Introduction to data modelling

Silvio Peroni
silvio.peroni@unibo.it – https://orcid.org/0000-0003-0530-4305 – @essepuntato

Data Science (A.Y. 2022/2023)
Second Cycle Degree in Digital Humanities and Digital Knowledge
Alma Mater Studiorum - Università di Bologna

DIPARTIMENTO DI FILOLOGIA CLASSICA E ITALIANISTICA

Summary of the previous lecture

A datum is a declarative statement subject-predicate-object that, through the predicate, either **attributes** a **literal** (i.e. a value such as a string, a number, etc.) to a subject **entity** *or* it **relates** such a subject **entity** with another **entity**

Each entity, being used either as subject or object of a statement, is characterised by **a unique identifier**

The **same entity** can be used as subject or object in one or more data

An attribute is intrinsically **part of** the entity to which it is associated

Modifying the value of an attribute affect **only** the entity to which it refers to

A literal **cannot be used** as subject in any datum

Any question about the previous lecture?

Predicates for attributes and relations

A predicate can be used either to define relations between a subject entity and an object entity or to specify attributes associated to an entity

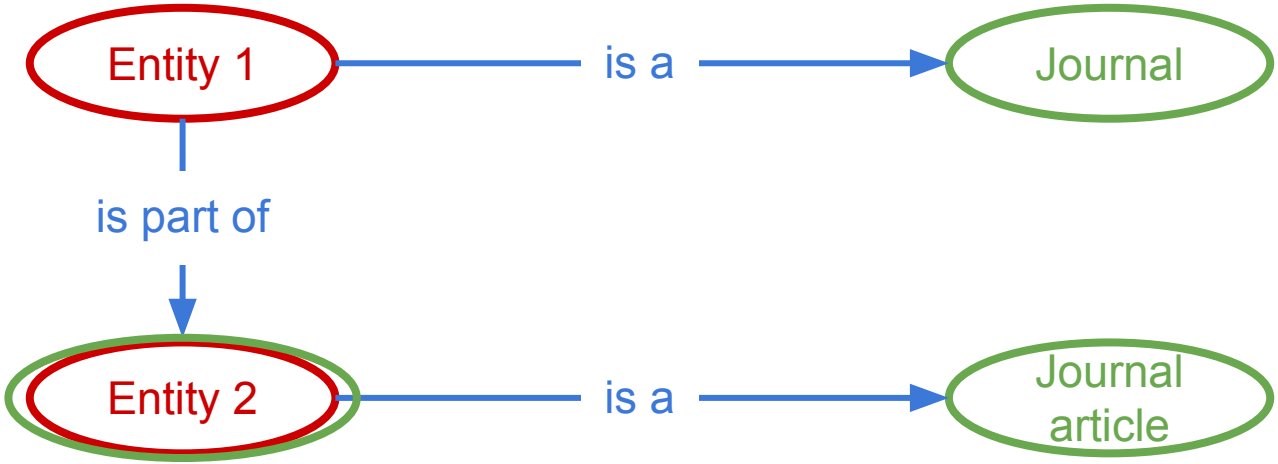
However, as anticipated, it is a good practice to use the same predicate only to express **one kind of data** (either only attributes or only relations) to avoid ambiguities that may become difficult to handle

Question 1
How can we explicitly specify how to use a certain predicate, so as to avoid such ambiguities?

Data correctness

Question 2

Is it possible to computationally claim that a collection of data is correct?



Data explanation

One of the activity we have to deal with when we have and handle a large collection of data is to explain to possible users and adopters what such data are about

Usual scenario: the person who was in charge of a information system and who handled the data of a company was fired, and has been substituted by a new employee. How can the newcomer to understand what such data are about?

Question 3
Can we explain the shape that data can have in a collection?

Data model

The answer to the three previous questions is that of defining a data model

A data model is an abstract and simplified representation of data of either a system or a real domain

In the context of this course, a data model is usually created through mathematical tools that define kinds of entities, their attributes and their relations

The adoption of such mathematical tools enables a formal definition of such a data model, so as to be used to assess data compliance against the data model through computational and automatic processes (e.g. to check the correctness of the data)

Types of entities

A data model enables the explicit definition the types of the entities described in the data – it is worth mentioning that each entity is assigned to **at least one type** (i.e. one **class** of entities)

A class of entities is the representation of a collection of entities compliant with the same description (i.e. set of attributes and relations) defined by such a class, for example:

- All of you who are attending this lecture (the entities) are people (the class)
- All the people who are enrolled in a course at any Italian university (the entities) are students (the class)

A data model permits the **specification of such classes**, their attributes and the relations they have with other classes (and, consequently, between the **occurrences of such classes**, that are the entities having those particular types)

Creating data models with UML

There are various languages that can be used to define a data model that facilitate also its consumption by humans

Among these languages, the Unified Modeling Language (UML) is one of the best well-known examples widely used in Informatics, in particular the class diagram which can be used to define classes, attributes and relations

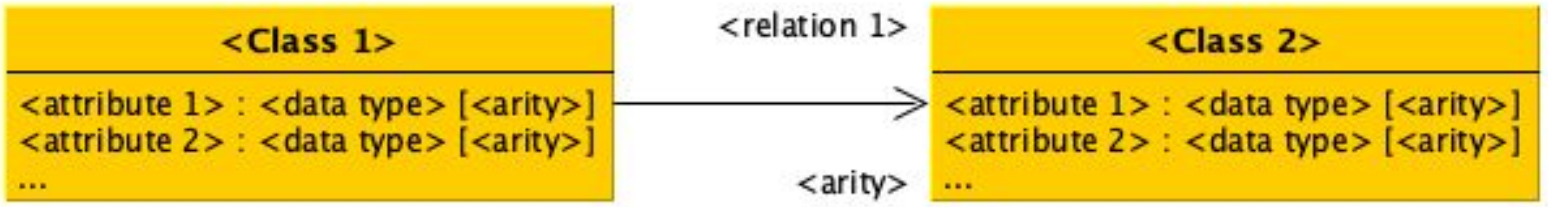
UML is supported by several visual editors – among those, the yEd graph editor is a viable, flexible and free alternative which is supported by Mac, Windows, and Linux, and already includes a simple palette for UML diagrams

Widgets used in UML class diagrams

Classes are represented by rectangles with an header which represent the name of the class

Such boxes also include (if any) the attributes associated to the related classes, which are defined specifying their name, the data type of the values attributed and the arity of such attributes: exactly one (1), zero or more (0..*), one or more (1..*), zero or one (0-1)

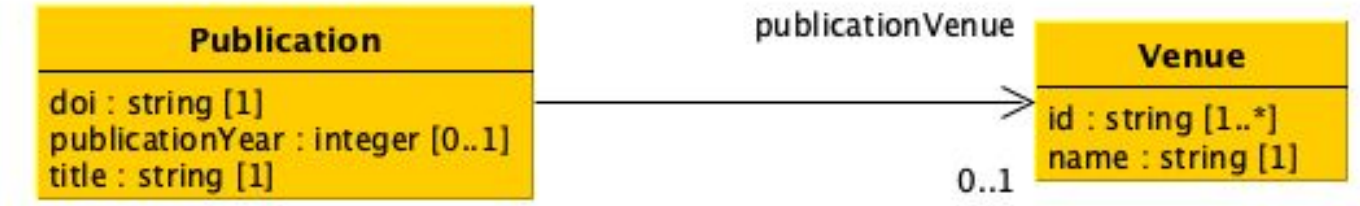
Relations are defined using an arrow from the source class to the target class, accompanied by a label (i.e. the name of the relation) and the related arity



A first example

All publications are identified necessarily by one and only one Digital Object Identifier (DOI), and are characterised by their title and, if known, their publication year. In addition, they may have been published within a particular venue.

Each venue is identified by at least one identifier and have always specified one name

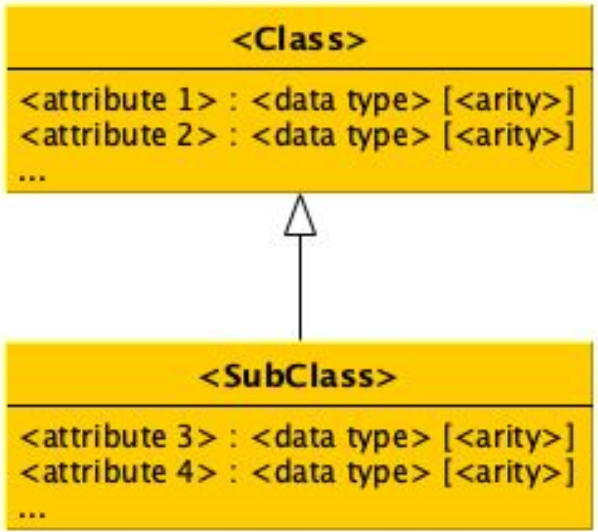


Defining subclasses

It is possible to specify more concrete specification of a given class, i.e. subclasses, by using a special arrow without any label specified on it

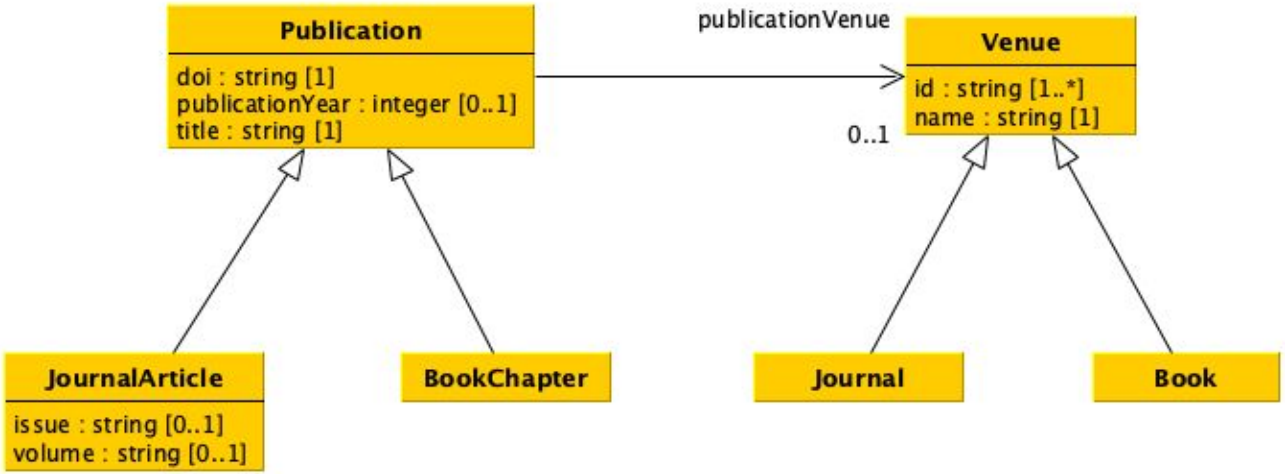
The subclass inherits all the attributes and relations defined in the parent class – meaning that they are defined also for all the entities that are occurrences of the subclass

In addition, the subclass may add new attributes and relations that apply only to it



Final example

Among the publications, there are journal articles and book chapters, published in journals and books respectively. Journal articles also may specify the issue number and the volume number in which they have been published.



Laboratory

Material:

- An A4 sheet of paper
- A pen

(alternatively, https://www.yworks.com/yed-live/ or similar)

Create the data model that enable one to describes the following scenario:

The article entitled “OpenCitations, an infrastructure organization for open scholarship” was published by the journal Quantitative Science Studies in 2020. The authors of this article, i.e. Silvio Peroni and David Shotton, also co-authored another conference paper entitled “The SPAR Ontologies”, that was published in the Proceedings of the Seventeenth International Semantic Web Conference, in 2018.

End

Introduction to data modelling

Silvio Peroni
silvio.peroni@unibo.it – https://orcid.org/0000-0003-0530-4305 – @essepuntato

Data Science (A.Y. 2022/2023)
Second Cycle Degree in Digital Humanities and Digital Knowledge
Alma Mater Studiorum - Università di Bologna

DIPARTIMENTO DI FILOLOGIA CLASSICA E ITALIANISTICA