I-INTELLIGENCE

Building a Data Model

Chris Pallaris | Working with Big Data | June 2022

Taxonomies

Taxonomies

- Definitions
  - A taxonomy is a formal, structured system for labelling or classifying objects
  - Taxonomies are typically composed of controlled vocabularies, an organised list of preferred terms and phrases that are intended to help users index content or navigate an information service
  - The two terms are distinct but closely related. For the purpose of this module, taxonomy will be our preferred term

Taxonomies

- Theory
  - Think of your library or closet. How did you decide where the individual items go? Put differently, how did you decide the order of things?
  - In order to arrange anything, we must choose a method for ordering and classifying content in a way that supports our needs
  - The more structured your approach to ordering, the easier it is to:
    - Communicate that order to others
    - Maintain that order over time

Taxonomies

- Theory
  - All systems tend toward chaos. To elaborate, the more books a library has, the more time and effort is needed to maintain order
  - Similarly, the more objects in database, the better your ordering mechanism needs to be
  - Taxonomies are one tool for organising the data, information and knowledge we accumulate in the course of our work

Taxonomies

- Theory
  - The process of establishing and maintaining a taxonomy can generate the insights needed to furnish knowledge and enable action. Thus:
    - Is there a simpler or smarter way to organise your library?
    - Are there entire categories of books you should dedicate space to?
    - Does the order of your books reflect the current state of your knowledge?
    - Is your ordering system designed to accommodate future branches of knowledge?

Taxonomies

- Theory
  - It often takes multiple taxonomies to define, order, or make sense of a single data object. For example, a book can have:
    - A contents page
    - Page numbers (Roman and Latin)
    - Section headings
    - An alphabetical glossary of terms
    - An index
    - One or more subject classifications
  - The ability to develop and navigate such taxonomies is a defining quality of information managers and knowledge workers

Taxonomies

- Purpose
  - Simply put, taxonomies inform almost every facet of our existence. They are used to:
    - Order our experience and understanding of the world
    - Structure and label the physical and digital environments we inhabit
    - Define our place and identity
  - We are defined by the taxonomies we subscribe to, as well as the taxonomies are assigned to us

Taxonomies

- Purpose
  - From an information management perspective, taxonomies are commonly used to:
    - Organise our thinking
    - Organise objects into groups
    - Indicate natural relationships
    - Characterise explicit concepts
    - Enable the capture, management and presentation of information
    - Provide structure to unstructured information
    - Enable findability

Taxonomies

- Purpose
  - More specifically, taxonomies are used for:
    - Indexing
    - Classifying
    - Categorisation
    - Filtering
    - Linking
    - Browsing / navigation
    - Content management
    - Content discovery

Taxonomies

- Process
  - All data management involves some measure of taxonomy development. Why?
  - Because all data management involves the ordering of records, information, knowledge, etc., into clearly defined categories
  - Arguably, the most challenging taxonomies to develop are subject taxonomies. Consequently, this module will focus on this activity

Taxonomies

- Process

    1. Define the subject and scope of your taxonomy

    2. Define the “top-level” categories of your taxonomy

        - These are the major subject branches that you will organise your terms under

        - They should be clear and easily understood

Taxonomies

- Process

3. Identify the terms you would like to include in your taxonomy

    - Brainstorm terms together with your colleagues
    - Identify existing taxonomies, lexicons if these already exist
    - Research academic sources, Wikipedia, etc. for additional inputs
    - Where possible, organise these using parent / child relationships
    - Ensure you apply the right degree of specificity and detail
    - Use Excel or a mind mapping tool to help you

Taxonomies

- Process

4. Group related terms and concepts together

    - Use the same parent / child relationships to give your taxonomy structure

    - Be as granular as you wish but try to go no more than three levels deep

    - Ensure that each term has only one “home”. Thus, do not list a keyword under multiple top level headings

    - Format these in a consistent fashion. Thus, for example:

        – Use British spelling consistently
        – Organise the terms alphabetically or according to a consistent logic

Taxonomies

- Process

5. Identify and resolve issues of equivalence

    - Very often, there are multiple ways to define a single concept. Thus:

      Solar Energy = Solar Power = Photovoltaic Energy

    - In such instances you have to decide your preferred term (PT) and which are synonymous (or non-preferred) terms

    - Your choice - and the reasons for it - should be captured in the taxonomy itself

Taxonomies

- Process

6. Extend your taxonomy further to include:

    - Synonymous terms (i.e. non-preferred terms)
    - Variant spellings (e.g., British or American equivalents)
    - Scientific and popular terms (photovoltaic panel vs. solar panel)
    - Acceptable acronyms (e.g. North Atlantic Treaty Organisation, NATO, OTAN)
    - Related terms (i.e. “cousins” rather than children)
    - Terms that should not be used
    - Usage notes (typically detailed in the final column)

Taxonomies

- Process

7. Define your classification rules

    - If you are developing a hierarchical taxonomy, determine whether content tagged with a narrower term should also be tagged with the broader term. Thus, is every article about Nigeria also about Africa?

    - Describe the rules governing the relationships between terms

    - Describe whether document can be indexed under more than one top-level category

    - Communicate these rules to staff, including via the Data Dictionary

Taxonomies

- Process

8. Integrate the taxonomy into your information services

    - When you are comfortable with the structure and scope of your taxonomy, integrate it into the information service you are using to organise your content

    - Oblige staff to tag content using the subject and region-specific terms detailed in your taxonomy

    - If you wish, allow staff to tag content using their preferred terms. Folksonomies enhance the richness of your metadata, but user consistency is unlikely

    - Taxonomies and controlled vocabularies, therefore, are the preferred means of indexing content consistently

Taxonomies

- Process

9. Operationalise the process of taxonomy maintenance

    - The taxonomy should be updated at regular intervals to reflect the changing information needs of staff

    - Devise a process by which staff can recommend:

        – New terms

        – Changes to the preferred term

        – Changes to the existing structure

Taxonomies

- Process

  9. Operationalise the process of taxonomy maintenance

     - The number of change-requests is a strong indicator of the taxonomy's value and the evolution of staff knowledge

     - While major structural changes should become less frequent over time, the addition or removal of terms should be welcomed

Taxonomies

- Twelve Criteria for an Effective Taxonomy

1. It is intuitive: it is easy to navigate and use; it does not require training but reflects the way users think

2. It is unambiguous. All terms are clearly defined with no room for misunderstandings

3. It is natural: it uses the terms, vocabulary and logic common to the user

4. It is hospitable: it can accommodate new content / categories as necessary

Taxonomies

- Twelve Criteria for an Effective Taxonomy

5. It is consistent and predictable: it provide context and aids navigation
6. It is relevant: it reflects user and / or organisational needs
7. It is parsimonious: there is no redundancy or repetition
8. It is meaningful: it provides context. Category, sub-category and topic terms help users anticipate the content they will find

Taxonomies

- Twelve Criteria for an Effective Taxonomy

9. It is manageable: there are no more than three levels

10. It is balanced: there is a consistent level of detail / depth across the taxonomy

11. It is durable: there is no need for frequent structural changes

12. It is future proof: it addresses today's content needs and is sensitive to tomorrow's

Taxonomies

- Remember…
  - All terms in your taxonomy should be unique, clear and mutually exclusive
  - Each term should represent a single concept (or unit of thought)
  - A term has no value if it is applied to *every* document or content object in the database
  - Disambiguate terms open to conflicting interpretation
  - Ensure relationships between terms are simple, logical and easy to follow
  - We categorise as we think. But how we think may need to change to improve the classification of data

Taxonomies

- Summing Up
  - Taxonomies are the basis for organising information. How we organise, so we shall find
  - A good taxonomy is a labour of love. The more time you invest, the better your IM efforts will be
  - No taxonomy is ever complete or perfect. Always expect it to grow
  - Start somewhere, start small, but start!

Taxonomies

- Exercise I
  - Establish a taxonomy to support the fight against fraud in the EU
  - This taxonomy will be used in a database that aggregates data on:
    - Instances of fraud affecting the EU's financial interests
    - The many types of fraud affecting these interests
    - The provenance or sources of fraud
    - Relevant threat actors
    - Relevant costs and impacts

Taxonomies

- Exercise II

1. Define a provisional structure for your taxonomy, together with the top-level branches

2. If time allows, brainstorm the terms to be included under each branch

3. Organise your thinking in a spreadsheet or a sheet of paper

Metadata

Metadata

- Definition
  - Metadata is structured information that describes, explains, locates or otherwise makes it easier to manage, retrieve or use information
  - Simply put, metadata is data about data (or information on information)
  - The term has different meanings and can denote:
    - Machine readable information
    - The records that describe an electronic resource
    - The descriptors applied to a digital or non-digital object

Metadata

- Purpose
  - Metadata is commonly used to:
    - Facilitate the discovery of relevant information
    - Assist with the organisation of digital resources
    - Enable knowledge sharing
    - Facilitate data integration and normalisation
    - Facilitate system interoperability
    - Support archiving and preservation
    - Locate a resource
    - Enable resource discovery (finding resources relevant to your work but which you are unaware of)

Metadata

- Three Types of Metadata
  - Broadly speaking, there are three types of metadata:
    1. Descriptive metadata - describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, keywords, etc.
    2. Structural metadata - indicates how compound objects are put together, e.g. chapters are ordered to form a book
    3. Administrative metadata - provides information to help manage a resource such as the creation data, file type, access controls and other technical data

Metadata

- Creating Metadata
  - Metadata is created manually, automatically, or both. Indeed, a single object may have multiple metadata schemas. For example:
    - Word auto-generates metadata every time you create a document
    - The document control page of a report obliges the author to complete additional metadata descriptors by entering data into pre-defined fields
    - These reports go in a folder which has its own metadata (e.g. file path)
    - The reports in a series are then added to a departmental file plan

Metadata

- Exercise
  - Returning to the Fraud Database identified earlier:
    1. Define the content objects you wish to include in the register, e.g.
       - The incident of fraud
       - The actors involved
       - The funds / projects affected
       - The impacts
       - Etc.
    - List these objects in an Excel spreadsheet or on a sheet of paper

Metadata

- Exercise

  2. Define the metadata attributes for each of these objects. Thus, a fraud actor might have:

     - An ID number
     - A title
     - A first name
     - A last name
     - An address

     - List these attributes next to each object in your spreadsheet

Metadata

- Exercise

    3. Identify any other taxonomies you will need to develop to index your data properly

Data Modelling

Data Modelling

- A First Thought Experiment
  - Consider your desktop of laptop computer. What information have you stored on this device?
  - Note that your computer is a database. It is a store of information relevant to you that you have developed over time
  - Unless organised effectively, this database is not very helpful except as a means of short-term storage
  - How might we make the content of this database more useable? How do we relate the different files and folders to one another?

Data Modelling

- A Second Thought Experiment
  - Let us imagine we want to develop a tackle instances of fraud and corruption affecting the EU's financial interests
    - What information do we need to do so?
    - How should we organise this information?
    - What entities or objects do we need to organise that information around?
    - What attributes or properties should these objects have?
    - How do we link these objects together?

Data Modelling

- From Unstructured Data to Structured Information
  - *Data* becomes *information* when it is *organised*
  - Organising your data typically involves entering it into a “structured” database
  - How you structure your data constitutes your *data model*

Data Modelling

- Definition
  - Data models are logical descriptions of the things we are interested in or want to collect information on
  - More concretely, they are pictorial representations of the objects in a database system.
  - A good data model therefore defines:
    - The objects in the database
    - The attributes of these objects
    - Their relationship to one another

Data Modelling

- Purpose
  - Data models are commonly used to:
    - Define the data requirements of an information system or database
    - Describe the data that will be contained within it
    - Create a blueprint of how the system will be built
    - Minimise changes during its development
    - Support the development of a Data Dictionary

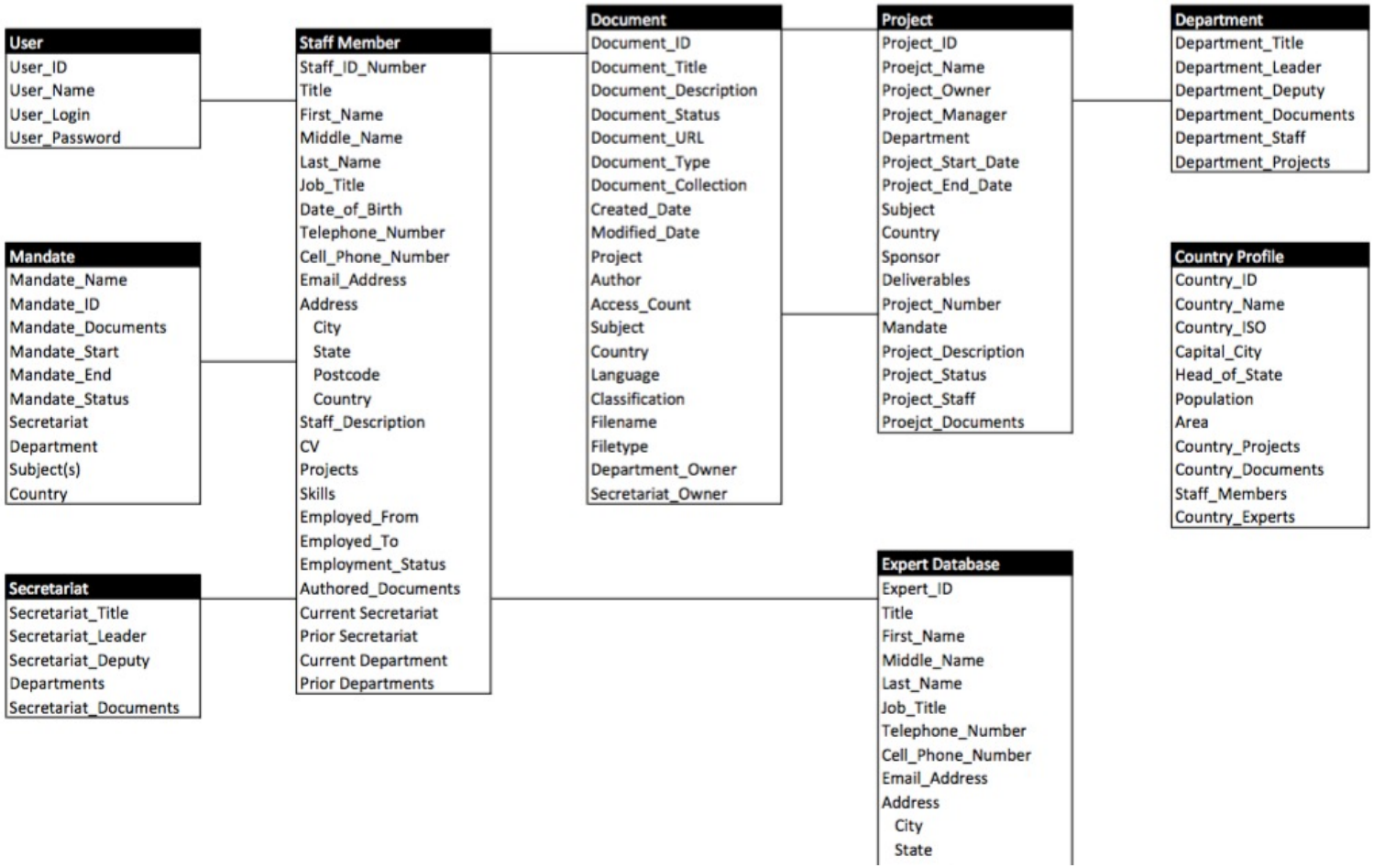
Data Modelling

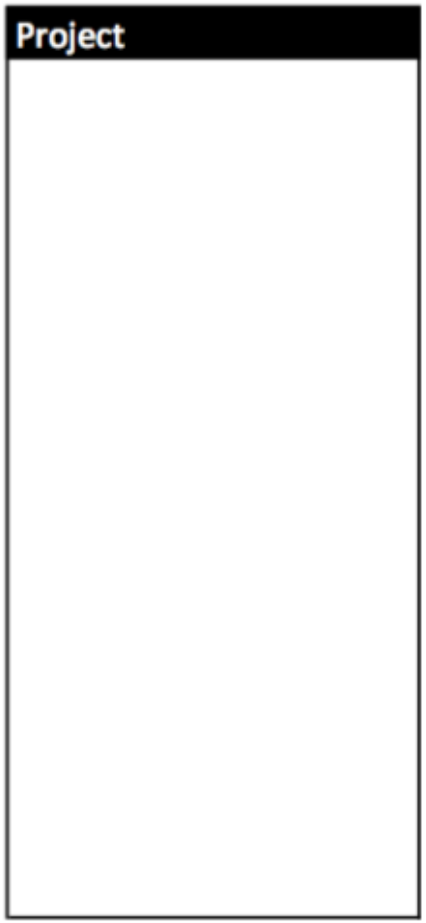
- Purpose
  - For non-technical staff, data models can be used to:
    - Communicate with developers and senior managers
    - Identify opportunities for systems integration
    - Shape the business architecture
    - Understand how information systems enable or constrain the flow of work

Data Modelling

- Entity Relationship Models
  - Entity Relationship Modelling (ERM) is the simplest and most common approach to data modelling, and the one we will use
  - ERMs have three components:
    - Entities (or objects)
    - Attributes (or descriptors)
    - Relationships

Data Modelling

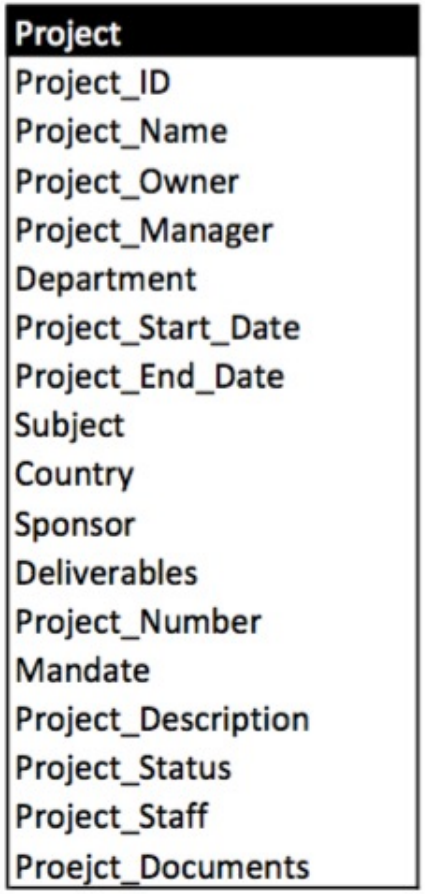




Entities

- An entity is something about which you wish to store information, e.g.:
  - A risk
  - A project
  - A member of staff
- These are the principal data objects on which information is collected
- Entities are represented by labeled rectangles. The label is the name of the entity. Entity names should be singular nouns

Data Modelling



- Attributes
  - Attributes describe the entity they are associated with. Thus, a project might have:
    - A title
    - A leader
    - A description
    - A donor
    - A start date
  - Attributes are listed inside the entity rectangle. Attributes which are identifiers are underlined. Attribute names should be singular nouns

- Relationships
  - Relationships represent an association between two or more entities. An example of a relationship would be:
    - Employees assigned to a project
    - Projects belonging to a department
  - Relationships are classified in terms of degree, connectivity, cardinality, and existence

- Relationships
  - A relation can be:
    - One to one - 1:1
    - One to many - 1:M / 1:N
    - Many to many - M:M
  - Relationships are represented by a solid line connecting two entities. The name of the relationship is written above the line as a verb
  - Remember: every relationship is an assertion that can be tested for accuracy

- One to One Relationship
  - A one-to-one (1:1) relationship is when, at most, one instance of an entity A is associated with one instance of entity B. Thus:
    - A project has a single leader

- One to Many Relationship
  - A one-to-many relationship (1:M) is when for one instance of entity A, there are zero, one or many instances of entity B. However, for one instance of entity B, there is only one instance of entity A. Thus:
    - A department has many employees
    - An employee is assigned to a single department

- Many-to-Many Relationship
  - A many-to-many (M:N) relationship, is when for one instance of entity A, there are zero, one, or many instances of entity B (and vice versa)
    - Employees can be assigned to more than two projects
    - A project must have at least two employees
  - Many-to-many relationships cannot be adequately represented in relational tables but should instead be transformed into two or more one-to-many relationships using associated entities

- Process
  - The process of developing an ERM can be rather convoluted
  - We have simplified this as far as possible and removed much of the technical jargon that accompanies the discipline
  - The more involved you become with information systems development, the more sophisticated your knowledge will need to become

Data Modelling

- Process

1. Requirements Gathering

    - The process of building your model begins by identifying users' requirements: what information would they like the system to have?

    - This information can be gleaned through interviews, questionnaires, use cases, a review of existing systems or similar instruments

    - These requirements should be catalogued and evaluated. Those that qualify as relevant should be featured in the data model

- Process

2. Entity Identification

   - Identify the entities (or objects) to be included in your data model

     – Make your entities as precise as possible

     – Every entity should be distinguishable from other objects

     – Every entity has to be related to at least one other. If not, there is no way to navigate to it in a relational database

- Process

3. Attribute Identification

    - Identify the attributes of each entity

        – Attributes should be “atomic”, or present as a single fact

        – Complex attributes (names, addresses) are best decomposed

        – An attribute may subscribe to an industry standard (e.g. a zip code)

Data Modelling

- Process

4. Relationship Identification

    - Identify the relationships between the entities

    - Ensure that all relationships are indicated by a verb connecting the entities. Thus:

        – Employees are “assigned” to projects

        – Projects “have” an end date

- Process

5. Apply Naming Conventions

    - The names ascribed to your entities and attributes must:

        - Be unique
        - Be intuitive
        - Have meaning for the end user
        - Are ascribed consistently
        - Use Title Case format (e.g. Project Leader)
        - Be singular (i.e. Report not Reports)
        - Not use acronyms or abbreviations
        - Contain the minimum number of words needed to be unique and accurately describe the object

Data Modelling

- Process

6. Develop your Entity and Attribute Matrices

    - Your data modelling efforts can yield two matrices:

        – An Entity-Entity Matrix can be used to map relationships between entities

        – An Entity-Attribute Matrix is used to indicate the assignment of attributes to entities

Data Modelling

	Person	Project	Document	Bank Account	Transfer
Person					
Project					
Document					
Bank Account					
Transfer					

Data Modelling

- Process

7. Iterate, refine and improve

    - No data model is ever complete. Expect it to evolve over time
    - The more you know about your organisation's data requirements the richer your model will become
    - Thus, submit yours for testing and review to technical and non-technical stakeholders and iterate accordingly
    - Remember: a flawed data model will always result in a flawed database

- Process

8. Update your data dictionary

    - When your data model has been fixed and approved, be sure to include the list of entities and their attributes in the data dictionary

Data Modelling

- Remember…
  - A good data module should give you:
    - A visual representation of the way entities and attributes are related
    - Information on the entities, their attributes and their relationships
    - A blueprint for the development of a database or information system
    - A means of interacting with end users

- Exercise
  - Consider an information asset you wish to develop
    - Organise the entities identified in Excel or on paper
    - List the attributes for each entity underneath the title
    - Draw arrows connecting these entities and / or their specific attributes
    - Clarify the relationships between these entities in an entity matrix

Thank You

Chris Pallaris
Director

i-intelligence

+41 (0) 44 243 3849 | Skype: chrispallaris | c.pallaris@i-intelligence.eu

www.i-intelligence.eu | @i_intelligence