Metadata at Statistics Canada

Canadian Metadata Forum

September 19-20, 2003

Corporate metadata at Statistics Canada

- Integrated Metadatabase (IMDB)
  - Collection of facts about each of Statistics Canada's 400+ surveys
  - Aimed at helping human users interpret statistical data
    - Survey description
    - Methodololgy
    - Concepts and variables measured
    - Data quality

Project status

- Data base implemented in November 2000
  - covers survey description, methodology and data quality
- Published on STC website, with daily updates
- Extensive efforts in improving metadata quality

What is the IMDB based on?

ISO 11179 Specification and Standardization of Data Elements

1. Framework for specification and standardization
2. Classification
3. Basic attributes
4. Rules and guidelines for formulation of definitions
5. Naming and identification
6. Registration

… What is the IMDB based on?

- Model for the Corporate Metadata Repository (CMR) of US Census Bureau by Dan Gilman
- Earlier work by Bo Sundgren of Statistics Sweden

Administered Component

- Any item that is defined and may be reused or shared
- Any item requiring registration

Administered Component

Organization

Contact

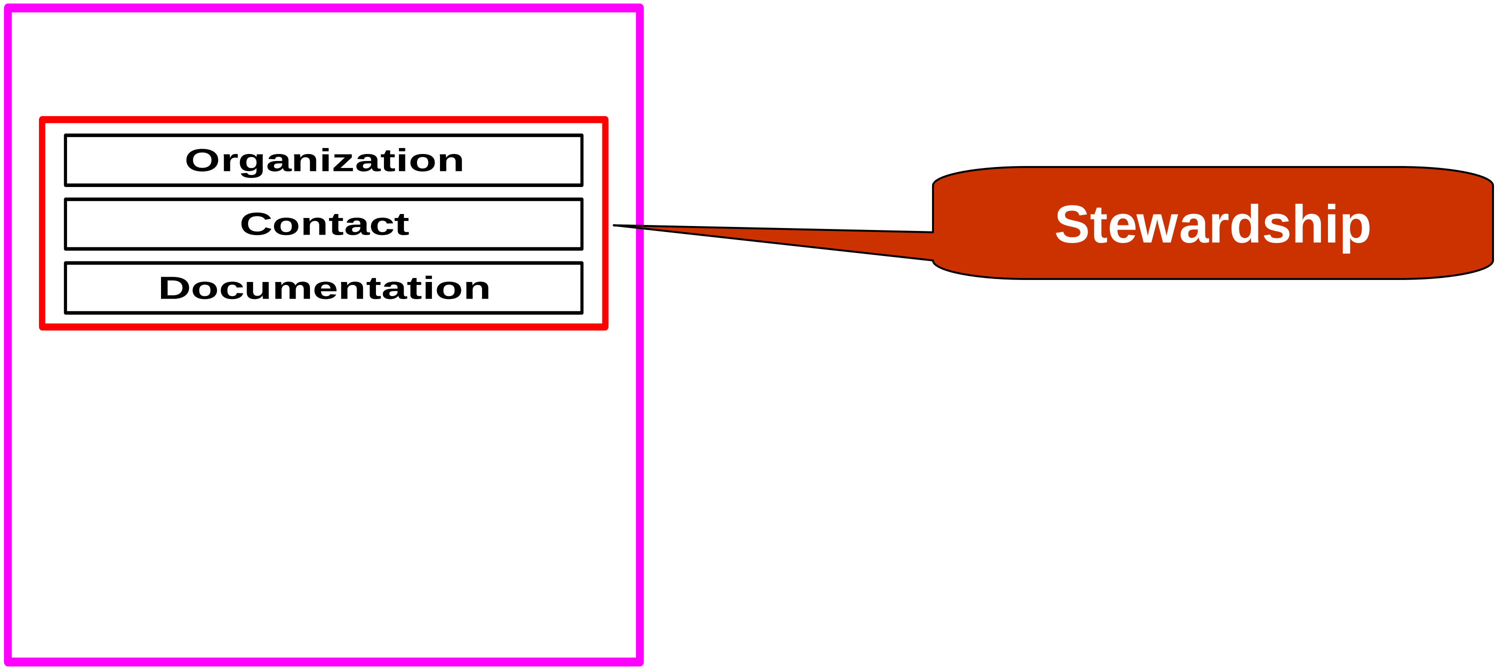
Documentation

Identification

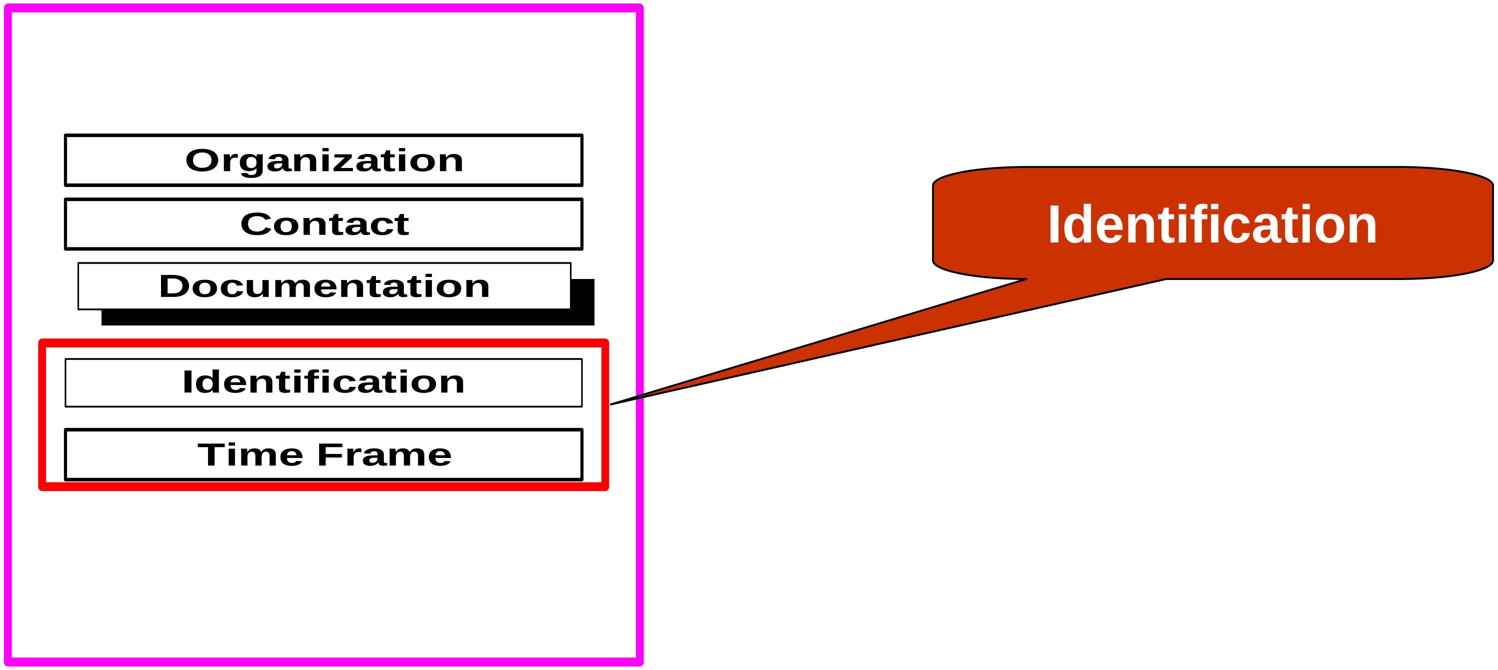
Time Frame

Keyword

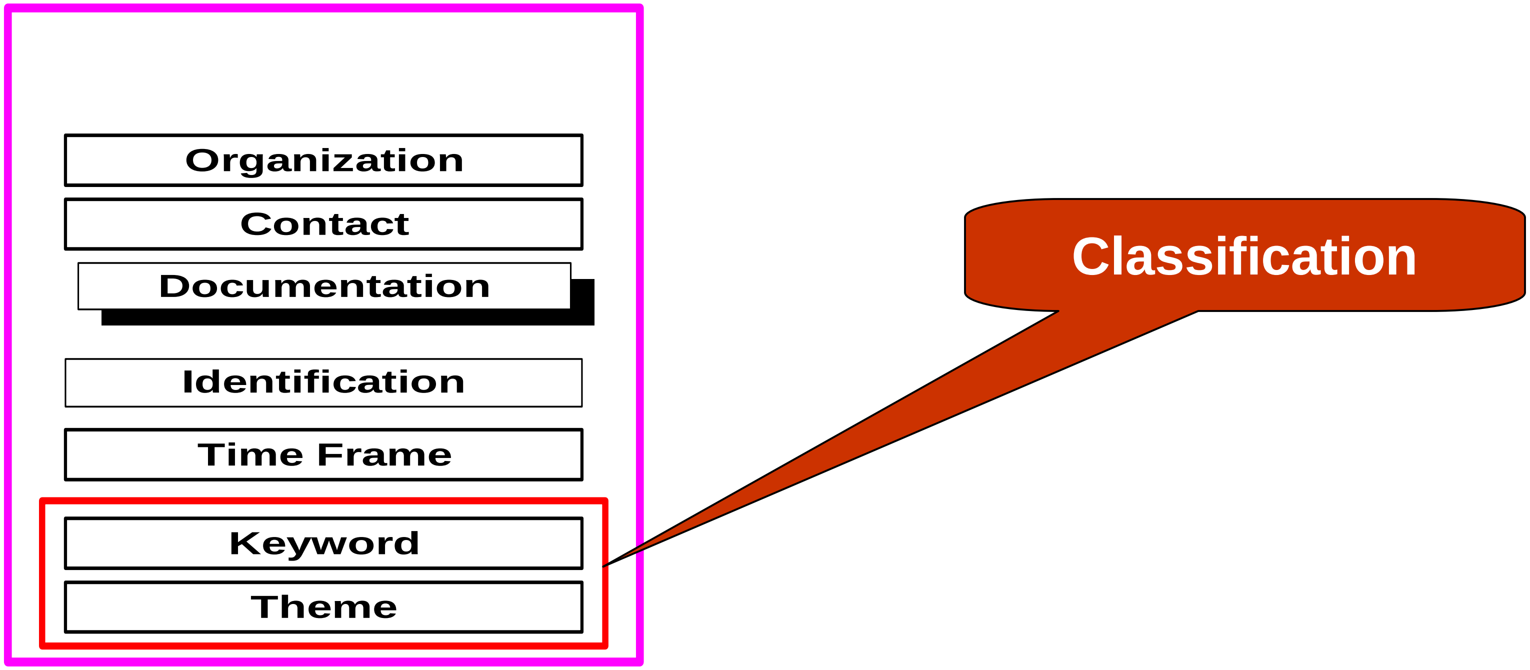
Theme



- Administration aspect of the component



- Naming and identification



Organization

Contact

Documentation

Identification

Time Frame

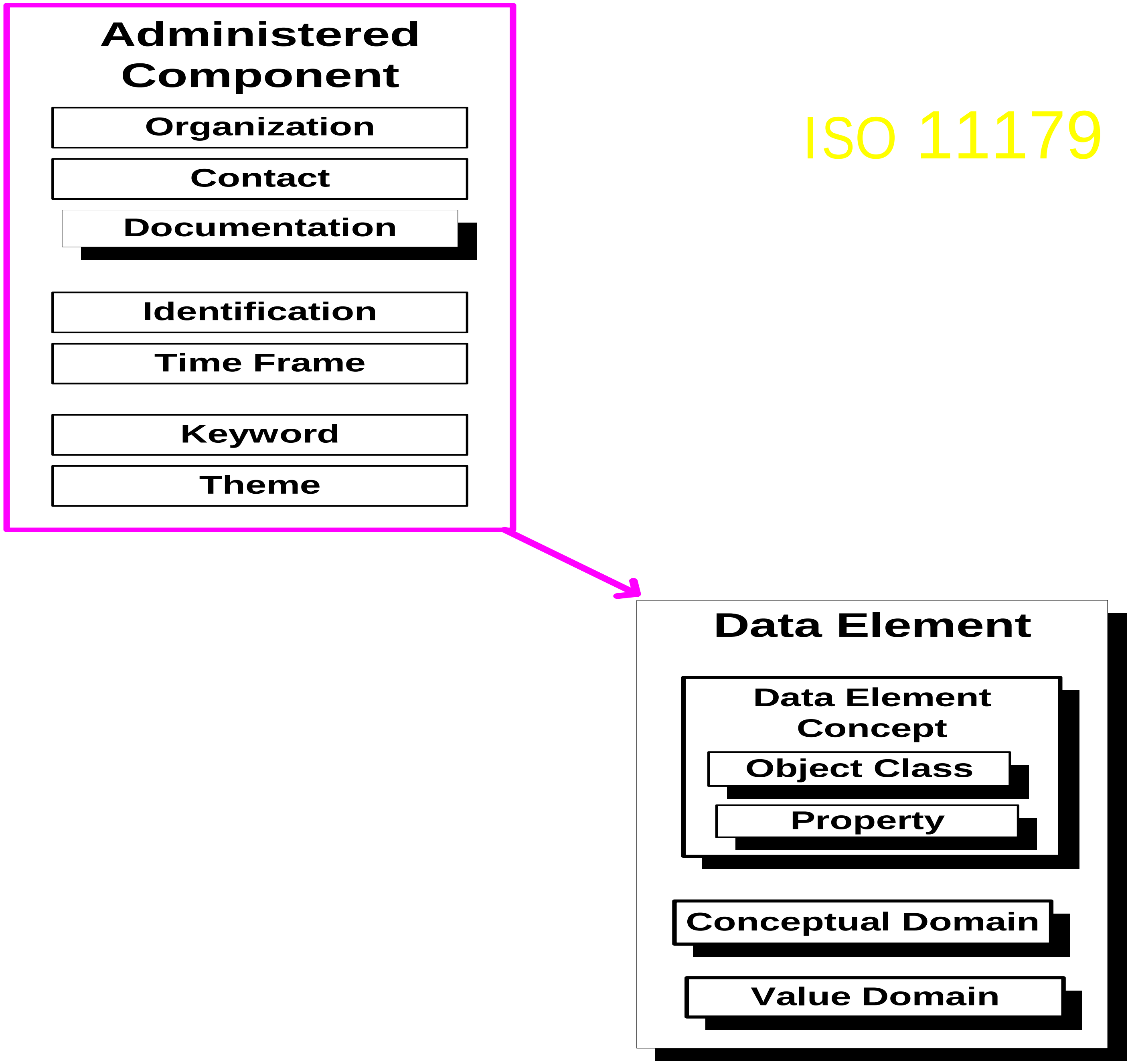
Keyword

Theme

Classification

- Management of classification schemes

ISO 11179



Administered Component

Organization

Contact

Documentation

Identification

Time Frame

Keyword

Theme

Data Element

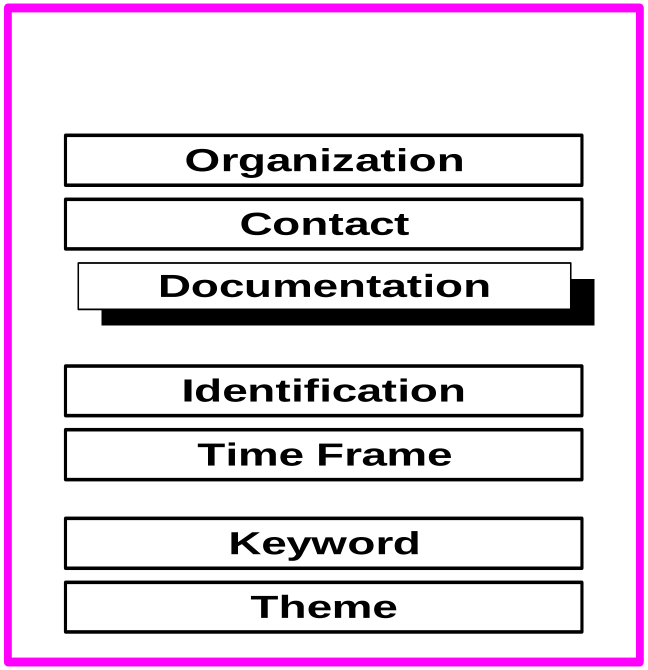
Data Element Concept

Object Class

Property

Conceptual Domain

Value Domain



Organization

Contact

Documentation

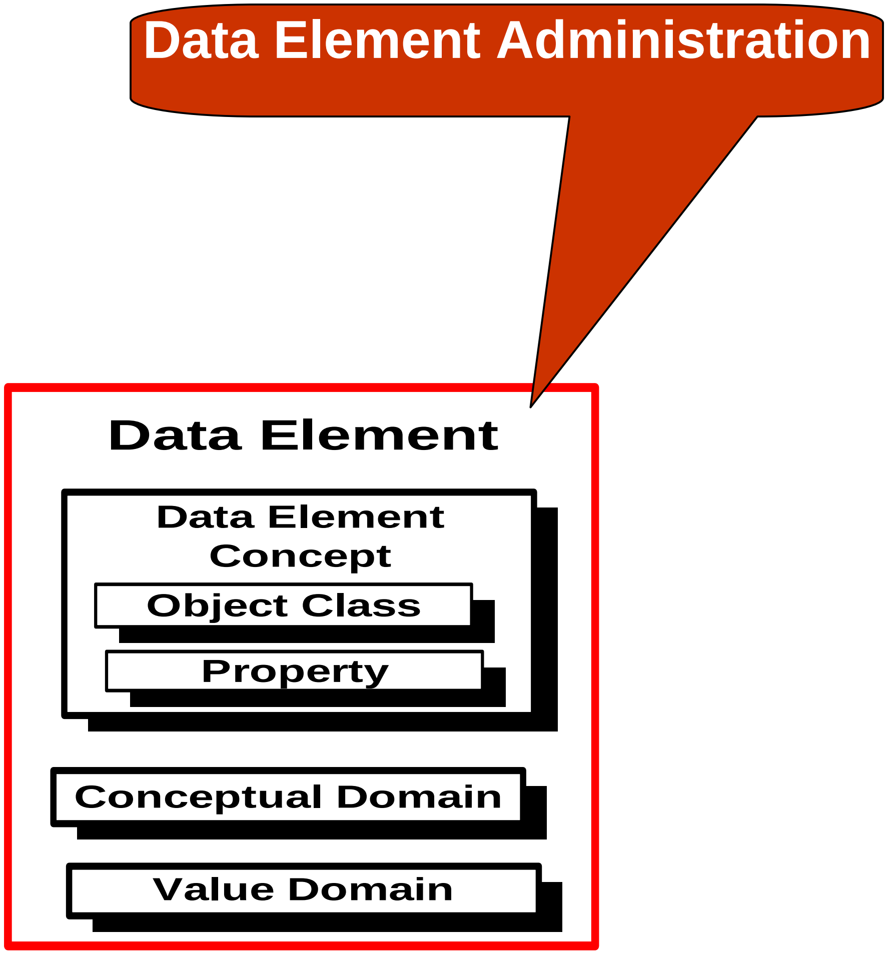
Identification

Time Frame

Keyword

Theme

- definition
- representation
- permissible values



Data Element Administration

Data Element

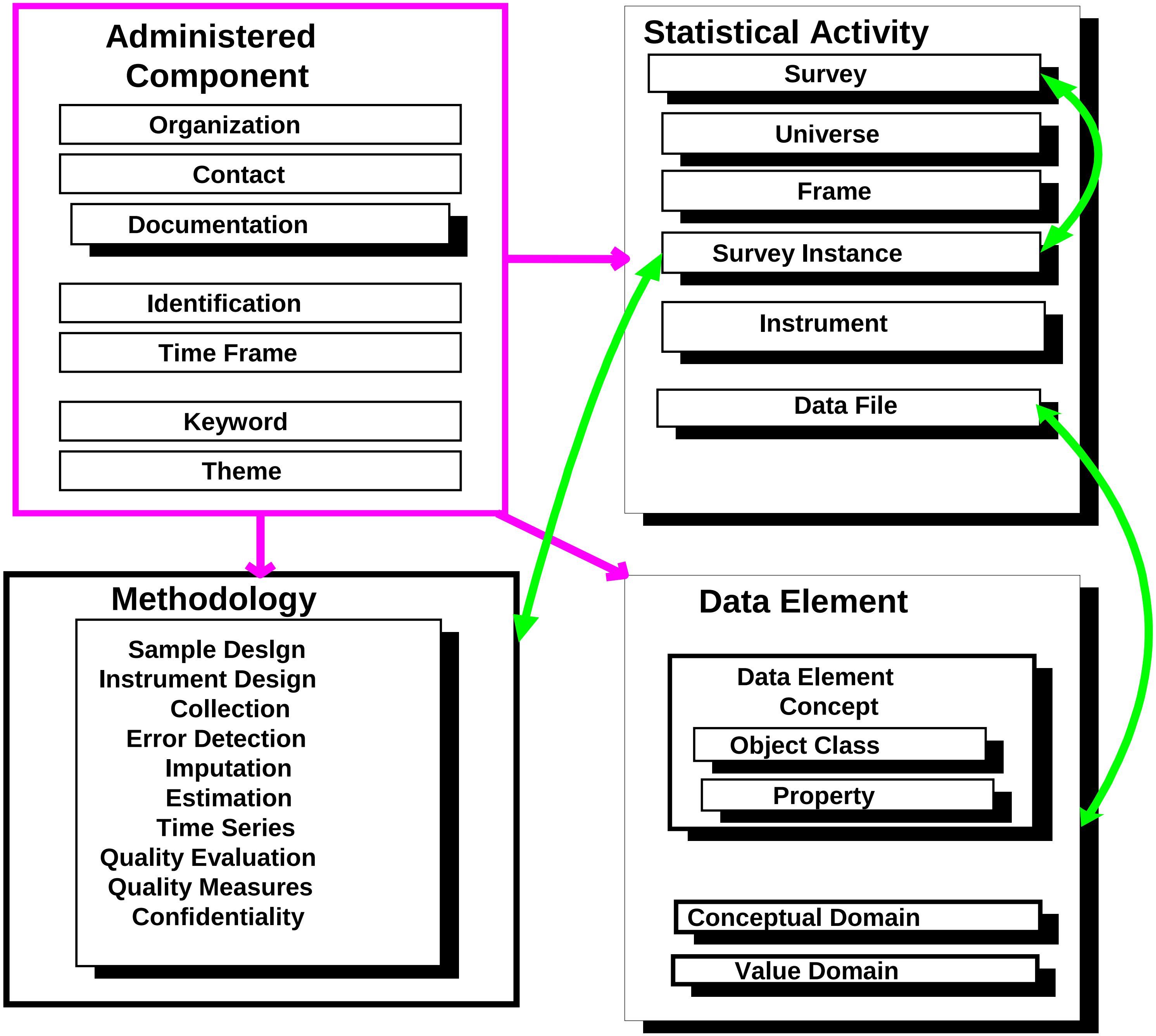
Data Element Concept

Object Class

Property

Conceptual Domain

Value Domain



Administered Component

Organization

Contact

Documentation

Identification

Time Frame

Keyword

Theme

Statistical Activity

Survey

Universe

Frame

Survey Instance

Instrument

Data File

Methodology

Sample DesIgn
Instrument Design
Collection
Error Detection
Imputation
Estimation
Time Series
Quality Evaluation
Quality Measures
Confidentiality

Data Element

Data Element Concept

Object Class

Property

Conceptual Domain

Value Domain

Next Phase of IMDB

- Extending the content of the IMDB database to include the concepts, variables and classifications published for every STC survey
- Focus first on data published through CANSIM

Expected benefits

- Most frequently cited “missing” metadata in recent market research
- Fulfill requirements of Policy on Informing Users of Data Quality and Methodology

Additions to STC web site

- For every survey, list of variables published, with hyperlinks to definitions, classification used and source of on-line data (CANSIM, Daily, Canadian Statistics table)
- On Statistical Methods page, searchable list of all variables, with links to definitions, classifications and source of on-line data
- New hyperlinks in CANSIM to definitions stored in IMDB

Implications

- To be stored in IMDB, information on variables must be consistently structured
- To be listed in web pages, variables must be meaningfully named
- To be most effectively searchable, variables must be consistently named

Structure of information on variables in IMDB

- *Statistical unit* + property + representation = *Variable*
- Statistical unit is agent, event or item about which data are produced
- Property is characteristic of statistical unit being measured
- Representation is form given to resulting data, e.g. Quantity, Value, Type

Naming convention

- All three elements used to create name of variable
  - Value of sales of establishment
  - Type of assets of establishment
  - Name of geographic location of person
  - Type of occupation of person
  - Value of GDP of economy

IMDB Phase III Data Element Model

Object Class

A set of ideas, abstractions, or things in the real world that can be identified with explicit boundaries and meaning whose properties and behavior follow the same rules.

•At STC = statistical unit

•Can be an agent, event or item

Macro Statistical Units

- In order to comprehensively cover the data and information published by Statistics Canada, different views are accommodated within the framework.
- Four Macro Statistical Units were chosen to provide four different views within the framework.

The four views

- The Macro Statistical Units are:
  - **People**
  - **Economy**
  - **Environment**
  - **The State**
- The four views divide the framework into four different sections

Fundamental Statistical Units: Definition

- Fundamental Statistical Units are defined as those that are *not* types of any other unit and can *not* be derived as grouping of any other unit.
- Fundamental Statistical Units keep the model simple and robust by limiting and organising the number of Statistical Units.

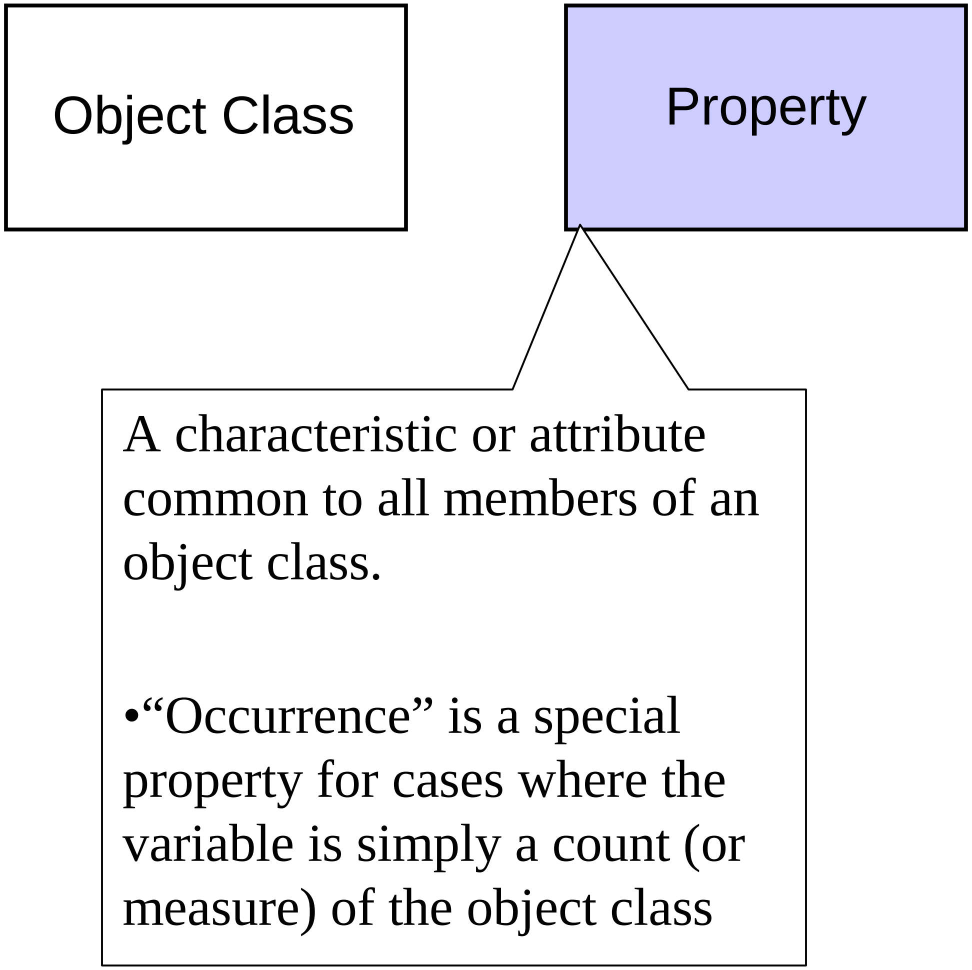
Fundamental Statistical Unit: Types

- Agents: Statistical Units that operate and whose operations are reported on by Statistics Canada .
- Events: Statistical Units that represent the actions of (or by) Agents as reported by Statistics Canada. Events are defined as occurrences that are discrete in time (occur in time period) and finite (can be counted).
- Items: Other Statistical Units reported on by Statistics Canada that are generally created by Agents.

Commonly Used Derivations of Fundamental Statistical Units

- Subclasses based on inherent characteristics
- Roles
- Aggregations

IMDB Phase III Data Element Model



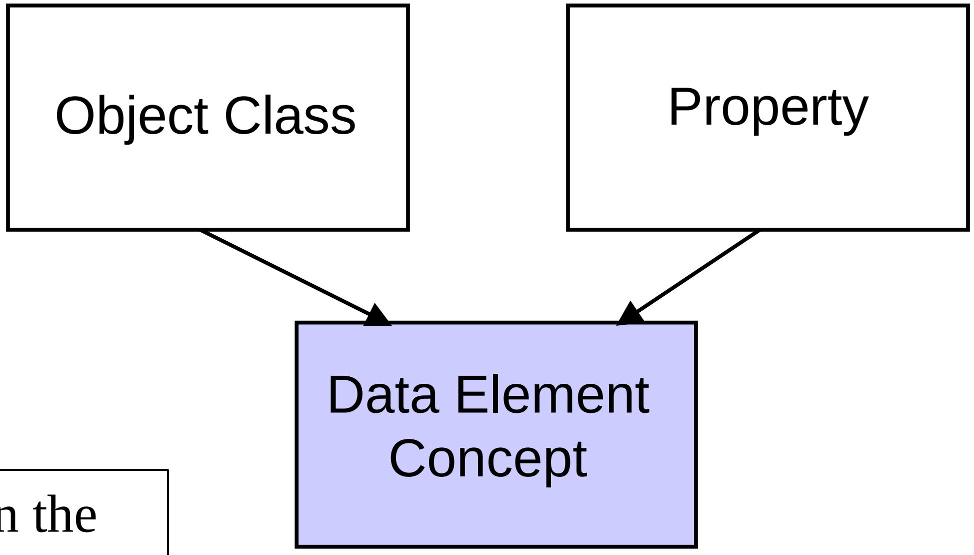
Object Class

Property

A characteristic or attribute common to all members of an object class.

- “Occurrence” is a special property for cases where the variable is simply a count (or measure) of the object class

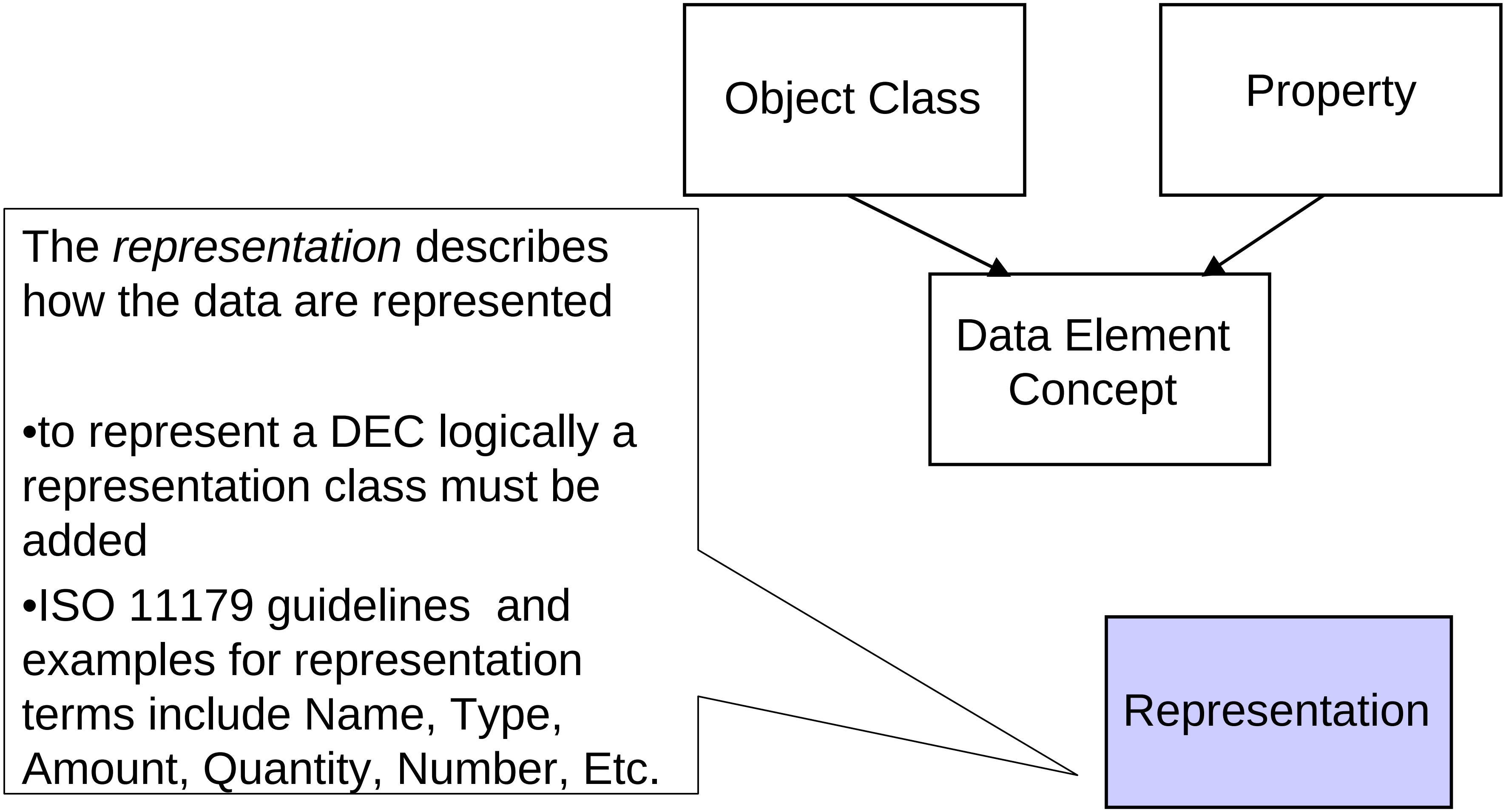
IMDB Phase III Data Element Model



A concept that can be represented in the form of a data element, described independently of any particular representation.

•amalgamation of the object class and the property.

IMDB Phase III Data Element Model



Object Class

Property

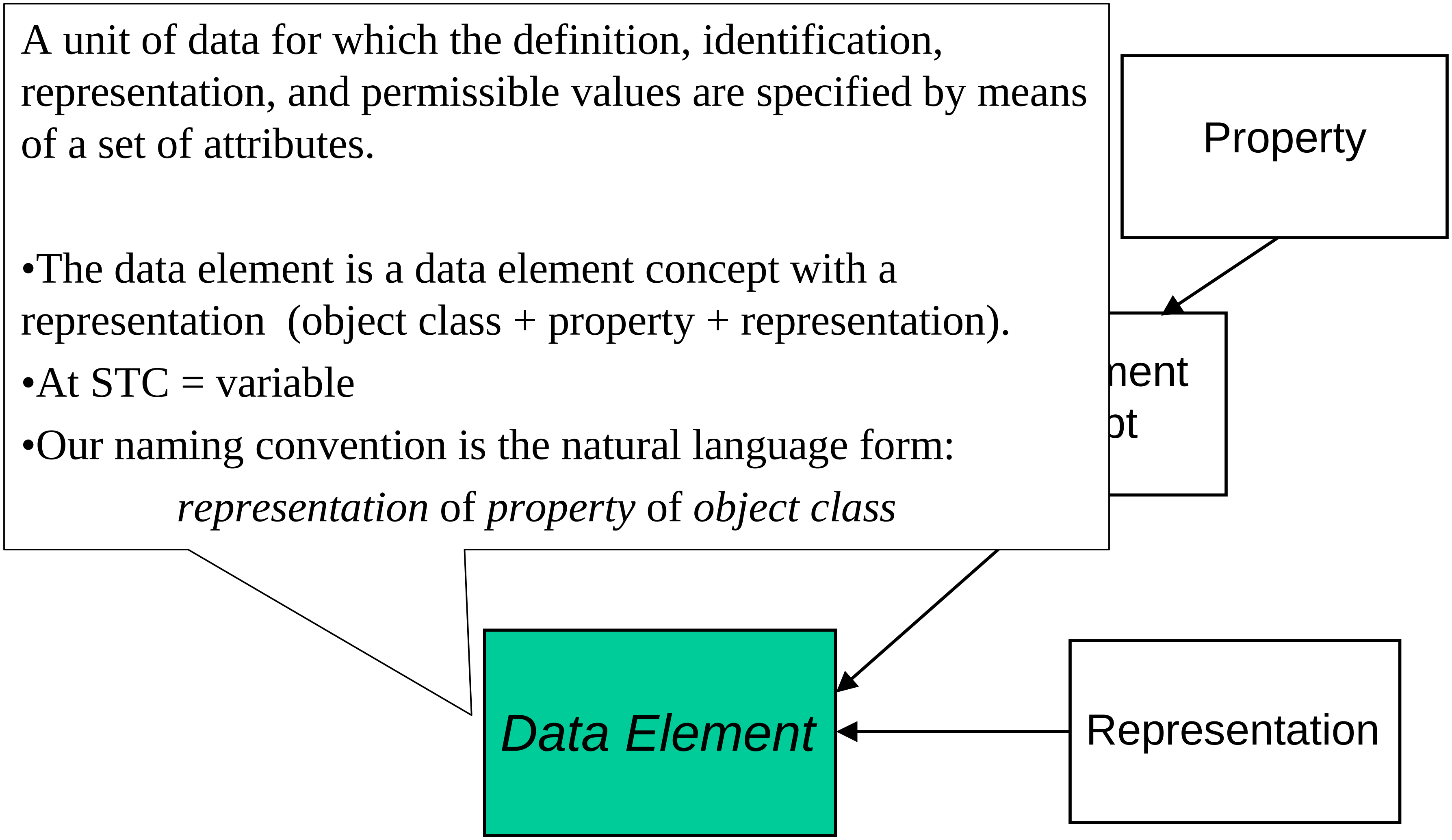
Data Element Concept

Representation

The *representation* describes how the data are represented

- to represent a DEC logically a representation class must be added
- ISO 11179 guidelines and examples for representation terms include Name, Type, Amount, Quantity, Number, Etc.

IMDB Phase III Data Element Model



A unit of data for which the definition, identification, representation, and permissible values are specified by means of a set of attributes.

- The data element is a data element concept with a representation (object class + property + representation).
- At STC = variable
- Our naming convention is the natural language form:

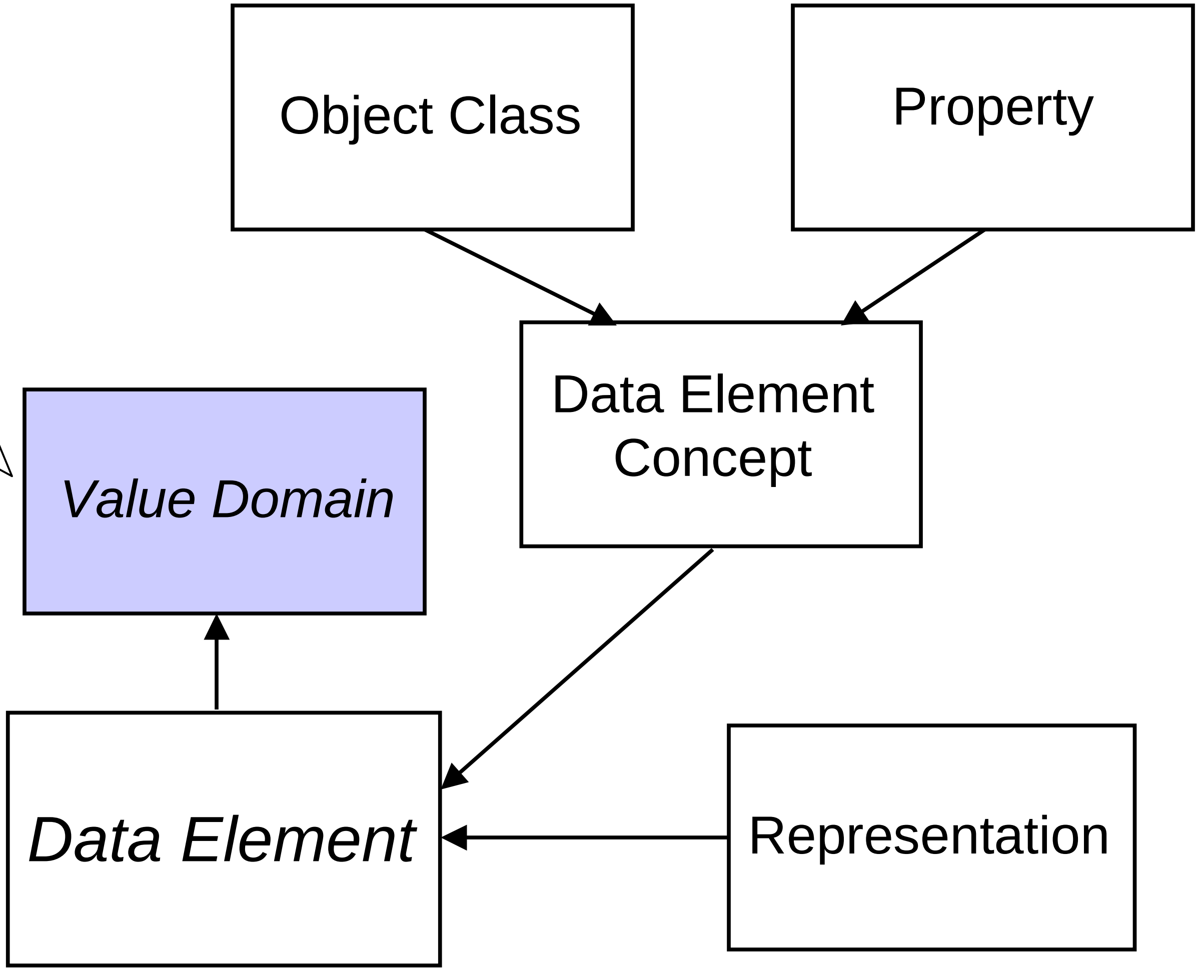
representation of *property* of *object class*

Data elements in a statistical agency

- Data most often presented in tabular form
- Data elements are dimensions of statistical tables
- Data element thus defined can have many value domains
- Data element has one and only one value domain in the context of a given datafile = Data element - value domain - table map

IMDB Phase III Data Element Model

- Set of permissible values and their associated meanings
- AT STC = classification
- Can have several value domains per DE
- Can be enumerated or non-enumerated



Value domains and classifications

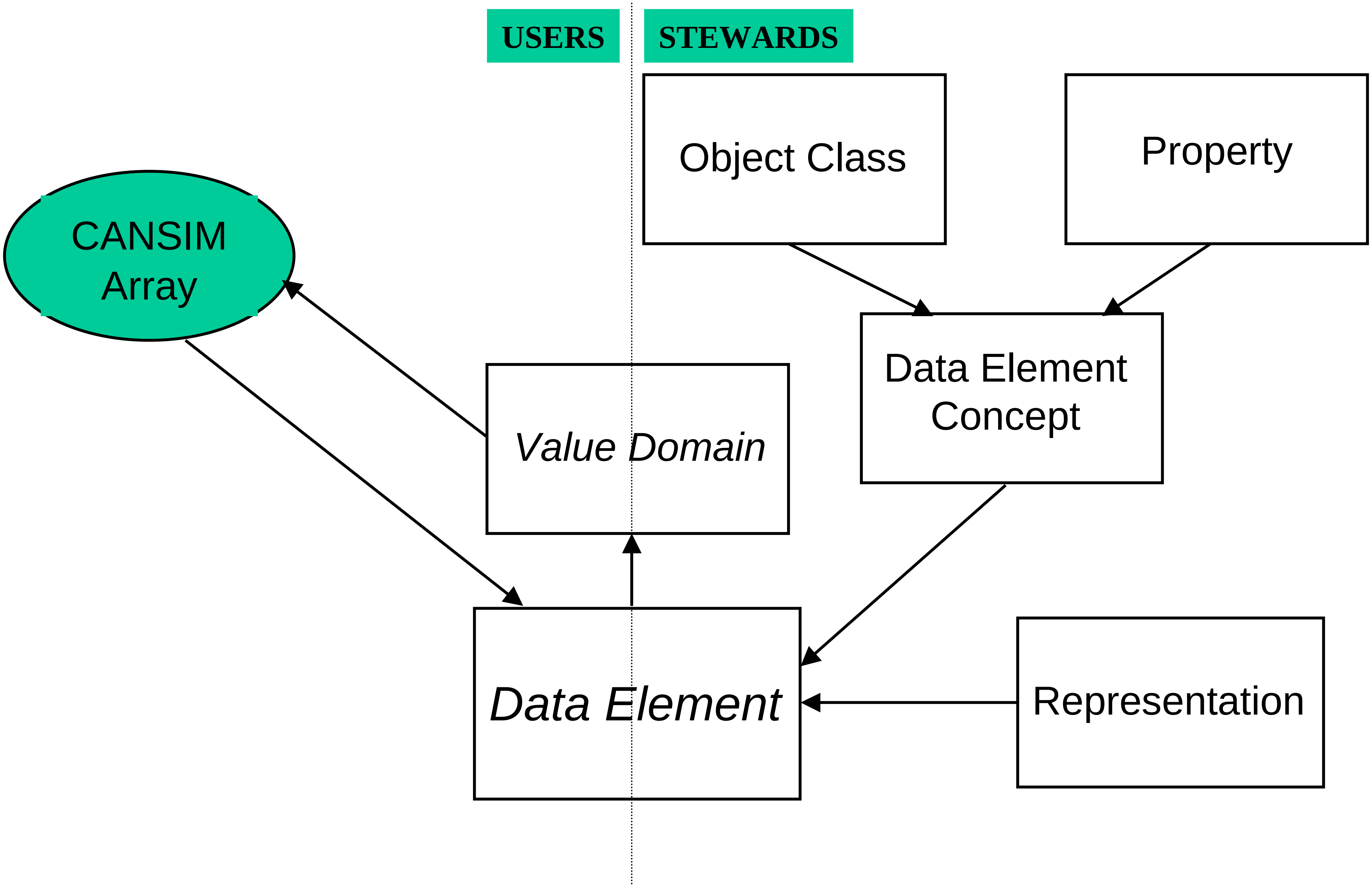
- Hierarchically related value domains structured as classifications or taxonomies
- Assign levels to value domains and parent-child relationships between levels and between permissible values

Value Domain – Standard Classification

Table 1

Current Account (Standard Classification)									
Goods	Services				Investment Income			Current Transfers	
	Travel	Transportation	Commercial Services	Government Services	Direct	Portfolio	Other	Private	Official

IMDB Phase III Data Element Model



USERS

STEWARDS

CANSIM Array

Object Class

Property

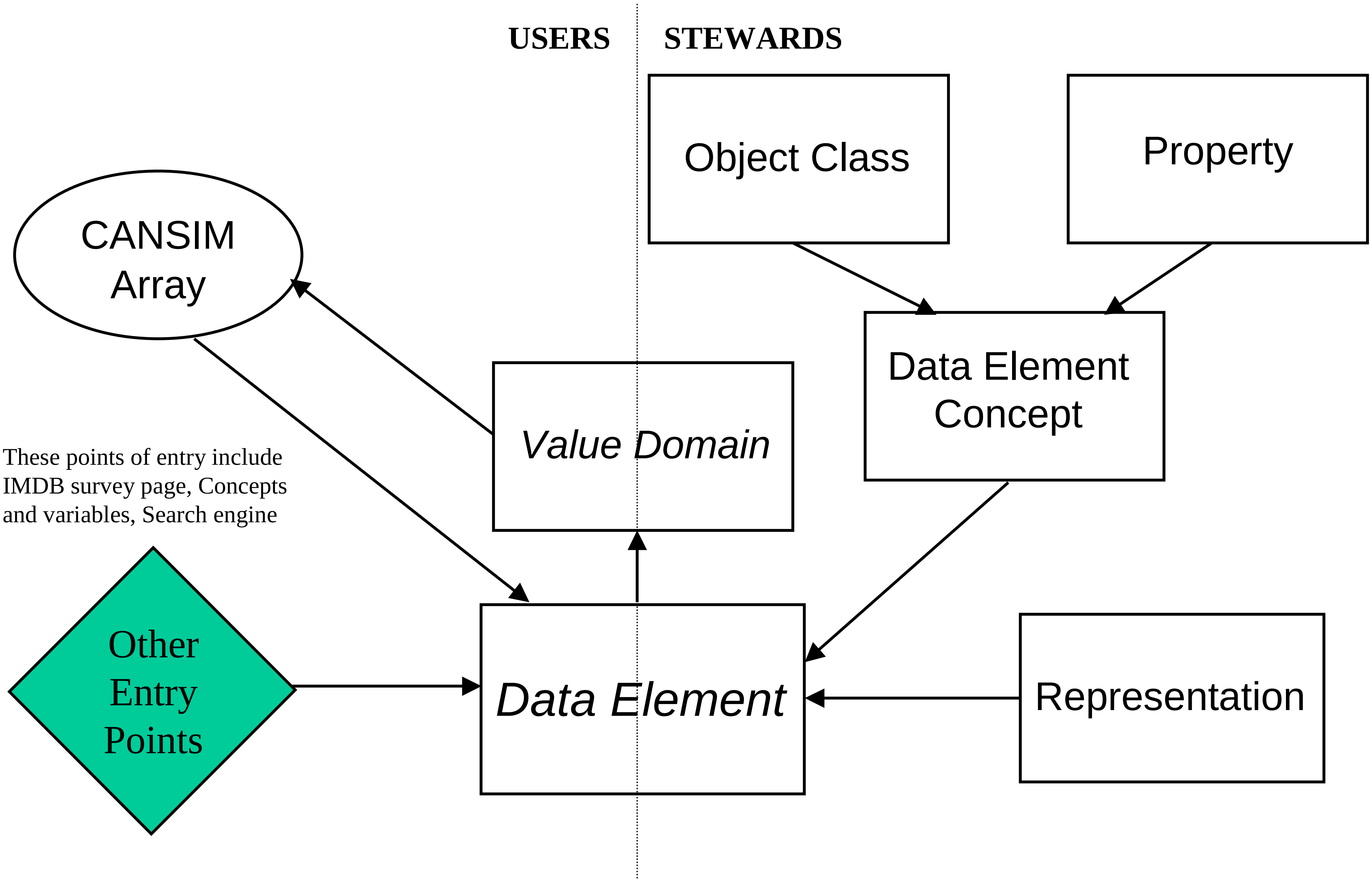
Data Element Concept

Value Domain

Data Element

Representation

IMDB Phase III Data Element Model



These points of entry include IMDB survey page, Concepts and variables, Search engine

Next steps

- Produce lists of variables and definitions from IMDB test environment for review and discussion with subject-matter areas
- Activate on STC Intranet for trial period
- Roll-out into production

Conclusion

- IMDB is a significant corporate infrastructure for Information Management
- Comprehensiveness and quality are continuously improving, with strong management support
- Will provide an additional tool to help users find and interpret data published by Statistics Canada