Data as a Service

Telling Canada's
story in numbers

CANADA 150

The role of international standards in integration, sharing, and safeguarding
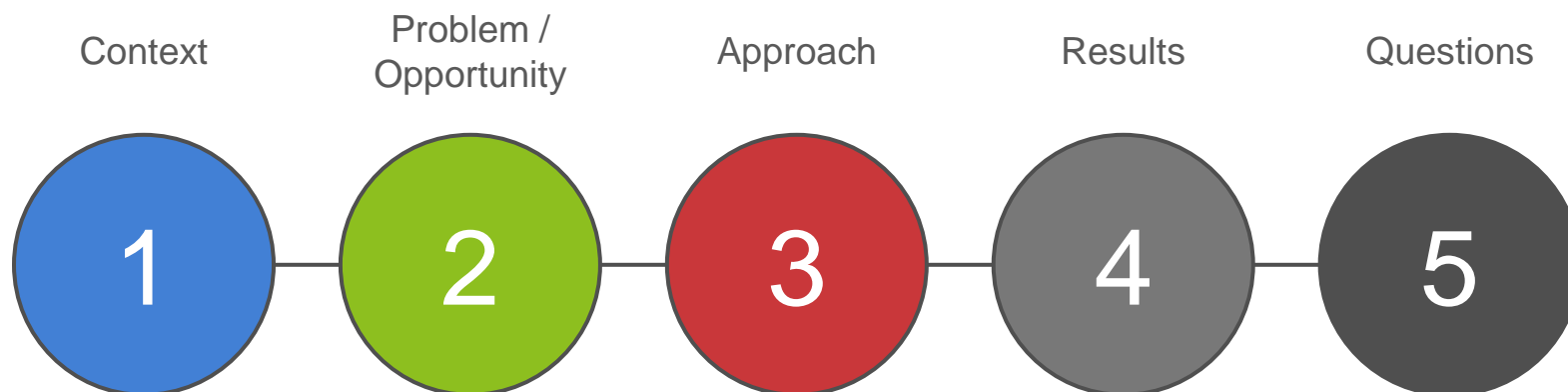Robert McLellan
Strategy, EA, and Innovation Group

Jan 29, 2018

Canada

Pathway

Context	Problem / Opportunity	Approach	Results	Questions
1	2	3	4	5

Business context

GC Digital Principles

- Understand users and their needs
- Iterate and improve frequently
- Build the right team
- Build a service-oriented culture
- Work in the open
- Integrate proportionate security and privacy from the outset
- Build in an open and interoperable way

- Use the right tools for the job
- Design and deliver transparent and ethical services
- Be inclusive and provide support for those who need it
- Know your data
- Be accountable to Canadians
- Develop open and innovative partnerships
- Spend money wisely

@ Jan 8, 2018

Statistics Canada Context

Statistics Canada Context

10 V's of data use

Volume — Size of Data

Velocity — The Speed at which Data is Generated

Variety — Different type of Data

Veracity — Data Accuracy

Value — Useful Data

BigData

Validity — Data quality, Governace, Moster Data Management on Massive

Variability — Dynamic, Evolving Behavior in Data Source

Venue — Distributed Heterogeneous Data from Multiple Platforms

Vocabulary — Data Models, Semantics that describes data Structure

Vagueness — Confusion over Meaning of BigData and Tools used

Source: Xenonstack.com

What's your strategy?

Business Strategy

By 2026, we will implement a register-based or -augmented Census.

We will shift surveys to be more administrative data based.

We will implement a core set of 4+ registers.

We will use a governed bimodal (IT) strategy to achieve agility and speed while maintaining quality and reliable production.

We will leverage non-traditional sources of data such as web-scraping and IoT devices to deliver relevant statistical products.

Business Outcomes

Agility - responsive answers to the tsunami of surveys

Flexibility - new methods and approaches to creating statistical products

Cost effective - alternative sources of data and e-self-response replace more costly collection approaches

Innovation - new statistical products to meet emerging needs

Talent acquisition - as an innovative, progressive employer we attract leading data scientists, MA, CS, EC

Relevant - we continue to be the trusted source of traditional statistics and we are at the vanguard of new statistics such as Census of Environment

Develop new data sources

- Enhanced legal framework
- New data source discovery
- Data Quality, cleaning, Preparation
- Data Security, Privacy
- Source data, metadata management
- "Internet of Things" (IoT) device exploration
- New sensor sources (web-scraping, drone survey, …)

Governance and Portfolio Management

- New Statistical Product Introduction Process
- Bimodal management of "Idea to Product" Value Chain
- Agile investment and experimentation review processes

Create Data Science R&D

- Linked Data, Metadata
- Data design and modeling
- Develop visual literacy
- Data visualization
- Natural Language Processing
- Algorithm exploration, validation
- Machine Learning
- Agile experimentation (Mode 2)
- Data Science Lab / Sandbox

Develop flexible IT enablers

- Information and Data Architecture
- Evolve database platforms - SQL, NOSQL, Graph DB, In-Memory
- Languages - R, Python, SAS, JavaScript, C++, SQL, .NET, Hive, Pig, Spark
- Distributed computing - SAS Grid, Hadoop
- Cloud Services acquisition and management – agile capacity on demand
- Data Analytics, Lakes, ETL, and Warehousing

Management Questions

How do we prioritize investments in the people, processes, information, technology against production (Mode 1) activities?

What is the availability of relevant data sources and what are the challenges to acquisition?

Notes

Requires completion of procurement path for cloud services (SaaS, PaaS, IaaS) at the GC level.

Skillset expansion to include languages, approaches, analytic skills, critical though, visual literacy, new database techniques.

Human Capital and Knowledge Development

- Design thinking
- Grow skills in critical thinking, modeling, hypothesis formulation, experimentation
- Build Talent Pool - recruit, train, assess, inspire
- Curated knowledge management and sharing
- People, project, team, subject, collaboration, communications spaces
- External collaboration and partnered development – GC, Academia, International

Success Measures

Reduce new survey "idea to production" cycle by x%

"Idea to Pilot" average interval reduced to 6 months or less

Data science community established (target population: 50)

Census 2026 Blueprint by 2019

Administrative Data Management covering 100% of data sets and instances

Reduce CAPI / CATI interviewing by 20%

New statistical products refined by or created from Big Data sources

Initiatives

Census 2026 White Paper

Innovation Lab

Cloud Services and roadmap

Big Data Community of Practice (CoP)

Data Science co-ops

SAS Grid

Data Visualization / Visual Analytics

Big Ideas Pilots

Integrated Methodology, IT, EC R&D

ISN Person Register

Web Scraping

Picasso

Business Intelligence and Analytics Technology Centre (BIATC)
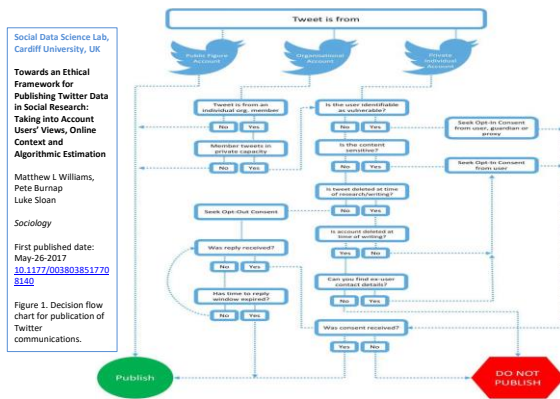
Sample strategy for big data

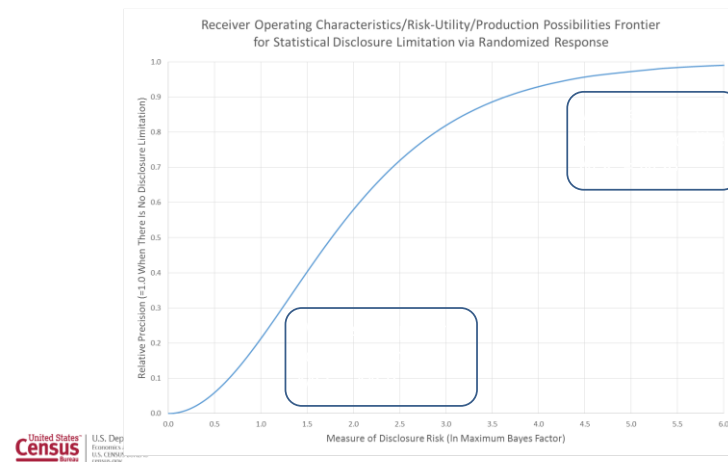# Social License to operate - ethics, privacy, transparent use

Data Ethics and the Public Good

National Statistician's Data Ethics Advisory Committee (NSDEC)

Source: Presentation by Peter Fullerton, ONS. Oct 2017

Source: The Challenge of Scientific Reproducibility and Privacy Protection for Statistical Agencies. Sept 15, 2016

Social Benefit vs Privacy Loss (cost)
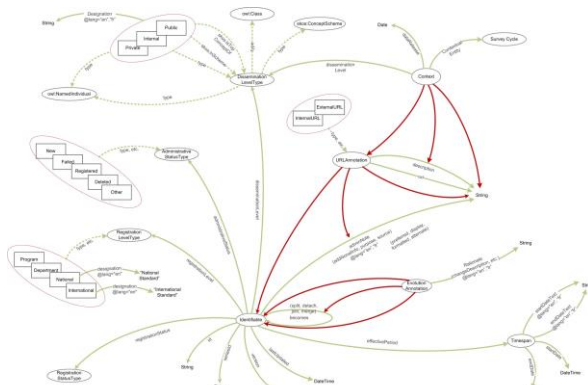
Current Activities – International community

Principles

Architectures

Knowledge representation –
RDF (semantic web)

Provenance

ModernStats
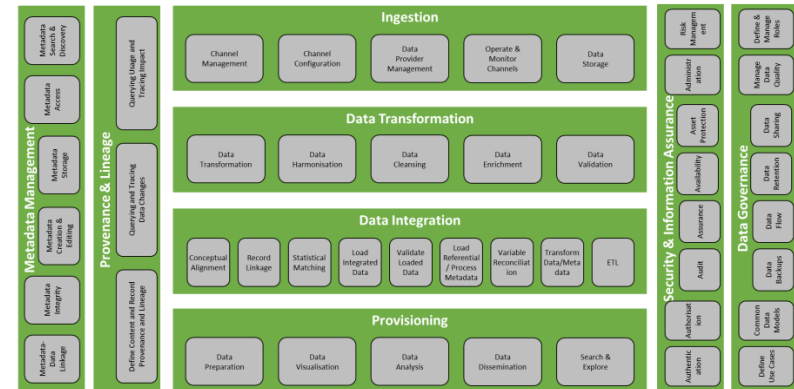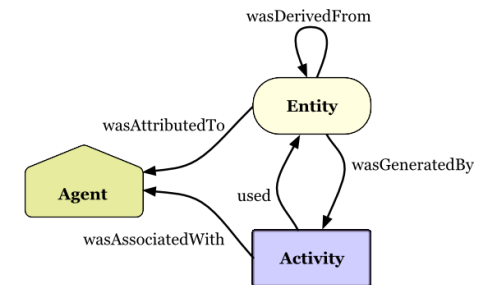
EIA Context

Enterprise Information Architecture Conceptual Components in Context (v.0.7)

Business vision
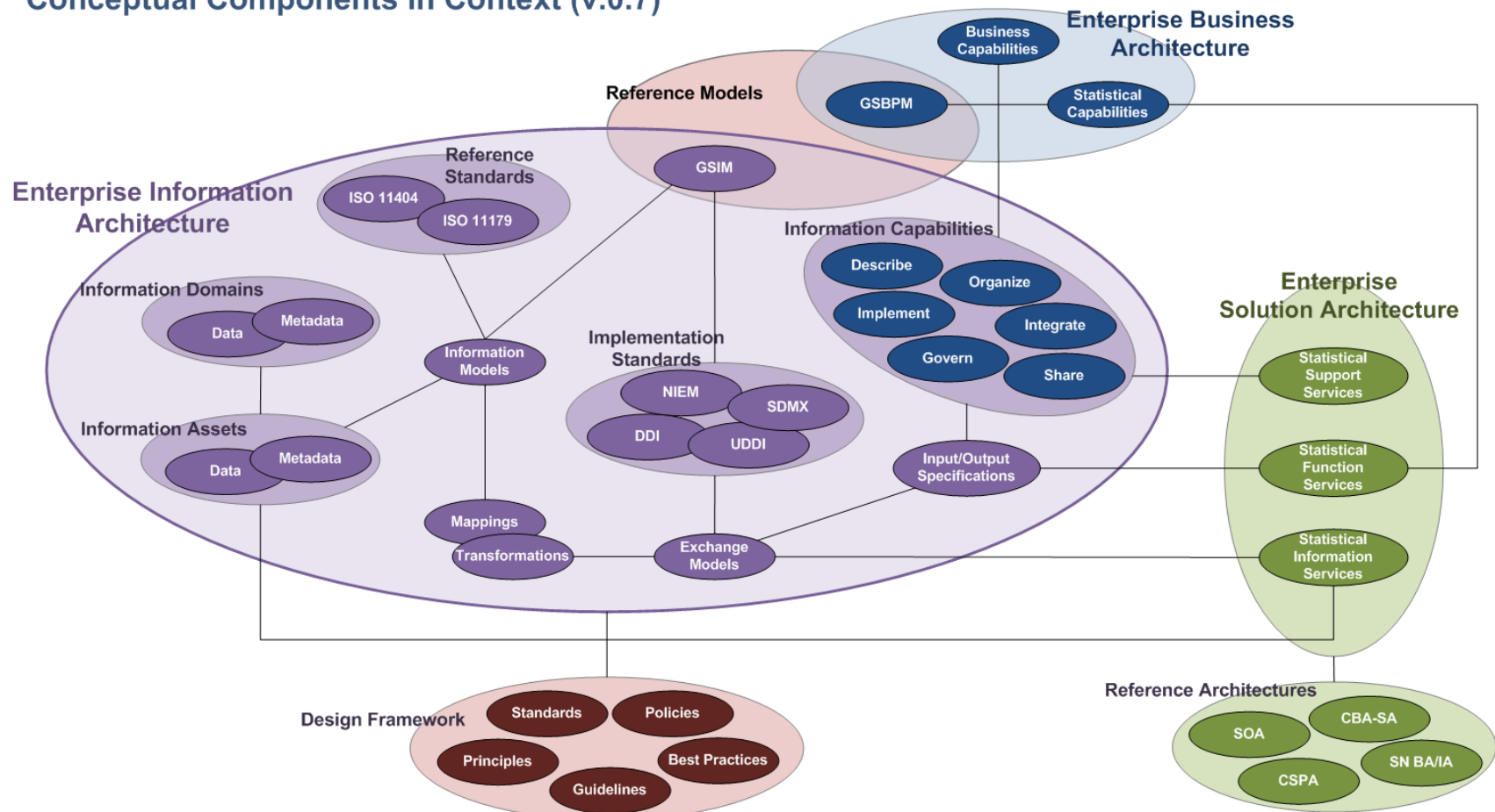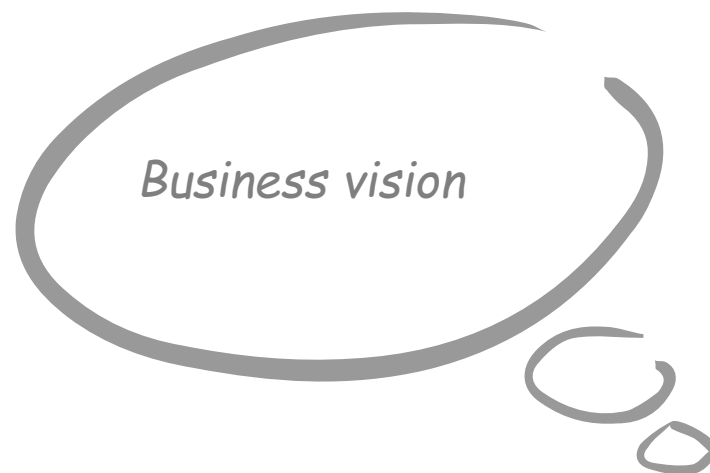
Who benefits from efficient sharing and safeguarding? How?

The Value Proposition Canvas
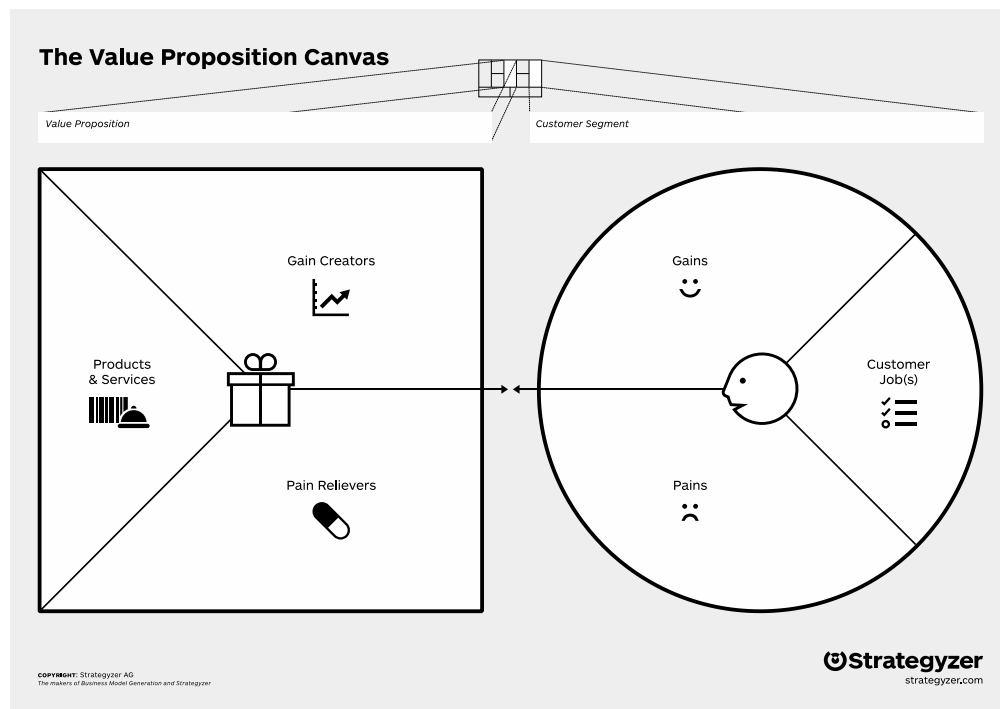
- Who are the beneficiaries?
  - Citizens
  - Partners
  - Peers
  - Data producers and consumers
  - Value-added data integrators
  - Policy and Service Delivery
- Gain delivery vs Pain relief
- The value of frictionless services and experiences
- What makes this happen?
- What is the cost? (net benefit)

What is your business model for sharing?

The Business Model Canvas

Designed for: | Designed by: | Date: | Version:

- Key Partners
- Key Activities
- Key Resources
- Value Propositions
- Customer Relationships
- Channels
- Customer Segments
- Cost Structure
- Revenue Streams

- Who are your partners ?
  - Other public sector
  - Private sector
  - Crowdsourcing
  - Data aggregators
- What are your key consumer, provider relationships?
- Producer / consumer segments?
- What are your input / output channels?
- Cost structure, cost sharing, value delivery

What do people want? – Virtual Data Labs

ANALOG – TODAY	DIGITAL – TOMORROW
• Data discovery – manual process, phone / email / in-person dialogue • Researcher access – manual application process (email / non-digital) • Analysis activities – physical locations only (28 across Canada) – local desktop machine, servers – travel required if not in location • Analysis tools – limited (no BYO-apps) • Data integration – limited ability to BYO data, access additional data • Scalability – limited by local physical capacity • Time-to-value – limited by process inertia, access overhead, review-to-release process • Policy framework – access to social, not commercial data only	▪ Data discovery – rich data, metadata catalogue with search / discover – self-serve ▪ Researcher access –digital workflows ▪ Analysis activities – secure remote access to virtual data labs (cloud hosted) ▪ Analysis tools – rich catalogue of SaaS, PaaS offerings (cloud), open-source, BYO-apps ▪ Data integration – shared trust data integration, batch / interactive linking, exploration supporting BYO-data ▪ Scalability – on-demand through cloud provisioning model – flow-through costing for researcher self-serve ▪ Time-to-value – accelerated, with continuous process improvement through monitoring ▪ Policy enabled

Value delivery — **Transparent research facilities supporting analysis from "idea" to "output", with a self-serve business model, anytime / anywhere access, and self-managed costing**

Components

EXPERTISE
On-demand access to experts (subject matter, methods, IT)

DATA
Self-serve search, discovery, secure delivery and use

COMPONENTS
Data OS (Lakes), SaaS/PaaS cloud parts, open-source

ENVIRONMENT
Public (or Hybrid) cloud base
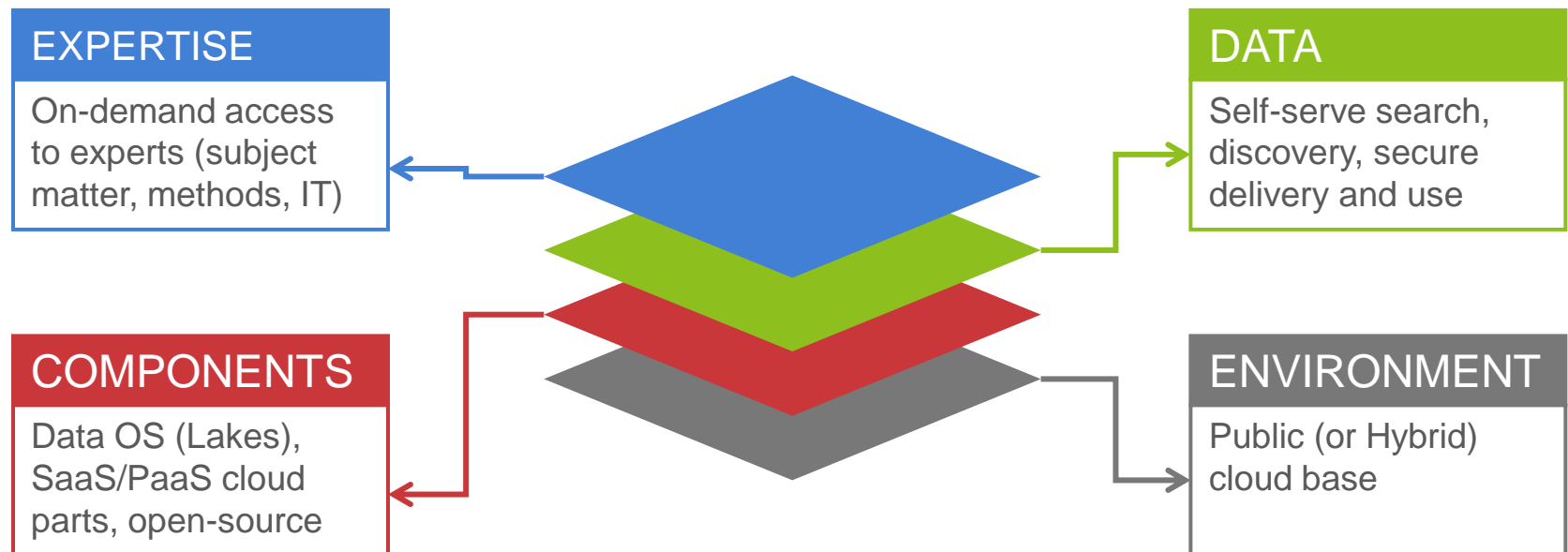
Real world example –
Metadata
Management

Pathway – "illustrative"

New metadata mgmt, data asset mgmt

Business capabilities

Conceptual information models

Solution architecture

Integration – services

Semantic relational hybrid.

Core solution

Future evolution

END
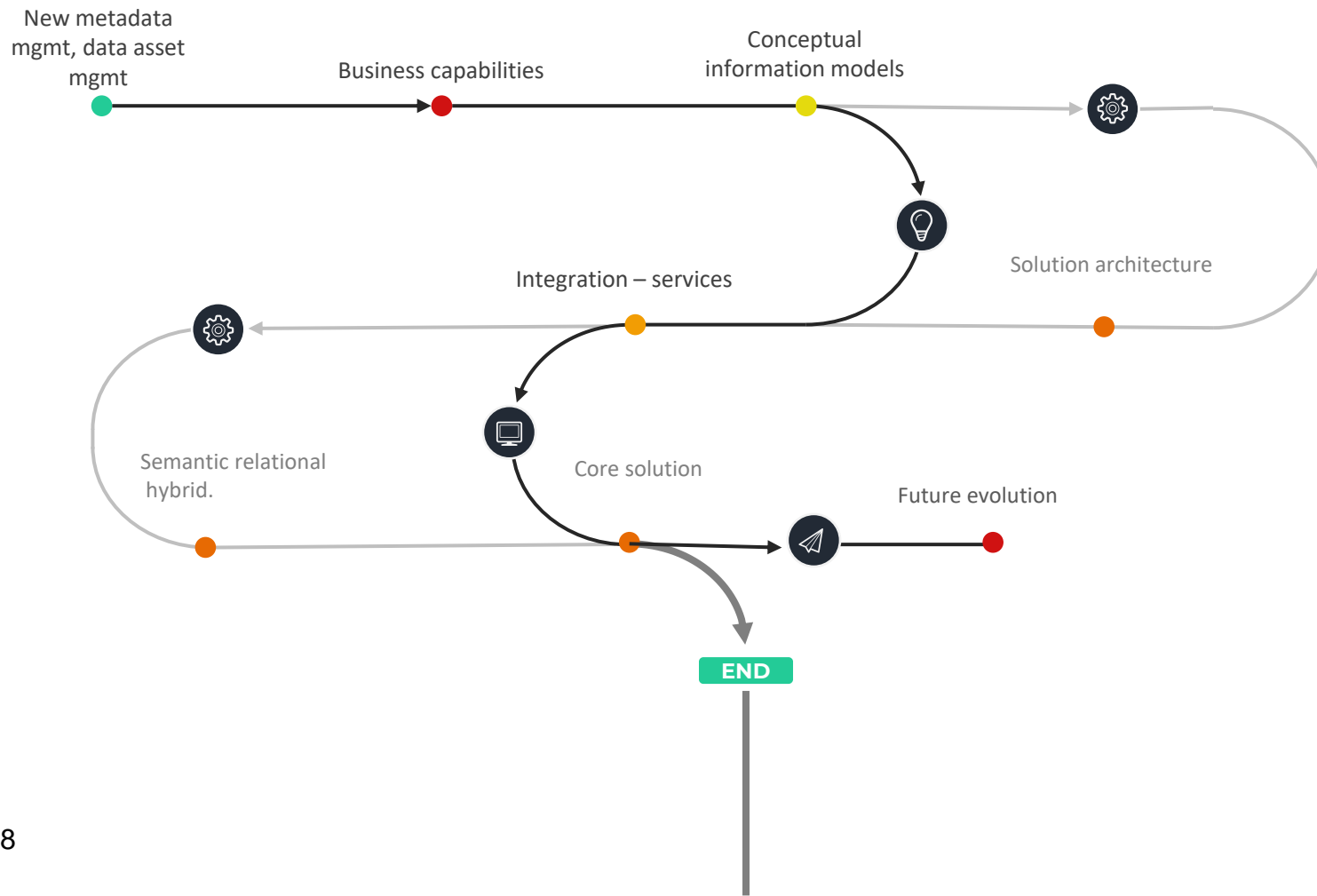
"Design your business" – Business Capabilities

- End user people roles
  - Metadata reviewers, approvers
  - Stewards
  - Producers, Consumers
- Processes
  - Metadata creation, update, delete
  - Metadata read, search
- Application / Technology
  - Metadata management solution
  - Metadata services (Q,S,C,V)
- Standards
  - GSIM, LIM
  - DDI
  - StatCan service standards

Business Capability

PEOPLE
- •Skills
- •Roles
- •Competence

METHODS
- •Algorithms
- •Procedures
- •…

PROCESSES
- •Tasks
- •Collaboration
- •Business rules

Application/Technology
- •Information systems
- •Information repositories
- •Communication networks

Standards
- •Service interface
- •Information models

19
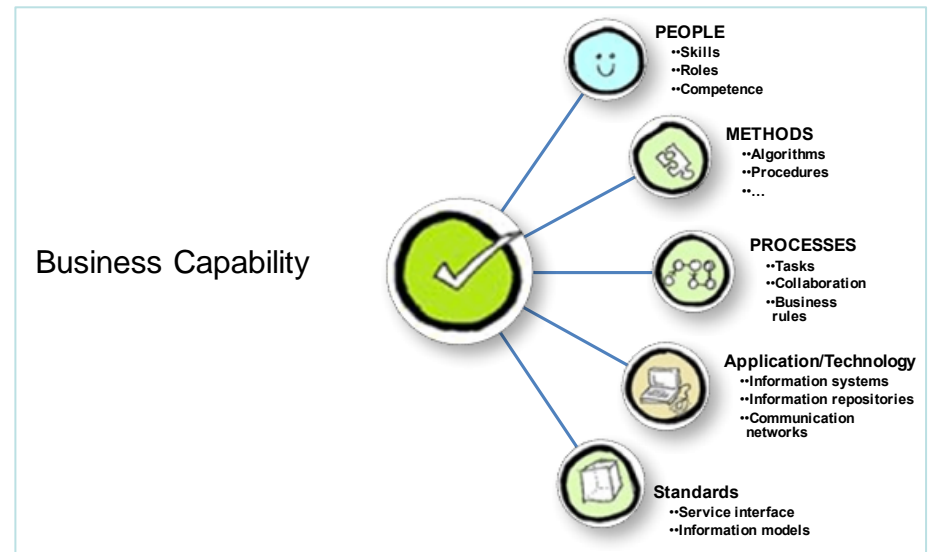
Conceptual Components implicated

Source Types

- Business Data
- Geo/Imagery
- Scientific Data
- Multimedia Data
- Data Streams
- Crowdsourced Data
- Big Data

Integrated Data Logistics

- Data Streaming
- Data Acquisition
- Data Ingestion
- Data Cleansing
- Data Movement

Data Consumption and Value Extraction

Data Mining	Data Prospecting	Data Visualization	Communities of Practice
High Performance Processing	Data Analysis	Machine Learning	Data and Software Collaboration

Data Management and Provisioning

Data Exploration	Search / Discovery	Governance	Auditing / Reporting
Data Transforming	Data Delivery / Subscriptions	Data Security	Privacy Management
Indexing / Cataloguing	Data Integration	Data Linkage	Data Supplier Management
Metadata Management	Data Enrichment	Data Management	Identity & Authorization

Data Lake (heterogeneous data formats and types)

Hybrid Cloud / Enterprise Data Centre Infrastructure

Data Users

Data Platform

Strategy Management

| Mainten. & consol. of strat. relations | Strategic Planning | Policy definition | PPM and budget def. |

Statistical Production Management

New statistics development	Statistical Design
Statistical Data collection	Information Resources Management
Statistical Processing	Statistical Analysis
Statistical Dissemination	Quality Assessment, Control & Improv.

Corporate Support Management

Quality Mgt.	Legal framework Mgt.
Security Mgt.	IT Management.
Administrative Information Mgt.	Human Resource Mgt.
Procurement Mgt.	Financial Mgt.

Source: Applying CSPA in the context of the ESS Vision 2020, Madrid 2016
Jean-Marc Museux, Eurostat

Statistical Production Management

New statistics development

- New data sources exploration
- Legislative work participation
- Statistical product innovation
- Methods and tools devevelopment for new statistics
- Identify user needs

Statistical Design

- (Re)Design statistical outputs
- Process & workflow design
- Process methods design
- Design production system, service and rules

Information Resources mgt

- Statistical Data mgt
- Metadata mgt.

Statistical Data Collection

- Provision agreement management
- Primary data collection
- Secondary data collection
- Metadata collection
- Statistical Registers Management

Statistical Processing

- Statistical data preparation
- New variables and units derivation
- Calculation and finalization of output

Statistical Analysis

- Statistical output analysis

Quality assessment, control & improv.

- Quality assessment
- Quality control mechanisms mgt.
- Quality improvement mgt.

Statistical Dissemination

- Release mgt.
- Products and services promotion
- Flexible data access provisioning
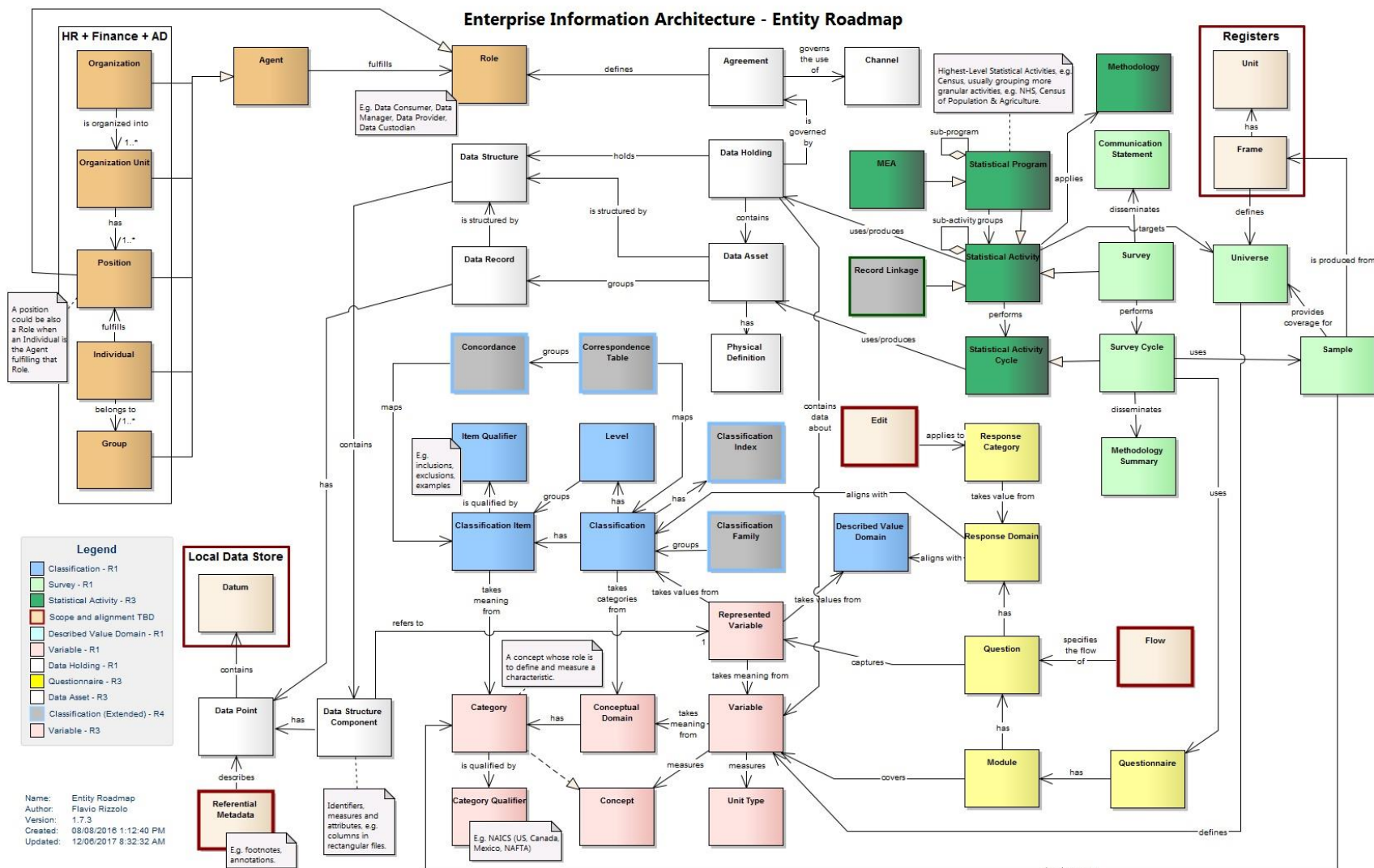- Statistical content mgt.

Generic Statistical Information Model
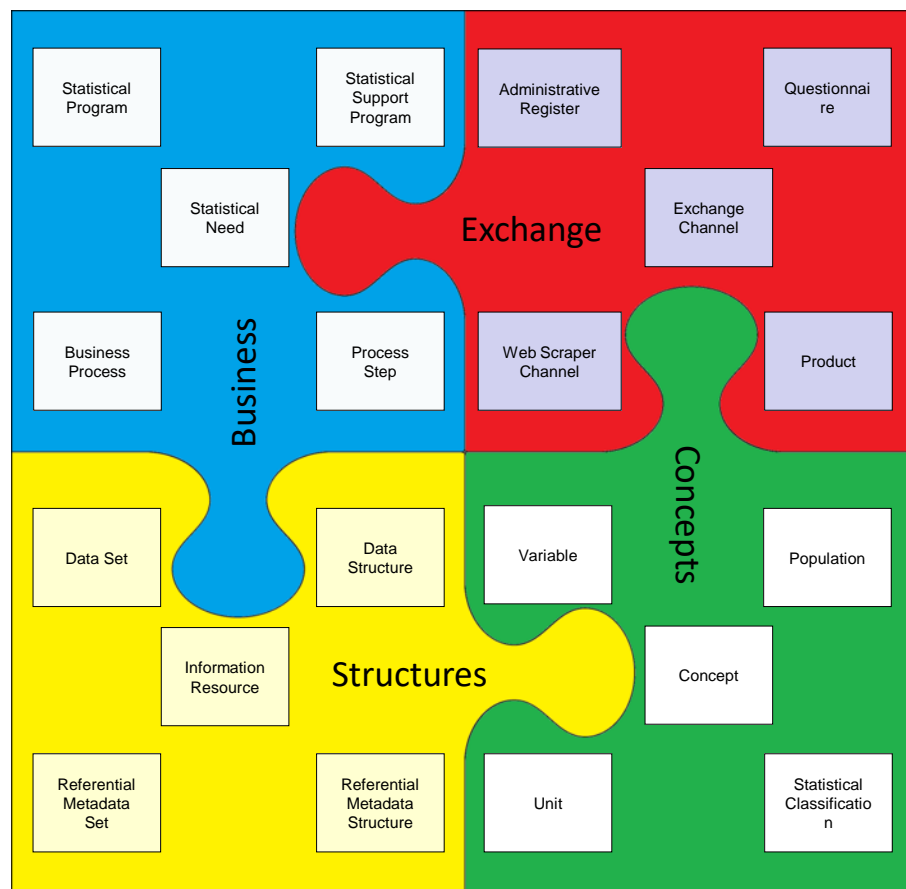
Information Architecture – Entity Roadmap (GSIM-based)

Enterprise Information Architecture - Entity Roadmap

Logical Models

Solution landscape

Local Survey Infrastructures
- Survey DB's
- Survey files
- Satellite Registers

Census Infrastructure
- Data Integration, Processing, Tabulation
- Census files
- Census Data

IBSP Infrastructure
- Data Integration, Processing, Tabulation
- Survey files
- IBSP DB

Common Tools Infrastructure
- PFM / SSPE
- Survey files
- SSME DB

National Accounts Infrastructure
- Data Integration, Analysis
- Survey files
- MEA Hub

Input Channels
- Web Scraping
- Sensor (IoT)
- Admin Files

Survey Collection
- MCS
- CMP
- Listing

Data Preprocessing
- Reception
- Coding
- QA & Validation
- Capture
- Transform

Processing
- Edit & Imputation
- Record Linkage
- Data Integration
- Data Exploration & Analysis
- Link Agreements

Core Register Infrastructure
- People
- Activity
- Structure
- Business / Institution
- Statistical Metadata (Picasso)
- Agency Link Register

Statistical Design
- Frames & Sampling
- Survey Design
- Questionnaire Design

Statistical Analysis
- Analysis (SAS)
- SAS
- Survey files

Innovation
- Data Science(R, etc.)
- Various
- Survey files

Analytical Studies
- Microsimulation, analysis
- Various
- Survey files

Dissemination (NDM)
- Web Presentation
- CODR
- OMR

Open Data Portal
- Web Presentation
- CKAN
- Data

Stakeholder Services
- StatCan Data Services Portal
- StatCan Reporting & Audit

Data Provisioning
- BI and Analytics
- Data Lakes
- Data Warehouses

Core Information Management Services
- Provision Agreements (LADMS)
- Access Authorizations (CARS)
- Authentication & Authorization (ActiveDirectory)
- EDRMS (GCDocs)
- Communities, Spaces & Collaboration (Confluence, STCWiki, GCConnex / Forum, ICN)

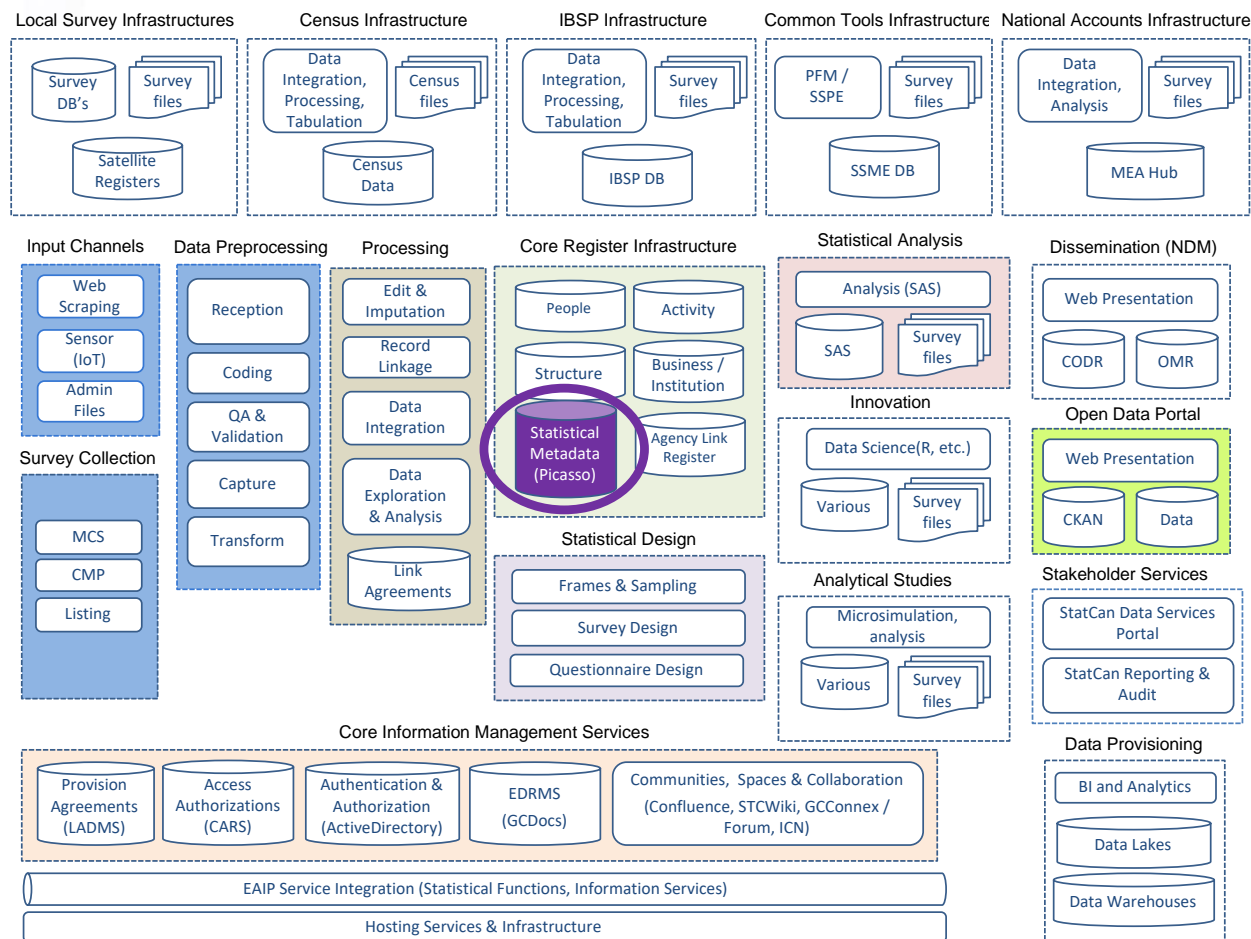EAIP Service Integration (Statistical Functions, Information Services)

Hosting Services & Infrastructure

Conversations and User Stories

- Subject Matter discussion – what metadata objects are being commonly used
    - Centres of Responsibility to standardize, direct the stewardship process definition
- Business analysis and user stories – metadata lifecycle
- User stories and UxD – how do people interact with the metadata?
- Client (presentation layer) design
- Service definition and implementation
- Data access layer definition and implementation
- Underlying persistence mechanisms (database)
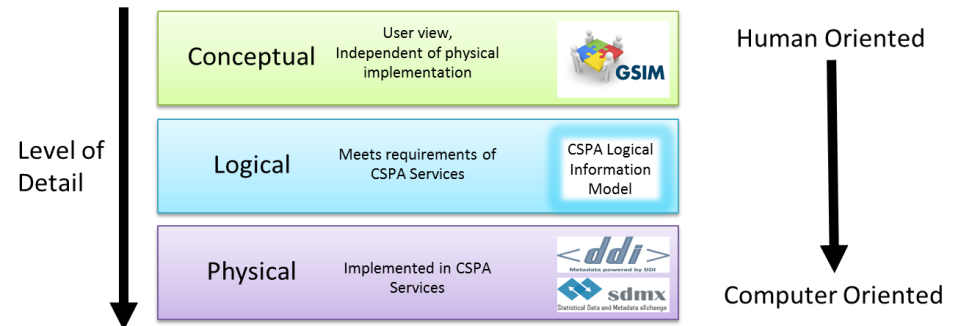
Service Catalogue

- Services implemented to standards in our Integration Technology Centre

- Enterprise Service Bus enabled

- Service Registry and Repository

- Service monitoring and logging

QUESTIONS ?

Contact

✉ robert.mclellan@canada.ca

Gccollab, gcconnex
LinkedIn
Slack

170 Tunney's Pasture Driveway
Jean Talon Building
Ottawa