# IPRs and their data analysis

## Module 6 – Patstat

Jacopo Staccioli, PHD[†‡]

[†] *Università Cattolica del Sacro Cuore, Milan*

[‡] *Scuola Superiore Sant'Anna, Pisa*

a.y. 2023-24

UNIVERSITÀ CATTOLICA del Sacro Cuore

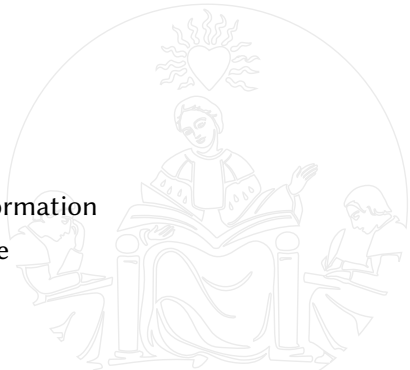
# Outline

# What is PATSTAT

- PATSTAT contains *bibliographic* and *legal status* patent data from leading industrialised and developing countries
  - ≈100 million patent records
  - ≈90 patent issuing authorities
  - mid-19th century up to today
- consists of 2 individual products

  Global : worldwide coverage[1]

  EP Register : EP patents with additional procedural information
- *snapshot* of the source databases at a single point in time
- we will focus on PATSTAT Global
- UNICATT version is "Spring(?) 2021"



[1] More information on coverage here: https://public.tableau.com/app/profile/patstat.support/viz/CoverageofPATSTAT2021SpringEdition/CoveragePATSTATGlobal

# Outline

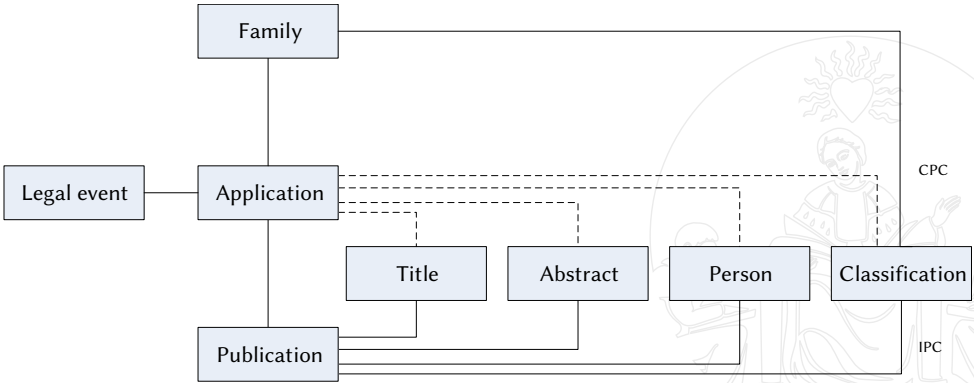
# Domain model – application

- core domain object is the application
  - i.e. a request for patent protection for an invention
- most other domain objects are related to applications
- during the life of a patent, various *publications* are issued
  - every application has at least one publication
- every application belongs to exactly one *simple family* and one *extended family*

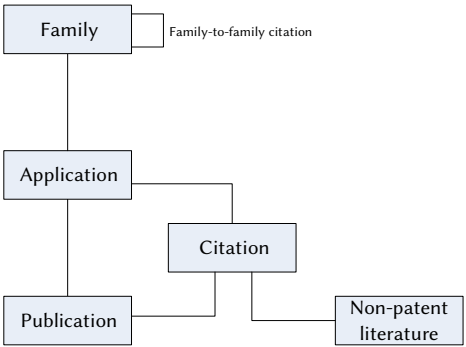
- strictly speaking, title, abstract, persons and classifications are part of the publication
- in PATSTAT these domain objects are *not* related to the individual publication, but to the application that gives rise to the publication

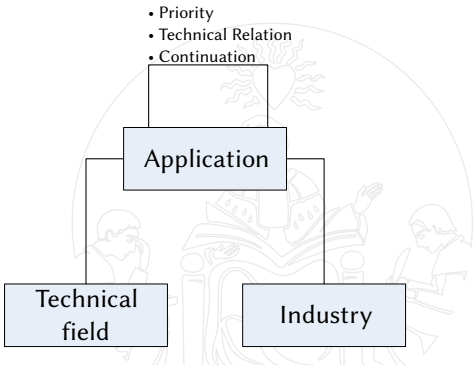
PATSTAT Global Data Catalog: http://documents.epo.org/projects/babylon/eponot.nsf/0/225F09FAA60945C2C125855F002797C2/$File/PATSTAT_DataCatalog_Global_v5-15.pdf

# Domain model diagrams



# Domain model diagrams (cont'd)



# Domain model – family

- applications which cover the same or similar invention are grouped into families
- each family contains one or more applications as family *members*

  *simple* : (AKA DOCDB)
  - group applications with the same *priority*
  - technical content of family members is regarded as (almost) identical
  - associated publications are sometimes called “equivalent”

  *extended* : (AKA INPADOC)
  - group applications linked to the same root priority application
  - applications are typically related to the same technical invention
  - but their individual content may differ
- for each type of family, each application belongs to exactly one family of that type
- extended family is potentially “broader” than the simple family
- each extended family contains the applications of one or more simple families

# Domain model – publication

- there are several types of publications
  - an application is typically published 18 months after filing date or priority date
  - granted patent specification is published when patent protection has been granted
  - corrections or publications of search reports, limitations…
- there is at least one publication for each application
  - or the application would still be confidential and would not be accessible in any database
- a publication typically consists of
  - a front page, which contains bibliographic data, the abstract, and a representative image
  - following pages with detailed description of the invention, the claims and the drawings

# Domain model – title, abstract

- Patstat reports the title of the invention, as shown on the front page of a publication
- in Patstat titles are related not to the individual publication, but to the application
- titles can be in any language
- Patstat contains only one title per application
- titles in English are given preference over titles in other languages

- the very same considerations also apply to abstracts

# Domain model – classification

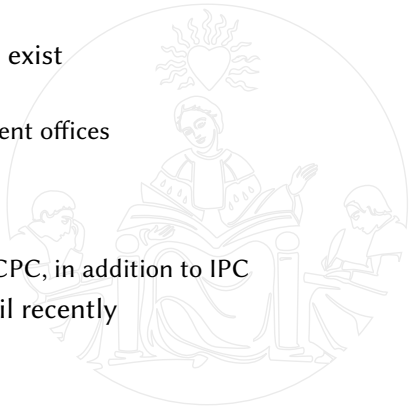
- applications are classified according to their technical content by some symbol or code to facilitate searching
- multiple, hierarchically structured classification systems exist

  IPC : *International Patent Classification*
  - maintained by WIPO and used by all patent offices

  CPC : *Cooperative Patent Classification*
  - created in 2013 as an extension of IPC
  - maintained by EPO and USPTO
  - many major offices are nowadays using CPC, in addition to IPC

  USPC : used by the US office for classification until recently

- other legacy national classification systems

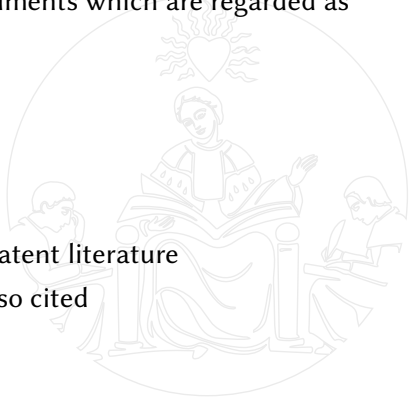


# Domain model – person

- persons may be *legal* or *natural*
- Patstat covers both the roles of
  - applicants : the person(s) who filed the patent application
  - inventors : in this case they are necessarily *natural* persons
- an application may have multiple applicants/inventors
- these may also change over time
- only applicants are mandatory for an application
- the same person can have multiple roles for the same application
  - e.g. a person can be applicant as well as inventor

# Domain model – citation

- citations are references from patent publications to documents which are regarded as relevant for the patent procedure
- identified in various stages
  - by the applicant before application
  - during search and examination by the patent office
  - during an opposition procedure
  - …
- patent publications typically cite other patents or non-patent literature
- less often applications (i.e. pre-grant publications) are also cited



# Domain model – other objects

## industry

- the European Union uses NACE rev. 2 to identify industries
- Statistical Classification of Economic Activities in the European Community
- PATSTAT assign NACE codes to applications by means of a IPC-NACE *crosswalk*

## legal event

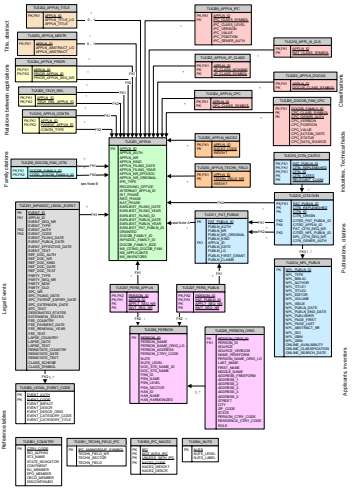
- procedural actions which change the (legal) status of an application or a granted patent
- e.g. refusal of an application, grant, change of address, attorney, person…

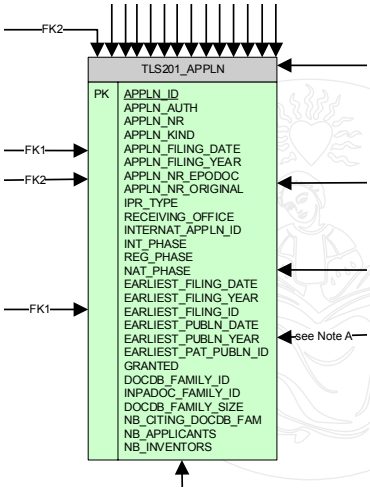
## technical field

- 35 technical fields assigned to applications, with reference table based on IPCs
- defined by WIPO and useful for some coarse statistical analysis

# Outline

# Logical model diagram

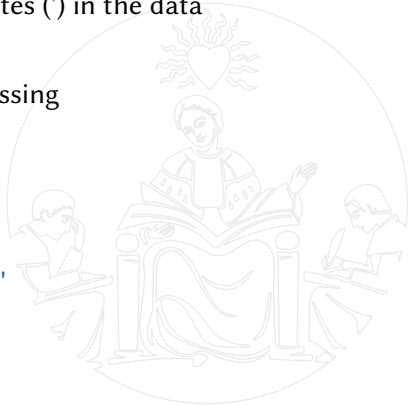




# Outline

# Design principles

- double quotes (") are consistently replaced by single quotes (') in the data
- line breaks within strings are replaced by `\n`
- for several documents, usually old ones, some data is missing
- PATSTAT does *not* contain any `NULL` values
  - all attributes satisfy a `NOT NULL` constraint
- PATSTAT represents missing values as *default* values
  - missing dates are represented as `'9999-12-31'`
  - missing literals are represented as a zero-length string `''`
  - missing numerics are represented as a `0`



# Design principles (cont'd)

## a word of caution

- suppose you want to SELECT all publications after a certain date, say 30th June, 2008
- consider a query with the following WHERE clause

```
. . . WHERE pub_date > '2008-06-30' . . .
```

- this does *not* simply return patents published later than 30th June, 2008
- publications with missing date are assigned the default value 9999-12-31 > 2008-06-30
- you need to explicitly exclude the default value

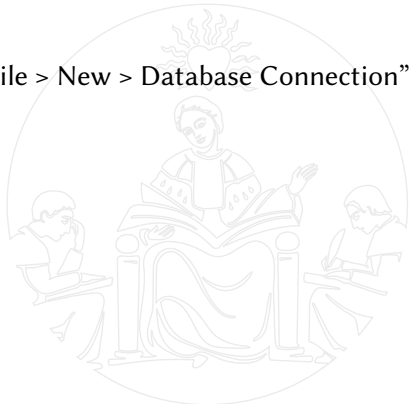
```
. . . WHERE pub_date > '2008-06-30' AND pub_date < '9999-12-31' . . .
```

# Housekeeping

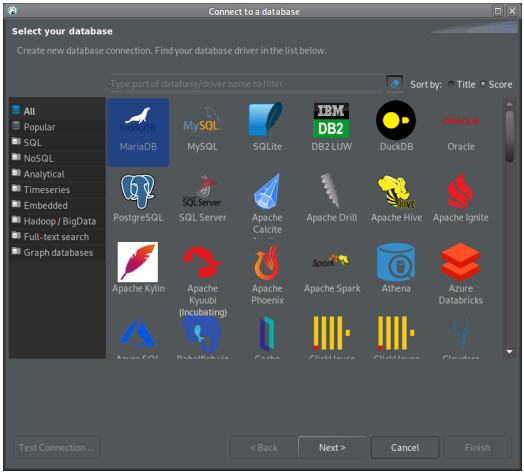
instructions to access Patstat

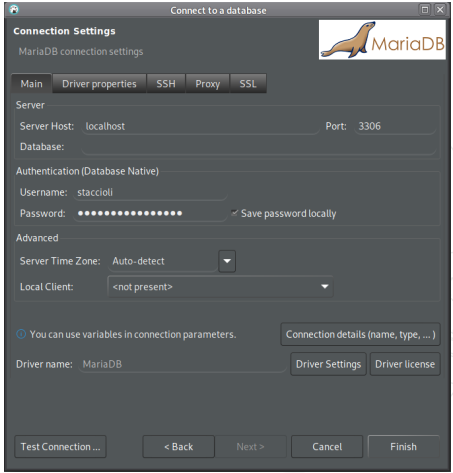
1. connect to the GUI of the `miced` server
2. open the DBeaver database client
3. click on the "New Database Connection" button  or "File > New > Database Connection"
4. select MariaDB and click "Next >"
5. fill in relevant parameters and credentials

   host: `localhost`
   port: `3306`
   database: `patstat2021`
   username: (received by email)
   password: (received by email)
6. leave everything else as is and click "Finish"



# Housekeeping (cont'd)



# Homework

1. Which patent office saw the largest number of applications filed in 2015?
2. What are the 10 most cited applications ever filed in Great Britain?
   Retrieve the number of citations, the application `id`, the whole (concatenated) application number, and the filing date.
3. Get all `A1` publications published by the USPTO within Q1/2009.
   Retrieve the whole publication number and the publication date.
   Also, count how many such publications are there.