Unifying the Information Management Disciplines

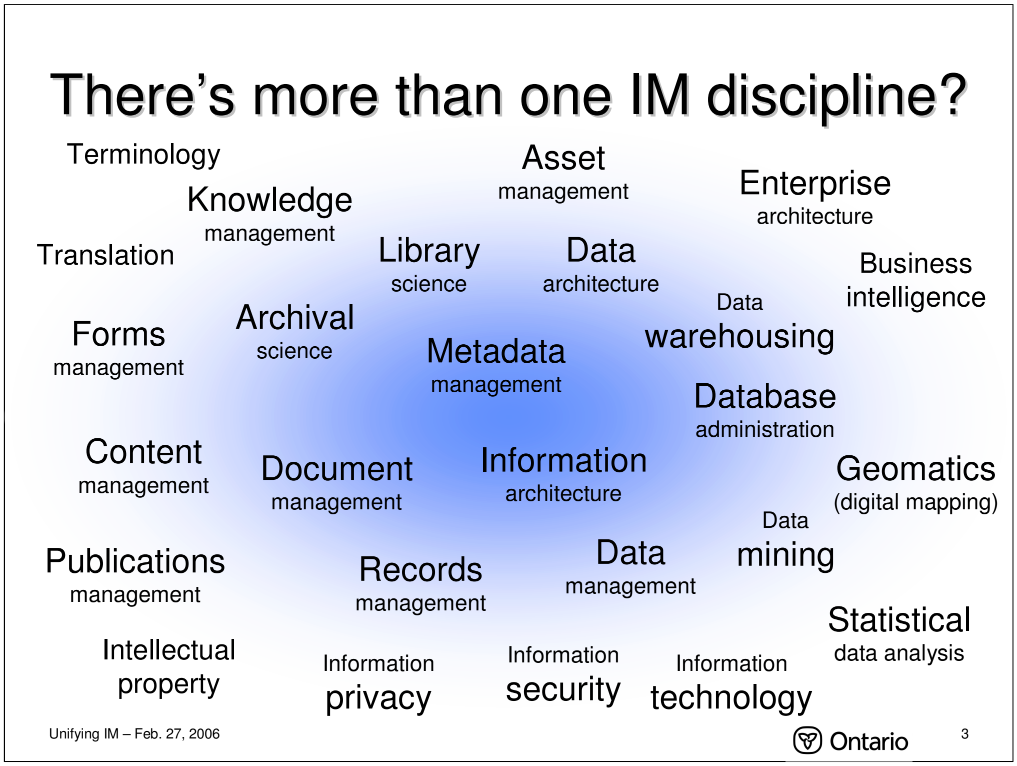
Managing Government Information
Ottawa, February 27, 2006

Alana Boltwood, Information Architect
Ontario Ministry of Government Services

Ontario

Hypothesis

- Many IM disciplines, with common goals
- Each IM technique is based on a pattern or structure for organizing information
- The existing structures are fundamentally similar, but expressed with different technical terms
- A common vocabulary and structure for organizing information would:
  - Improve communication among IM disciplines
  - Improve consistency among IM technologies
  - Help make IM a unified and mature discipline that really can "manage information as an asset"



Barriers to communication among the IM disciplines:

Many definitions of "information management"

Different training for practitioners

Separate units of the organization

Separate professional associations

Different technical vocabularies

Different standards, tools, software

Shared goals for all public sector IM practitioners

- Collect and store information **efficiently**
  - Minimize cost of acquiring information
  - Minimize storage space (physical or digital)
- Make information **accessible** and **useful**
  - Easy to find and interpret the information
  - Maintain sufficient information quality
- Protect **privacy** & confidentiality
  - Secure access, storage and destruction
  - Users aware of conditions of information use

Unifying IM – Feb. 27, 2006 Ontario 4

Make information accessible and useful

Ensure information is maintained to sufficient quality (updates)

Ensure information with evidentiary/historical value is preserved long into the future

Ensure vital/critical information is available quickly in an emergency/disaster/crisis

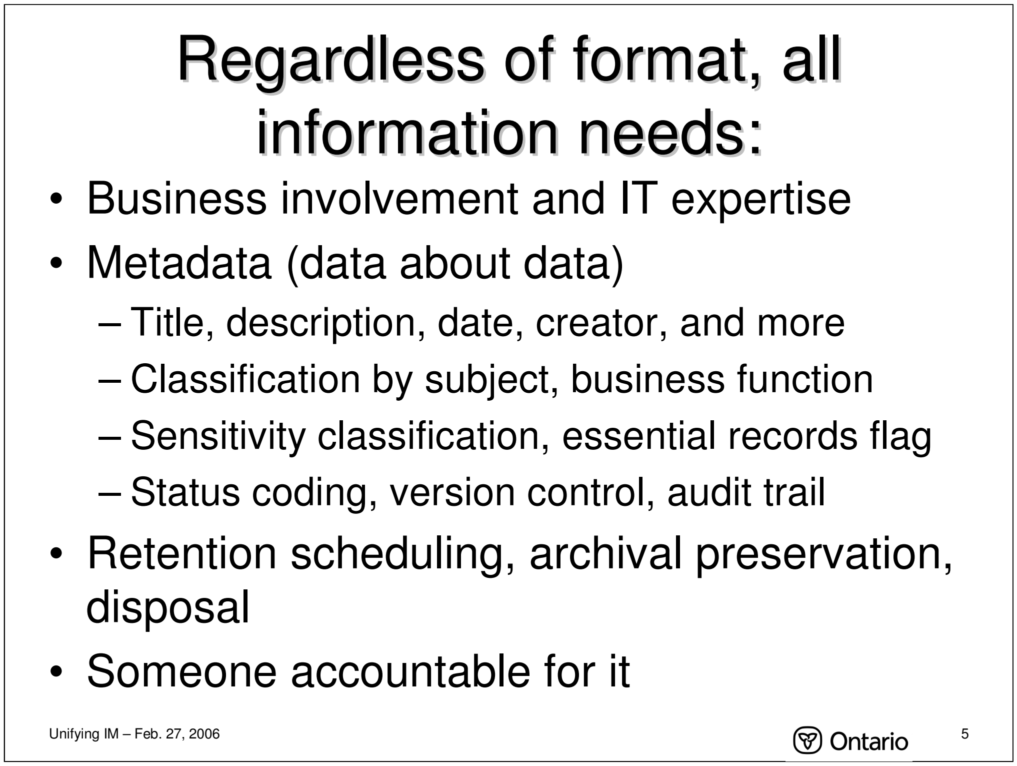
Ensure information for operational & policy decision-making is easily found, understandable, consistent and accurate

Ensure information can be retrieved for further transformation (e.g. writing a summary report, statistical analysis)

Ensure the public can access information, for good service delivery and accountability

Ensure information can be evaluated for quality (recency, accuracy) and relevance to the user's need

Ensure the information subject can read their own information



Solutions: Information needs structure and metadata

Convert unstructured documents to structured information wherever feasible

Parse information to a small granularity (break down into small reusable pieces)

Document structure & definitions in a model

Reassemble it into different formats

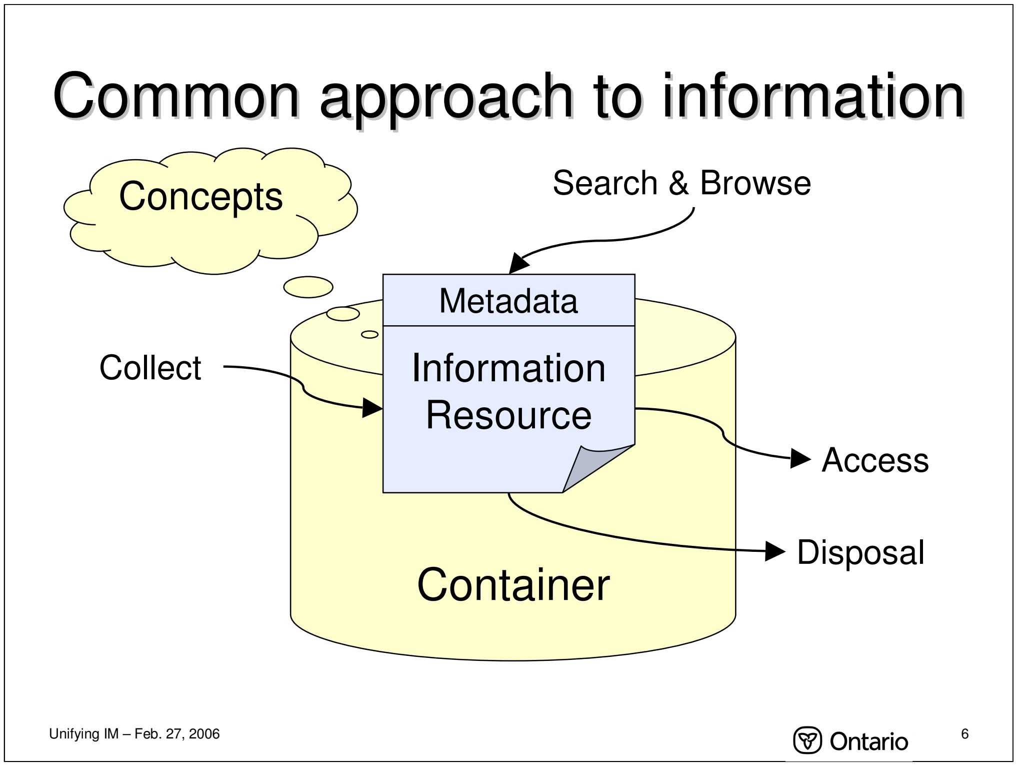
Analyze structured information to get full value from it

Add metadata to all levels of information structure

E.g. subject keywords for database columns make it easier to find the information you need

Aim for consistent information across tools & repositories (e.g. with a metadata standard)

Maximize use of the information and metadata you collect: put it in a shareable, searchable repository (unless it's personal or sensitive!)



All IM disciplines do the following basic things:

Information is the explicit recording of **concepts** (thoughts or knowledge).

Information can be **collected** by research, measurement, interviewing, forms, analysis of other information, or transfer from others.

All information is stored as **information resources**: documents, images, messages, database records, sound recordings, etc.

Information resources are stored in **containers** such as filing cabinets, databases, file servers, content management systems.

Metadata (data about data) can be applied to describe each information resource. (The metadata may be stored in the information resource itself (e.g. email header), wrapped around it (e.g. file folder), in a separate part of the container, or in another container.)

Users should be able to discover the information resource by **browsing or searching** for its metadata.

Authorized users should be able to **access** the information resource (read it or modify it).

Information resources should be **disposed of** when they no longer have operational or historical value.

How do we organize information?

Library	Dewey Decimal; MARC cataloguing
Archives	Finding aid; record series; *fonds*
Relational database	Table; column; row; code; **index**; third normal form; primary key
Thesaurus	Broader term; literary warrant; **indexing**
Data model	Entity; relationship; **attribute**; domain
XML	Schema; element; **attribute**; enumeration
Statistics	Dataset; **variable**; observation
Data warehouse	Mart; star; fact; dimension; cube
Document management	Document; version; Dublin Core metadata; file classification plan

Unifying IM – Feb. 27, 2006 Ontario 7

Each IM discipline has its own way of organizing information, and has specialized terminology to describe its techniques. Some of these terms have synonyms in other disciplines. Others are unique, and reflect an idea worth sharing with other disciplines.

For some concepts & terms worth sharing across IM disciplines, see Loretta Mahon Smith's presentation "Can We Share?" at DAMA 2005.
http://www.wilshireconferences.com/MD2005/Orlando-2005-Trip-Report.pdf

Metadata for books

Title	*Author*	*Subject*
A brief history of time: from the big bang to black holes	Hawking, Stephen W.	1. Cosmology.
Gödel, Escher, Bach: an eternal golden braid	Hofstadter, Douglas R.	1. Meta-mathematics. 2. Symmetry.
Life, the universe and everything	Adams, Douglas	1. Milky Way Galaxy– History.

Unifying IM – Feb. 27, 2006 Ontario 8

From the Cataloguing in Publication data (with one exception).

Librarians do in-depth cataloguing of the whole book, so that you can find lots of books on Cosmology or Symmetry. A library catalogue doesn't often break a book down into smaller components, because books have a rather flexible structure. It would be enlightening to compare Hofstadter, Hawking and Adams' books, but comparing Chapter 10 of each would be less natural:

Hofstadter: *Levels of Description, and Computer Systems*

Hawking: *The Unification of Physics*

Adams: *Chapter 10* (in which the Universe shatters into a million glittering fragments, but it's all an Informational Illusion)

How is information structured?

Structured — Design info structure to collect many similar records. Broken into small, predictable pieces.

Unstructured — Placement of info within the record is unpredictable. Flexibility to create unique organizations of info.

Manipulate info

Specific searches

Analyze patterns in records

Full-text search

Interpret each record separately

Database

Email headers

Spreadsheet

Email body

Photograph

Unifying IM – Feb. 27, 2006

Ontario 9

Structured databases are managed separately from unstructured documents & messages. The management methods have different strengths & weaknesses.

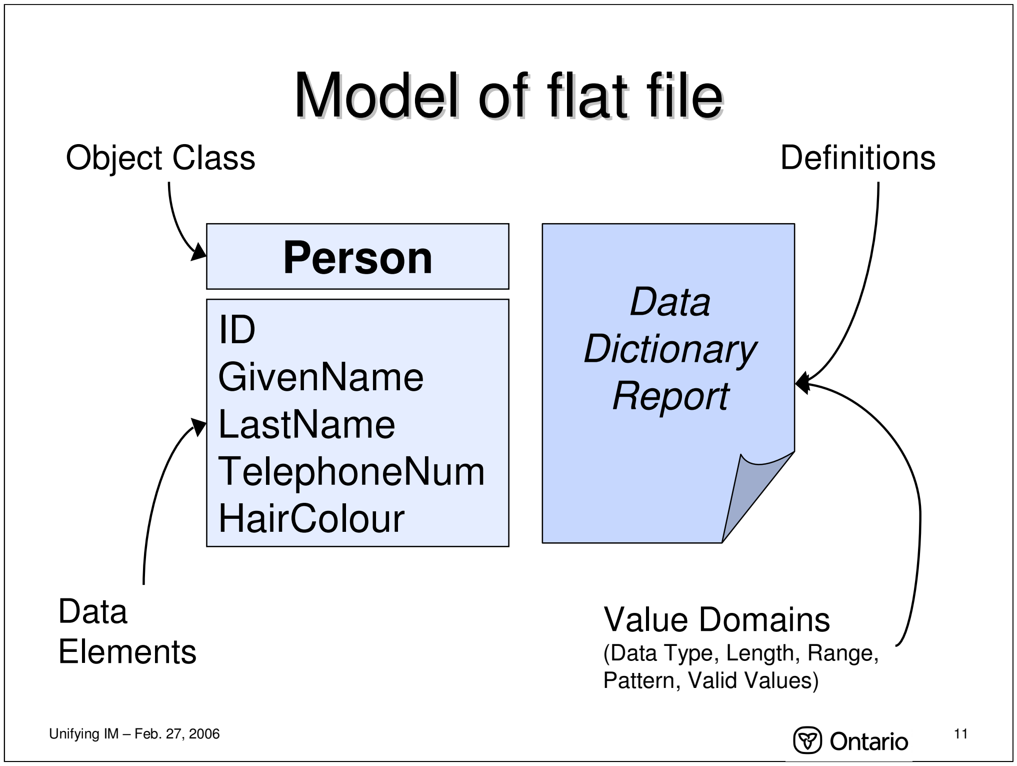
Simple information structure (flat file)

ID	*GivenName*	*LastName*	*TelephoneNum*	*HairColour*
23	J. Claude	Tremblay	819-456-7890	Blond
24	Alana	Boltwood	613-726-9292	Brown
25	Sarah	McKay Jones	613-726-9292, 819-765-9876	Red
26	Alana	Boltwood	416-327-6771	Blue

A good information structure breaks down information into small, well-defined data elements. For example, GivenName and LastName are separate, so the data can be sorted by either given or last name, and rearranged (e.g. "Tremblay, J. Claude"). For our purposes, it's OK to have two words in a name (e.g. "McKay Jones").

This information structure has a problem: it doesn't handle multiple phone numbers for a person, or sharing phone numbers between people.

Another problem: historical information (such as Alana's 613 phone number and hair colour) is being stored with current information, but not marked with validity dates.

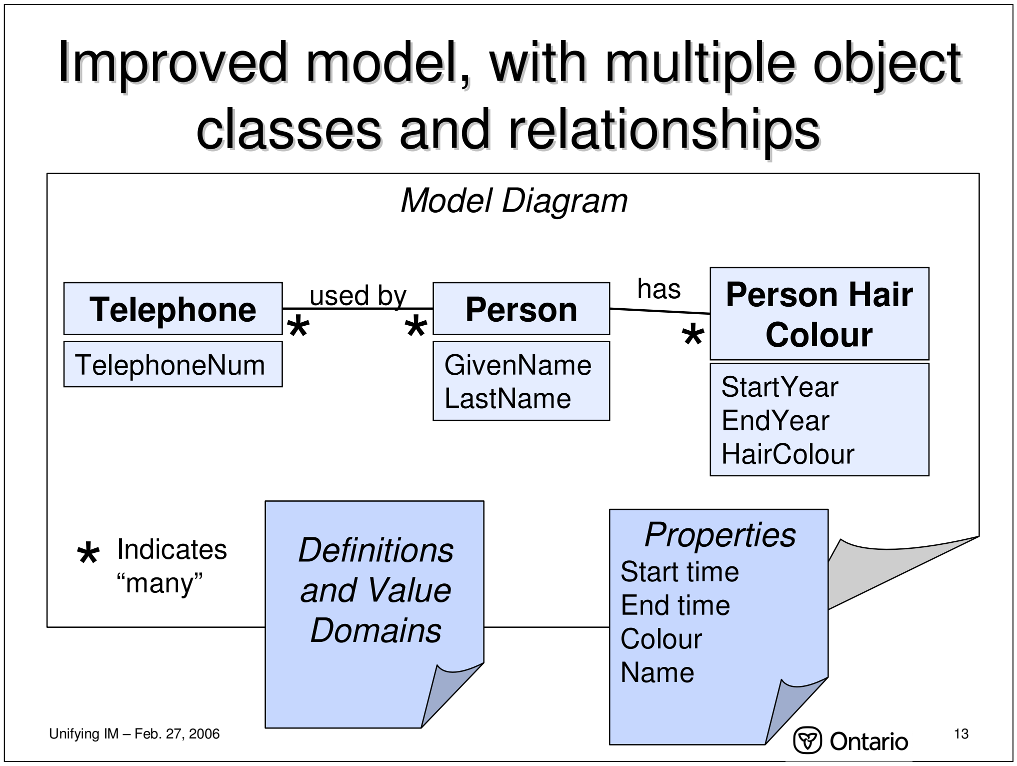

Model of flat file
Object Class
Definitions
Person
ID
GivenName
LastName
TelephoneNum
HairColour
Data Dictionary Report
Data Elements
Value Domains
(Data Type, Length, Range, Pattern, Valid Values)
Unifying IM – Feb. 27, 2006
Ontario
11

Improved information structure

TelephoneNum
819-456-7890
416-327-6771
613-726-9292
819-765-9876

ID	*Given Name*	*Last Name*
23	J. Claude	Tremblay
24	Alana	Boltwood
25	Sarah	McKay Jones

Start Year	*End Year*	*Hair Colour*
n/a	n/a	Blond
n/a	2001	Brown
2004	2006	Blue
1983	n/a	Red



Need relationships between object classes to escape the constraints of flat files.

Removing redundancy from information structures is called "normalization" in database design.

Some IM disciplines & tools don't have multi-object structures. Others have them with constraints (e.g. relational databases can't inherit).

Modeling terminology comparison

ISO 11179-3	Database	XML Schema
Object class	Table	Element
Data element	Column	Sub-element
Enumerated value domain	Codeset	Enumeration

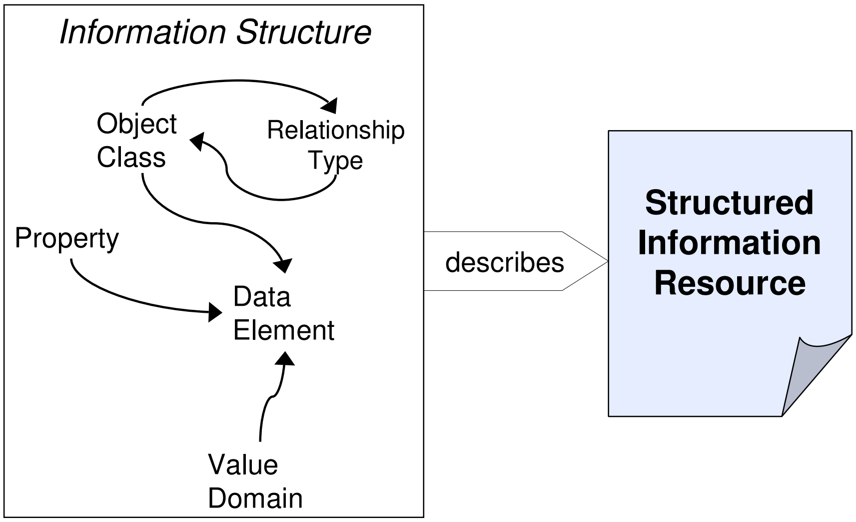
ISO 11179 is a standard for describing data elements.

The many variants of Entity-Relationship Modeling are used to design relational databases.

XML Schemas define the structure of XML data files, for information interchange.

These are just some of many modeling languages available for describing the structure of information.

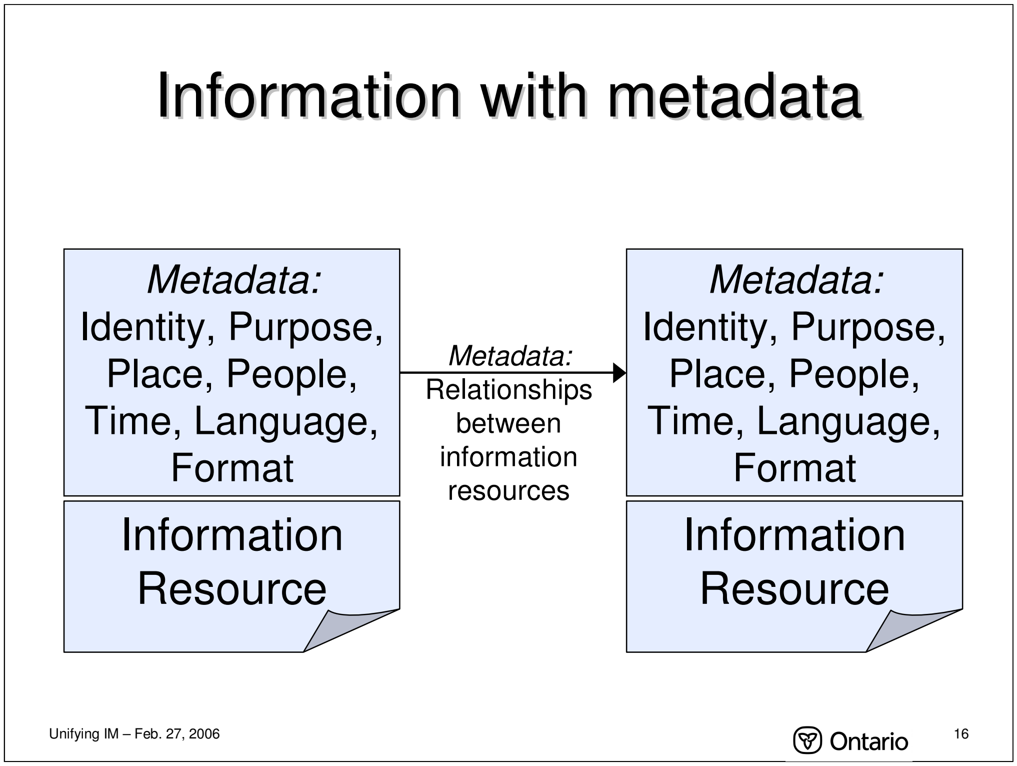
Generic information structure



Unifying IM – Feb. 27, 2006

Ontario

15



Example for books:

Identity = ISBN

Purpose = Audience (Children, Youth, Adult, Academic…)

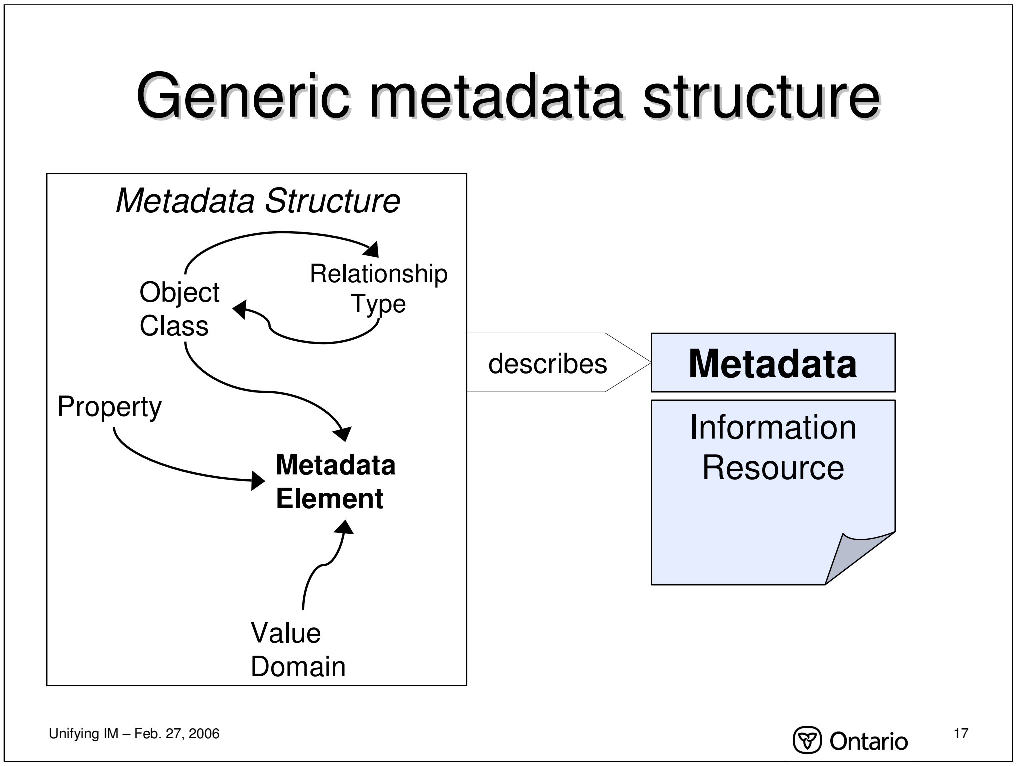
Place = City of publication

People = Author, Illustrator, Editor

Time = Copyright Year

Language = Language

Format = Text or pictures; dimensions of book

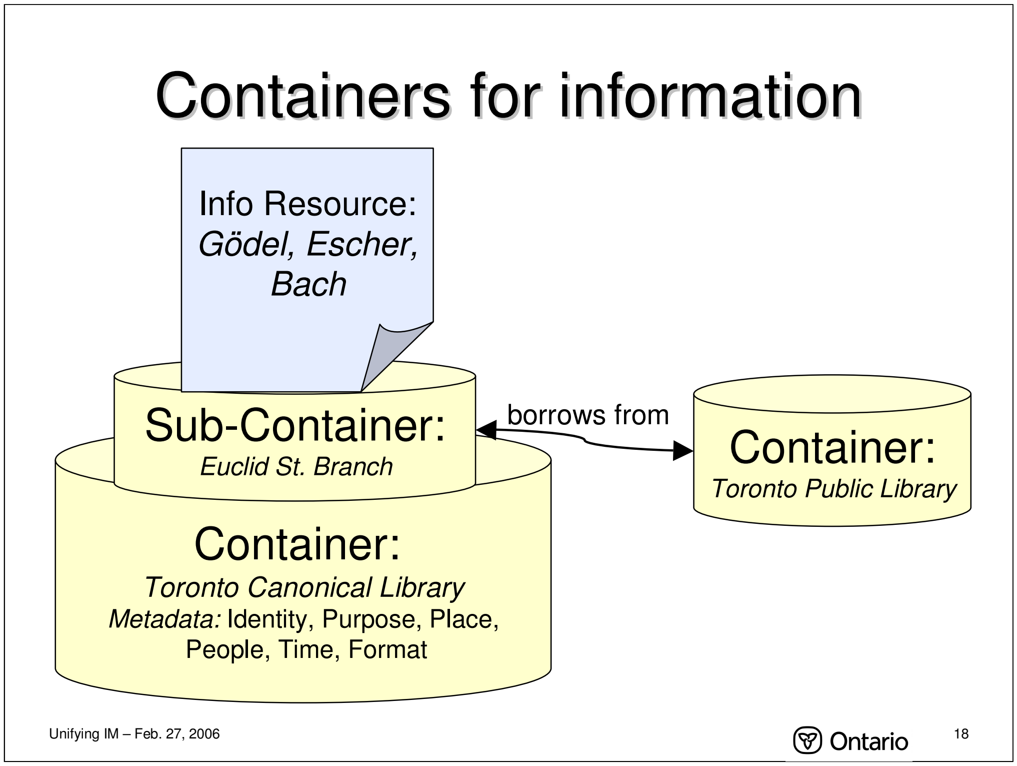


Metadata is structured the same way as any other information: with object classes (such as Book and Chapter) and relationships between objects (such as "*The Little Prince* is translation of *Le Petit Prince*"). The (meta)data elements (such as Title, Copyright Year, Dewey Decimal Classification) have value domains (such as "80 characters of text" or "Dewey classification codes".)

A property such as "Title" can apply to both Book and Chapter; it is a different metadata element for each object class.

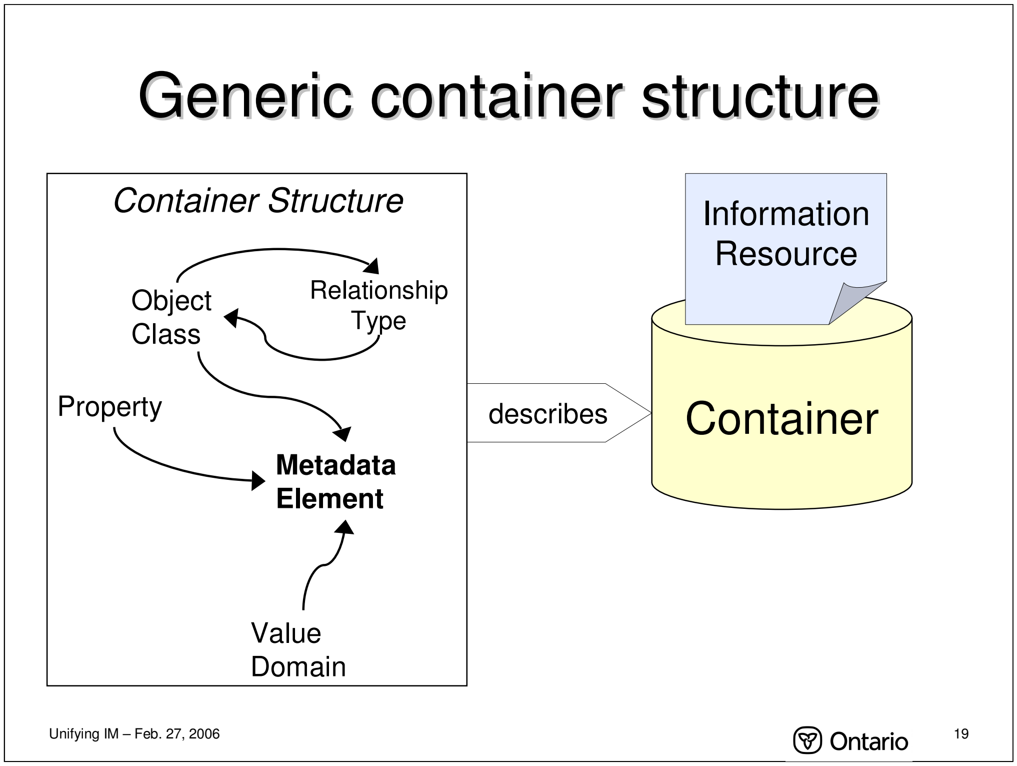
Book Title: *Gödel, Escher, Bach*

Chapter Title: *Consistency, Completeness and Geometry*

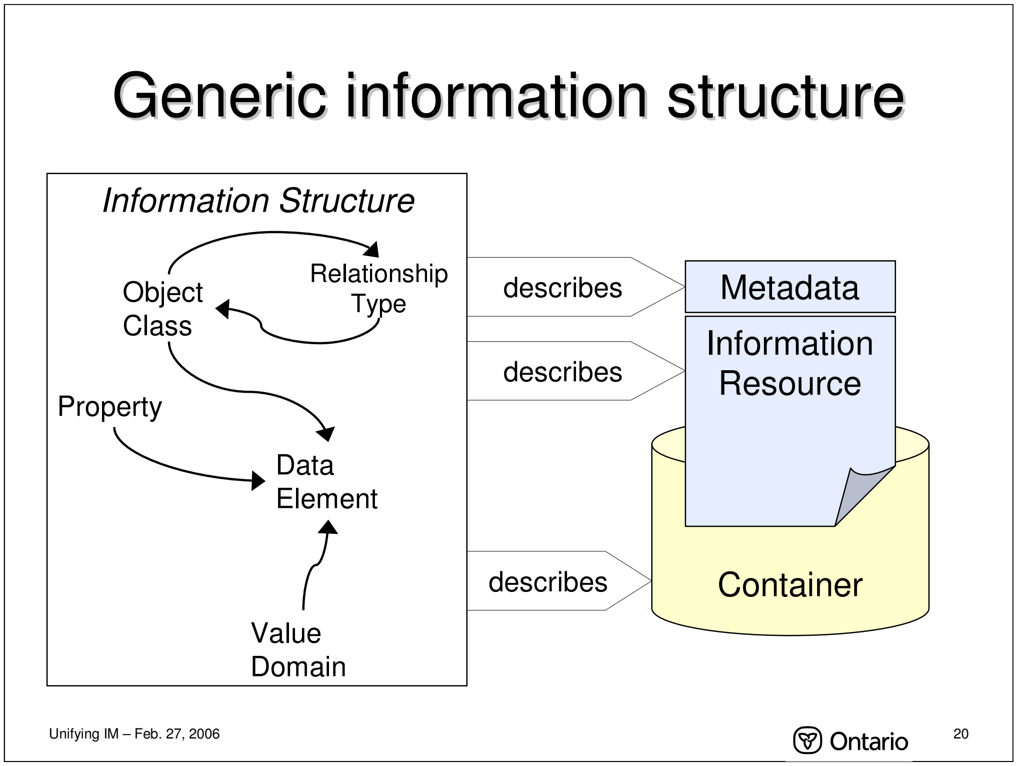


There can be relationships between containers, such as hierarchy.

Another example: a shared drive has a hierarchy of directories, each containing files and/or more directories.



Containers can be described with the same information structure as metadata and information resources: with object classes (types of container, such as Library, File Server, Filing Cabinet, Box) and relationships between containers (such as "part of"). The (meta)data elements (such as Name, Size, Location, Access Rights) have value domains (such as "30 characters of text" or "5 to 20000 square metres".)

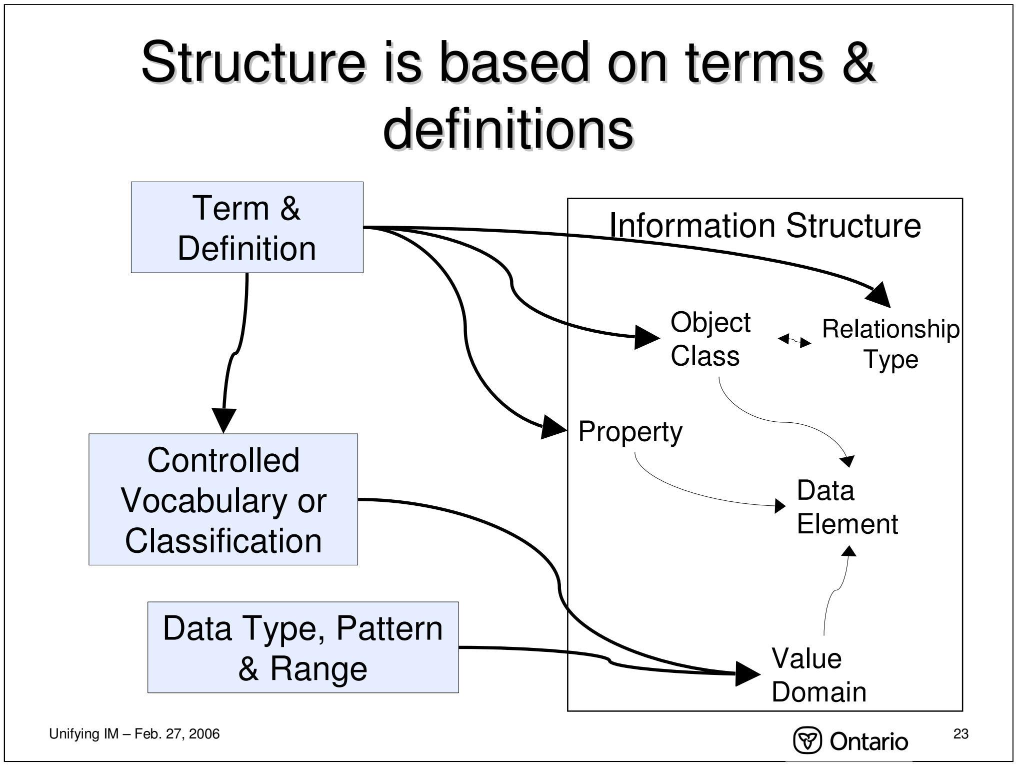

Generic information structure
Information Structure
Object Class
Relationship Type
Property
Data Element
Value Domain
describes
describes
describes
Metadata
Information Resource
Container
Unifying IM – Feb. 27, 2006
Ontario
20

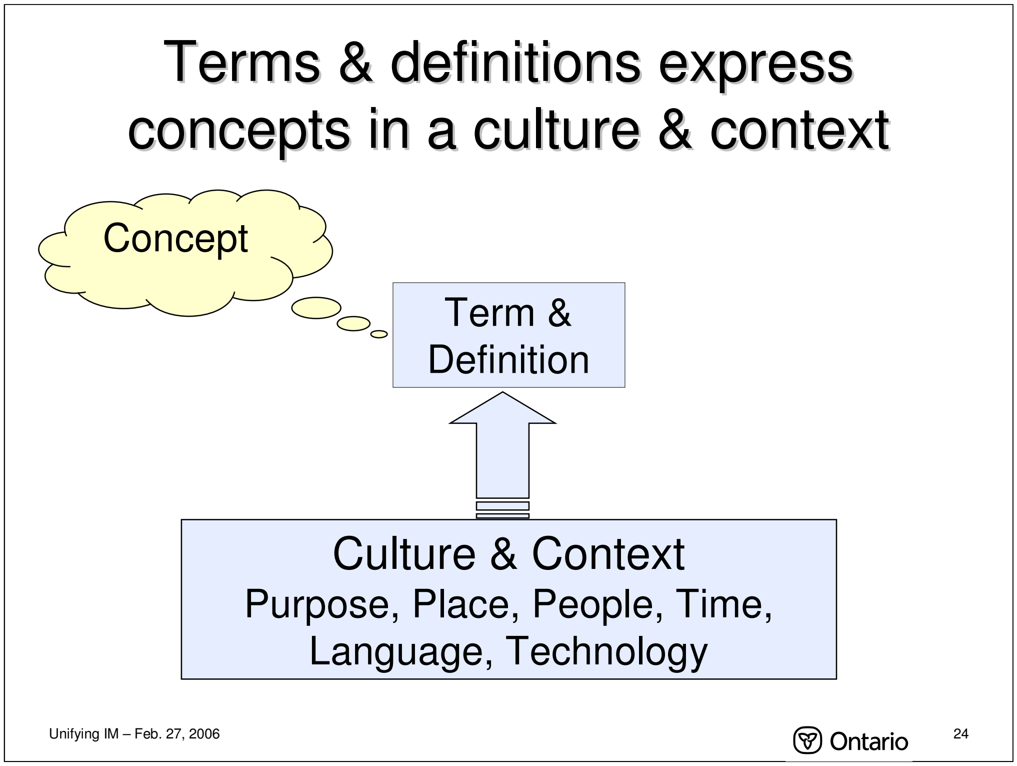
Information structure uses terms & definitions

	Term	*Definition*
Object Class	Person	A human being.
Property	Colour	The sensation produced on the eye by rays of light when resolved into different wavelengths.
Data Element	HairColour	The main colour of hair, as selected from the HairColour value domain.
Relationship Type	Person *has* Hair Colour	The Person's head hair is mainly the related Hair Colour, naturally or by dyeing.

Enumerated value domains use controlled vocabulary terms

Value of HairColour	*Definition*
Blond	Light-coloured (yellow or beige).
Brown	Medium or dark brown (the colour produced by mixing red, yellow and black).
Red	Red, orange or brown tones near red (the primary colour).
Grey	Mixture of white and blond, brown, black or red.


Structure is based on terms & definitions
Term & Definition
Controlled Vocabulary or Classification
Data Type, Pattern & Range
Information Structure
Object Class
Relationship Type
Property
Data Element
Value Domain
Unifying IM – Feb. 27, 2006
Ontario
23



Example: the term “extension” (the number dialed to reach a specific telephone set after dialing the main telephone number)

Time: 20^{th} century

Language: English

Technology: Telephone networks

Place: within a large organization

People: office workers and other telephone users (not telephone technicians)

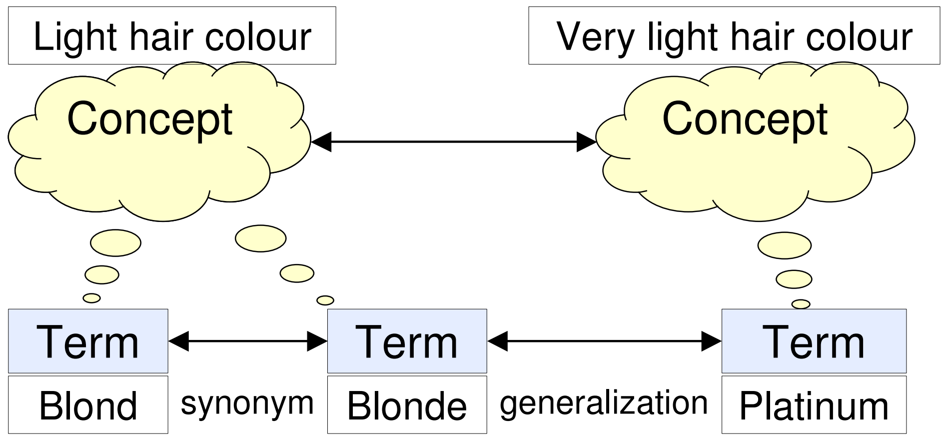
Purpose: To discuss part of a complete telephone number

French equivalent: *poste*

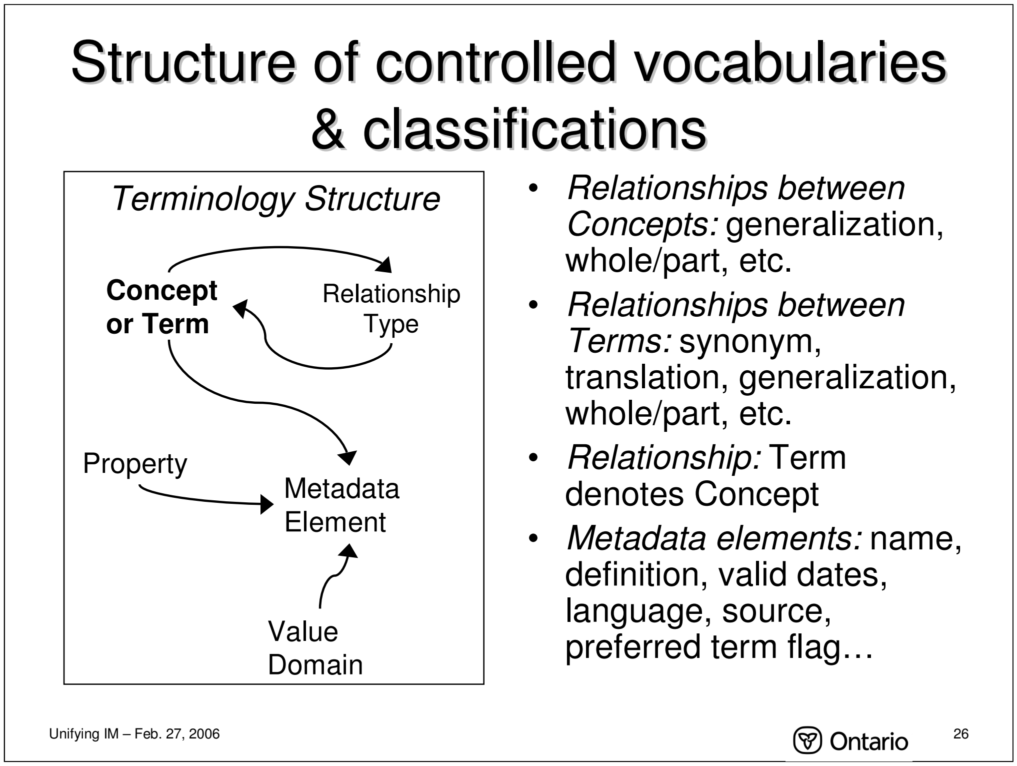
German equivalent: *Apparat*

Compare homonyms for “extension”: prolongation, enlargement, electric extension cord, extramural college courses, delayed essay deadline, building addition, group of things denoted by a term (logic & terminology theory), etc.

Controlled vocabularies & classifications



Unifying IM – Feb. 27, 2006 Ontario 25



Concept is an Equivalence Set of synonyms & quasi-synonyms & lexical variants & translations of Terms

Relationships between concepts:

Generalization (inheritance)

Generic / instance

Whole / part

Other associations

Relationships between terms for same concept:

Synonym

Quasi-synonym

Translation

Lexical variant

Preferred term

Relationships between terms for separate concepts:

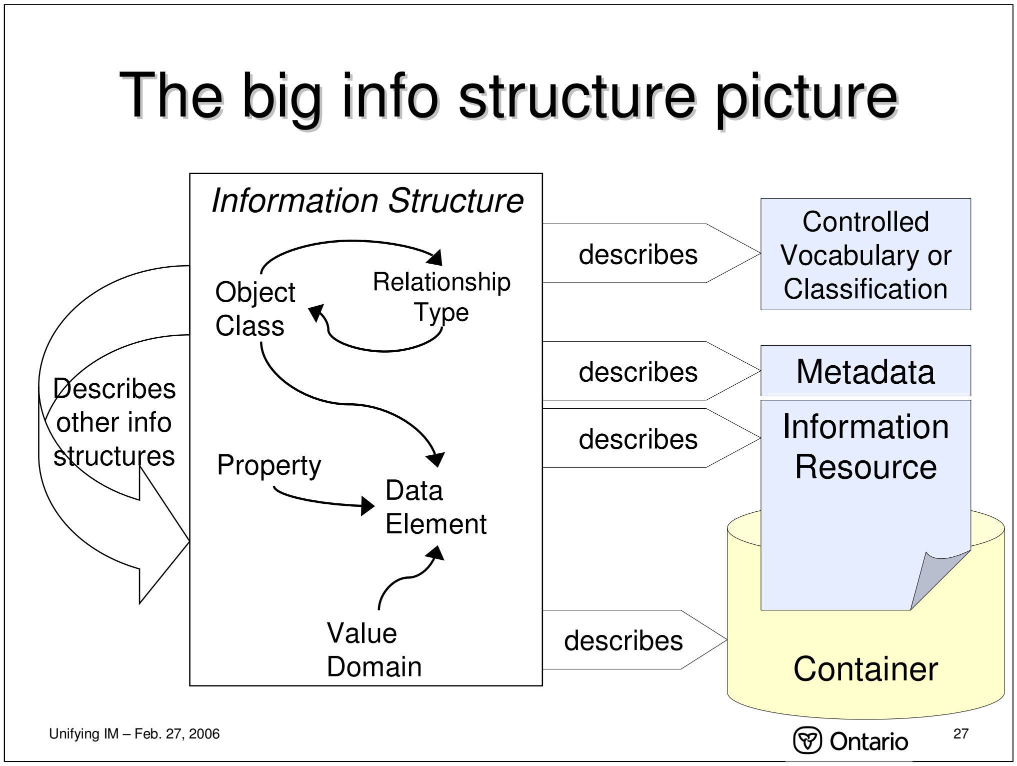
Generalization (inheritance)

Generic / instance

Whole / part

Other associations

example from immanent ontologies: Conditional probability of co-occurrence



This structure is based on ISO 11179, which focuses on defining the detailed structure of Data Elements and Value Domains. It begins with defining Object Classes and Properties. Many other standards define Relationship Types (such as generalization, association) between Object Classes or equivalent.

This information structure greatly simplifies ISO 11179. In particular, Object Classes and Properties are combined first into Data Element Concepts. With the addition of a Value Domain (derived from a Conceptual Value Domain), the Data Element Concept becomes a Data Element.

This structure also omits some important details & features of commonly used information structures. For example, an Object Class may have a Unique Identifier (known in database design as a "candidate key" or "primary key".) Some information structures are rich with features (such as ontology definition languages), others are more constrained (to make them easier for humans or computers to use).

Why is this important?

- IM disciplines have a lot in common, even if they use different terminology.
- Common information structure could harmonize IM:
  - Teach IM concepts & techniques with common terminology
  - Understand notation, features & limitations of particular information structures
  - Design software & repositories for information in any format
  - Avoid reinventing the wheel

"We thought of that in 1979 but we called it…"

Metamodels

- A "metamodel" rigorously defines how information is modelled or structured
  - As in this presentation, but formal & detailed
  - Terms such as "object class", "entity", "domain"
  - Permissible relationship types & other structures
  - What needs to be named, defined, specified in a model
  - Model diagram notation, report format
- Metamodels may be formalized in international standards, or be semi-formal (articles, textbooks), or be as implemented in software or repository

Some international standards

- Terminology theory: ISO 704
- Controlled vocabularies: ISO 2788, ISO 5964 and ANSI/NISO Z39.19
- UML (Unified Modeling Language)
- Entity-relationship modeling: IDEF1X
- Data elements, value domains: ISO 11179
- CWM (Common Warehouse Metamodel)
- Metadata: Dublin Core Abstract Model
- W3C XML Schema

Differences between metamodels

- Terminology differences
- Diagram notation differences
- Differences in portrayal of structural features
  - E.g. multiplicity *vs.* optionality & cardinality
- Limitations in structures
  - E.g.: Relational databases cannot directly implement inheritance (generalization)
  - Some limitations are realistic for implementation (e.g. binary relationships, not n-ary)

Unifying IM – Feb. 27, 2006 Ontario 31

Choosing a type of IM software implies choosing a metamodel.

Choosing a metamodel requires either:

Reading hundreds of pages of abstractions, or

Trusting a colleague or vendor’s advice, or

Using the standard or tool you already know.

More structural features that are only available in some metamodels:

generalization hierarchy (inherited data elements)

generalization hierarchy (inherited relationships)

multiple inheritance

containment & cascade delete (parent-child tables, or XML subelements)

hierarchy of attributes (as in Dublin Core)

one-many associations

many-many associations

n-ary relationships not just binary

multi-valued attributes

Attributes of relationships (not just of object classes)

International efforts to map metamodels

- ISO/IEC 19502 Meta Object Facility (MOF)
- ISO/IEC 19503 XML Metadata Interchange (XMI)
- ISO/IEC 19763 Framework for Metamodel Interoperability
  - Work in progress, covers ontologies
- OMG Ontology Definition Metamodel
  - Compares UML class models, Entity-Relationship models, RDF, OWL, Topic Maps, Common Logic

Unifying IM – Feb. 27, 2006 Ontario 32

More work is needed to:

Connect these metamodels to controlled vocabulary standards

Map terminology across metamodels

Write the metamodels in plainer language

Open Forum 2006 for Harmonization of Terminology, Ontology and Metadata

Kobe, Japan, March 20-22, http://www.tiu.ac.jp/org/openforum2006/

ISO/IEC 19763, Information Technology --
Framework for Metamodel Interoperability

The 19763 standard is a multipart standard that includes the following parts:

Parts

Description

19763-1

Part 1: Reference Model -- This part of ISO/IEC 19763 describes the concepts and an overall architecture of the metamodel framework standard to be applied in the development and the registration of the following individual metamodel frameworks.

19763-2

Part 2: Core Model -- This part of ISO/IEC 19763 specifies the core model which is required to describe metamodel items, and which may be used in situations where a complete metadata/metamodel registry is appropriate.

19763-3

Part 3: Metamodel for Ontology Registration -- This part of ISO/IEC 19763 specifies the metamodel that provides a facility to register administrative information about ontologies.

19763-4

Part 4: Metamodel for Model Mapping -- This part of ISO/IEC 19763 provides a normative metamodel for describing differences regarding formats and types of objects to be exchanged or shared. This metamodel framework also provides a capability for describing transformation rules between different objects in term of a metamodel instance.

ISO 11179 working group (ISO/IEC JTC1 SC32 WG02, http://metadata-standards.org/) is looking at its connection to ISO 19763-3.

OMG Ontology Definition Metamodel http://www.omg.org/docs/ad/05-08-01.pdf

Testing the hypothesis

- Ontario's *Enterprise Architecture Repository* will test the hypothesis that a common metamodel is possible and useful.
  - Combine architecture (data, process & other models), glossaries, classifications, standards, other metadata
  - Using this common information structure
  - Find resources *via* Dublin Core metadata, classification by subject, organization, function, and mappings of similar objects
  - Enable enterprise-wide reuse of information & IT resources, by any business, IM or IT staff

Unifying IM – Feb. 27, 2006 Ontario 33

Can the various IM structures be combined into one metamodel?

Can information from various software be converted to the common metamodel?

Will users trust the resulting combined information?

Will users still get the advantages of "special features" only found in some metamodels?

Can people accustomed to simpler metamodels still understand the information they find?

Will IM practitioners gravitate towards the common terminology and full-fledged common metamodel?

Ontario's Enterprise Architecture Repository

Custom metamodel for any objects & relationships

- Business, information, application, technology & security architecture models
- Metadata standards (adapt or map to Dublin Core)
- Taxonomies, thesauri, glossaries, codesets
- Standards, policies, project documentation

Relationships mapping any "aligned" objects

Enables re-use, integration across government

Currently in development

- Based on Metis software from Troux Technologies

What else does IM need?

- In addition to a common metamodel, a mature IM discipline needs harmonized:
  - Objectives
  - Performance benchmarks, evaluation tools
  - Policy, Standards, Principles
  - Lifecycle
  - Techniques, Software
  - Education and Training
  - Professions (roles) and associations
  - Organization units, services within large organizations

Unifying IM – Feb. 27, 2006 Ontario 34

Harmonized standards can be built around IM frameworks and common terminology.

Integrated software and tools will be built around the standards.

Coordinated organization units will be built around the software & tools.

Education, training and professional associations will become more comprehensive as the terminology, standards and tools become unified.

The immaturity of IM is a world-wide problem, manifested in the OPS, because it is a large and complex organization with a lot of information to manage. We must collaborate with the international IM community to develop common standards and terminology, which will enable the software industry to build common tools.

Conclusion: What can you do?

- Set up regular communications among the IM activities within your organization
- Learn about other IM disciplines
- Organize a presentation to your professional association, from another IM discipline
- Define the IM terms & structures you use, and compare them to other IM disciplines
- Ask educational institutions for training that crosses IM disciplines
- Participate on international standards bodies, and work towards common structures in all standards

Questions? Further thoughts?

Alana Boltwood
Information Architect
Office of the Corporate Chief Technology Officer
Ministry of Government Services
Government of Ontario
416-327-6771
Alana.Boltwood@mgs.gov.on.ca