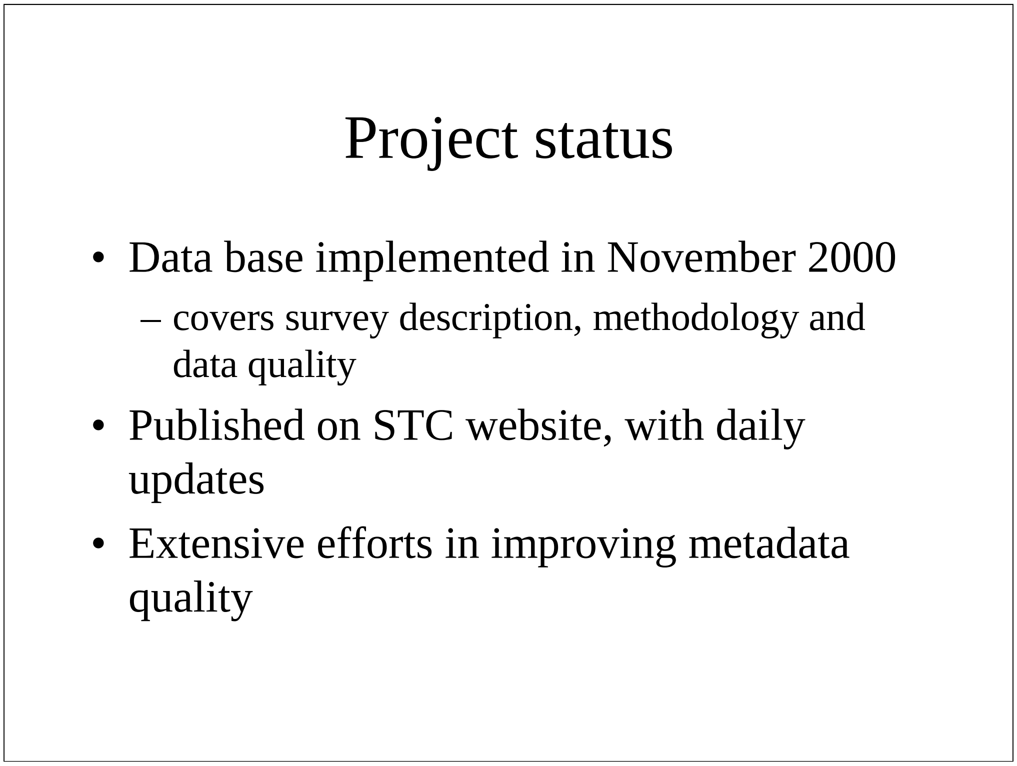
Metadata at Statistics Canada

Canadian Metadata Forum

September 19-20, 2003

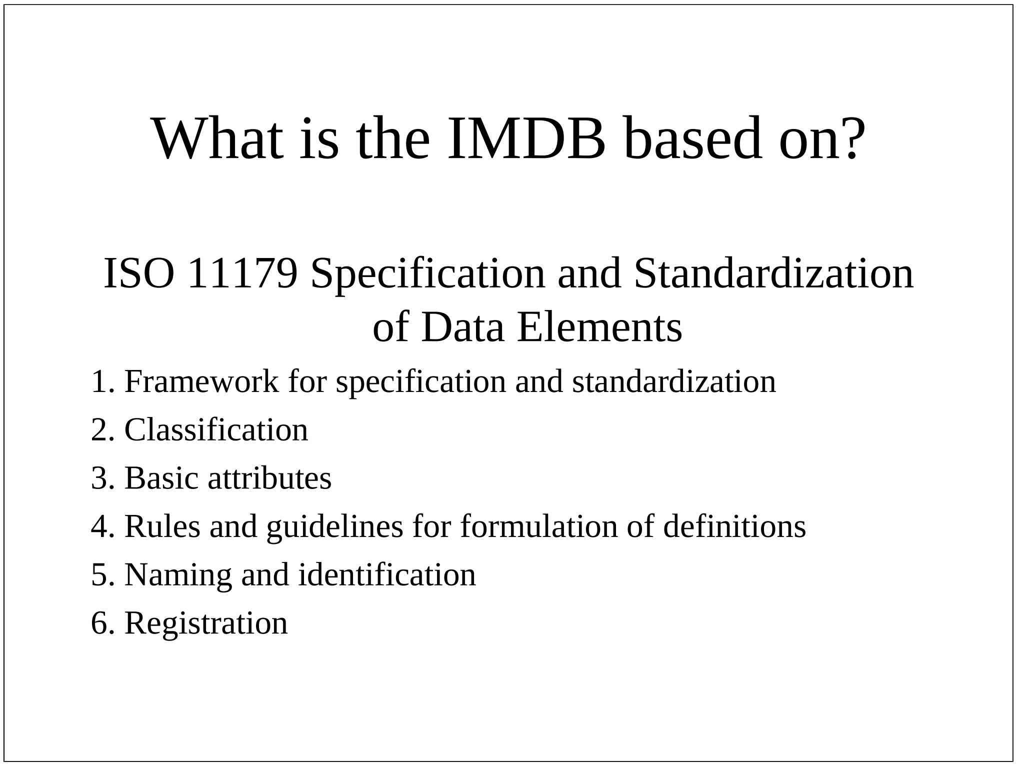
Corporate metadata at Statistics Canada

- 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?

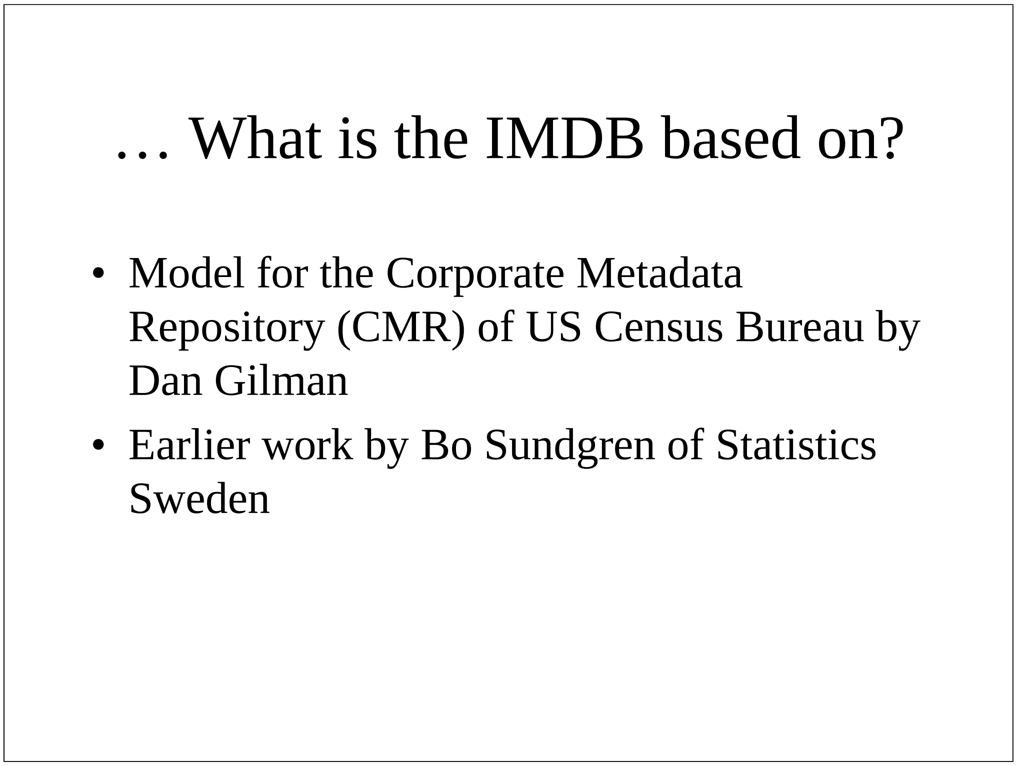
ISO 11179 Specification and Standardization of Data Elements

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

6 part Standard

Attributes include:

- identifying
- definitional
- relational (classification schemes, keywords)
- representational
- administrative



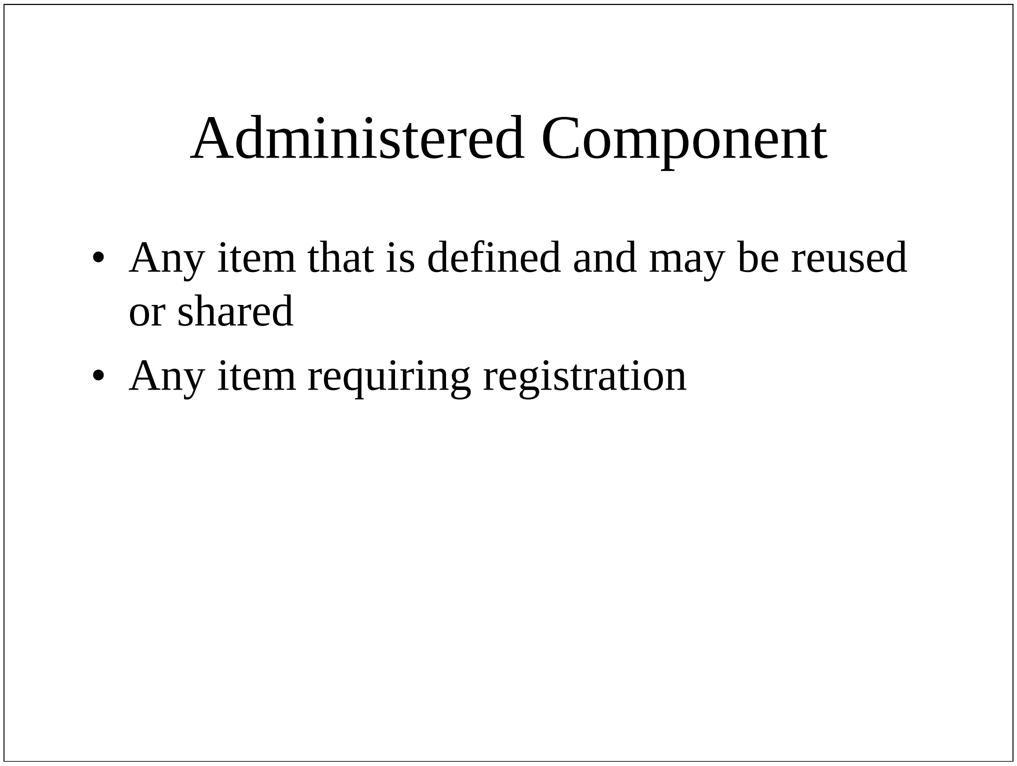
… What is the IMDB based on?

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

ANSI 3. shows the structure and relationship among parts 1 to 6 of ISO 11179.

What is ANSI 3.285?

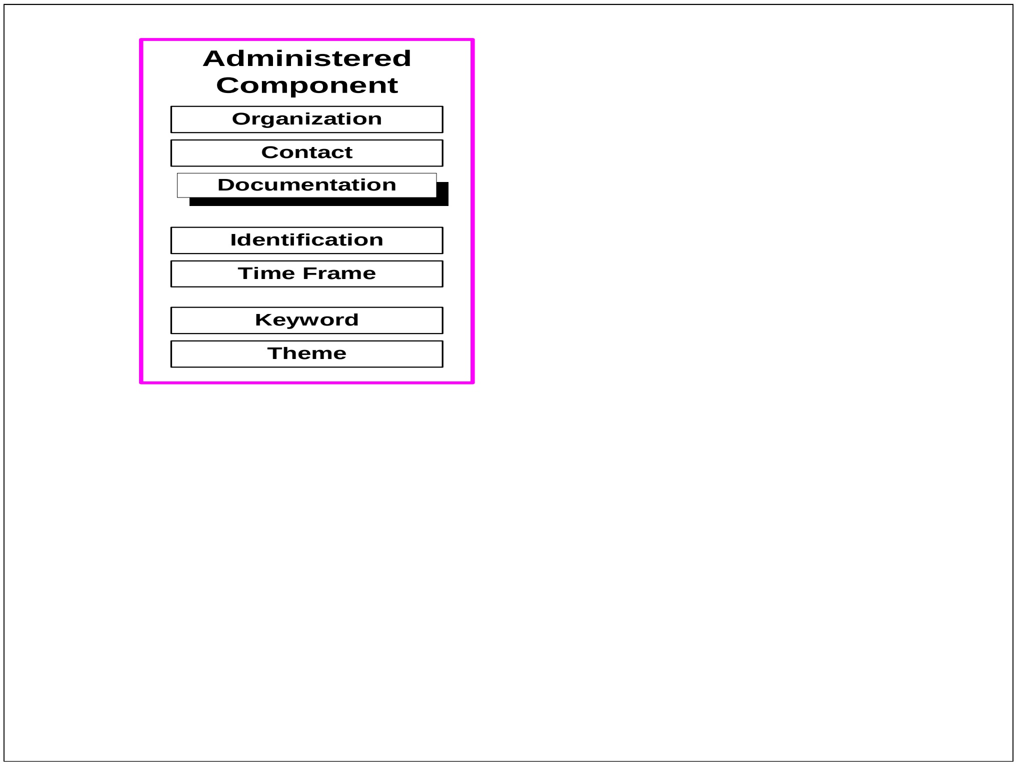
The ANSI 3.285 consists of 6 regions. I will describe each region.



Administered Component

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

Administered component = Any item that is managed, tracked and organised.



Administered Component

Organization

Contact

Documentation

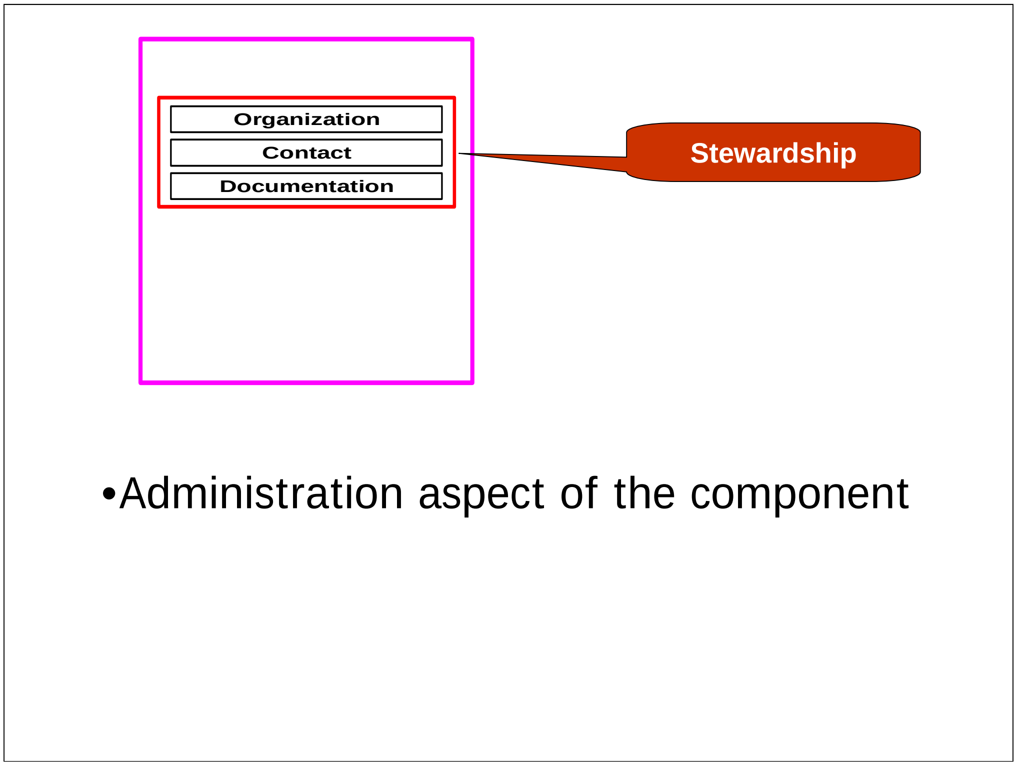
Identification

Time Frame

Keyword

Theme

These are the items that are required to administer a component - manage, track and organise.



Organization

Contact

Documentation

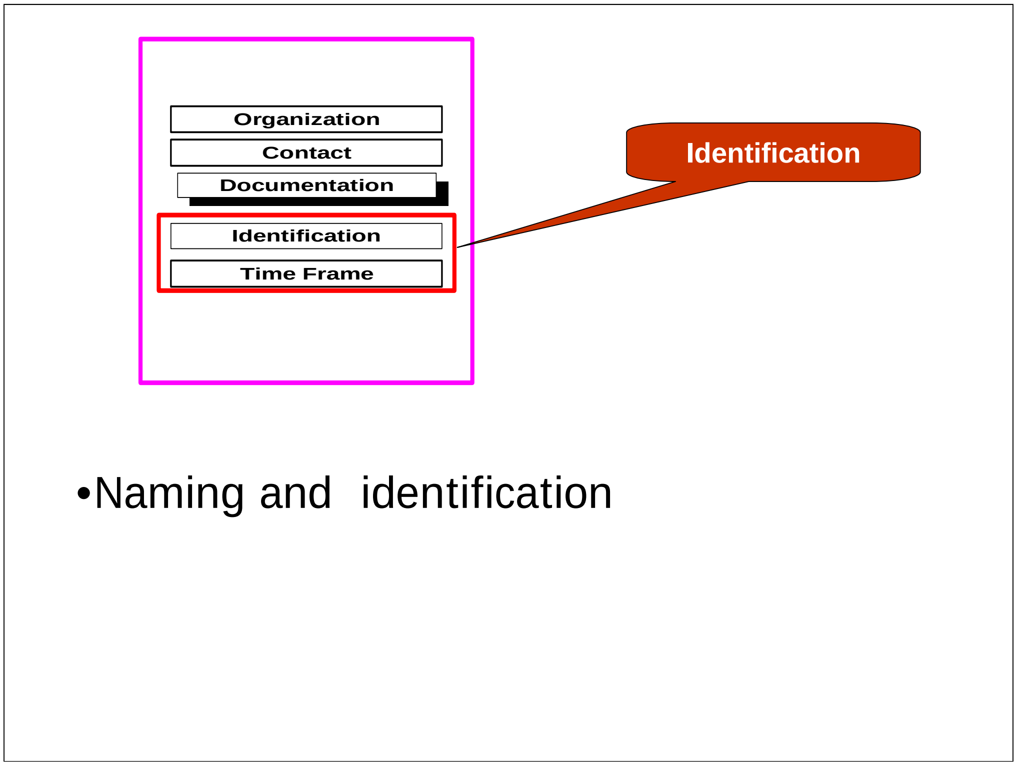
Stewardship

- Administration aspect of the component

The stewardship region supports the administration aspect of the components

- organisation that interact with the component and contacts within the organization
    * responsible organisation = owner
    * submitting organisation that submits the component for registration
    * registrar - the organisation who registers the component

- supporting documentation of the component



Organization

Contact

Documentation

Identification

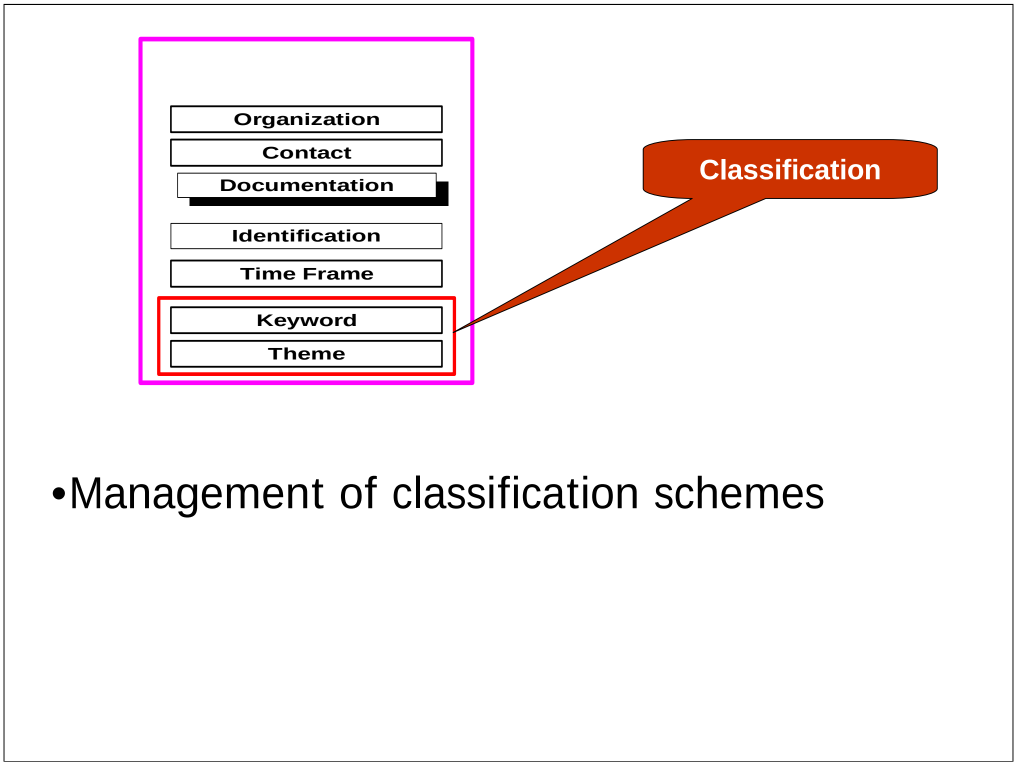
Time Frame

Identification

- Naming and identification

The identification region manages the names of the component.

Time Frame was added here to bring a time context for the component identified. The Time Frame can also be part of the Stewardship.



Organization

Contact

Documentation

Identification

Time Frame

Keyword

Theme

Classification

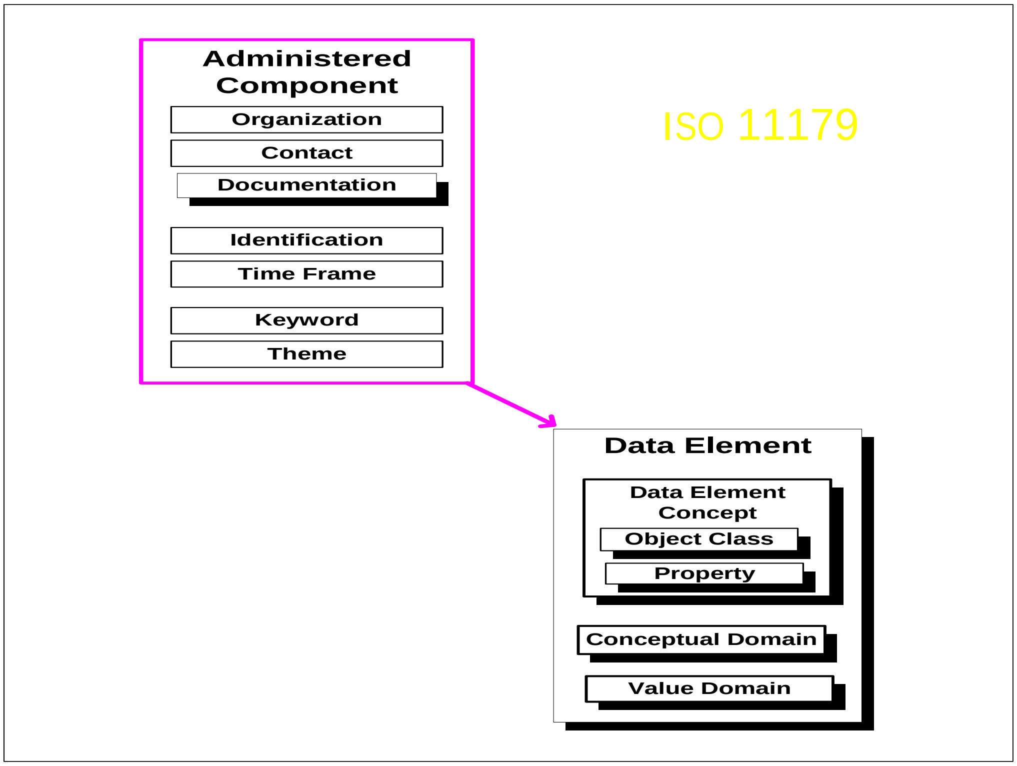
- Management of classification schemes

Classification region is the

- management of classification schemes
- and management of components that are classification schemes.

Keywords which is a list of preferred terms from the LIBRARY’s thesaurus.

Themes list from the LIBRARY.



With respect to the ISO11179, it is the management, organisation and tracking of the DE and each of the components that form it.

The shadowed components are AC.

They are defined, registered and can be reused/shared.

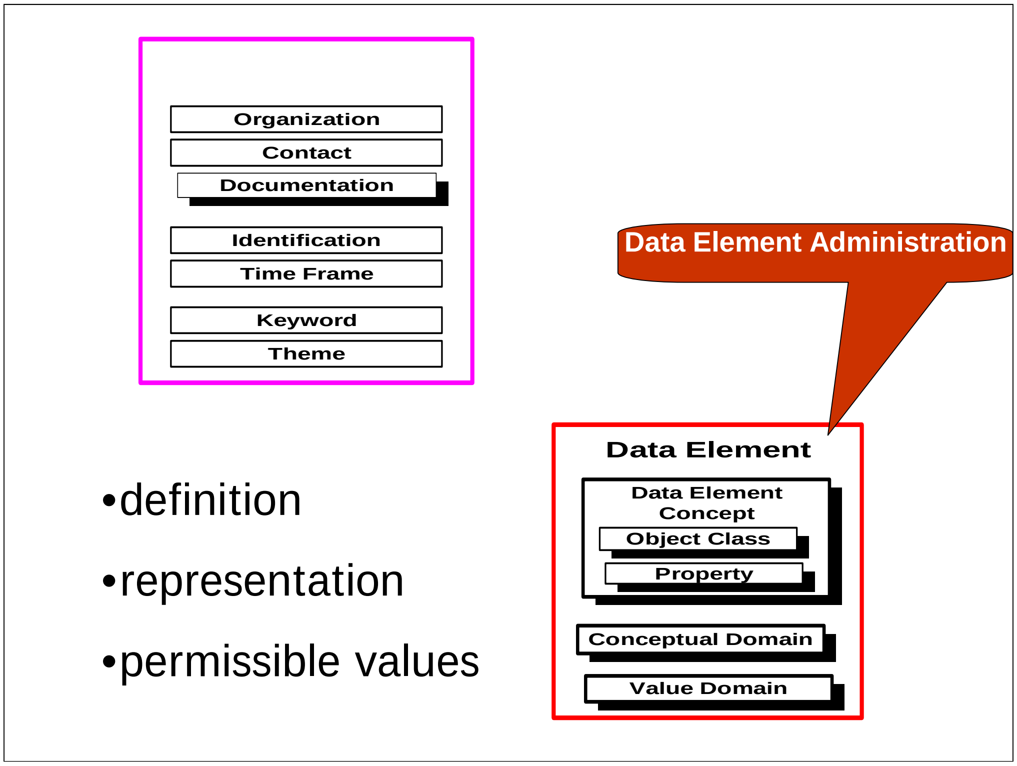
They are sub-types of the AC and therefore share all the relationships of the AC.

The ORG, CONTACT, DOCUMENTATION…etc

As an example the DEC AC in addition to its relationship with OBJECT CLASS and PROPERTY, it has relationships with ORG, CONTACT, DOC and the remaining items in the AC box.

Note that Documentation which is part of the Stewardship region, is also an AC.

Earlier I had mentioned that the ANSI 3.285 was extend to include the statistical activities. What does this mean?



Organization
Contact
Documentation
Identification
Time Frame
Keyword
Theme

Data Element Administration

- definition
- representation
- permissible values

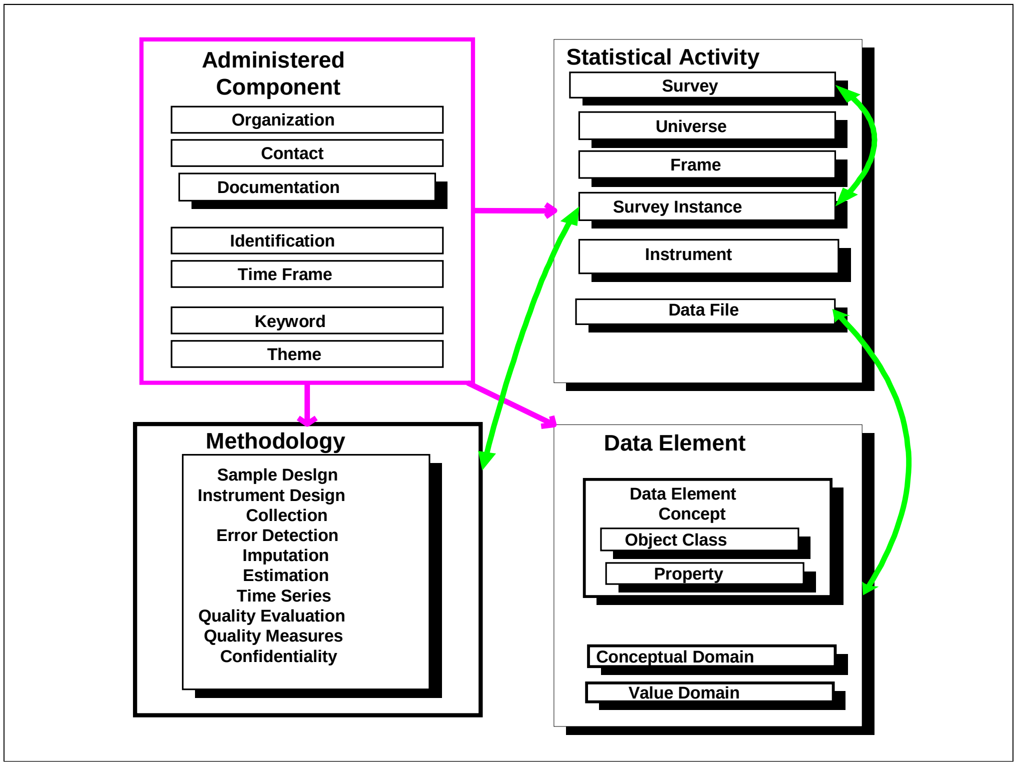
Data Element
Data Element Concept
Object Class
Property
Conceptual Domain
Value Domain

Data Element Administration region covers the management of a DE.

DE is a unit of data with
def’n attributes
identification attributes
representation attributes
permissible values attributes

DEC is Object Class + Property -- No representation

DE is DEC + CD + VD (DEC + Representation)
These are the components that are required to fully define a shareable DE.



It means that all the components of the statistical activities are also managed, organised and tracked by the administered component.

In addition to the relationship among the components themselves, they also have a relationship with AC.

STAT ACTIVITY AC = SURYEY, UNIVERSE etc

Methodology = Sample Design, Instrument Design etc…

Each are AC's, but because of space constraints, they have been grouped together.

Some of the relationships are displayed in the green arrows:

Survey ….

Examples of Data Element links are:

Questions are asked for Data Elements

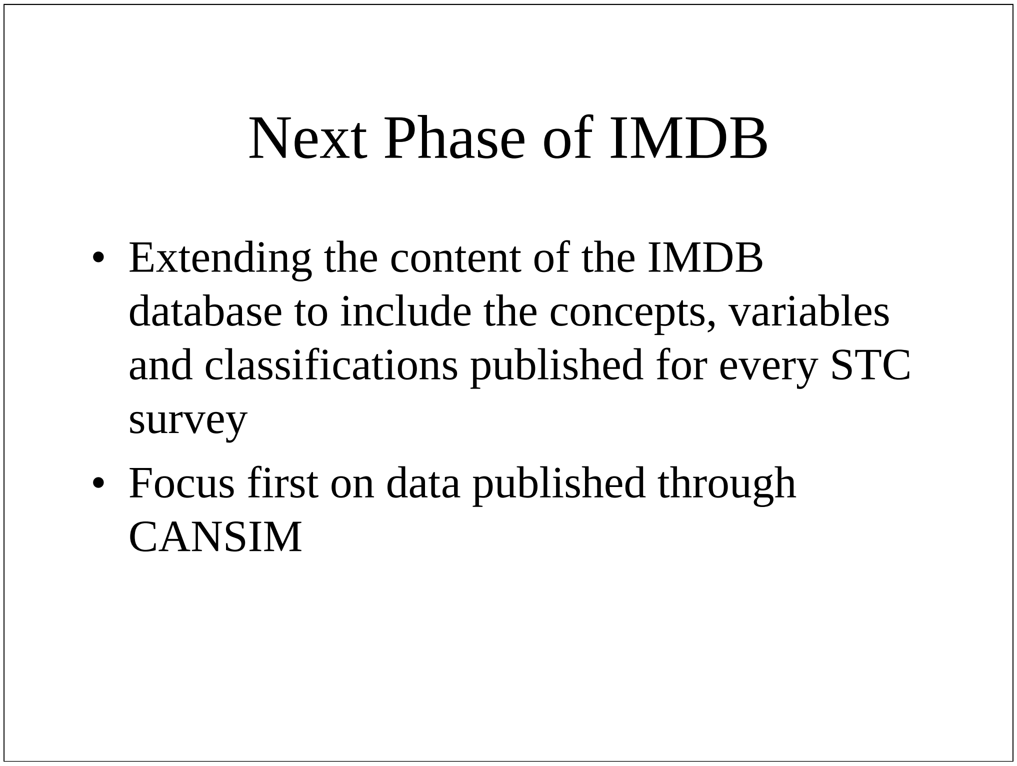
DataFile contain Data Elements

Universe is defined by DEC.

Frames are defined by DEC.

This concludes the overview of the ANSI 3.285 and ISO 11179.

(shows the structure and relationship among parts 1 to 6)



Next Phase of IMDB

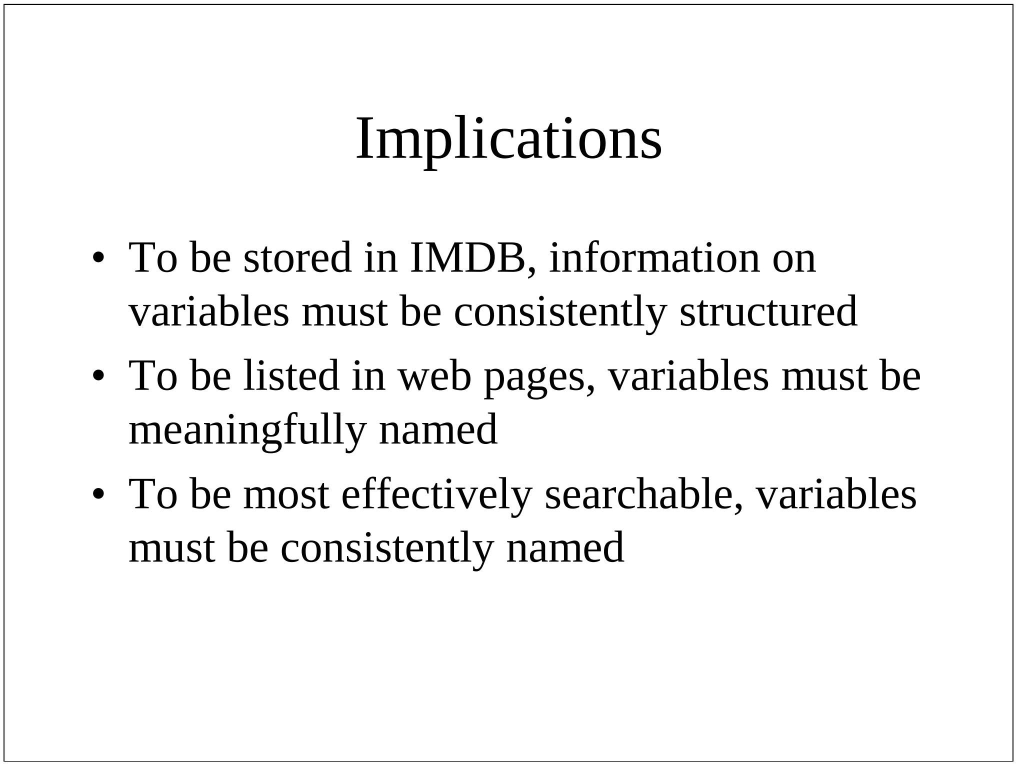
- Extending the content of the IMDB database to include the concepts, variables and classifications published for every STC survey
- Focus first on data published through CANSIM

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

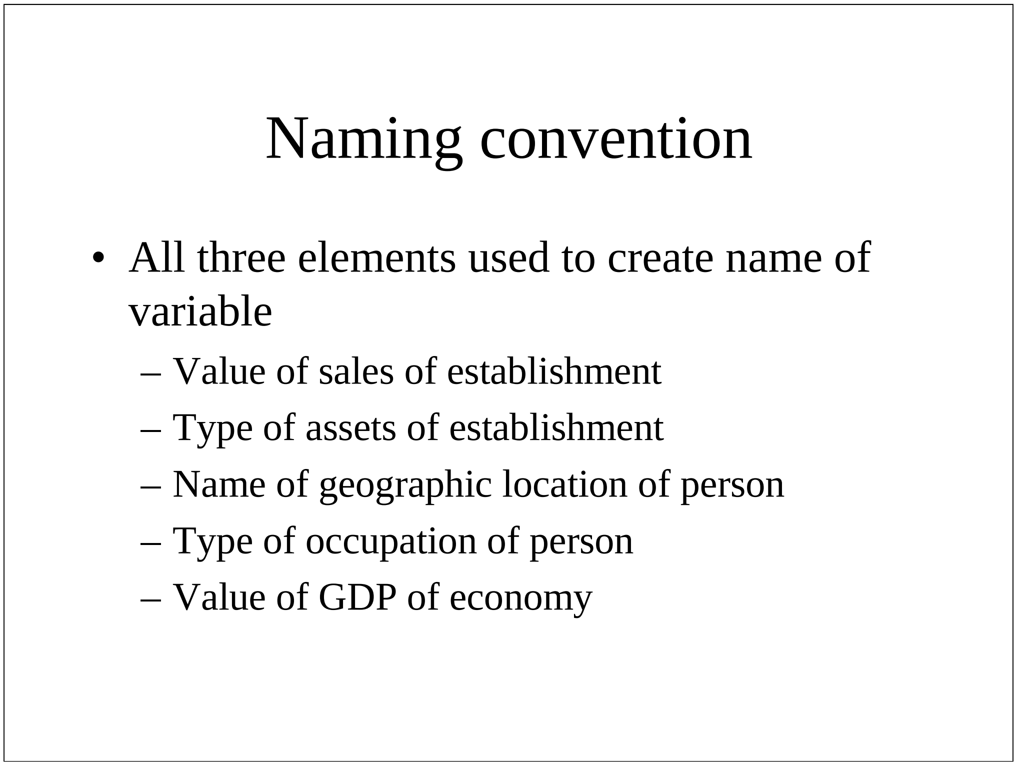


Implications

- 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

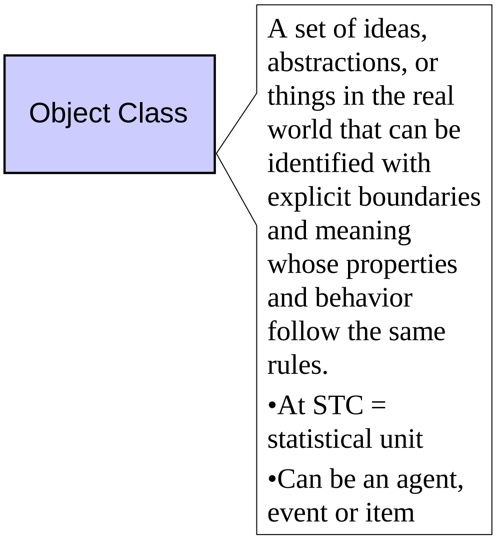
- *Statistical unit* + property + representation = *Variable*
- Statistical unit is agent, event or item about which data are produced
- Property is characteristic of statistical unit being measured
- Representation is form given to resulting data, e.g. Quantity, Value, Type



Naming convention

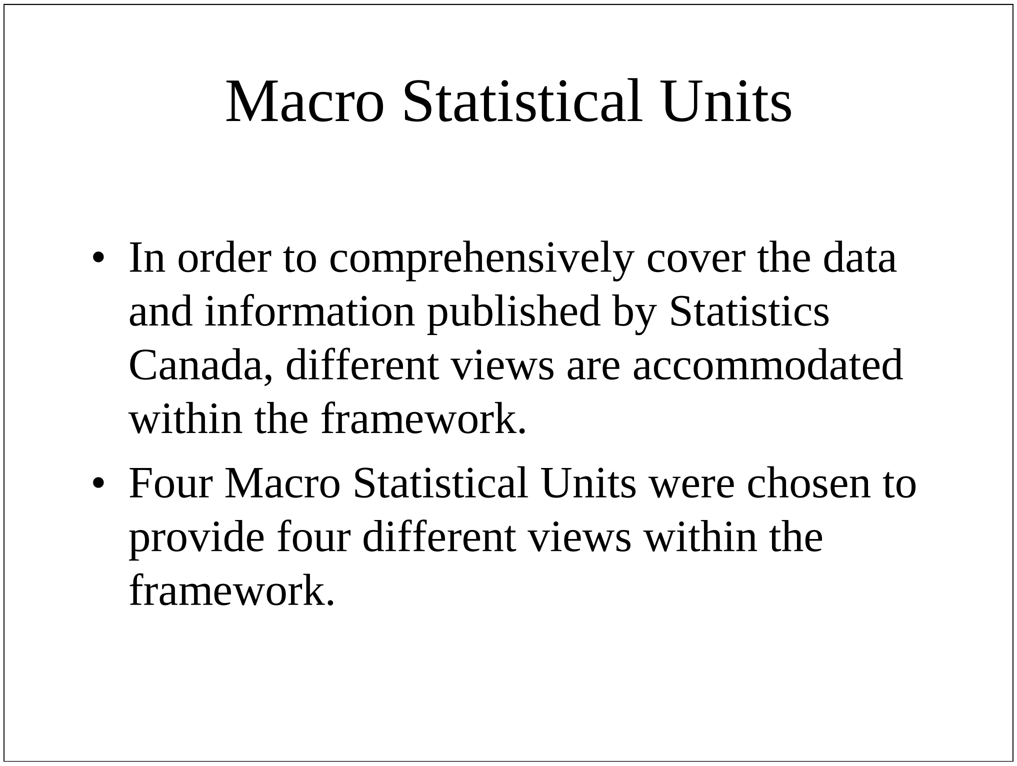
- All three elements used to create name of variable
  - – Value of sales of establishment
  - – Type of assets of establishment
  - – Name of geographic location of person
  - – Type of occupation of person
  - – Value of GDP of economy

IMDB Phase III Data Element Model



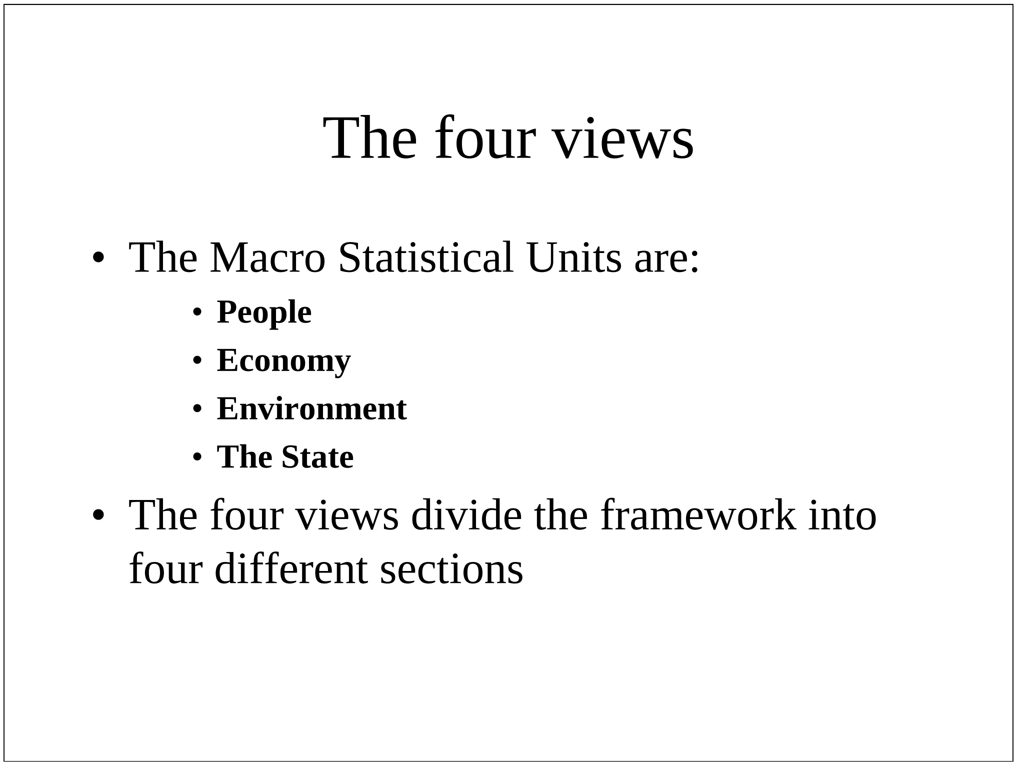
A set of ideas, abstractions, or things in the real world that can be identified with explicit boundaries and meaning whose properties and behavior follow the same rules.

- 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.

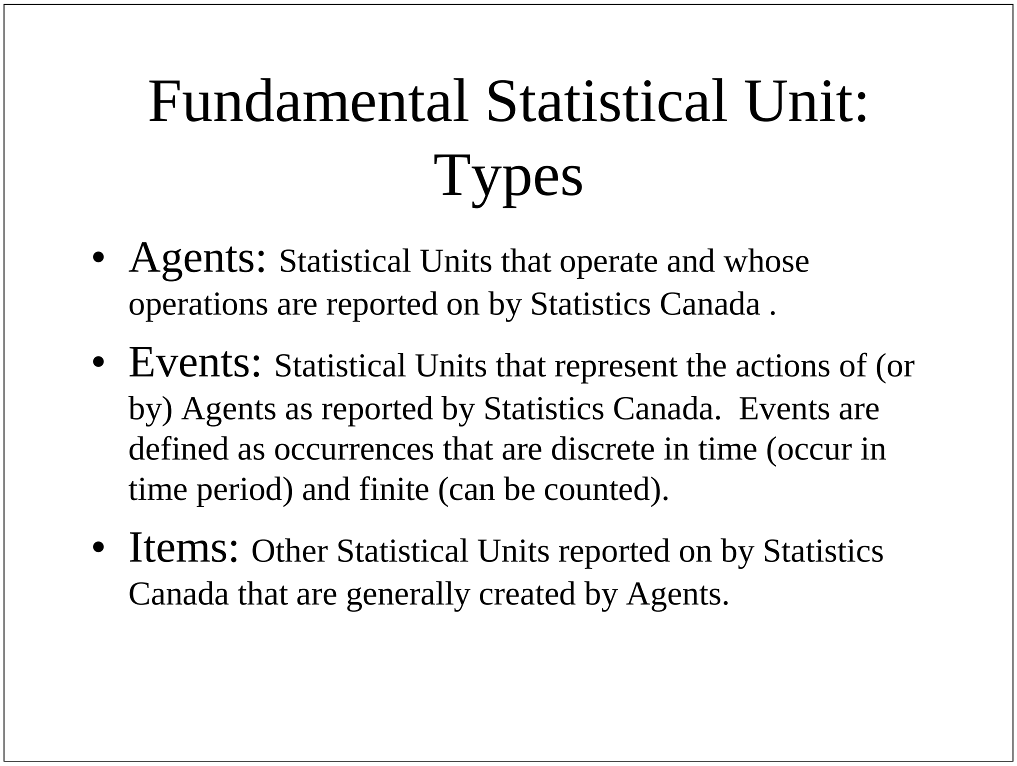


The four views

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

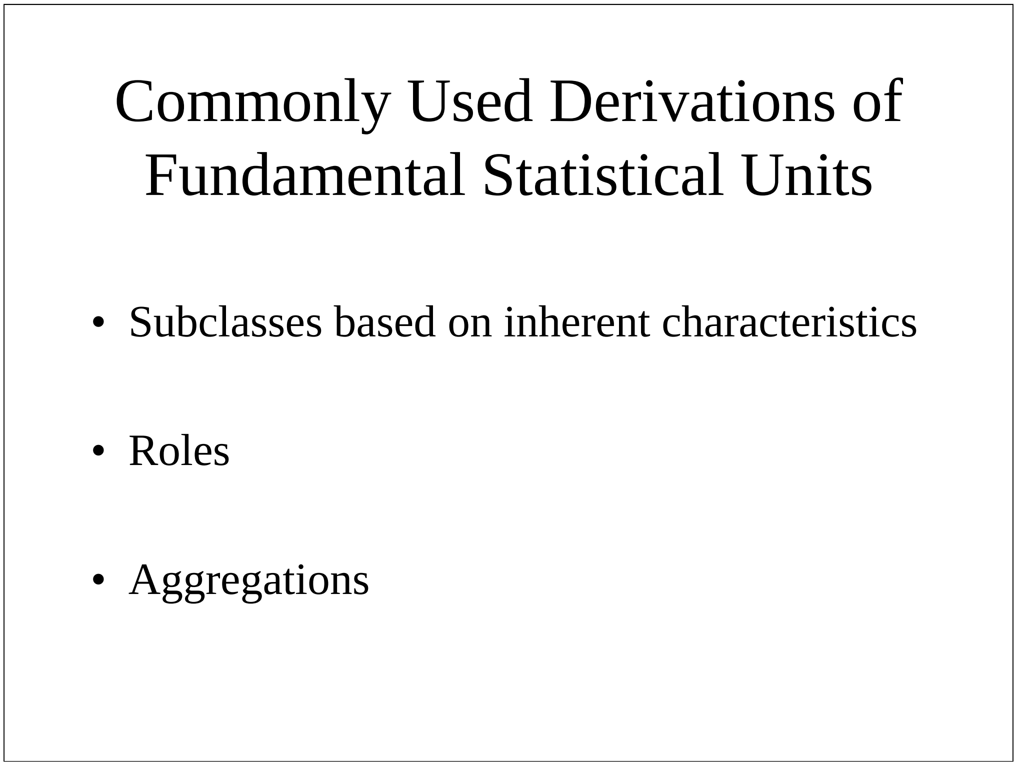
Fundamental Statistical Units: Definition

- 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.



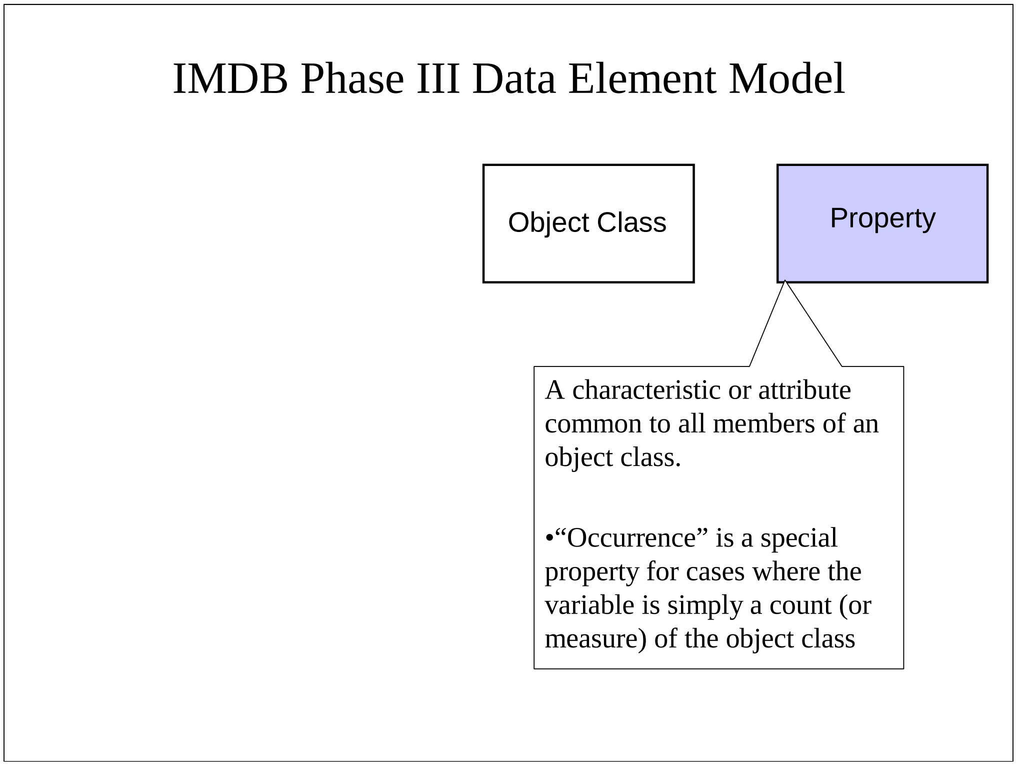
Fundamental Statistical Unit: Types

- 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

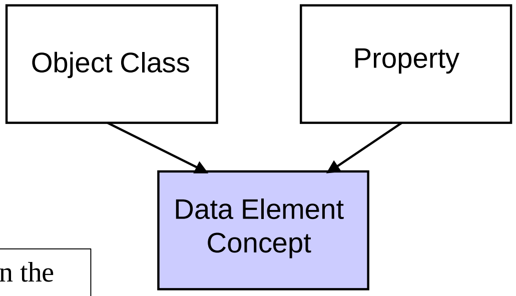


IMDB Phase III Data Element Model

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

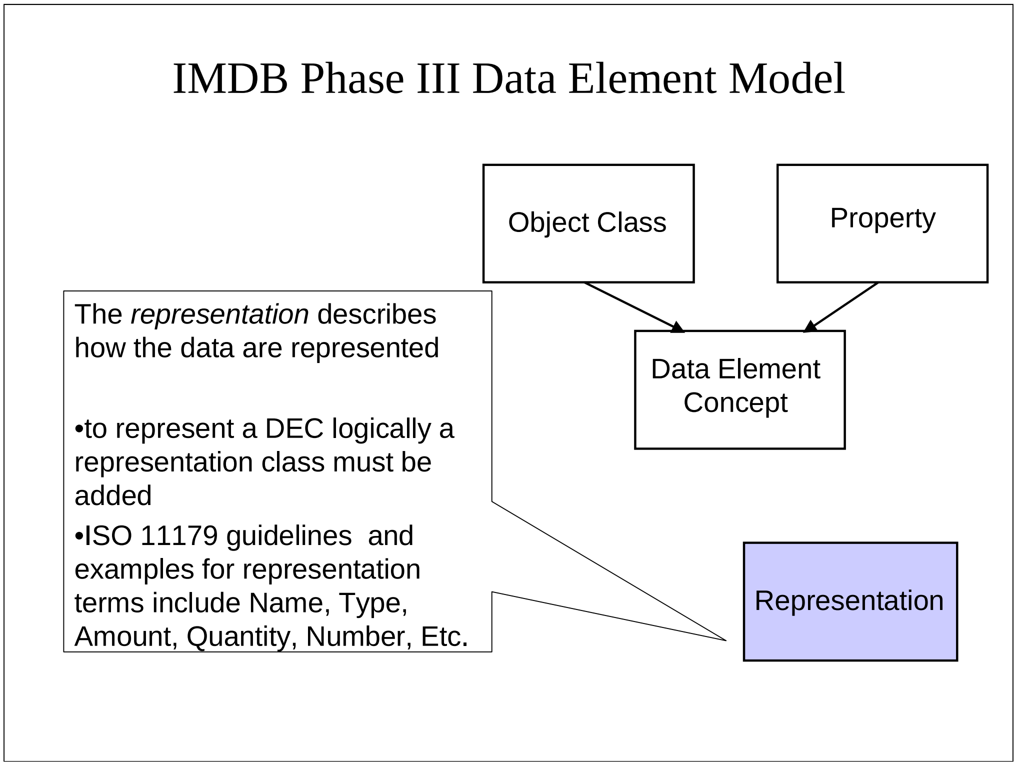
•“Occurrence” is a special property for cases where the variable is simply a count (or measure) of the object class

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.

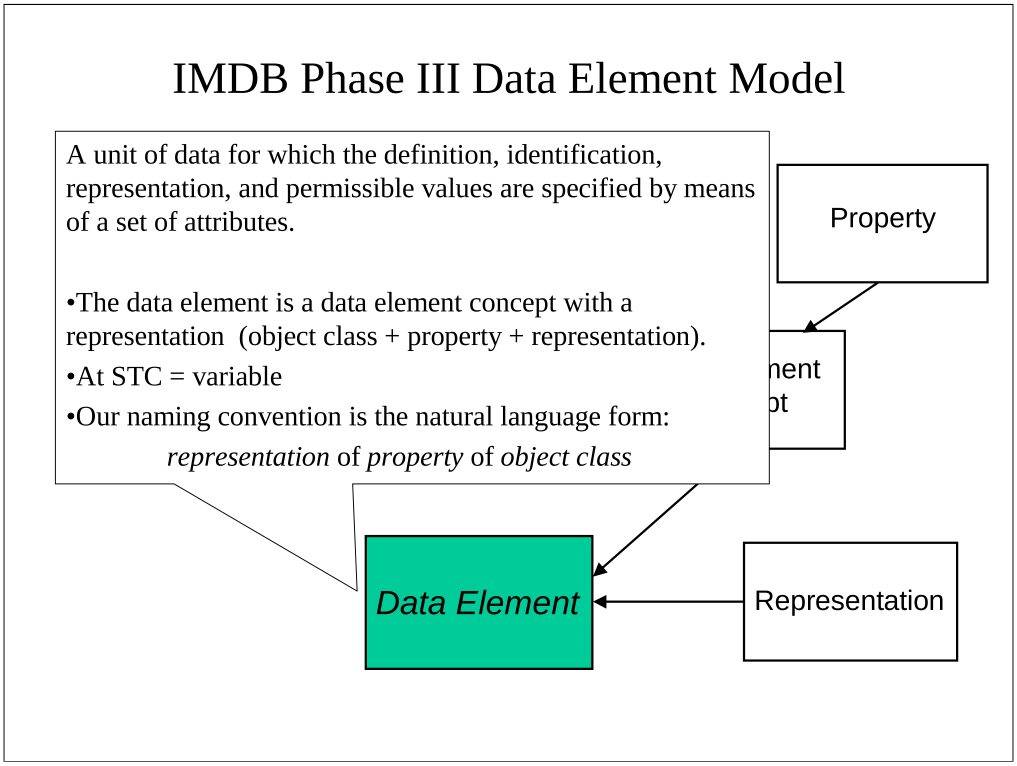


IMDB Phase III Data Element Model

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*

Property

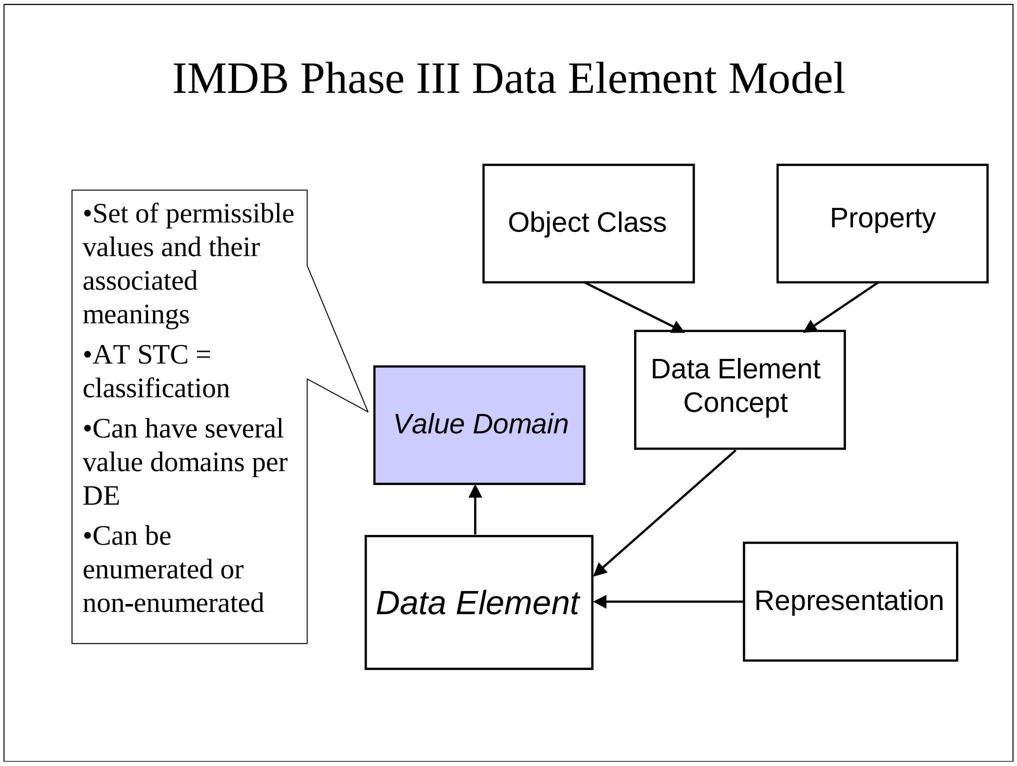
nent
pt

Data Element

Representation

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

Object Class

Property

Data Element Concept

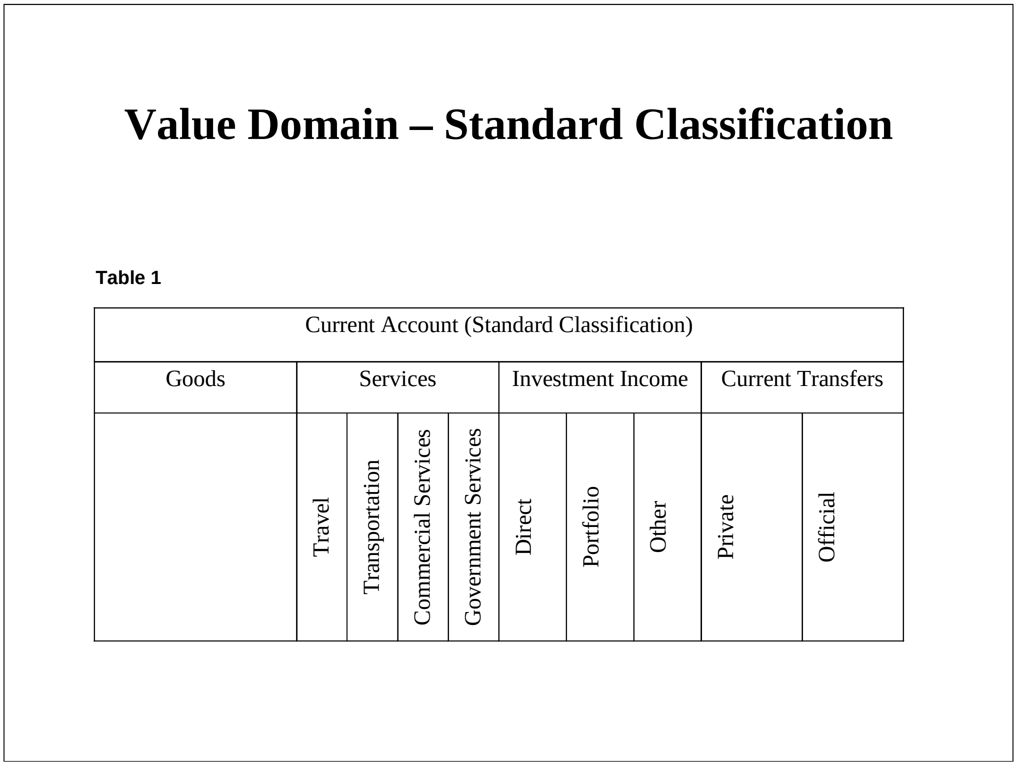
Value Domain

Data Element

Representation

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

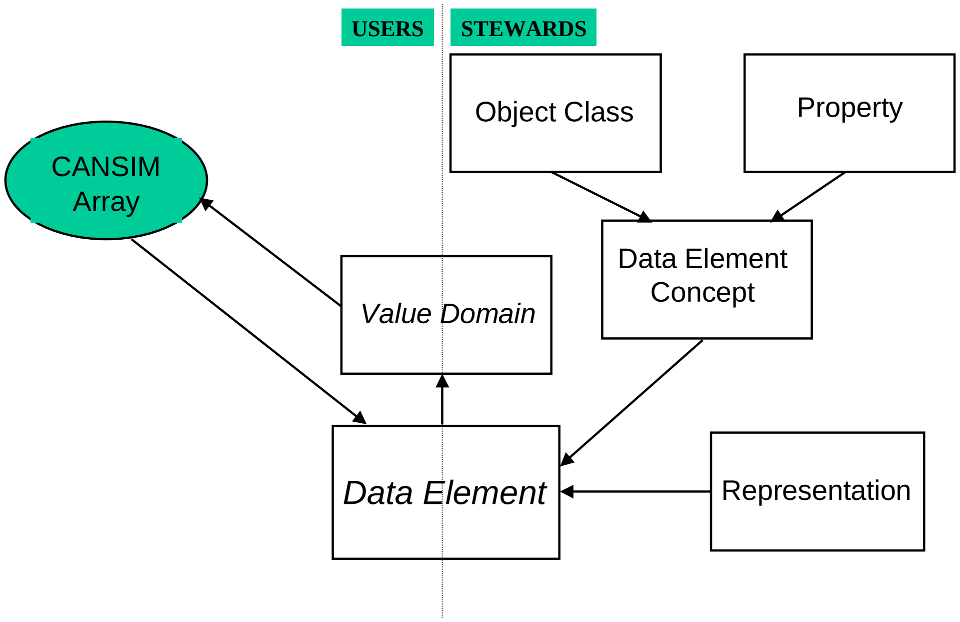


Value Domain – Standard Classification

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

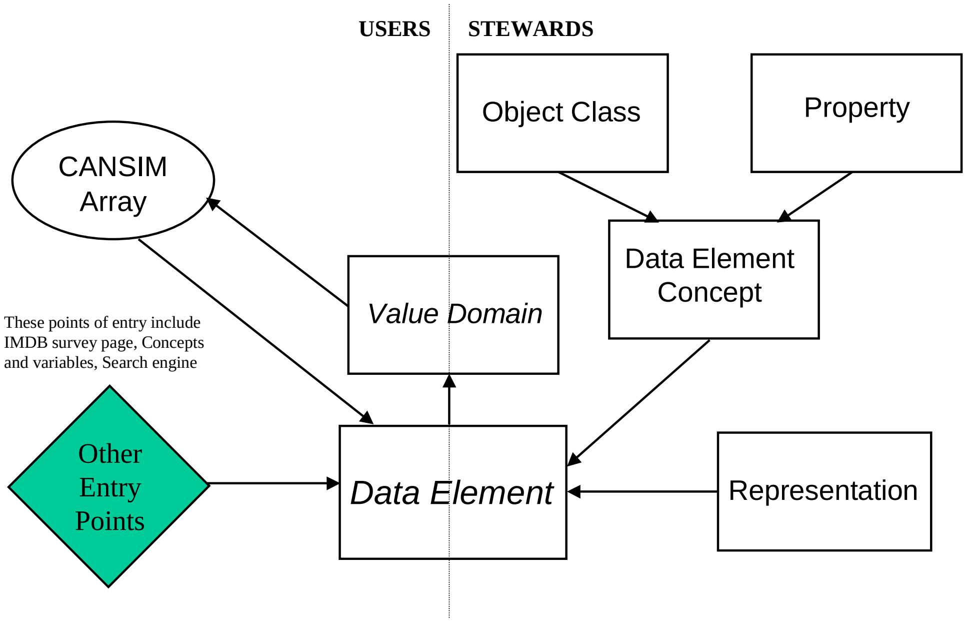
Data Element Concept

Value Domain

Data Element

Representation

IMDB Phase III Data Element Model



USERS

STEWARDS

Object Class

Property

CANSIM Array

Data Element Concept

Value Domain

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

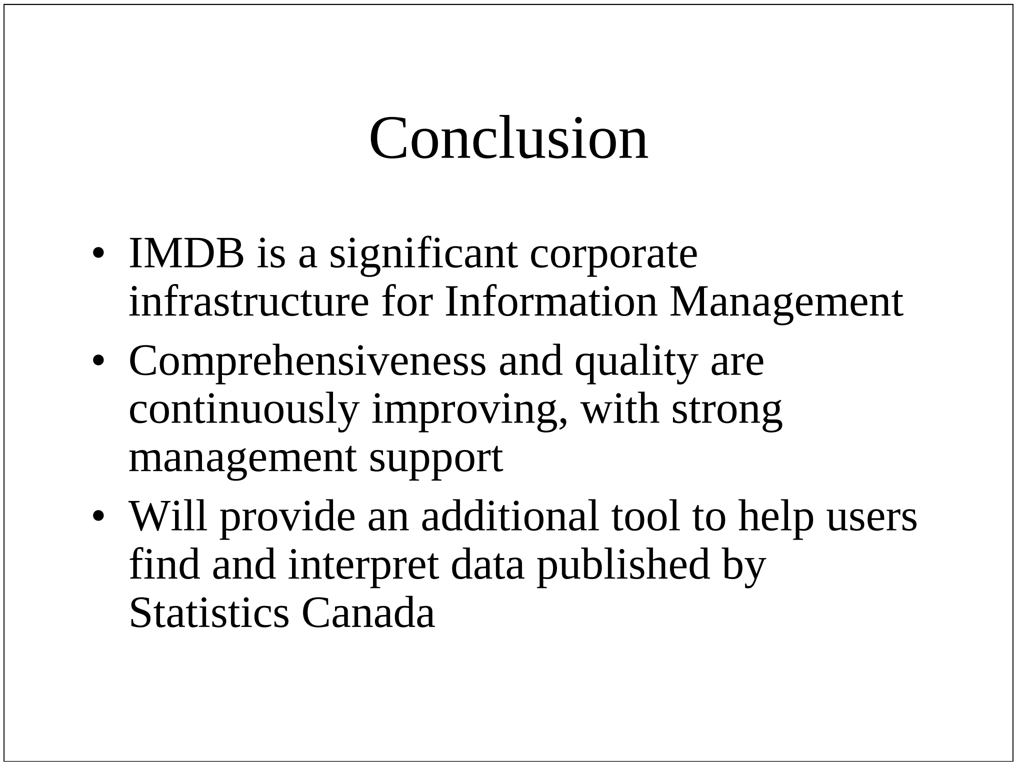
Other Entry Points

Data Element

Representation

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