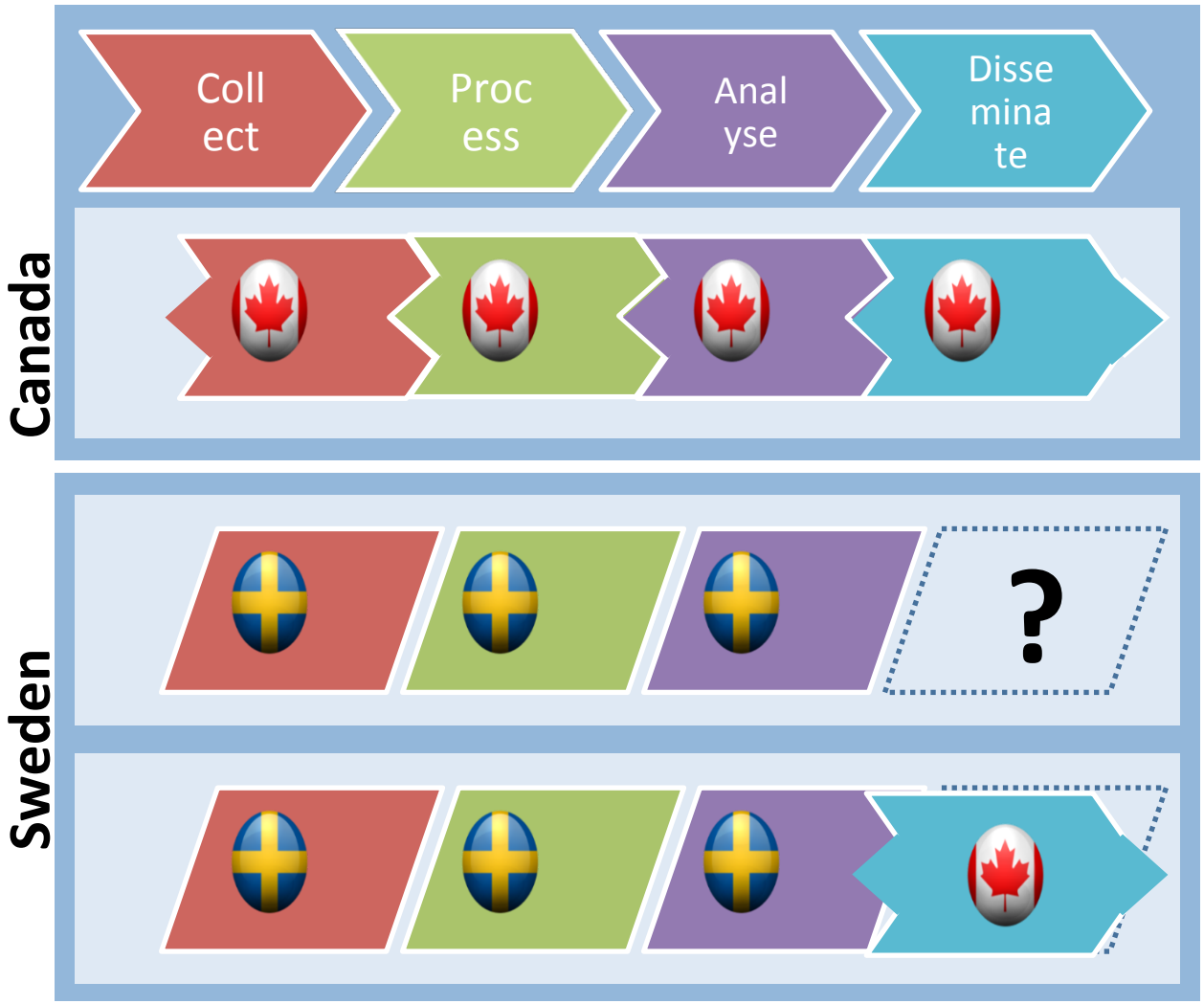
Resources

- There is an easy-to-use "Clickable GSIM" available on the UN/ECE site:
  - http://www1.unece.org/stat/platform/display/GSIMclick/Clickable+GSIM
  - Users can add "views"
- Specification is available
- Includes some profiles of DDI for implementing GSIM (more to come)

Statistical Architecture

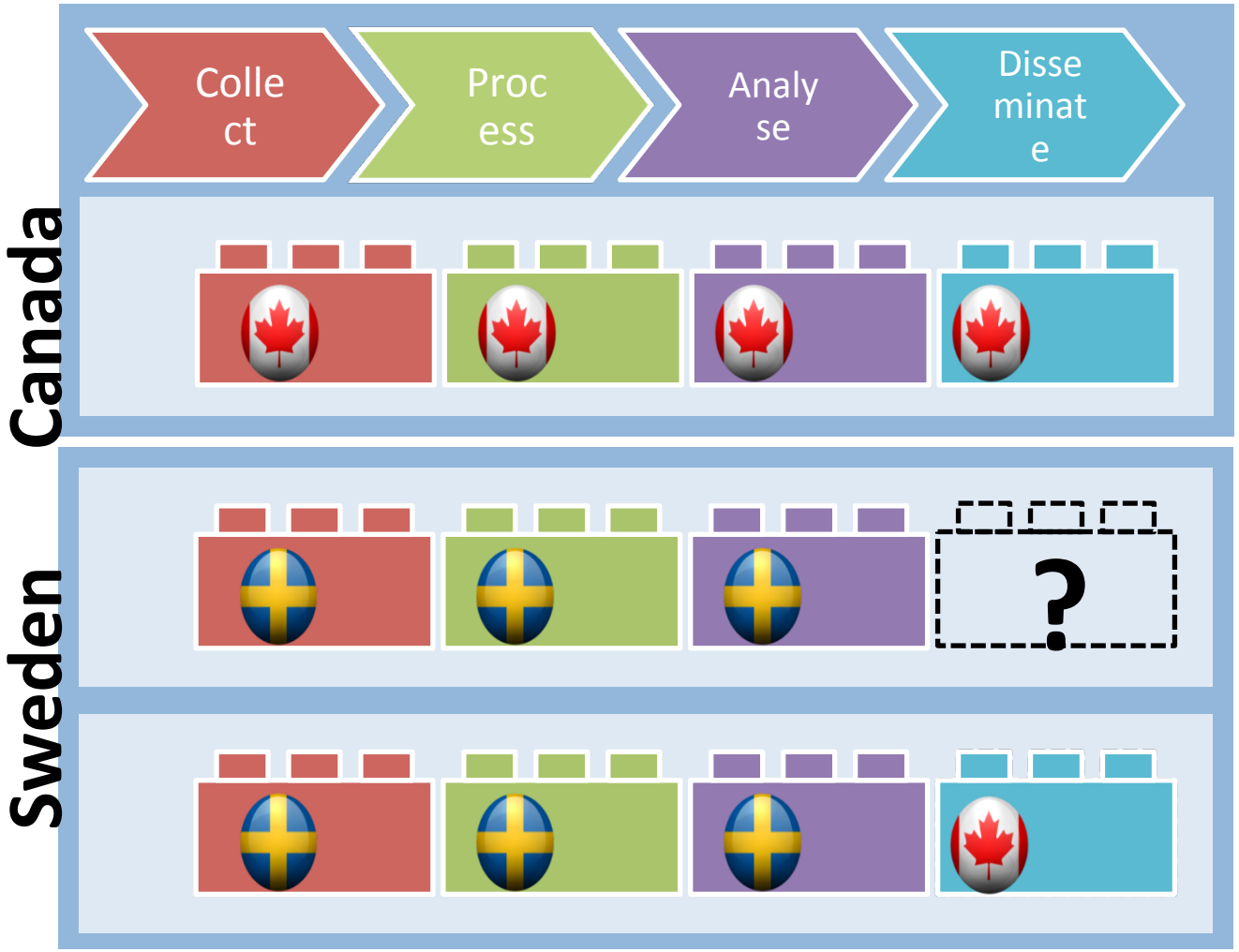
- To increase efficiency, statistical agencies want to share IT development resources by sharing services
- CSPA – Common Statistical Production Architecture
  - "Plug and play" services using standard interfaces
  - Standard architecture model was developed

This makes it hard to share and reuse!



…but if statistical organisations work together?

This makes it easier to share and reuse!



CSPA Prototype

- A number of shared services were developed in 2013
  - Based on DDI interfaces
  - “Wrapping” existing production proceses
- The prototype was a success!
  - Feedback to DDI is being incorporated into the DDI 4 development

Looking Forward

- HLG has now focused on two areas:
  - Big Data (from a statistical office perspective)
  - More CSPA services
    - Several are now being developed, based not only on DDI but also on SDMX
- DDI-related work is now being coordinated by the Modernization Committee on Standards
  - Emphasis is placed on having standard profiles and mappings from GSIM to DDI
  - Work now being organized for 2014
  - DDI Alliance is represented on this committee

Questions?