Metadata Driven Statistical Data Management

Pascal Heus, Metadata Technology North America
pascal.heus@mtna.us | http://www.mtna.us

September 2019

About Pascal

- Metadata Technology North America (2009-Present)
  - Focus: Statistical Data & Metadata Management Solutions / Tools
  - Work with: NSOs (SNZ, ABS, StatCan), NORC (US federal / state agencies), Europe
  - R&D: Modern Technology, Metadata Standards, Data Quality
- Metadata Technology (UK) / Open Data Foundation (2007-2009)
- World Bank / International Household Survey Network (2000-2007)
  - Focus: Statistical Capacity Building, Metadata (DDI / SDMX), Tools development (IHSN toolkit)
  - Work with: NSOs (Africa), International Organizations (WB, UN, WHO, …), DDI Alliance
- Early years (< 2000)
  - Embassy of Belgium (1990), Ministry of Foreign Affairs
- Education:
  - MS in Computational Science from GMU (HPC / Quantum Computing)
  - Graduate in IT from UCL in Belgium (1987)
- http://www.linkedin.com/in/pascal

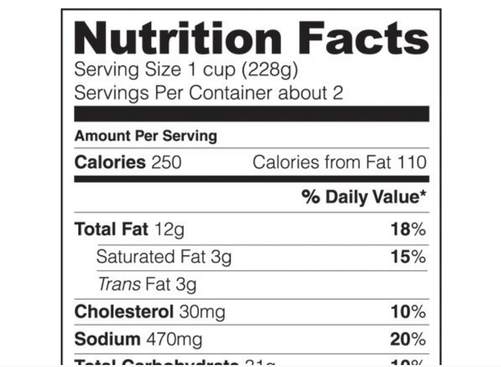
Agenda

- Part 1: Why Metadata? (10')
- Part 2: Metadata Models & Standards (15')
- Part 3: Architecture & Technologies (5')
- Part 4: Implementation (15')
- Q&A

Part 1: Why Metadata?

Everyday Metadata

100% NATURAL
Hunt's
FlashSteam FRESHNESS
Petite
Diced
NO Artificial Preservatives
Chef BOYARDEE
Spaghetti & Meatballs
Pasta and Meatballs made with Pork, Chicken and Beef in Tomato Sauce
RO★TEL
Original
DICED TOMATOES & GREEN CHILIES

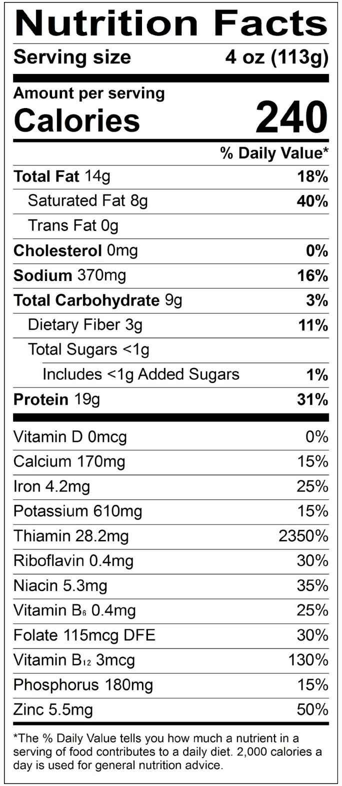


Nutrition Facts

Serving Size 1 cup (228g)
Servings Per Container about 2

Amount Per Serving

Calories 250	Calories from Fat 110
	% Daily Value*
Total Fat 12g	**18%**
Saturated Fat 3g	**15%**
Trans Fat 3g	
Cholesterol 30mg	**10%**
Sodium 470mg	**20%**
Total Carbohydrate 31g	**10%**

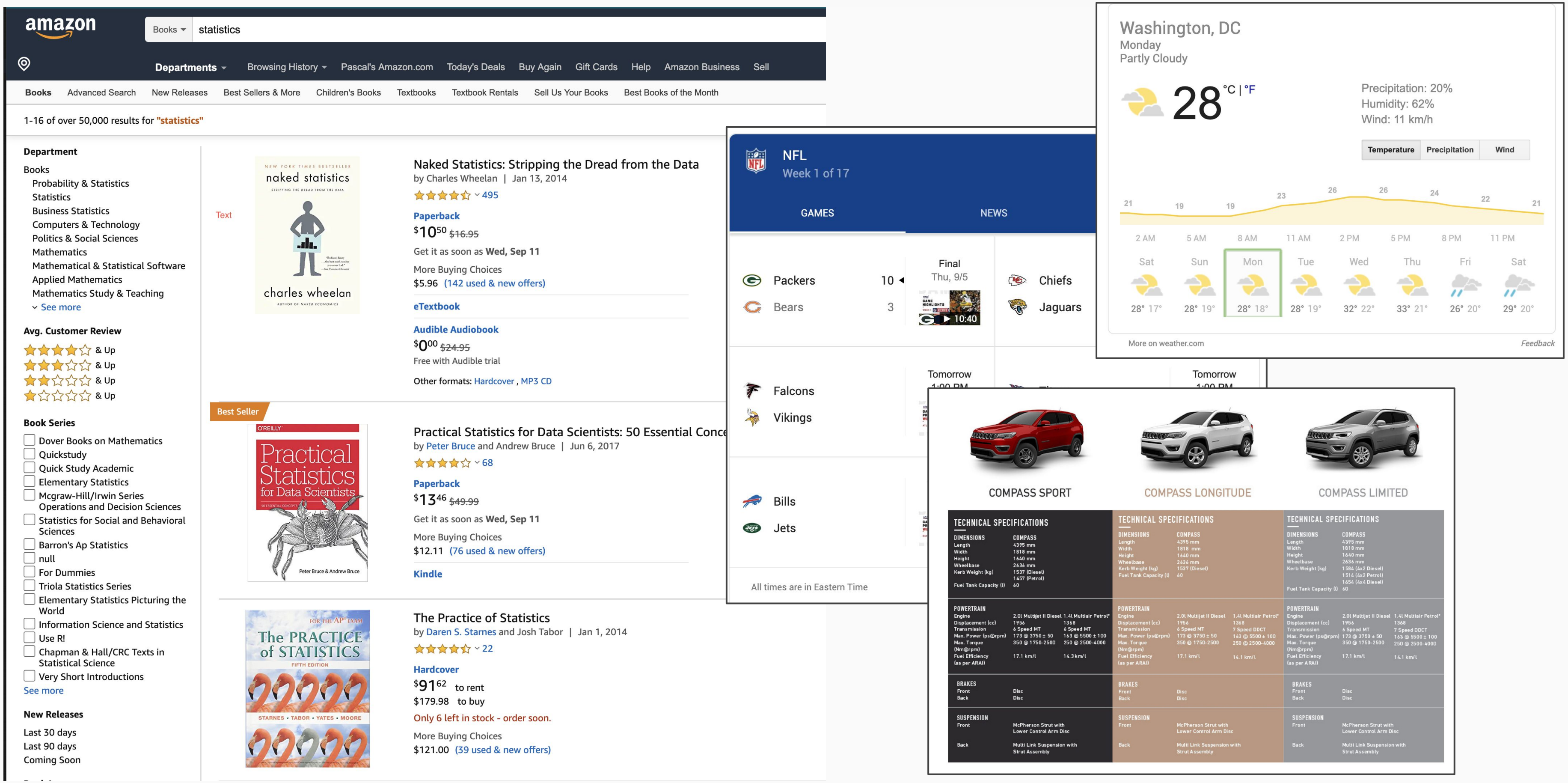


Nutrition Facts

Serving size	**4 oz (113g)**
Amount per serving	
Calories	**240**
	% Daily Value*
Total Fat 14g	**18%**
Saturated Fat 8g	**40%**
Trans Fat 0g	
Cholesterol 0mg	**0%**
Sodium 370mg	**16%**
Total Carbohydrate 9g	**3%**
Dietary Fiber 3g	**11%**
Total Sugars <1g	
Includes <1g Added Sugars	**1%**
Protein 19g	**31%**
Vitamin D 0mcg	0%
Calcium 170mg	15%
Iron 4.2mg	25%
Potassium 610mg	15%
Thiamin 28.2mg	2350%
Riboflavin 0.4mg	30%
Niacin 5.3mg	35%
Vitamin B_6 0.4mg	25%
Folate 115mcg DFE	30%
Vitamin B_{12} 3mcg	130%
Phosphorus 180mg	15%
Zinc 5.5mg	50%

*The % Daily Value tells you how much a nutrient in a serving of food contributes to a daily diet. 2,000 calories a day is used for general nutrition advice.

Everyday Metadata



amazon

Books statistics

Departments Browsing History Pascal's Amazon.com Today's Deals Buy Again Gift Cards Help Amazon Business Sell

Books Advanced Search New Releases Best Sellers & More Children's Books Textbooks Textbook Rentals Sell Us Your Books Best Books of the Month

1-16 of over 50,000 results for "statistics"

Department

Books
- Probability & Statistics
- Statistics
- Business Statistics
- Computers & Technology
- Politics & Social Sciences
- Mathematics
- Mathematical & Statistical Software
- Applied Mathematics
- Mathematics Study & Teaching
- See more

Avg. Customer Review

★★★★☆ & Up
★★★☆☆ & Up
★★☆☆☆ & Up
★☆☆☆☆ & Up

Book Series

- □ Dover Books on Mathematics
- □ Quickstudy
- □ Quick Study Academic
- □ Elementary Statistics
- □ Mcgraw-Hill/Irwin Series Operations and Decision Sciences
- □ Statistics for Social and Behavioral Sciences
- □ Barron's Ap Statistics
- □ null
- □ For Dummies
- □ Triola Statistics Series
- □ Elementary Statistics Picturing the World
- □ Information Science and Statistics
- □ Use R!
- □ Chapman & Hall/CRC Texts in Statistical Science
- □ Very Short Introductions

See more

New Releases

Last 30 days
Last 90 days
Coming Soon

Text

Naked Statistics: Stripping the Dread from the Data
by Charles Wheelan | Jan 13, 2014
★★★★☆ 495

Paperback
$10⁵⁰ ~~$16.95~~
Get it as soon as **Wed, Sep 11**
More Buying Choices
$5.96 (142 used & new offers)

eTextbook

Audible Audiobook
$0⁰⁰ ~~$24.95~~
Free with Audible trial

Other formats: Hardcover , MP3 CD

NEW YORK TIMES BESTSELLER
naked statistics
STRIPPING THE DREAD FROM THE DATA
charles wheelan
AUTHOR OF NAKED ECONOMICS

Best Seller

Practical Statistics for Data Scientists: 50 Essential Conce
by Peter Bruce and Andrew Bruce | Jun 6, 2017
★★★★☆ 68

Paperback
$13⁴⁶ ~~$49.99~~
Get it as soon as **Wed, Sep 11**
More Buying Choices
$12.11 (76 used & new offers)

Kindle

O'REILLY
Practical Statistics for Data Scientists
50 ESSENTIAL CONCEPTS
Peter Bruce & Andrew Bruce

The Practice of Statistics
by Daren S. Starnes and Josh Tabor | Jan 1, 2014
★★★★☆ 22

Hardcover
$91⁶² to rent
$179.98 to buy
Only 6 left in stock - order soon.
More Buying Choices
$121.00 (39 used & new offers)

NFL
Week 1 of 17

GAMES NEWS

Packers	10	Final Thu, 9/5	Chiefs
Bears	3	10:40	Jaguars
Falcons		Tomorrow 1:00 PM	
Vikings			
Bills			
Jets			

Tomorrow
1:00 PM

All times are in Eastern Time

Washington, DC
Monday
Partly Cloudy

28 °C | °F

Precipitation: 20%
Humidity: 62%
Wind: 11 km/h

Temperature Precipitation Wind

Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat
28° 17°	28° 19°	28° 18°	28° 19°	32° 22°	33° 21°	26° 20°	29° 20°

More on weather.com Feedback

COMPASS SPORT

TECHNICAL SPECIFICATIONS

DIMENSIONS	COMPASS	
Length	4395 mm	
Width	1818 mm	
Height	1640 mm	
Wheelbase	2636 mm	
Kerb Weight (kg)	1537 (Diesel) 1457 (Petrol)	
Fuel Tank Capacity (l)	60	
POWERTRAIN		
Engine	2.0l Multijet II Diesel	1.4l Multiair Petrol*
Displacement (cc)	1956	1368
Transmission	6 Speed MT	6 Speed MT
Max. Power (ps@rpm)	173 @ 3750 ± 50	163 @ 5500 ± 100
Max. Torque (Nm@rpm)	350 @ 1750-2500	250 @ 2500-4000
Fuel Efficiency (as per ARAI)	17.1 km/l	14.3 km/l
BRAKES		
Front	Disc	
Back	Disc	
SUSPENSION		
Front	McPherson Strut with Lower Control Arm Disc	
Back	Multi Link Suspension with Strut Assembly	

COMPASS LONGITUDE

TECHNICAL SPECIFICATIONS

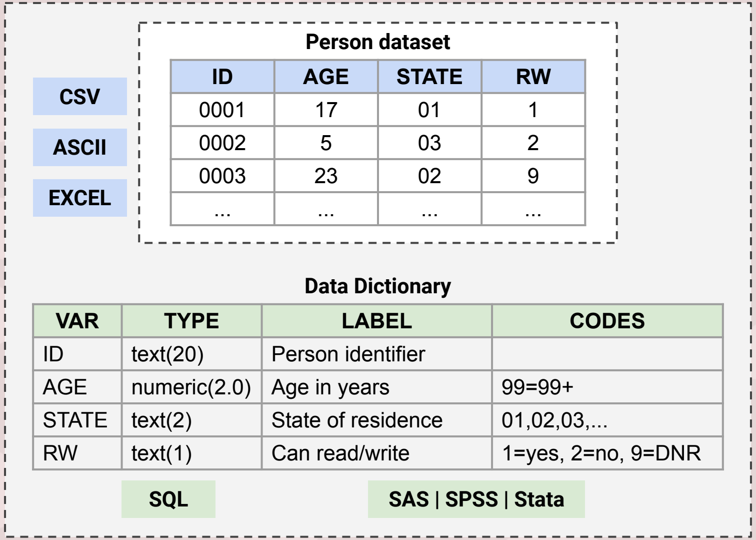
DIMENSIONS	COMPASS	
Length	4395 mm	
Width	1818 mm	
Height	1640 mm	
Wheelbase	2636 mm	
Kerb Weight (kg)	1537 (Diesel)	
Fuel Tank Capacity (l)	60	
POWERTRAIN		
Engine	2.0l Multijet II Diesel	1.4l Multiair Petrol*
Displacement (cc)	1956	1368
Transmission	6 Speed MT	7 Speed DDCT
Max. Power (ps@rpm)	173 @ 3750 ± 50	163 @ 5500 ± 100
Max. Torque (Nm@rpm)	350 @ 1750-2500	250 @ 2500-4000
Fuel Efficiency (as per ARAI)	17.1 km/l	14.1 km/l
BRAKES		
Front	Disc	
Back	Disc	
SUSPENSION		
Front	McPherson Strut with Lower Control Arm Disc	
Back	Multi Link Suspension with Strut Assembly	

COMPASS LIMITED

TECHNICAL SPECIFICATIONS

DIMENSIONS	COMPASS	
Length	4395 mm	
Width	1818 mm	
Height	1640 mm	
Wheelbase	2636 mm	
Kerb Weight (kg)	1584 (4x2 Diesel) 1514 (4x2 Petrol) 1654 (4x4 Diesel)	
Fuel Tank Capacity (l)	60	
POWERTRAIN		
Engine	2.0l Multijet II Diesel	1.4l Multiair Petrol*
Displacement (cc)	1956	1368
Transmission	6 Speed MT	7 Speed DDCT
Max. Power (ps@rpm)	173 @ 3750 ± 50	163 @ 5500 ± 100
Max. Torque (Nm@rpm)	350 @ 1750-2500	250 @ 2500-4000
Fuel Efficiency (as per ARAI)	17.1 km/l	14.1 km/l
BRAKES		
Front	Disc	
Back	Disc	
SUSPENSION		
Front	McPherson Strut with Lower Control Arm Disc	
Back	Multi Link Suspension with Strut Assembly	

And yet when it comes to data....



CSV

ASCII

EXCEL

Person dataset

ID	AGE	STATE	RW
0001	17	01	1
0002	5	03	2
0003	23	02	9
...	...	...	...

Data Dictionary

VAR	TYPE	LABEL	CODES
ID	text(20)	Person identifier	
AGE	numeric(2.0)	Age in years	99=99+
STATE	text(2)	State of residence	01,02,03,...
RW	text(1)	Can read/write	1=yes, 2=no, 9=DNR

SQL

SAS | SPSS | Stata

PDF / DOC / XSL

****** does not come with the data ******

****** hard to find for users ******

!!! out of reach of informations system !!!

Data: the missing knowledge problem

TIME POP GEO

2110 Survey of Settlers in Southern Mars Colonies

Variables
- Definitions and attributes
- Data Elements
- Changes over time

Survey
- Description
- Questionnaires / Instruments

Methodology / Technical Docs:
- Data Collection
- Data Processing
- Users Guides

Catalogs
- Producers, Archives, Portals

....

CSV

ASCII

EXCEL

Person dataset

ID	AGE	STATE	RW
0001	17	01	1
0002	5	03	2
0003	23	02	9
...	...	...	...

Data Dictionary

VAR	TYPE	LABEL	CODES
ID	text(20)	Person identifier	
AGE	numeric(2.0)	Age in years	99=99+
STATE	text(2)	State of residence	01,02,03,...
RW	text(1)	Can read/write	1=yes, 2=no, 9=DNR

SQL

SAS | SPSS | Stata

PDF / DOC / XSL

Concepts
- This is how we think or search
- Almost metadata element relates to concept(s)

Classifications
- Not only Code/Categories...
- Definition/Incl./Excl
- Versions / Changes over time
- Levels (concepts)
- Concordances
- Associations with concepts

Provenance
- Producers, Publisher
- Version

....

****** does not come with the data ******

****** hard to find for users ******

!!! out of reach for informations system !!!

Metadata / Knowledged Persistence

People friendly vs machine actionable

<u>PEOPLE</u>

- Happy with unstructured format (can read)
- Storage
  - Documents
  - Emails / Messages
  - Conversations
  - Brains
- Low to very high volatility
  - People move or retire...
- Often change mind or have different perspectives ("models")
- Not in the habit to share knowledge with information system

<u>INFORMATION SYSTEMS</u>

- Need semi-structured formats & models
- Storage (NOSQL)
  - XML
  - JSON
  - RDF
  - Machine Learning / AI
- Highly persistent
  - Ensure Institutional memory
- Works with stable information models (e.g. standards)
- Happy to deliver knowledge back to users in friendly / favorite formats (no loss)

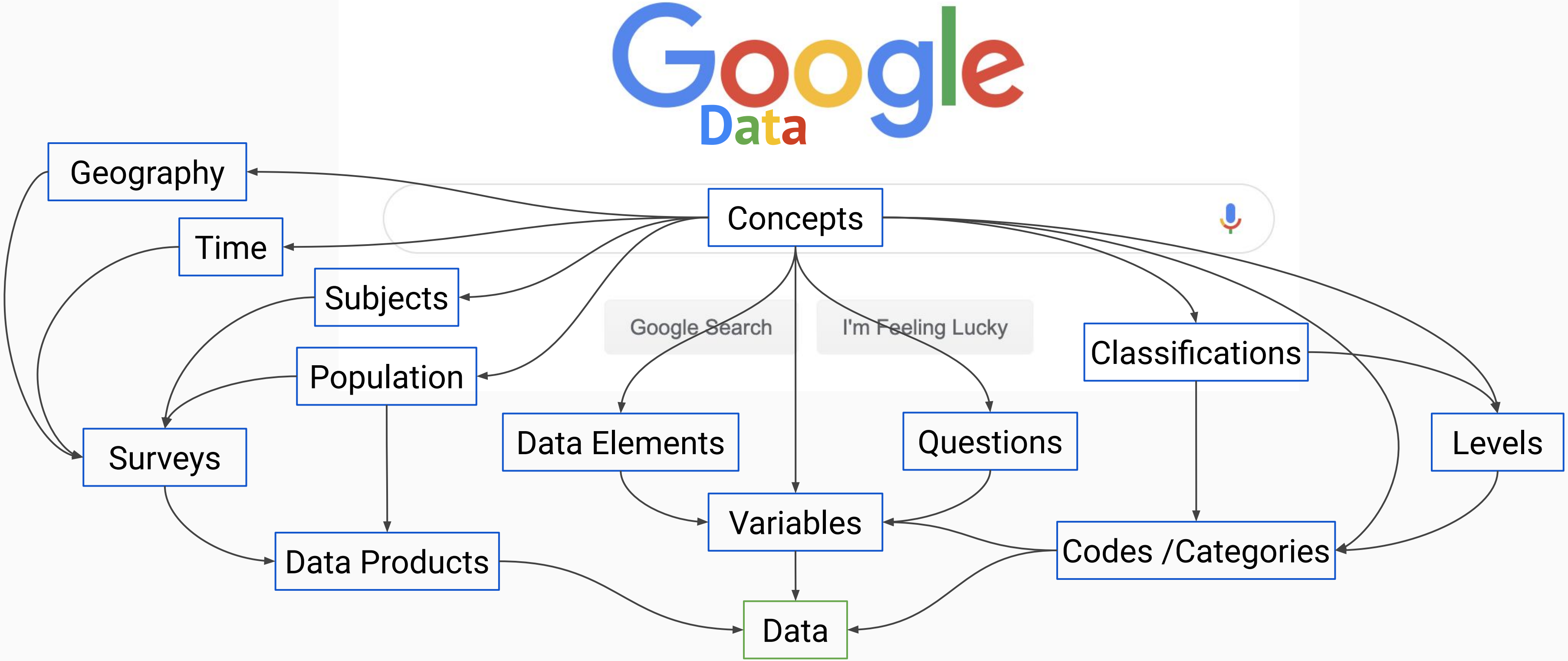
metadata / knowledge handover

Why empower information systems with metadata?

- Task automation
  - production, dissemination, analysis, repurposing, exploration
- Ensure and enhance quality
  - Data meets expectations (not the other way around, after the fact metadata)
- Enables search & discovery
  - Portals, Data Google
- Reduce data wrangling!
  - 70%+ of user time spend on data hunting, cleaning, formatting, transforming, linking,...
- Open data packaging
  - Ready to use, extract, formatting, conversion, etc.
- Achieve data/statistics as a service
  - Server both computers and users (researchers, data scientists, app developers)
- Security / privacy
  - Access control, Risk assessment and reduction (SDC), transparency
- Automate metadata capture / Knowledge inference
  - Create new metadata, profiling, machine learning
- Knowledged Preservation / Institutional memory
- (more)

We search / discover / think by concept

*** We need rich metadata to achieve this ***



Concepts

Geography

Time

Subjects

Population

Surveys

Data Products

Data Elements

Questions

Classifications

Levels

Variables

Codes /Categories

Data

domain standards

technology standards

Part 2: Metadata Models & Standards

Models & Standards

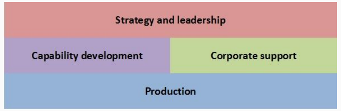
- Model
  - Defines how information is structured (elements, attributes, relationships, data types, etc.)
  - Needed by information system to ensure quality and understand meaning
  - Information Technology has standards to manage standards (UML, XML, RDF, JSON, ...)
- Why we need standards?
  - There can be many models, even about the same thing (books, cans of food, news, cars,...)
  - Common "language" / framework (so we know we are talking about the same thing)
- Many domains, many standards
  - Books, news, weather, cars, sports, aerospace, statistical data (complex)
  - Information technology standards
- Standards Pros:
  - Robust models, pack lots of expertise, enables standard based tools, common/best practices
- Standard Cons:
  - Can be too generic (not meet specific needs), can be slow moving / changing
- How to use?
  - Wisely, as a best practice / guidelines / reference
  - When exchanging information with others or public, to bridge disparate systems

High Level Group for Modernization of Official Statistics (HLG-MOS)

modernstats
BY HLG - MOS

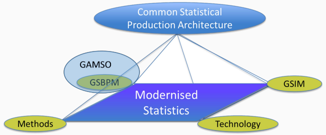
GAMSO

Generic Activity Model for Statistical Organizations



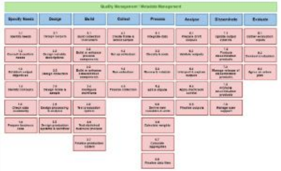
MMM

Modernisation Maturity Model



GSBPM

Generic Statistical Business Process Model



GSIM

Generic Statistical Information Model

CSPA

Common Statistical Production Architecture

Fostering Interoperability in Official Statistics: Common Statistical Production Architecture

<ddi>
Metadata powered by DDI

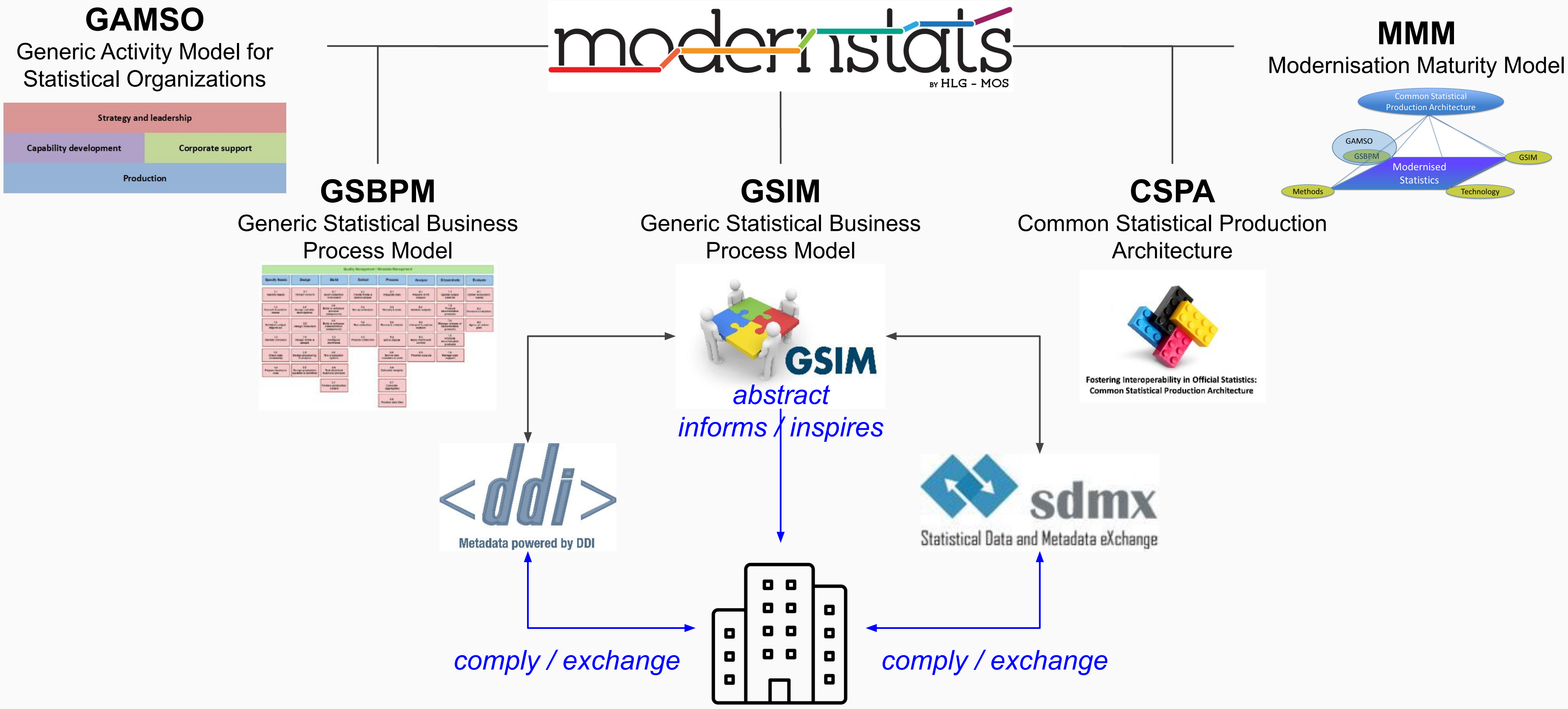
sdmx
Statistical Data and Metadata eXchange

microdata
raw data
analytical data
management / exchange

←-----------→

macrodata
indicators
time series
aggregated data
publication / exchange

Modernization



GAMSO
Generic Activity Model for Statistical Organizations

MMM
Modernisation Maturity Model

GSBPM
Generic Statistical Business Process Model

GSIM
Generic Statistical Business Process Model

CSPA
Common Statistical Production Architecture

GSIM
Fostering Interoperability in Official Statistics: Common Statistical Production Architecture
<ddi>
Metadata powered by DDI
sdmx
Statistical Data and Metadata eXchange

abstract
informs / inspires

comply / exchange

comply / exchange

adopt models & tools that meets institutional needs

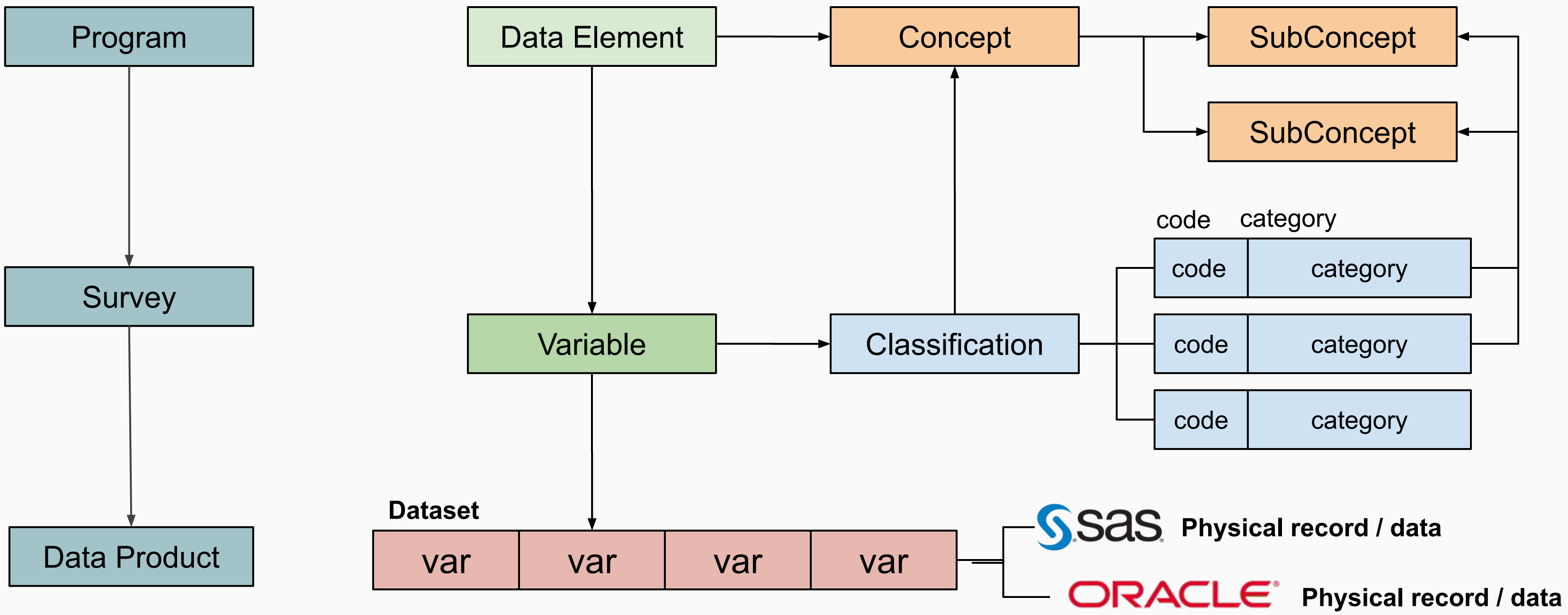
Part 2.1: High Level Model

Metadata Model

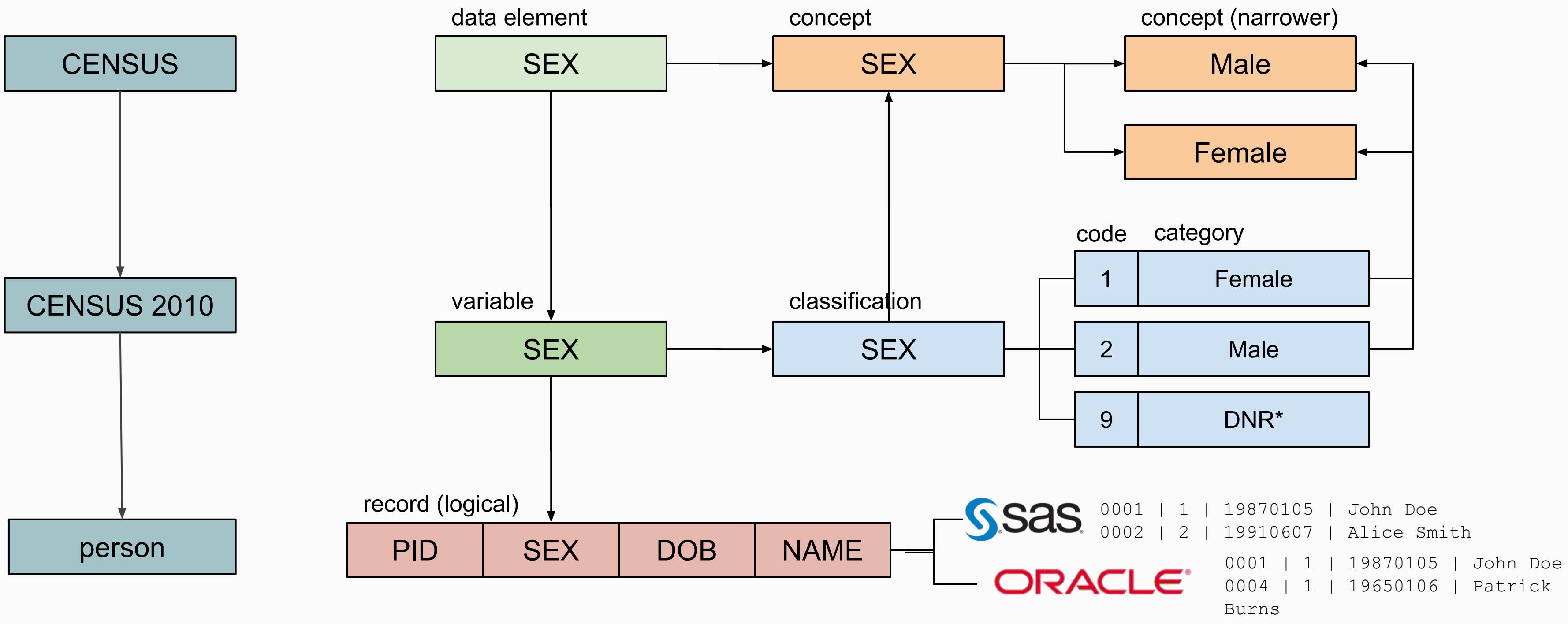
- Components
  - Programs, Surveys, Data Products, Datasets
  - Concepts
  - Questions, Questionnaires / Instruments
  - Classifications, Codes, Categories, Level, Concordances
  - Data Elements, Variables
  - Dataset, Record Layout, Dataset Instance
- Features
  - Identification & Referencing mechanisms
  - Core resource properties
  - Versioning
  - Properties: Types, Inheritance, Faceting, Extensions

High Level Model

warning: terminology varies widely across institutions! (GSIM can help)



High Level Model



CENSUS

CENSUS 2010

person

data element

SEX

concept

SEX

concept (narrower)

Male

Female

variable

SEX

classification

SEX

code	category
1	Female
2	Male
9	DNR*

record (logical)

PID	SEX	DOB	NAME

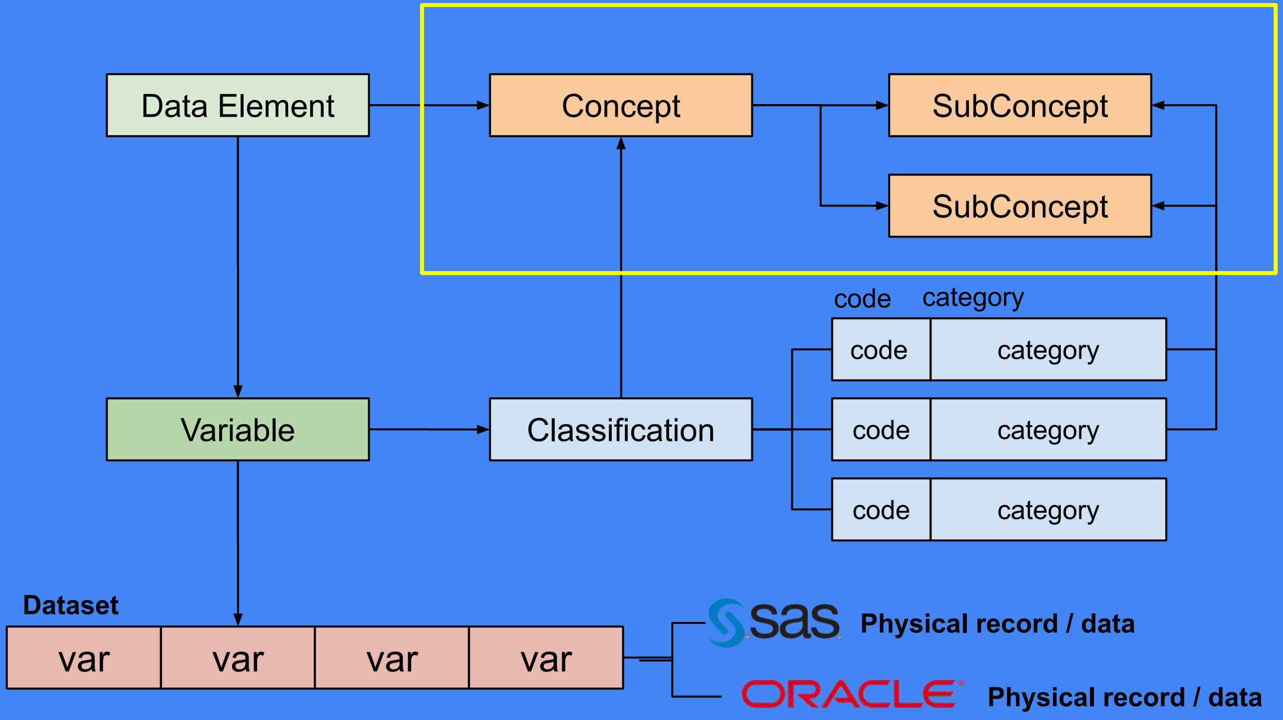
sas

```
0001 | 1 | 19870105 | John Doe
0002 | 2 | 19910607 | Alice Smith
```

ORACLE

```
0001 | 1 | 19870105 | John Doe
0004 | 1 | 19650106 | Patrick
Burns
```

Part 2.2: Concepts



Data Element

Concept

SubConcept

SubConcept

code category

code	category
code	category
code	category

Variable

Classification

Dataset

var	var	var	var

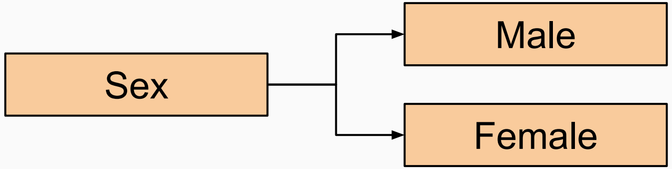
sas Physical record / data

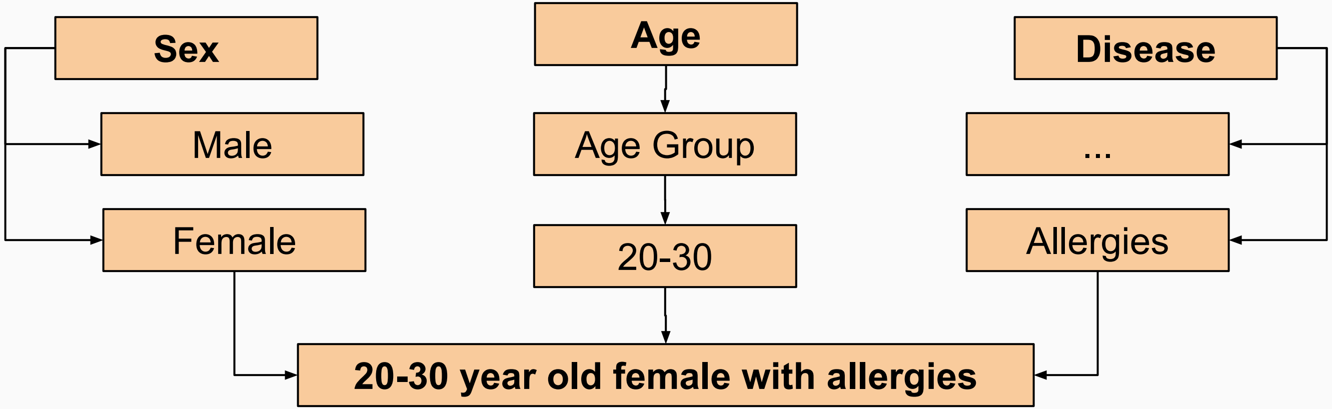
ORACLE Physical record / data

Concepts

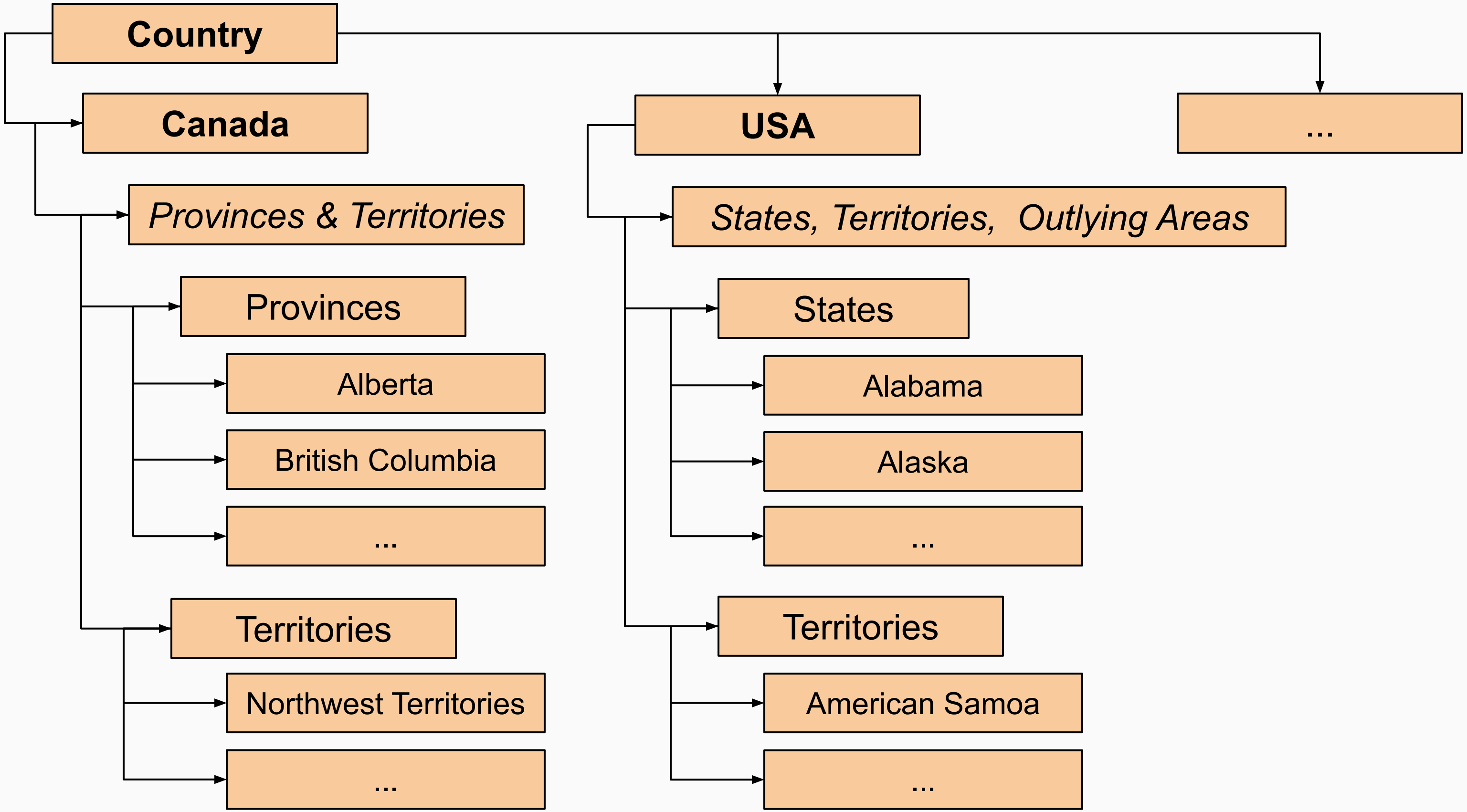
- This is how users search for information!
- Should be at the foundation of metadata management
  - Used by variables, classifications/levels/categories, units/population, etc.
  - Common denominator
- Relates to the semantic web / knowledge representation (not domain specific)
- A concept carries a set of descriptive/defining properties
- Concepts relate to each other to capture relationship or describe complex knowledge
- Technologies: RDF, OWL, SKOS, XKOS, etc.
- Generic tools are available to manage concepts
- Connected to each other using various semantic relationships
  - skos:narrower,broader,related
  - owl :unionOf, complementOf, intersectionOf, disjointWith, inverseOf
- Example: SNOMED

Concepts: examples

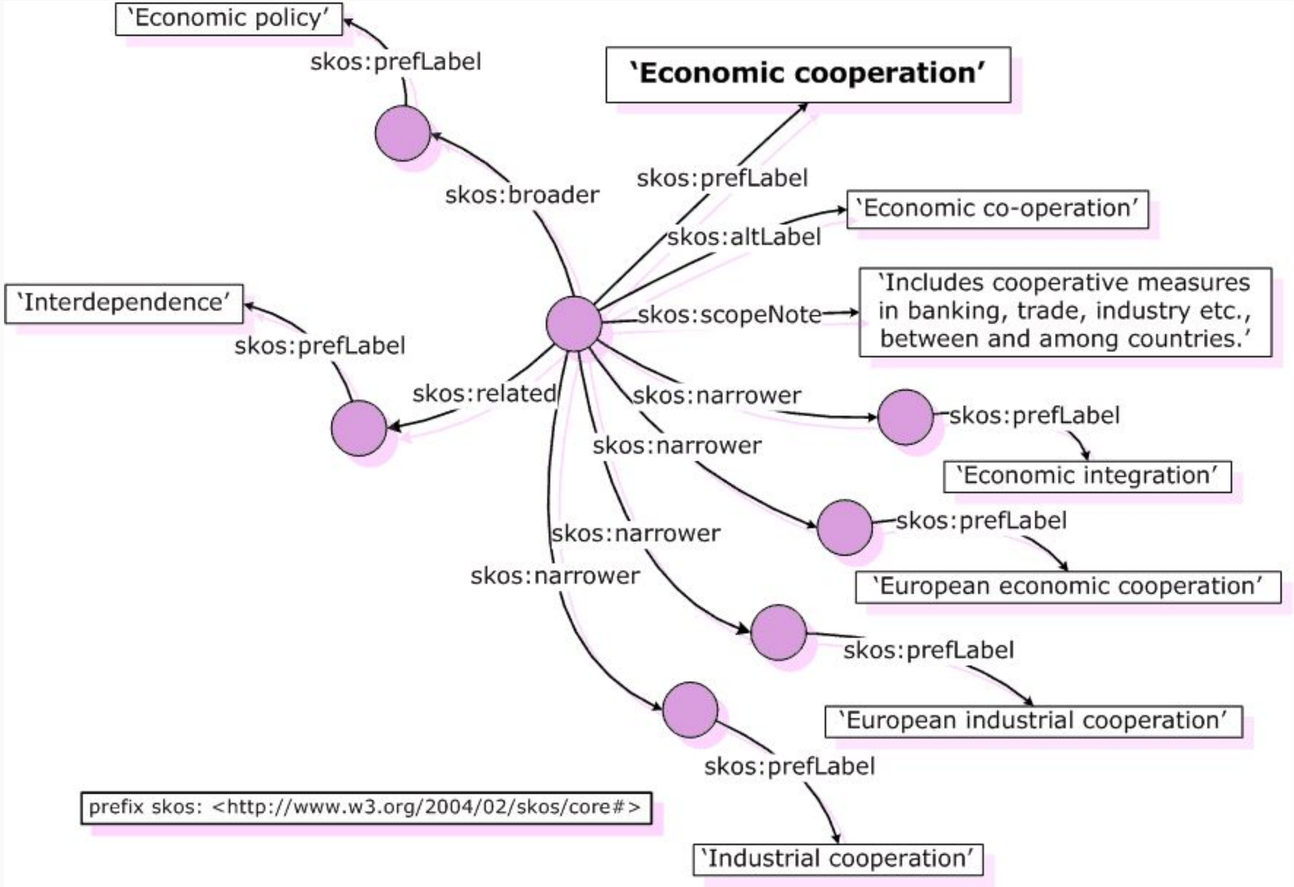




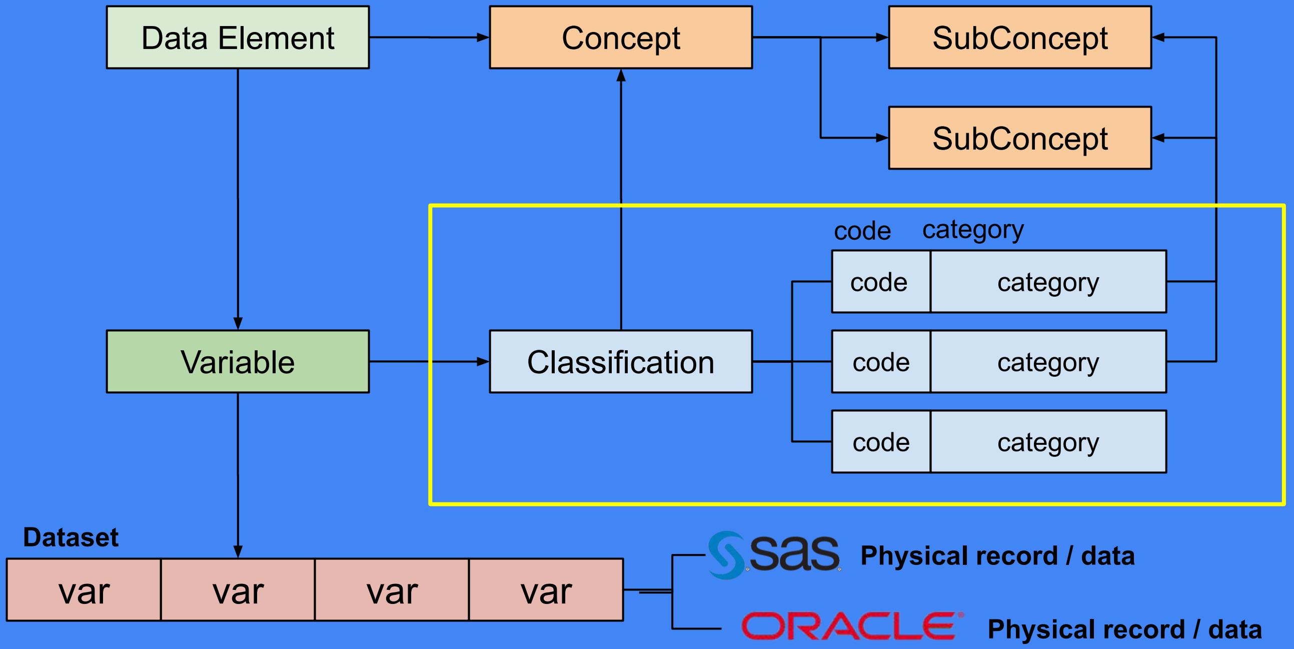
Concepts: narrower/broader and compound examples



Concepts: can have many relationships



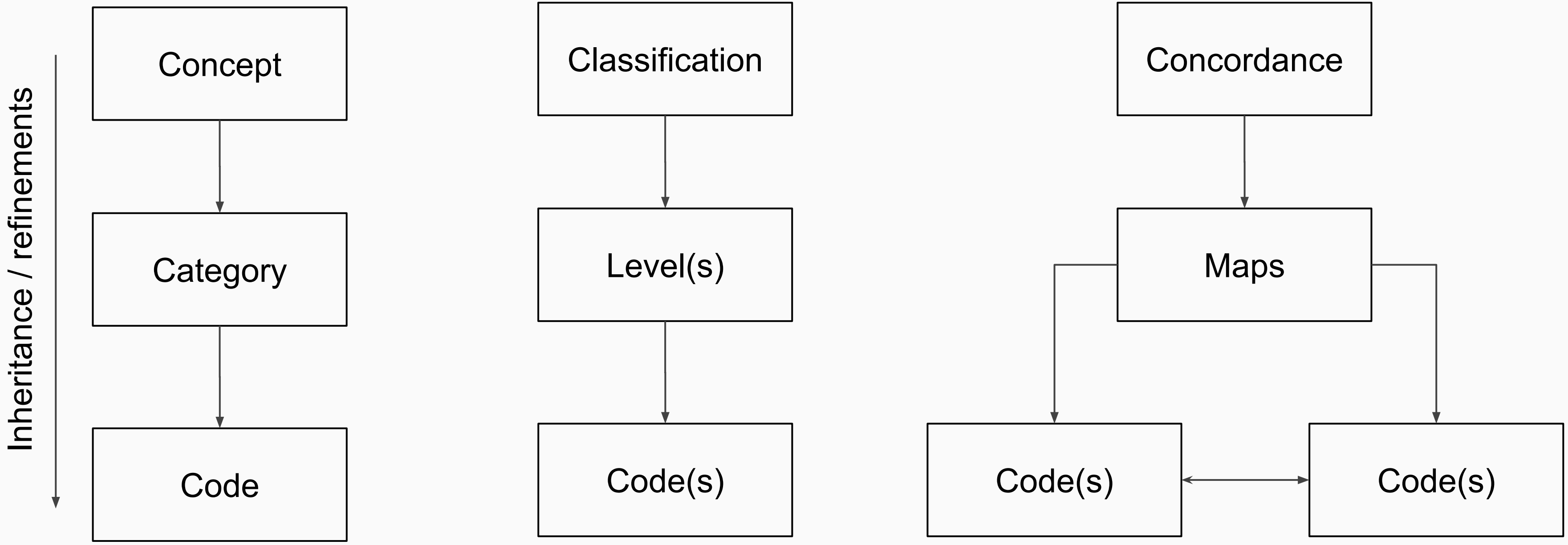
Part 2.3: Classifications



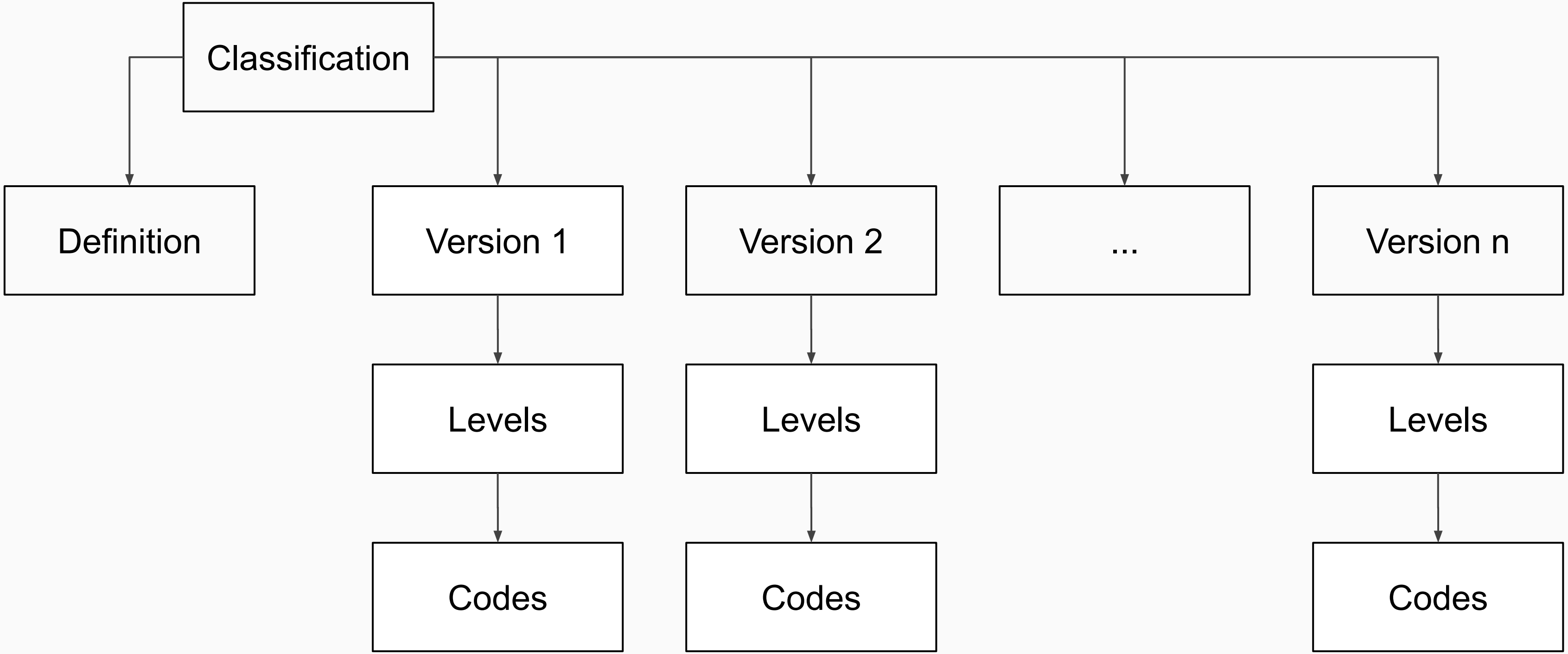
Classifications: overview

- Exists independently from the data (but often not maintained as such)
- Used by categorical variables, aggregated data dimensions, etc.
- Not just a code/label list. Composed of:
  - classification: definition and version
  - levels
  - codes
  - categories: the use of a concept for the purpose of coding
  - concordances (maps, x-walks)
- Ideal model should support:
  - stand alone / one-off classifications and formally maintained classifications
  - flat and hierarchical classifications
  - classification specific properties
  - multiple coding of same categories
  - view and derived classifications
  - versioning of classifications, levels, categories
  - synonyms
  - coding systems
  - serializing classifications in many formats (end users, standards, code)

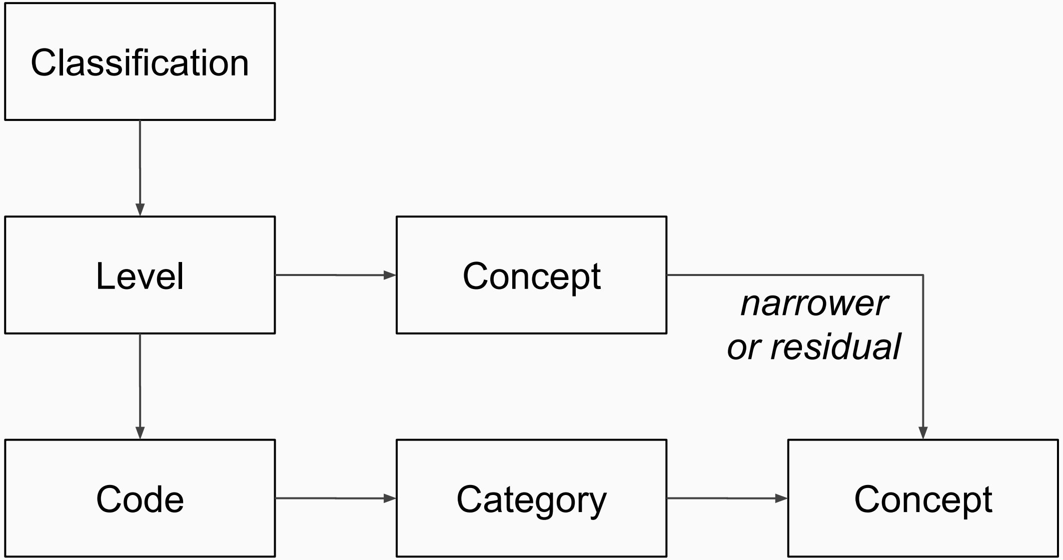
Classifications: High Level Model



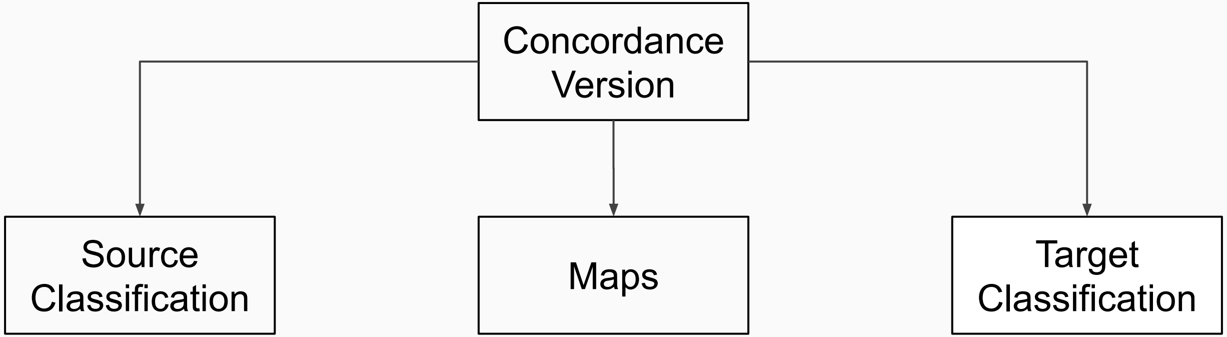
Classifications: Versioning



Classifications: Classifications and Concepts



Classifications: Concordances



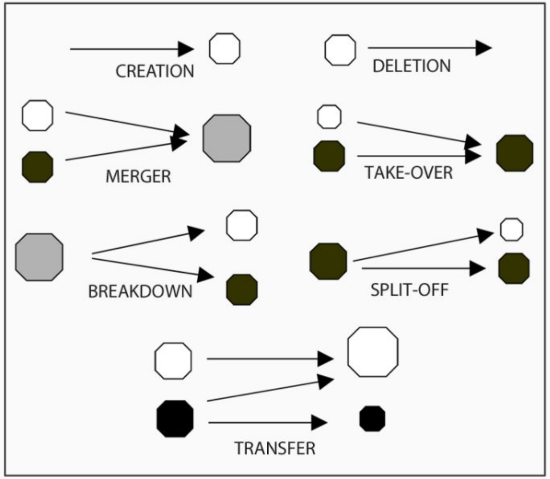
Concordance Version

Source Classification

Maps

Target Classification

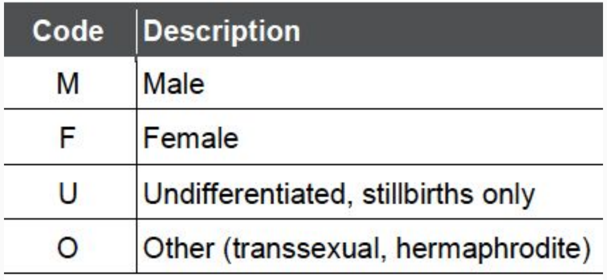
Informed by Neuchatel / GSIM



Can potentially be auto-generated

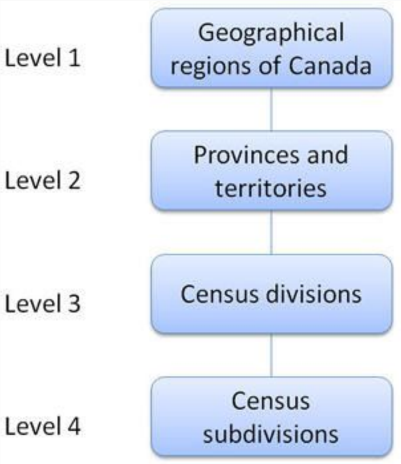
Classifications: examples

Gender



Code	Description
M	Male
F	Female
U	Undifferentiated, stillbirths only
O	Other (transsexual, hermaphrodite)

Hierarchical classification: Canada Standard Geographical Classification (SGC) 2016



Level 1 — Geographical regions of Canada

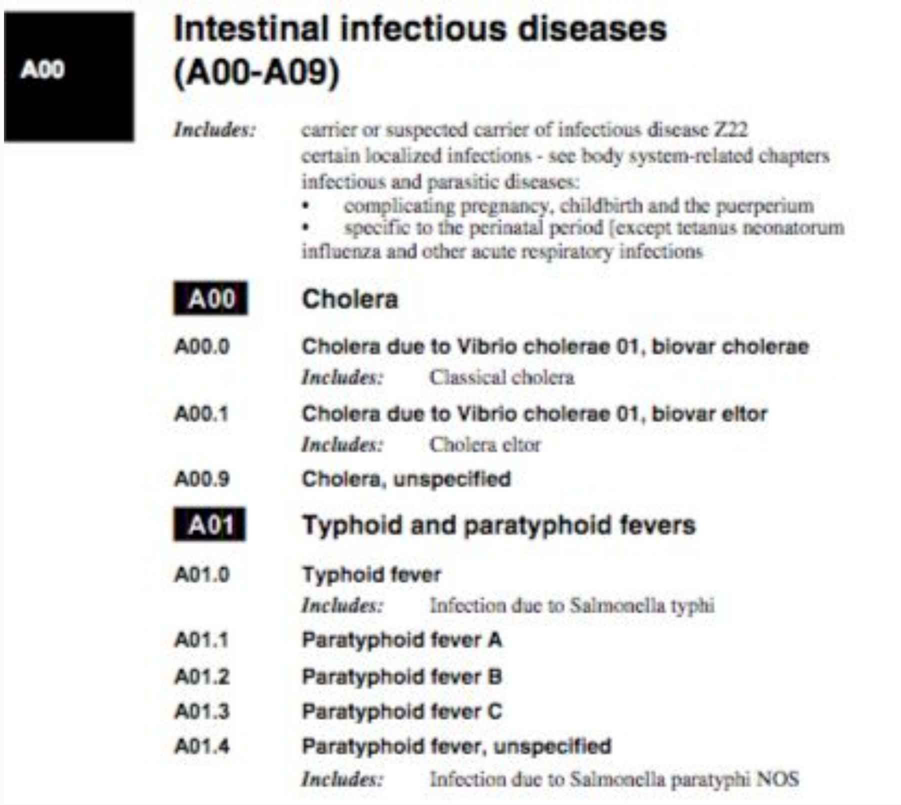
Level 2 — Provinces and territories

Level 3 — Census divisions

Level 4 — Census subdivisions

ICD-10

hierarchical, extended properties



A00

Intestinal infectious diseases (A00-A09)

Includes: carrier or suspected carrier of infectious disease Z22
certain localized infections - see body system-related chapters
infectious and parasitic diseases:
- complicating pregnancy, childbirth and the puerperium
- specific to the perinatal period [except tetanus neonatorum

influenza and other acute respiratory infections

A00 **Cholera**

A00.0 Cholera due to Vibrio cholerae 01, biovar cholerae
Includes: Classical cholera

A00.1 Cholera due to Vibrio cholerae 01, biovar eltor
Includes: Cholera eltor

A00.9 Cholera, unspecified

A01 **Typhoid and paratyphoid fevers**

A01.0 Typhoid fever
Includes: Infection due to Salmonella typhi

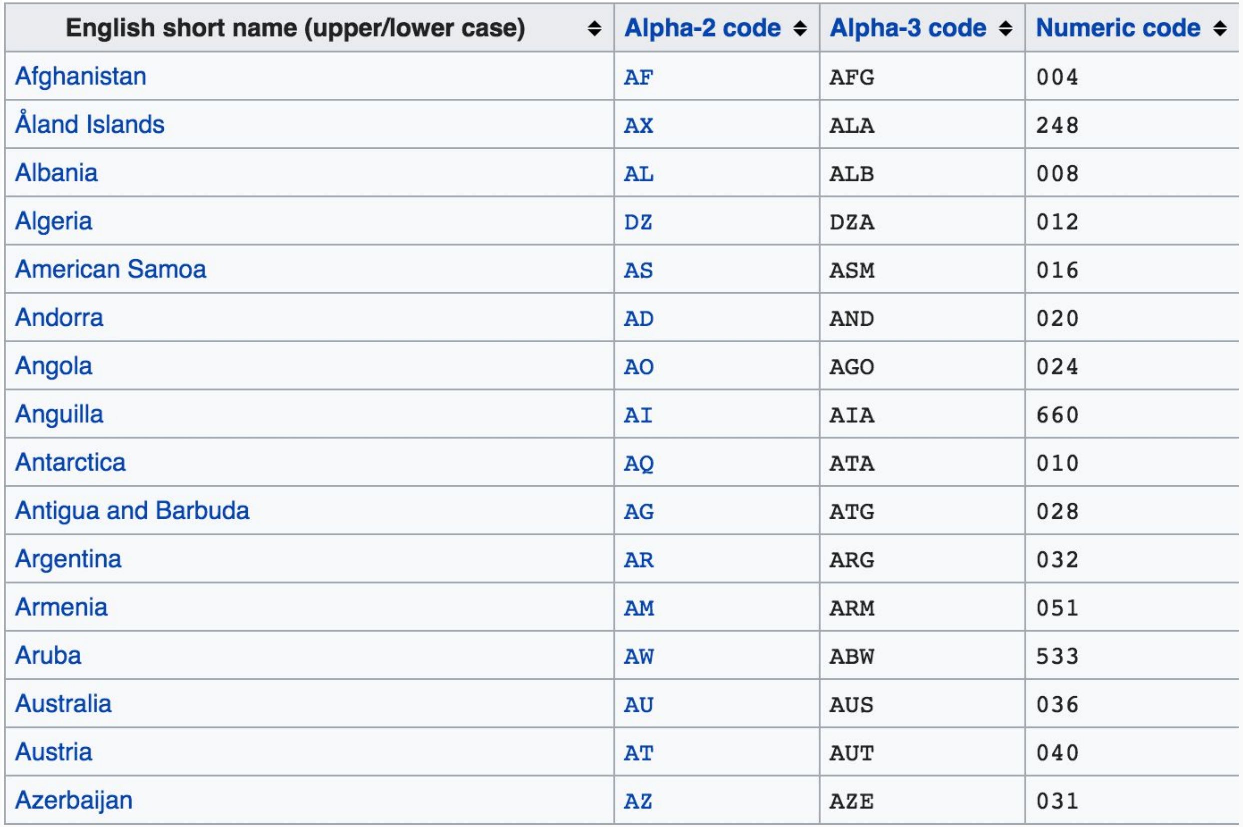
A01.1 Paratyphoid fever A

A01.2 Paratyphoid fever B

A01.3 Paratyphoid fever C

A01.4 Paratyphoid fever, unspecified
Includes: Infection due to Salmonella paratyphi NOS

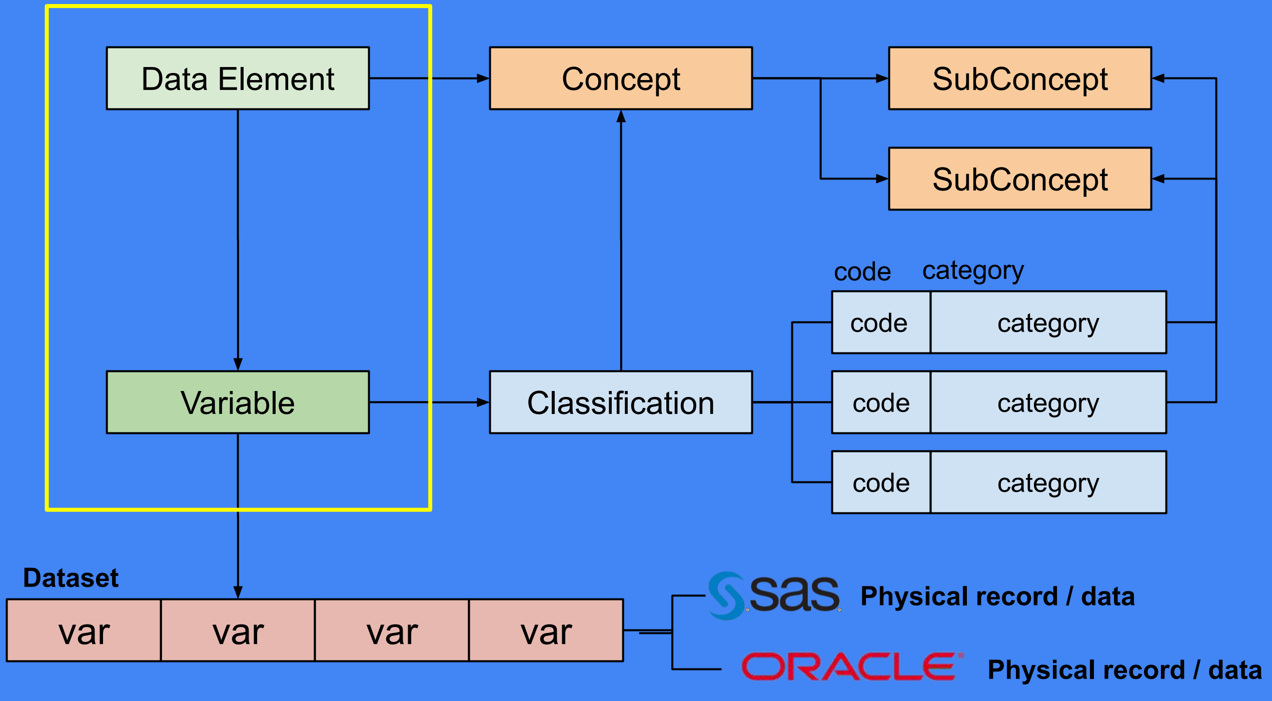
Classifications: ISO 3166-1



English short name (upper/lower case)	Alpha-2 code	Alpha-3 code	Numeric code
Afghanistan	AF	AFG	004
Åland Islands	AX	ALA	248
Albania	AL	ALB	008
Algeria	DZ	DZA	012
American Samoa	AS	ASM	016
Andorra	AD	AND	020
Angola	AO	AGO	024
Anguilla	AI	AIA	660
Antarctica	AQ	ATA	010
Antigua and Barbuda	AG	ATG	028
Argentina	AR	ARG	032
Armenia	AM	ARM	051
Aruba	AW	ABW	533
Australia	AU	AUS	036
Austria	AT	AUT	040
Azerbaijan	AZ	AZE	031

- 3 classifications sharing a common category set
- categories can version over time
- carry faceted names
    - english/french
    - full, short, upper/lower
    - local short
    - independent flag
    - currency

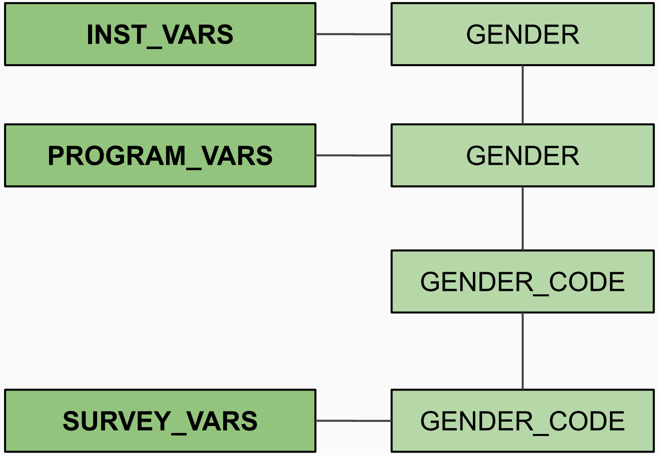
Part 2.3: Variables



Variables & Data Elements

- Variables evolve from a high level conceptual state to an actual field/column associated with data in physical a data file or table
  - At the highest level, can simply be the use of a concept
  - As we go to lower levels, the variable becomes more concrete (gains or refines its properties)
- The higher the level the higher the reusability
- GSIM distinguish between variables, represented variables, and instance variables
  - But technically can have any number of refinements
  - The set of properties or context can be used to categorize it from a GSIM perspective
- The term Data Element often used for high level / conceptual variable (cannot be used in a dataset)

Variables example



INST_VARS — GENDER

Institutional / conceptual definition of a gender element. Associated with a "gender' concept and a simple male/female classification.

PROGRAM_VARS — GENDER

Statistical program definition of the gender data element. May use specific classification adding a couple of codes

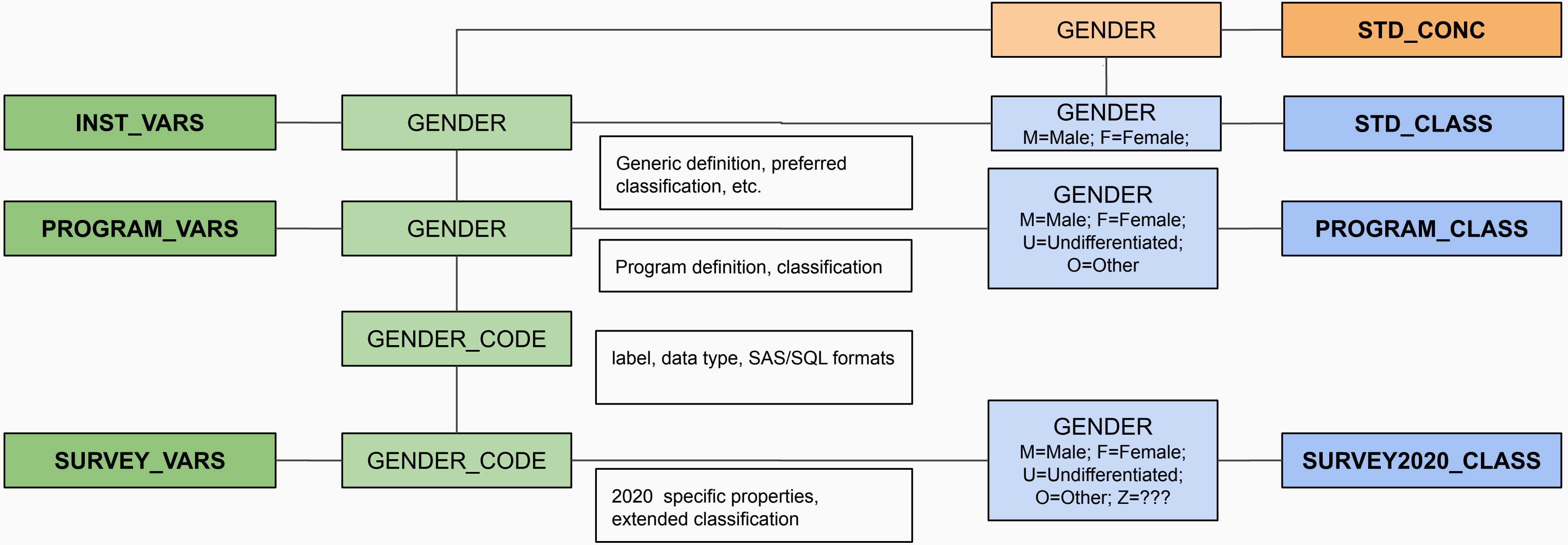
GENDER_CODE

Gains SAS and SQL specific formats, a different name/label, and possibly other attributes. Anything at this level is expected to be reused across multiple years, until a new version comes up.

SURVEY_VARS — GENDER_CODE

Adds any survey specific properties and an extended classification

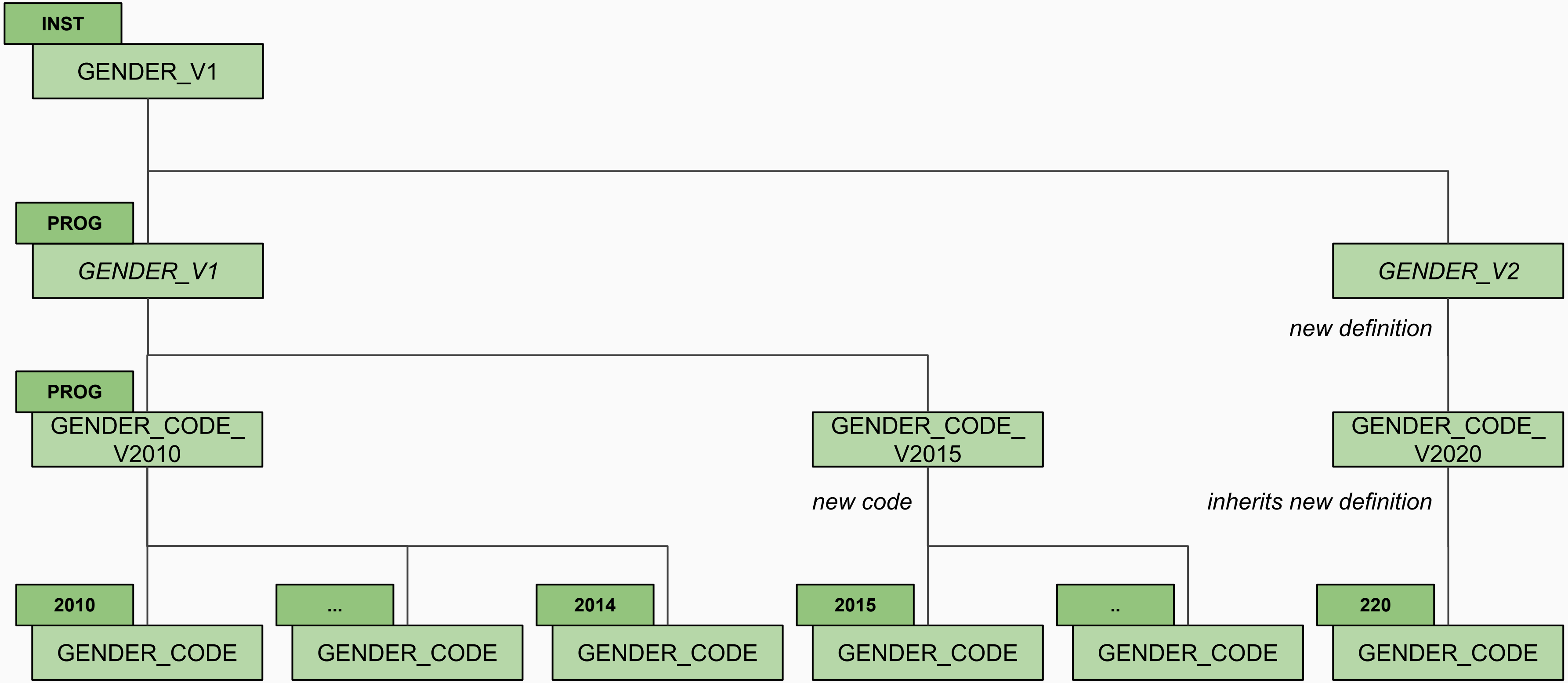
Variables example



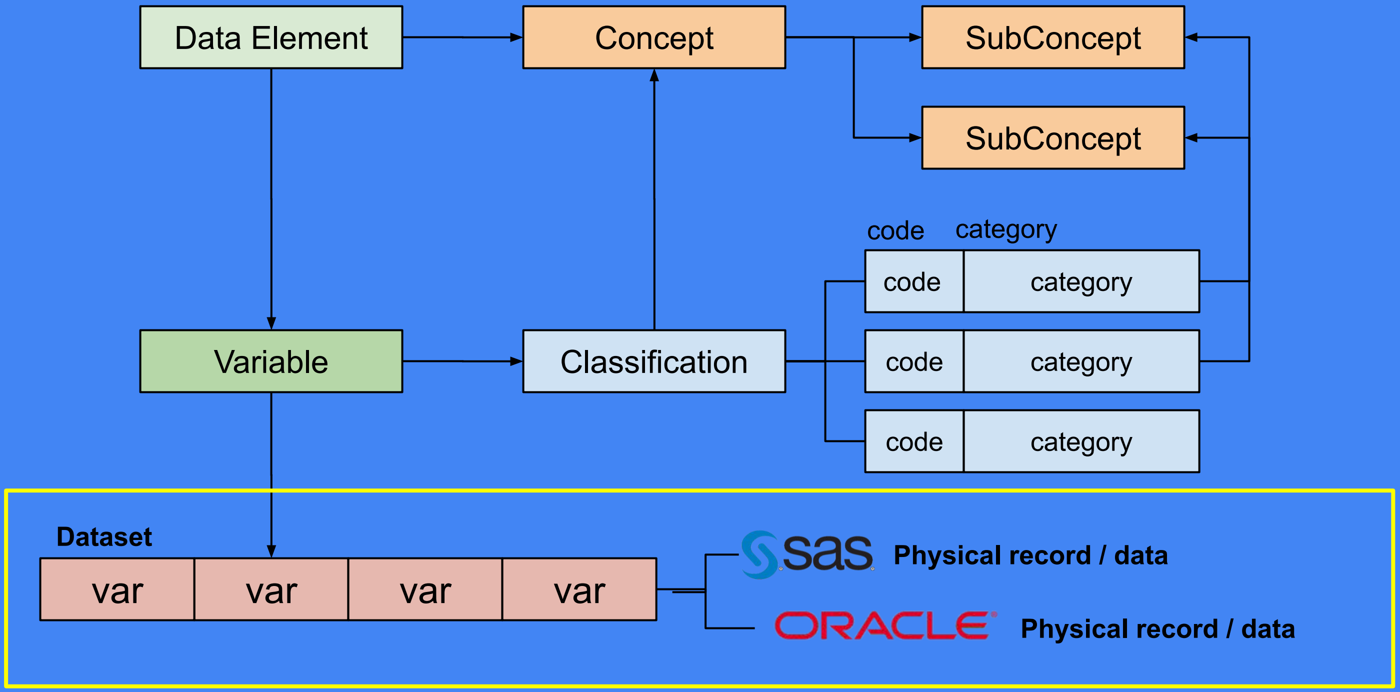
GENDER
STD_CONC
INST_VARS
GENDER
GENDER
M=Male; F=Female;
STD_CLASS
Generic definition, preferred classification, etc.
PROGRAM_VARS
GENDER
GENDER
M=Male; F=Female; U=Undifferentiated; O=Other
PROGRAM_CLASS
Program definition, classification
GENDER_CODE
label, data type, SAS/SQL formats
SURVEY_VARS
GENDER_CODE
GENDER
M=Male; F=Female; U=Undifferentiated; O=Other; Z=???
SURVEY2020_CLASS
2020 specific properties, extended classification

we will shortly see how this connects to datasets

Managing variable changes over time



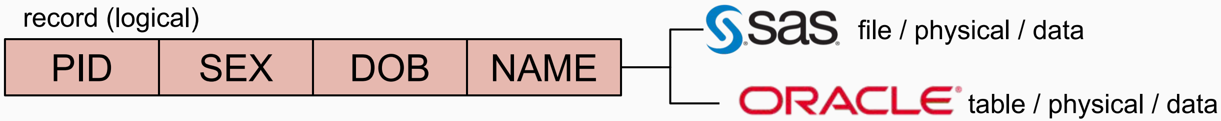
Part 2.4: Datasets



Dataset: Record, Layout, Instance

- Where data lives
- Stitches variables and data together (dictionary)
- GSIM: An organized collection of data
- An ordered collection of variables with which data can be associated (dictionary)
- We may distinguish between logical records and dataset instance
- Logical record:
  - how the data is organized (dictionary)
  - can also have primary keys, relationships to other records, and other attributes (name, description, etc.)
- Dataset instance
  - A file (e.g. SAS, ascii) or SQL table that contains actual data
  - GSIM: unit or dimensional dataset (same thing in the end)
- One logical record can be reused by many physical instances

Logical Record vs Physical instance



record (logical)

PID	SEX	DOB	NAME

sas file / physical / data

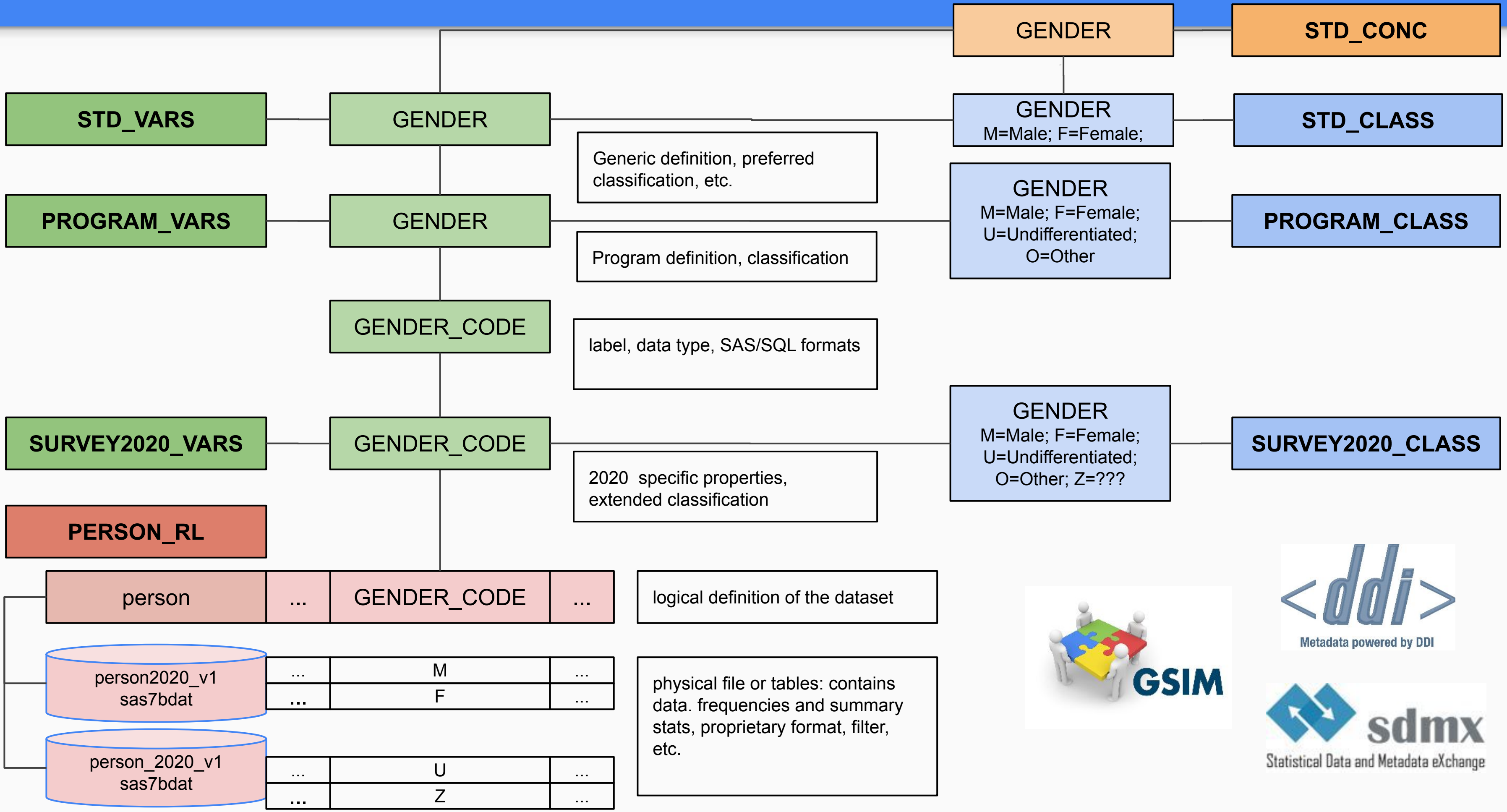
ORACLE table / physical / data

All the following physical instances can share the same logical record:

- A SAS file with all the person records
- An Oracle SQL table with all the person records
- A text, SPSS, Stata, R, Excel version of the data
- Multiple versions of the data (that changes over time)
- Multiple copies of the same file stored at different locations (e.g. backups)
- A data cut (filtered / subset) stored in any of the above formats (e.g. person age 50+, females only, etc.)

Note: summary statistics / frequencies are stored at the instance level

Variables & Datasets



GENDER

STD_CONC

STD_VARS

GENDER

GENDER
M=Male; F=Female;

STD_CLASS

Generic definition, preferred classification, etc.

PROGRAM_VARS

GENDER

GENDER
M=Male; F=Female;
U=Undifferentiated;
O=Other

PROGRAM_CLASS

Program definition, classification

GENDER_CODE

label, data type, SAS/SQL formats

SURVEY2020_VARS

GENDER_CODE

GENDER
M=Male; F=Female;
U=Undifferentiated;
O=Other; Z=???

SURVEY2020_CLASS

2020 specific properties, extended classification

PERSON_RL

person | ... | GENDER_CODE | ...

logical definition of the dataset

person2020_v1
sas7bdat

... | M | ...
... | F | ...

person_2020_v1
sas7bdat

... | U | ...
... | Z | ...

physical file or tables: contains data. frequencies and summary stats, proprietary format, filter, etc.

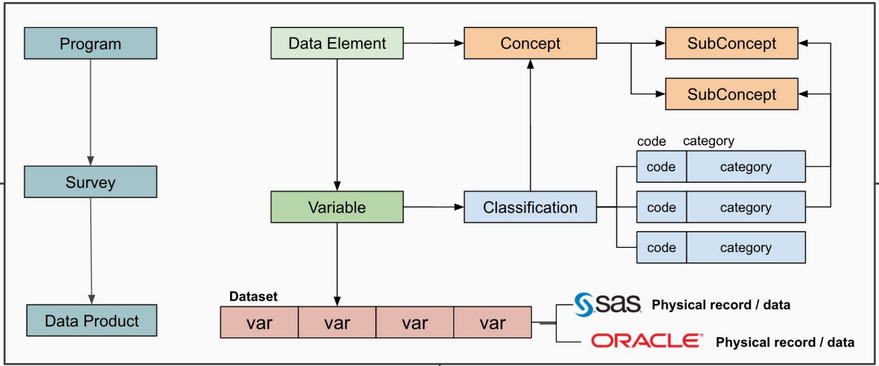
GSIM

<ddi>
Metadata powered by DDI

sdmx
Statistical Data and Metadata eXchange

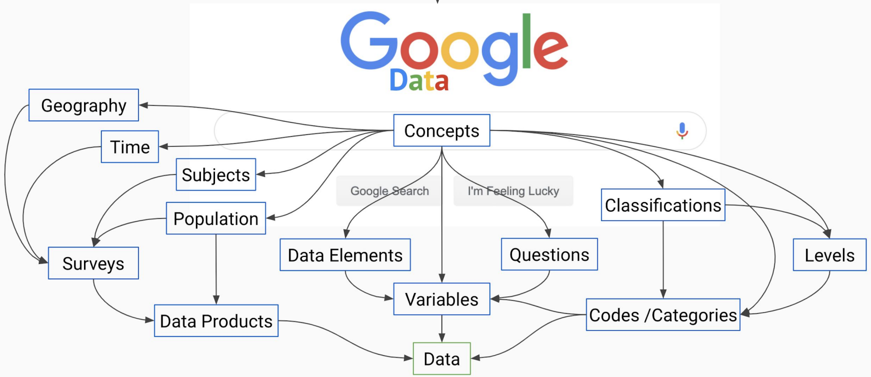
Metadata Driven Data Management

Caters to producers and users needs



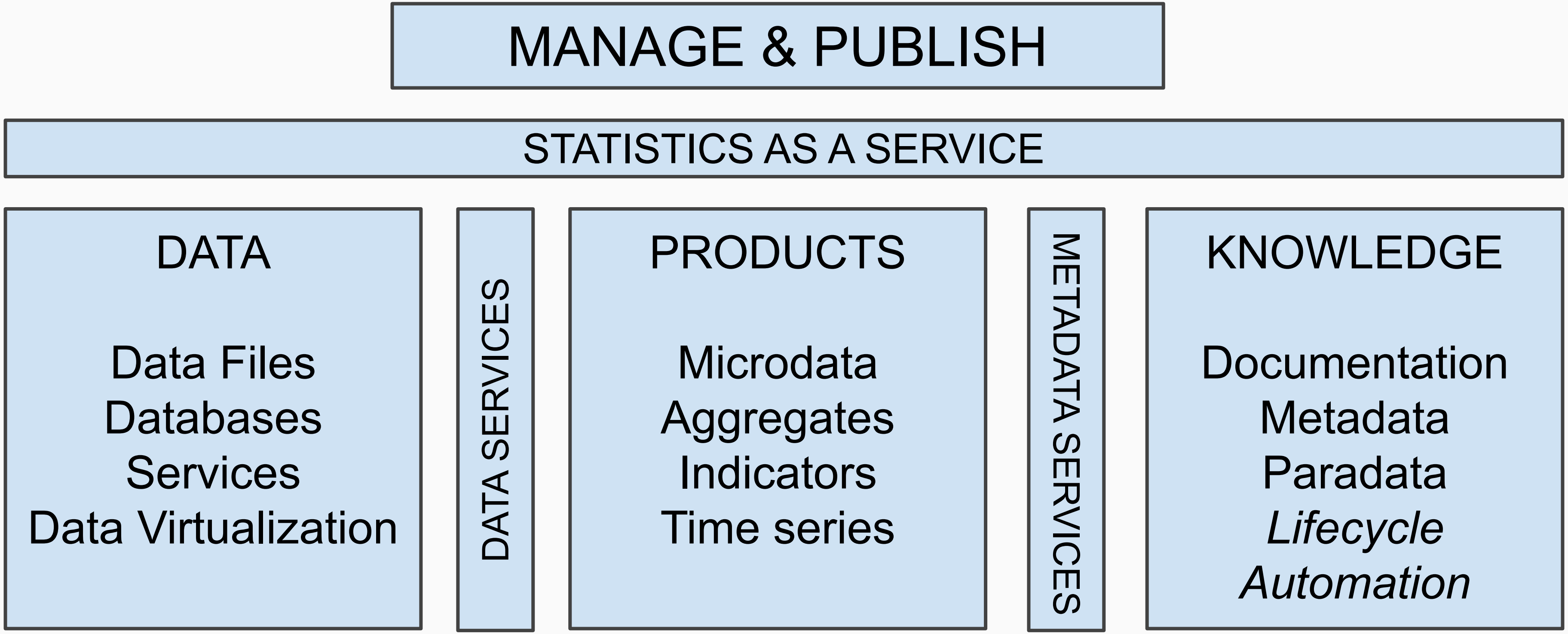
Caters to computer needs (machine actionable)

Enables innovation



Part 3: Technologies & Framework

Architectural Vision



MANAGE & PUBLISH

STATISTICS AS A SERVICE

DATA

Data Files
Databases
Services
Data Virtualization

DATA SERVICES

PRODUCTS

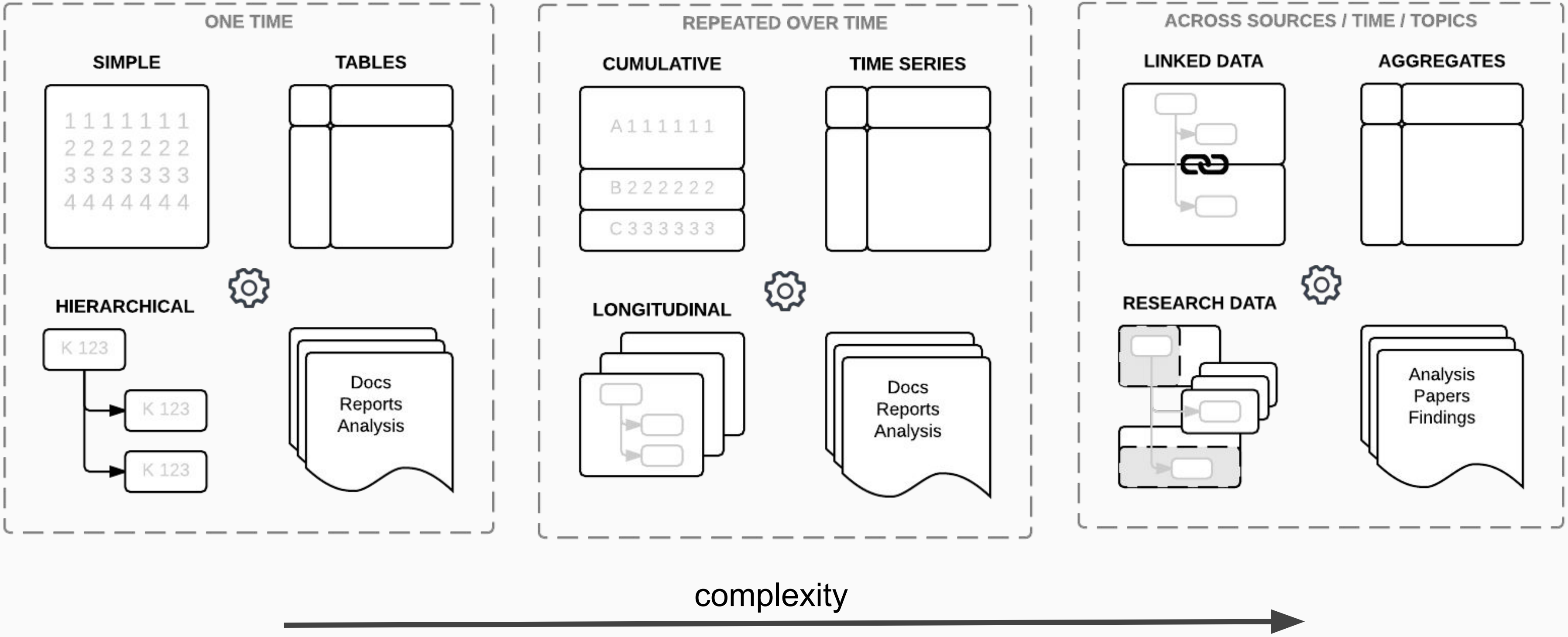
Microdata
Aggregates
Indicators
Time series

METADATA SERVICES

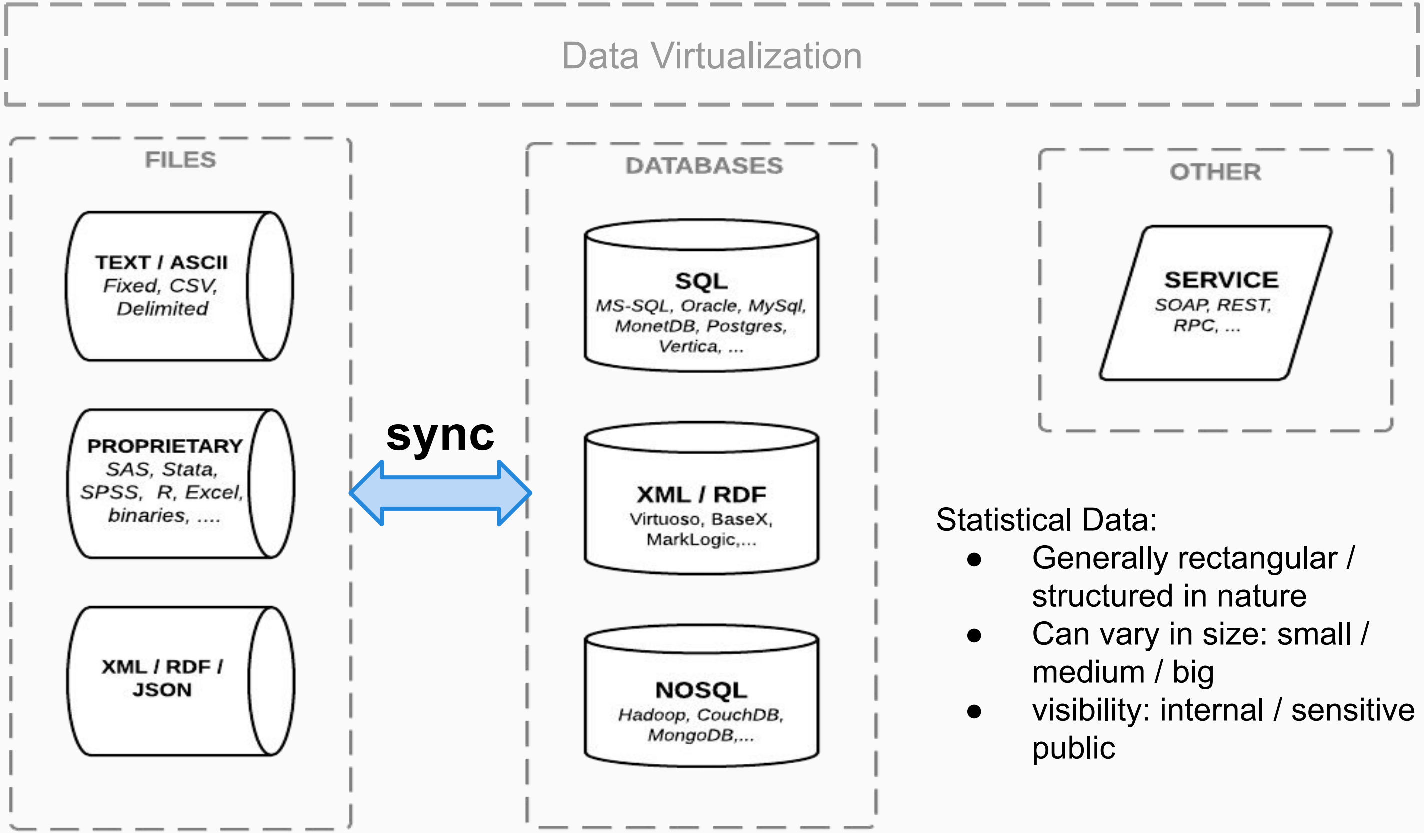
KNOWLEDGE

Documentation
Metadata
Paradata
Lifecycle
Automation

Data Products



Data



Data Virtualization

FILES

TEXT / ASCII
Fixed, CSV, Delimited

PROPRIETARY
SAS, Stata, SPSS, R, Excel, binaries, ...

XML / RDF / JSON

sync

DATABASES

SQL
MS-SQL, Oracle, MySql, MonetDB, Postgres, Vertica, ...

XML / RDF
Virtuoso, BaseX, MarkLogic,...

NOSQL
Hadoop, CouchDB, MongoDB,...

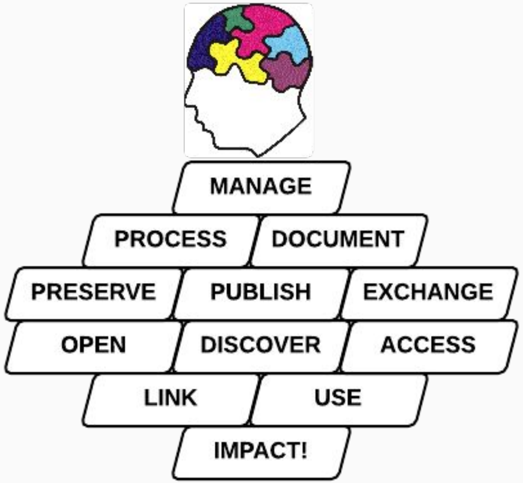
OTHER

SERVICE
SOAP, REST, RPC, ...

Statistical Data:

- Generally rectangular / structured in nature
- Can vary in size: small / medium / big
- visibility: internal / sensitive / public

Metadata



MANAGE

PROCESS DOCUMENT

PRESERVE PUBLISH EXCHANGE

OPEN DISCOVER ACCESS

LINK USE

IMPACT!

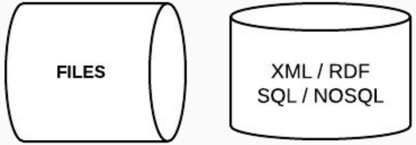
CONTENT

Management Process, Tasks, Administration, Business Rules	**Catalog** *Seties, Surveys, Data Products, Documentation, Citations*
Conceptual *Concepts, Classifications, Population, etc.*	**Content** *Methods, Questions, Groups Variables, Logical Structures*
System *Identifiers, Vocabularies, Subscription, Logs*	**Data** *Records, Relations, Files, Tables, Cubes, Instances*

REPRESENTATION

Technology XML / RDF / JSON SOAP / REST	**Standards** GSIM, DDI, SDMX, DCAT, ISO 11179, Dublin Core,...

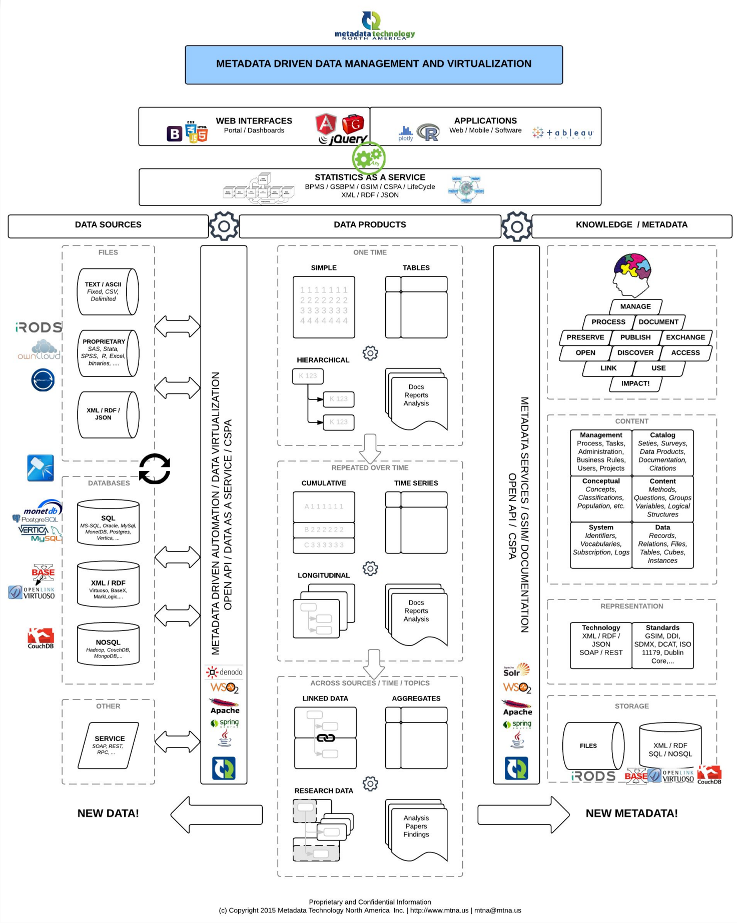
STORAGE



FILES

XML / RDF
SQL / NOSQL

Metadata Driven Data Management Framework



Management Process, Tasks, Administration, Business Rules, Users, Projects	**Catalog** *Seties, Surveys, Data Products, Documentation, Citations*
Conceptual *Concepts, Classifications, Population, etc.*	**Content** *Methods, Questions, Groups Variables, Logical Structures*
System *Identifiers, Vocabularies, Subscription, Logs*	**Data** *Records, Relations, Files, Tables, Cubes, Instances*

Technology XML / RDF / JSON SOAP / REST	**Standards** GSIM, DDI, SDMX, DCAT, ISO 11179, Dublin Core,...

Proprietary and Confidential Information

- Can be achieved incrementally
  - Lean / AGILE approach
  - Can't do it all at the same time
- Technology is available
  - Leverage what already exists
  - Leverage SOA, standards (reuse)
  - Not me biggest challenge
  - Complement / modernize traditional software
- Non-intrusive integration strategy
  - day to day business continues
- Managing the change is the key

Implementing / Solutions

IT / Data Expertise

OSS / COTS / SOA

Storage/Management

- File: FS, IRODS, …
- SQL: Oracle, MS-SQL, MySql, Vertica, MonetDB
- XML/RDF: BaseX, Virtuoso
- NOSQL: CouchDB, MongoDB, Hadoop,
- DV: Denodo, etc.

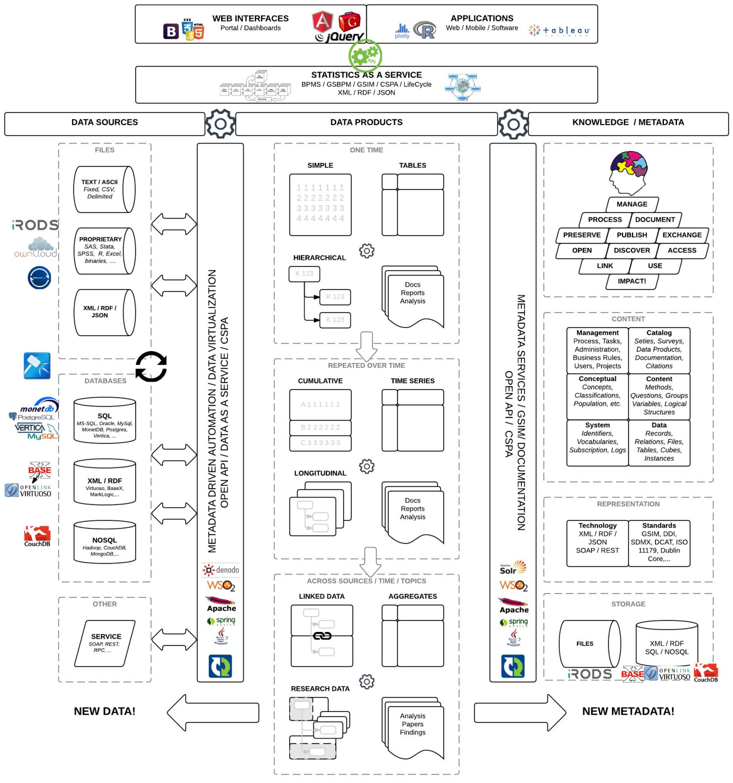
Data Tools

- SAS, Stata SPSS
- R, Python
- Custom Apps

Services/Webapps

- *SOAP/REST*
- *Java, Apache*
- *WSO2*
- *HTML5, Angular, GwT*

Statistics as a Service



WEB INTERFACES
Portal / Dashboards

APPLICATIONS
Web / Mobile / Software

STATISTICS AS A SERVICE
BPMS / GSBPM / GSIM / CSPA / LifeCycle
XML / RDF / JSON

DATA SOURCES

FILES
TEXT / ASCII *Fixed, CSV, Delimited*
PROPRIETARY *SAS, Stata, SPSS, R, Excel, binaries, ...*
XML / RDF / JSON

DATABASES
SQL *MS-SQL, Oracle, MySql, MonetDB, Postgres, Vertica, ...*
XML / RDF Virtuoso, BaseX, MarkLogic,...
NOSQL *Hadoop, CouchDB, MongoDB,...*

OTHER
SERVICE *SOAP, REST, RPC, ...*

METADATA DRIVEN AUTOMATION / DATA VIRTUALIZATION
OPEN API / DATA AS A SERVICE / CSPA

NEW DATA!

DATA PRODUCTS

ONE TIME
SIMPLE
TABLES
HIERARCHICAL
Docs
Reports
Analysis

REPEATED OVER TIME
CUMULATIVE
TIME SERIES
LONGITUDINAL
Docs
Reports
Analysis

ACROSS SOURCES / TIME / TOPICS
LINKED DATA
AGGREGATES
RESEARCH DATA
Analysis
Papers
Findings

METADATA SERVICES / GSIM/ DOCUMENTATION
OPEN API / CSPA

NEW METADATA!

KNOWLEDGE / METADATA

MANAGE
PROCESS DOCUMENT
PRESERVE PUBLISH EXCHANGE
OPEN DISCOVER ACCESS
LINK USE
IMPACT!

CONTENT

Management Process, Tasks, Administration, Business Rules, Users, Projects	**Catalog** *Seties, Surveys, Data Products, Documentation, Citations*
Conceptual *Concepts, Classifications, Population, etc.*	**Content** *Methods, Questions, Groups Variables, Logical Structures*
System *Identifiers, Vocabularies, Subscription, Logs*	**Data** *Records, Relations, Files, Tables, Cubes, Instances*

REPRESENTATION

Technology XML / RDF / JSON SOAP / REST	**Standards** GSIM, DDI, SDMX, DCAT, ISO 11179, Dublin Core,...

STORAGE
FILES
XML / RDF
SQL / NOSQL

Proprietary and Confidential Information

IT / Metadata Expertise

COTS Statistical Data/Metadata

MTNA: Rich Data Services

- Enterprise solution
- Flexible core model
- Standards compliant (DDI, GSIM, etc.)
- 60-70% COTS
- Can integrate with o/platforms

Colectica

- DDI based, Windows

Nesstar (legacy)

Proprietary System

…

Part 4: Implementation

Why is it challenging?

- Common Data Tools Limitations
  - Data tools (databases, stats packages, etc.) are metadata poor and not suited for managing knowledge
- Our metadata models are more complex
  - Unlike books, cars, etc., statistical metadata models are solid but can be heavy
- We don't have good metadata habits
  - We are generally not used or taught to focus on metadata and we don't have tools
  - We document after the fact (which also result in loss of knowledge, document before or during)
- Metadata is not a budgeted activity
  - The term metadata is often sparse in budget proposals
  - We don't invest in dissemination / packaging
  - We need better data marketing
- Some of the above can be solved with technology, other require change management.....
  - Good news is that technology is available.
- We need to establish an environment that fosters/facilitates the use of metadata
- short term: We need to demonstrate the benefits

Plan

- Technology + Standards/Model + Resources ($ + leadership) + Change Management
- Assessment + Roadmap
    - Infrastructure (inexpensive)
    - Enhance data management platform as needed
    - Adopt Metadata / Knowledge Management Platform
    - Training: producers, librarians, collectors, IT (but not users)
    - Migration: data / metadata / possibly scripts programs
    - Maintenance
- Establish 2-5 years Strategy:
    - Incremental / non-intrusive (day to day business must continue)
    - Start with dissemination, focus on most popular dataset
    - Change management!
    - Change the way you produce new data; Migrate existing data over time (popularity driven);
- In the meantime: go for quick wins / fixes
    - No need to wait + demonstrates that above is realistic and beneficial

Some common situations and quick fixes

No or limited institutional coordination in terms of managing metadata standards and best practices	The "what is metadata" question? Internal external data harmonization / linking issues; Limited documentation and institutional knowledge;	Need committee; Get top management support; Provide guidelines, training; Establish institutional repository (start with key classifications, data elements; concepts); Require use of standard entities (e.g. classifications);
Data only available in ASCII/CSV and/or stats packages as static download	Does not cater to all user needs (what is SAS some are asking...); Does not cater to applications / developer / web; Costly custom extracts preparation;	Data as a service; Expose both for dynamic queries and static extraction / tabulation tools; A DWH is all it takes to start (enhance metadata over time);
Minimalistic public or internal catalog (e.g. HTML pages)	Data is hard to find/discover; Can't look at variable level;	Establish proper catalog, inventory; Expose as a service using standards (DCAT, DDI); Use IHSN Catalog; Past & ongoing surveys; Event notification;

Common situations and quick fixes

We don't have metadata, and have 20+ years of data to document;	Don't know where to start; It feels overwhelming;	Let information systems do the initial work (convert, scan, profiling, metadata inference); Start with new data; Document most popular data; Outsource metadata capture;
Thousands of files stored on shared network drives (data and docs); Accumulated over many years;	It's chaotic; Don't know which file is where / what or who is owner is; Too many files;	Use scanning/profiling tools to assess; Establish file management guidelines and institutional repository; Use tools to monitor file systems or adopt intelligent file system (iRODS);
Mix of SAS and custom / proprietary programs, and/or traditional SQL databases for data management;	Stuck in legacy code; High licensing fees; Does not handle medium/big/emerging data well;	Leverage R & Python; & modern database (Column SQL, hybrid, JSON data types); Use code generators; Transition away from 20th century technology; Let nextgen data scientists take over;

Common situations and fixes

The useful data is sensitive / disclosive	Don't know how to provide access; Hesitant to disseminate;	Establish virtual data enclave (e.g NORC DE); Leverage all SDC method and available tools; Automate SDC steps; *Expose public metadata!*
Data quality varies greatly and QA procedures consumes significant resources	We have a hard time assessing or improving the quality of data	Leverage metadata to ensure data meets expectations; Automate QA steps.
We struggle collecting data from disparate / diverse sources or providers	Putting this together takes significant resources; Consolidation / Harmonization is challenging;	Use metadata driven data ingestion and QA; Can automate a significant portion of the process; Provide data collectors with metadata specifications / submission guides;
<INSERT YOUR USE CASE HERE>	<INSERT YOUR PAIN POINTS HERE>	<ASK METADATA / IT EXPERTS>

Costs & ROI

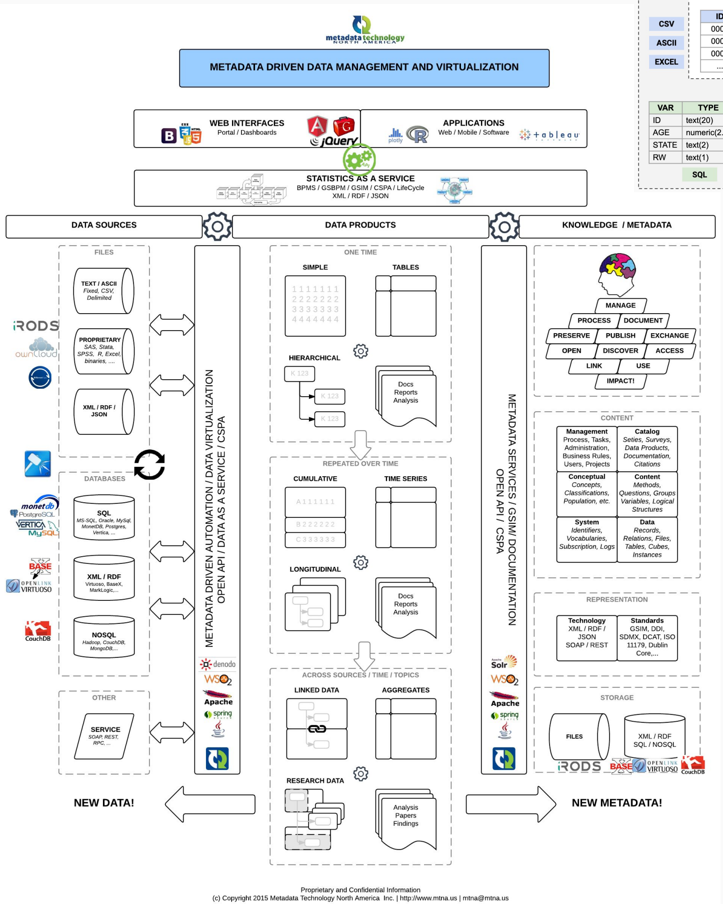
- Costs
  - Must typically be examined on a case by case basis (no magic wand, different institutions have different needs / capacity / priorities). But <$1M can go a long way.
  - Why not start by committing 1,2,3,4,5%+ of the budget? Ask yourself:
    - If I would be a book publisher, how much of my budget would I invest in packaging & marketing?
    - What are the costs of not having metadata driven environments? What is the cost of users (external or internal) searching and recreating the same metadata over and over again?
  - People time investment is a significant portion of the costs. Need champions / leaders.
- ROI
  - Monetary: operational and research cost reduction
  - Resources: Reduce burden on both producers and users (researchers, developers)
  - Quality of: data, service, research, policy/decision making, user satisfaction
  - Leadership: Demonstrate / encourage good practices (statistics and IT)
- Keep in mind at in the end, making and managing change is the biggest challenge
  - Need top level management support and champions

Conclusions

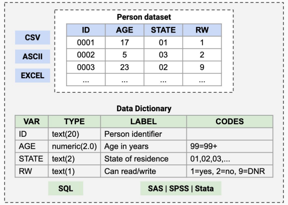
Where to go from here?

- Technological **solutions** and metadata models are **available today**
- Need to put metadata on the road map and **empower information systems**
- Go for the **quick / easy wins**
  - Start on the dissemination side (easier and more visible) and with new data
  - Must be concrete actions and outputs (ideally a "wow!" story)
    - to showcase the benefits and go beyond early adopter (crossing the chasm)
- Perform **assessment** and 2-5 year **roadmap**
- Establish **institutional** board / repository / **standards** (agency and/or inter-agency)
- When possible, do it as a group, not per agency
- Encourage / **promote change** (don't fear it)
- Put the **young generation** in charge (they are born with this)
- **Educate** people about the benefits of metadata (making life easier, minimize the fear of change)
- **Budget** metadata (make sure this can actually happen)

Q&A



metadata technology NORTH AMERICA
METADATA DRIVEN DATA MANAGEMENT AND VIRTUALIZATION
WEB INTERFACES
Portal / Dashboards
APPLICATIONS
Web / Mobile / Software
STATISTICS AS A SERVICE
BPMS / GSBPM / GSIM / CSPA / LifeCycle
XML / RDF / JSON
DATA SOURCES
DATA PRODUCTS
KNOWLEDGE / METADATA
FILES
TEXT / ASCII
Fixed, CSV, Delimited
PROPRIETARY
SAS, Stata, SPSS, R, Excel, binaries, ...
XML / RDF / JSON
DATABASES
SQL
MS-SQL, Oracle, MySql, MonetDB, Postgres, Vertica, ...
XML / RDF
Virtuoso, BaseX, MarkLogic,...
NOSQL
Hadoop, CouchDB, MongoDB,...
OTHER
SERVICE
SOAP, REST, RPC, ...
NEW DATA!
METADATA DRIVEN AUTOMATION / DATA VIRTUALIZATION
OPEN API / DATA AS A SERVICE / CSPA
ONE TIME
SIMPLE
TABLES
HIERARCHICAL
Docs Reports Analysis
REPEATED OVER TIME
CUMULATIVE
TIME SERIES
LONGITUDINAL
Docs Reports Analysis
ACROSS SOURCES / TIME / TOPICS
LINKED DATA
AGGREGATES
RESEARCH DATA
Analysis Papers Findings
METADATA SERVICES / GSIM/ DOCUMENTATION
OPEN API / CSPA
MANAGE
PROCESS
DOCUMENT
PRESERVE
PUBLISH
EXCHANGE
OPEN
DISCOVER
ACCESS
LINK
USE
IMPACT!
CONTENT
Management
Process, Tasks, Administration, Business Rules, Users, Projects
Catalog
Series, Surveys, Data Products, Documentation, Citations
Conceptual
Concepts, Classifications, Population, etc.
Content
Methods, Questions, Groups Variables, Logical Structures
System
Identifiers, Vocabularies, Subscription, Logs
Data
Records, Relations, Files, Tables, Cubes, Instances
REPRESENTATION
Technology
XML / RDF / JSON
SOAP / REST
Standards
GSIM, DDI, SDMX, DCAT, ISO 11179, Dublin Core,...
STORAGE
FILES
XML / RDF
SQL / NOSQL
NEW METADATA!
iRODS
ownCloud
monetdb
PostgreSQL
VERTICA
MySQL
BASE
OPENLINK VIRTUOSO
CouchDB
denodo
WSO2
Apache
spring
Solr
Proprietary and Confidential Information
(c) Copyright 2015 Metadata Technology North America Inc. | http://www.mtna.us | mtna@mtna.us



CSV
ASCII
EXCEL

Person dataset

ID	AGE	STATE	RW
0001	17	01	1
0002	5	03	2
0003	23	02	9
...	...	...	...

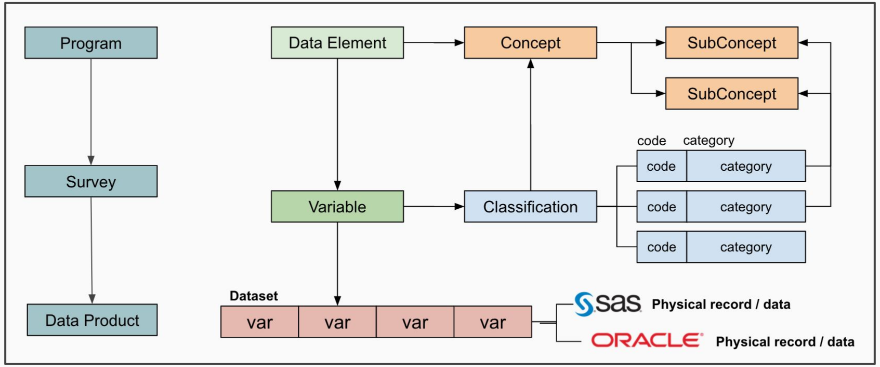
Data Dictionary

VAR	TYPE	LABEL	CODES
ID	text(20)	Person identifier	
AGE	numeric(2.0)	Age in years	99=99+
STATE	text(2)	State of residence	01,02,03,...
RW	text(1)	Can read/write	1=yes, 2=no, 9=DNR

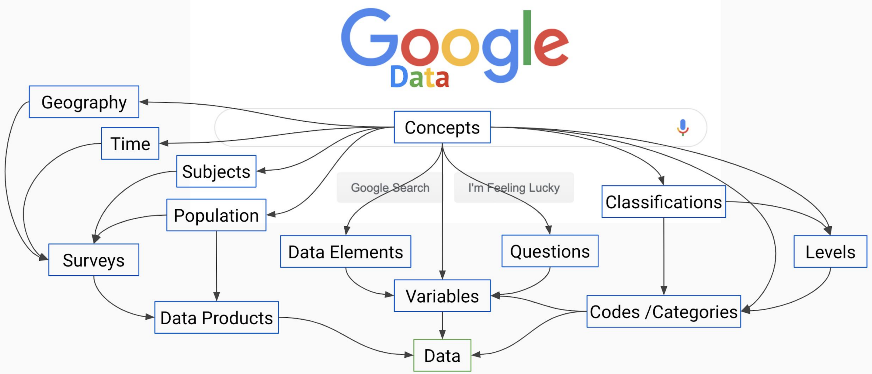
SQL

SAS | SPSS | Stata

DCAT-AP FOR DATA PORTALS IN EUROPE · GSIM · ddi Metadata powered by DDI · sdmx Statistical Data and Metadata eXchange · JSON-LD · Dublin Core Metadata Initiative · JSON · RDF · XML · W3C



Program
Survey
Data Product
Data Element
Variable
Concept
Classification
SubConcept
SubConcept
code
category
code
category
code
category
code
category
Dataset
var
var
var
var
sas
Physical record / data
ORACLE
Physical record / data



Google Data
Geography
Time
Subjects
Population
Surveys
Concepts
Google Search
I'm Feeling Lucky
Classifications
Data Elements
Questions
Levels
Variables
Data Products
Codes /Categories
Data