IFRIS-Patstat database

Patents course : from EPO Patstat database to IFRIS patents database

08/10/2015

General introduction

- * Data sources
- * Data coverage

Technical and conceptual frameworks

- * What is a relational database ?
- * Architecture server-client
- * Data model patstat
- * Conceptual model (application)

Attributes and tables

- * Main different type of patents
- * What are the main analytical dimensions ?
- * Main tables and examples
- * Focus on specific relations : how to catch inventor locations ?
- * Live demo (sql queries) and results

What is Patstat IFRIS ?

- * Cleaning country code and adding classification
- * Problem of Silent :
  - * State of the coverage of the addresses
  - * enriching : regpat / inp
  - * addresses propagation
  - * artificial : what are they, how do we complete them
- * How to characterized technology : IFRIS technology classification
- * Some other attributes to facilitate the selection of patents
- * Live demos (sql queries) and results

Open discussion and links with Risis datasets

PATSTAT, also known as the EPO Worldwide Patent Statistical Database:

- It contains about 30 tables with bibliographic data, citations and family links...
- 70 million applications of about 90 countries.

Europäisches Patentamt
European Patent Office
Office européen des brevets

Patstat is a bianual snapshot of the EPO master documentation database (DOCDB, weekly updated by data provided by national offices). So what is not in DOCDB will not be available in PATSTAT !

Minor exceptions :

- with regards to dummy applications that have been created to compensate for un-linkable (unknown) applications (publications);
- also extra address information has been added from the EPO register and the USPTO register.

We will present here the IFRIS-patstat september 2011 version

We will present here the IFRIS-patstat september 2011 version

------

General introduction

* Data sources
* Data coverage

Technical and conceptual frameworks

* What is a relational database ?
* Architecture server-client
* Data model patstat
* Conceptual model (application)

Attributes and tables

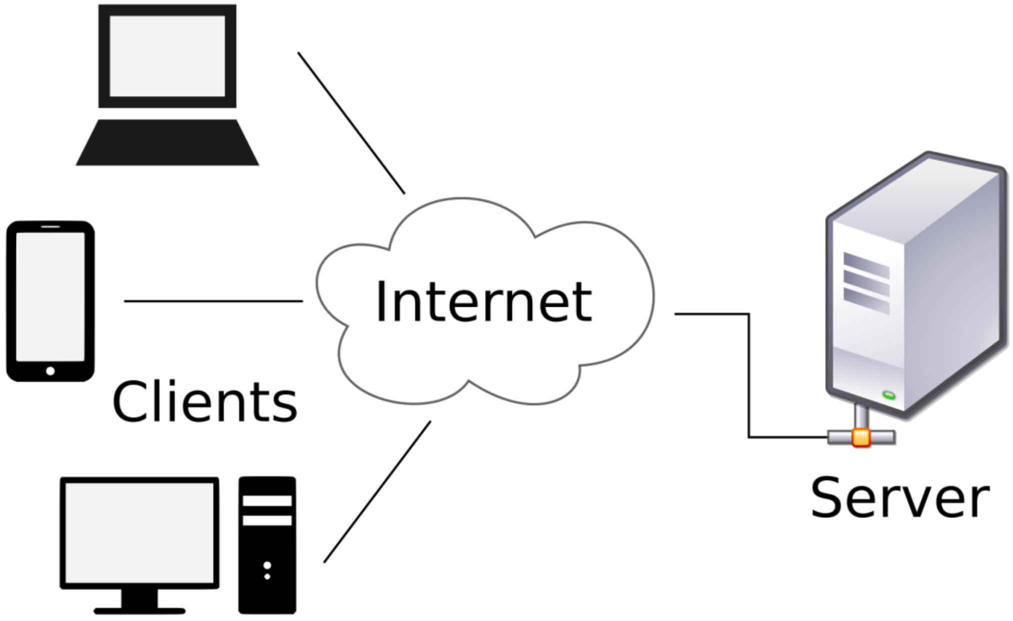
* Main different type of patents
* What are the main analytical dimensions ?
* Main tables and examples
* Focus on specific relations : how to catch inventor locations ?
* Live demo (sql queries) and results

What is Patstat IFRIS ?

* Cleaning country code and adding classification
* Problem of Silent :
    * State of the coverage of the addresses
    * enriching : regpat / inp
    * addresses propagation
    * artificial : what are they, how do we complete them
* How to characterized technology : IFRIS technology classification
* Some other attributes to facilitate the selection of patents
* Live demos (sql queries) and results

Open discussion and links with Risis datasets

Client-server architectur



One location for the data, with an efficient software (MySQL) for a remote access with :

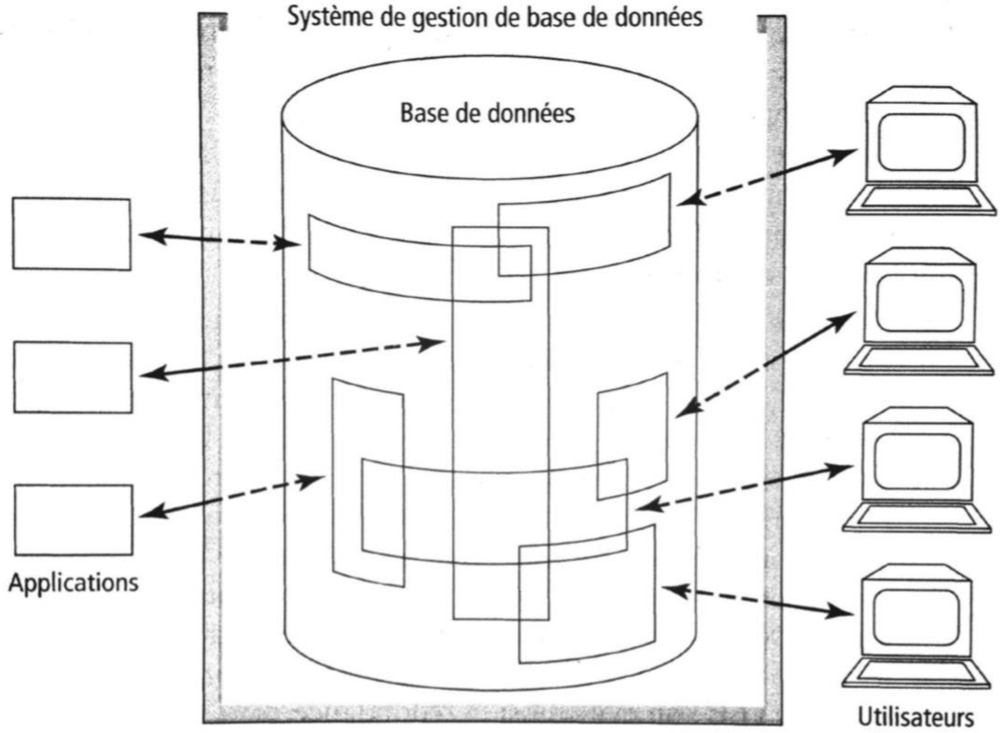
- Security of the data (back up strategy)
- Management of the concurency of the queries

Differents type of uses :

- Users with a direct access (command line, MySQL Workbench)
- Sotwares and application (Web Applications, and softwares for statistical analysis like R or SPSS)

Differents tables with :

- attributes (variables)
- links between them



troduction aux bases de données, Chris Date, 8ème édition, Vuibert (2004), p. 7

What is a relational database ?

A relational database is an softawre and hardare infrastructure where numerical information are strored.

It is like a collection of excel spreadsheet (tables), with variables (attributs), but with relations between some specific variables (the keys).

Theese keys make able to cross analytical dimensions throught spreadsheets.

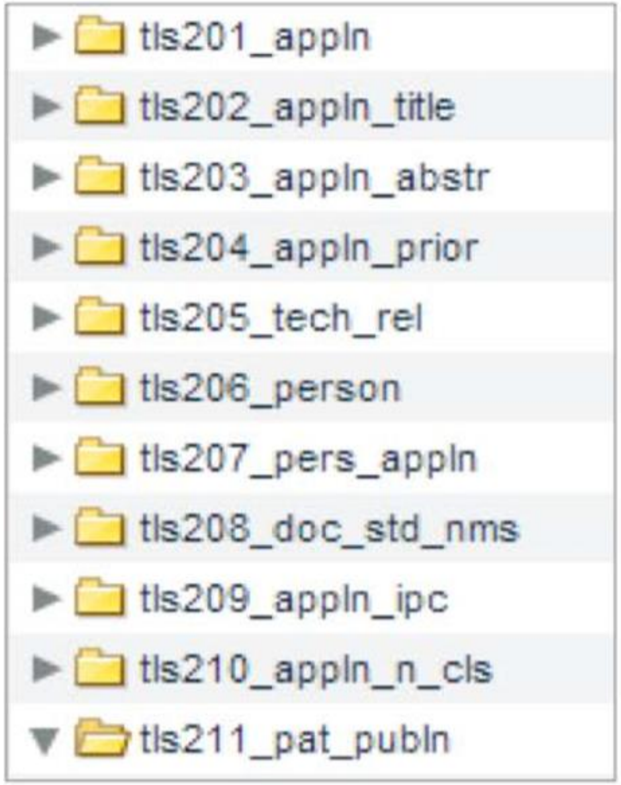
Values of attributes are stored in rows. You can ask complex questions (queries) to the data system and you can do some descriptive analysis.

Tables, Rows, Columns

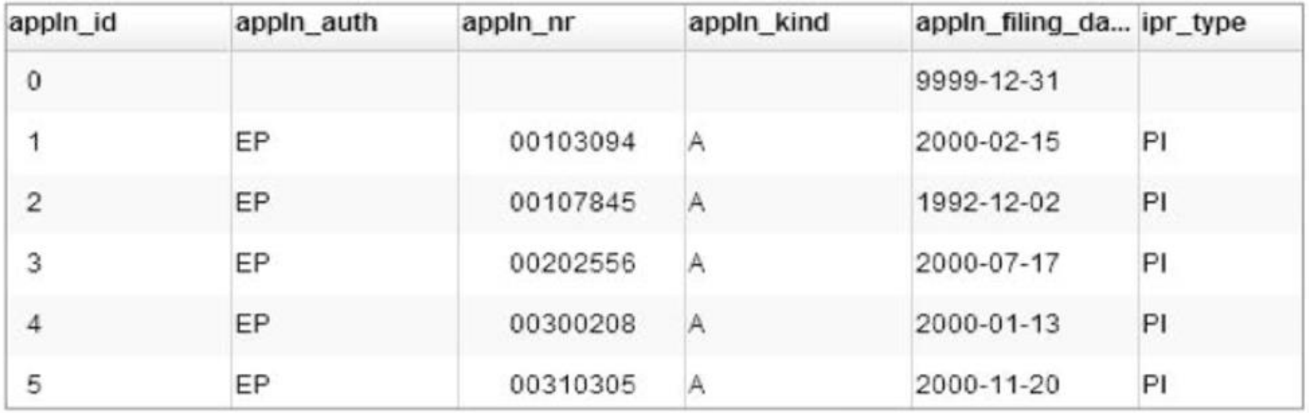
A relational database is a collection of tables.

A table consists of columns and rows.

The cells contain the data.



- ▶ tls201_appln
- ▶ tls202_appln_title
- ▶ tls203_appln_abstr
- ▶ tls204_appln_prior
- ▶ tls205_tech_rel
- ▶ tls206_person
- ▶ tls207_pers_appln
- ▶ tls208_doc_std_nms
- ▶ tls209_appln_ipc
- ▶ tls210_appln_n_cls
- ▼ tls211_pat_publn



appln_id	appln_auth	appln_nr	appln_kind	appln_filing_da...	ipr_type
0				9999-12-31	
1	EP	00103094	A	2000-02-15	PI
2	EP	00107845	A	1992-12-02	PI
3	EP	00202556	A	2000-07-17	PI
4	EP	00300208	A	2000-01-13	PI
5	EP	00310305	A	2000-11-20	PI

Relational Database concepts

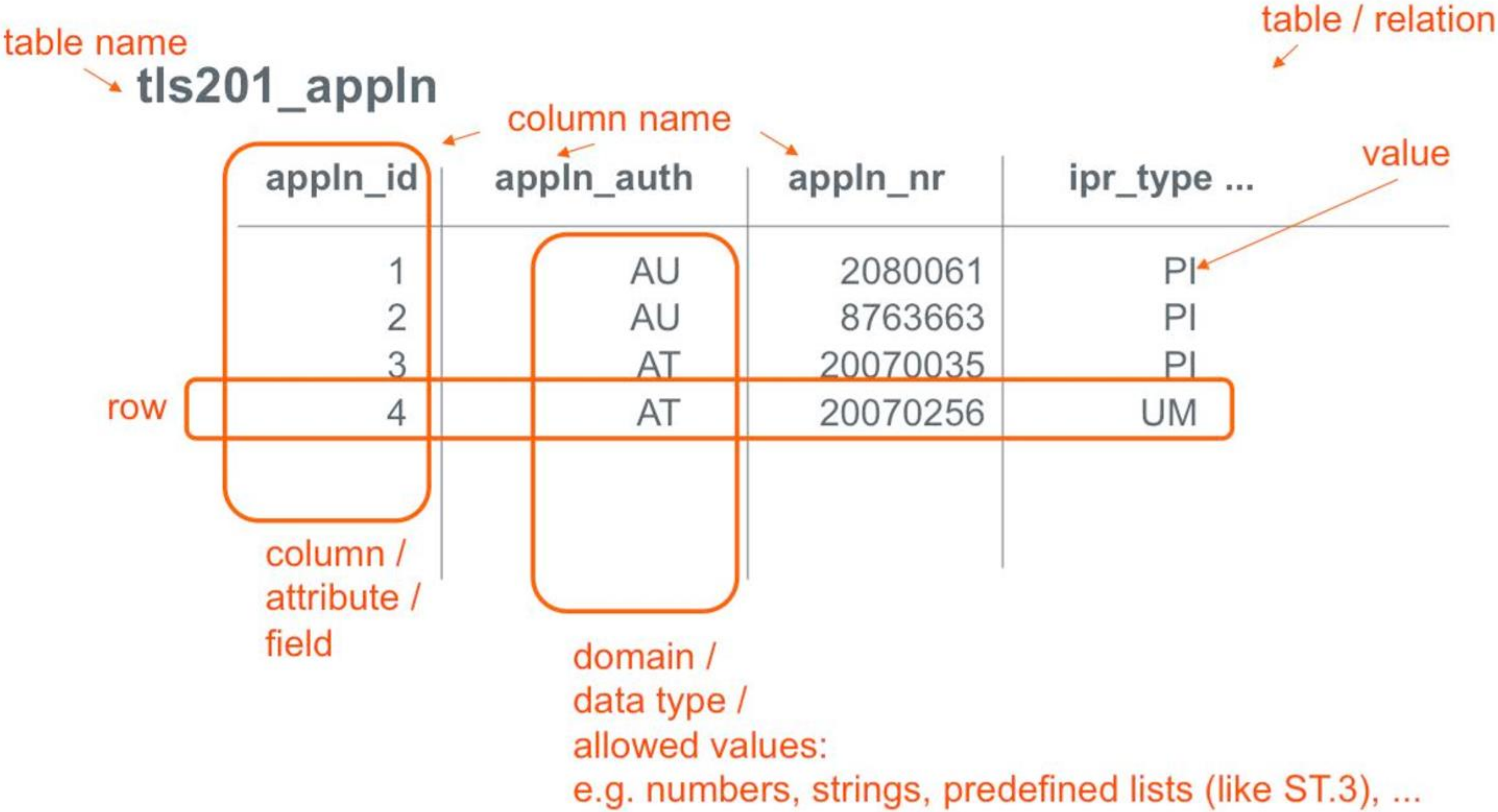


table name

tls201_appln

table / relation

column name

value

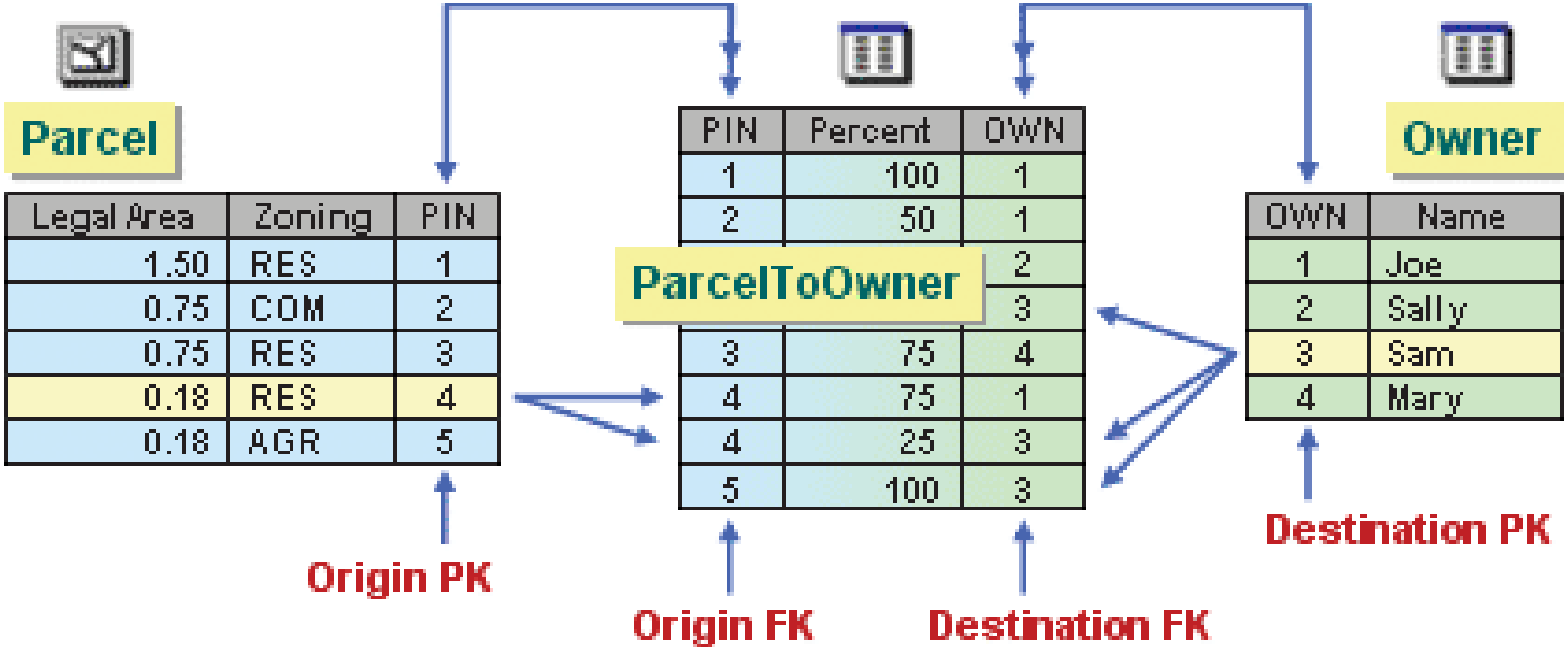
appln_id	appln_auth	appln_nr	ipr_type ...
1	AU	2080061	PI
2	AU	8763663	PI
3	AT	20070035	PI
4	AT	20070256	UM

row

column /
attribute /
field

domain /
data type /
allowed values:
e.g. numbers, strings, predefined lists (like ST.3), ...

Exemple with three tables for land occupation and owners



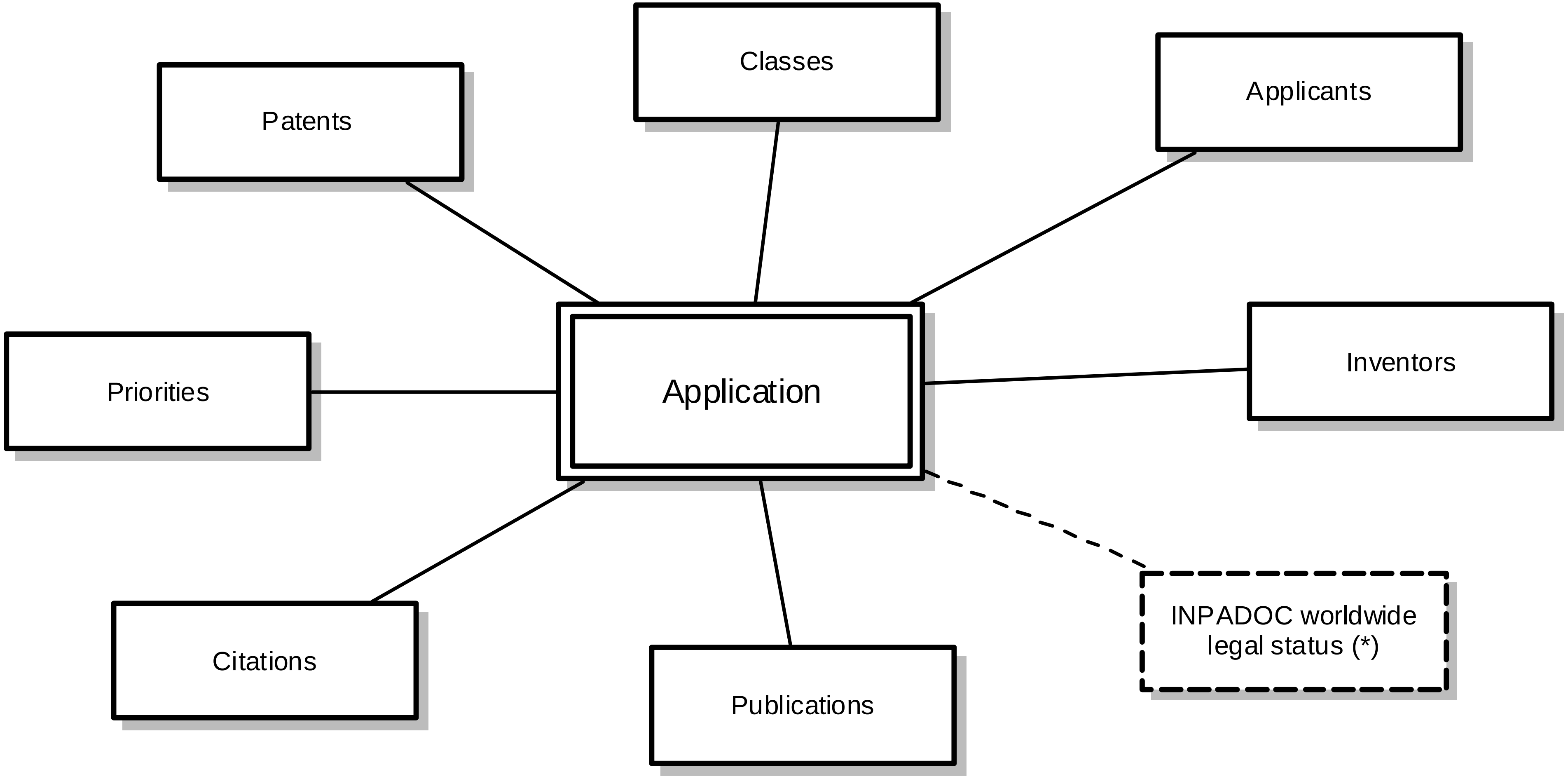
Legal Area	Zoning	PIN
1.50	RES	1
0.75	COM	2
0.75	RES	3
0.18	RES	4
0.18	AGR	5

PIN	Percent	OWN
1	100	1
2	50	1
		2
		3
3	75	4
4	75	1
4	25	3
5	100	3

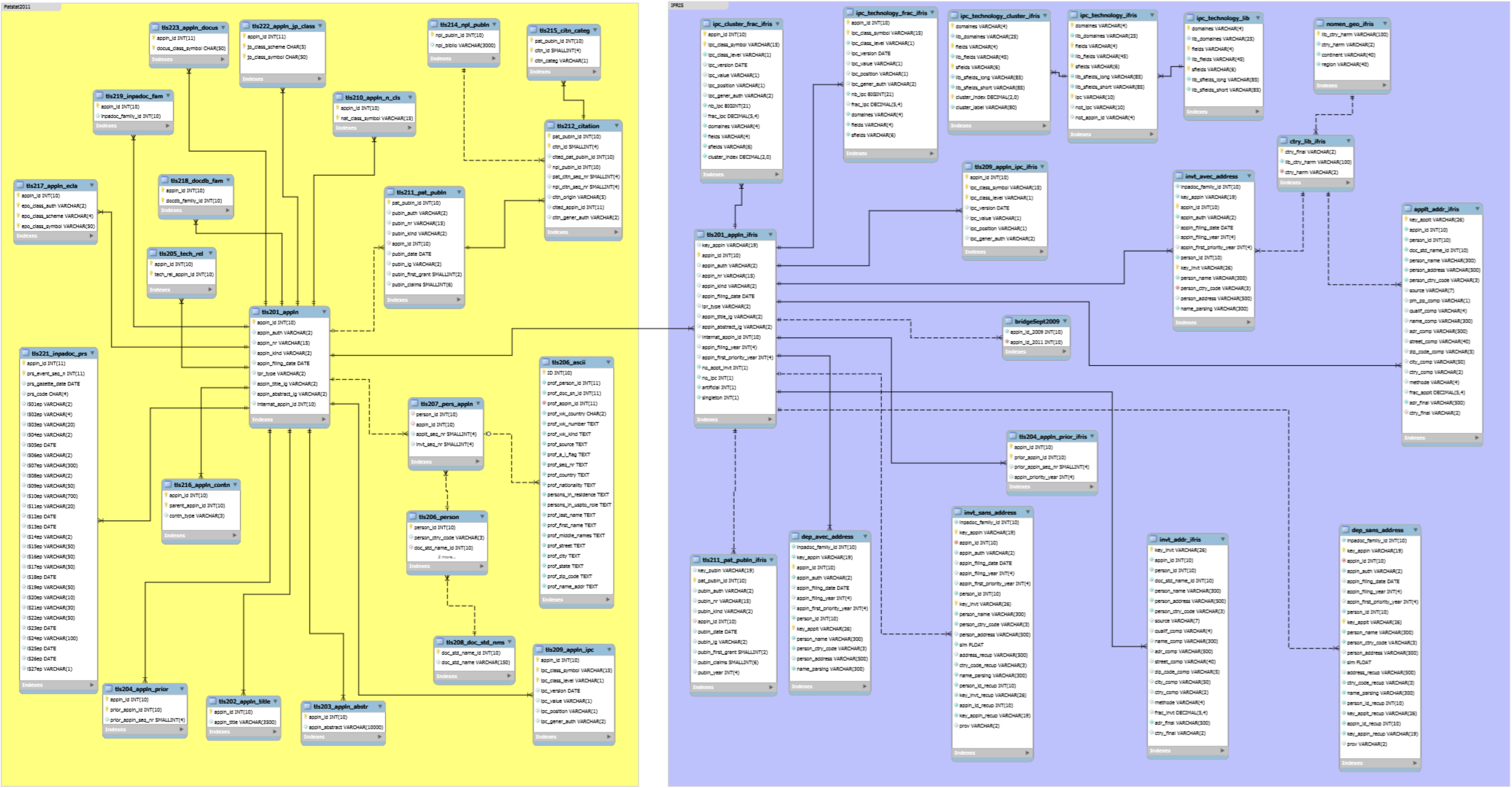
OWN	Name
1	Joe
2	Sally
3	Sam
4	Mary

One to many relation or one to one relation !

Conceptual relational model of Patstat



Central position of the application table (tls_201)


Patstat2011
tls223_appln_docus
appln_id INT(11)
docus_class_symbol CHAR(50)
Indexes
tls222_appln_jp_class
appln_id INT(11)
jp_class_scheme CHAR(5)
jp_class_symbol CHAR(50)
Indexes
tls214_npl_publn
npl_publn_id INT(10)
npl_biblio VARCHAR(3000)
Indexes
tls215_citn_categ
pat_publn_id INT(10)
citn_id SMALLINT(4)
citn_categ VARCHAR(1)
Indexes
tls219_inpadoc_fam
appln_id INT(10)
inpadoc_family_id INT(10)
Indexes
tls210_appln_n_cls
appln_id INT(10)
nat_class_symbol VARCHAR(15)
Indexes
tls212_citation
pat_publn_id INT(10)
citn_id SMALLINT(4)
cited_pat_publn_id INT(10)
npl_publn_id INT(10)
pat_citn_seq_nr SMALLINT(4)
npl_citn_seq_nr SMALLINT(4)
citn_origin VARCHAR(5)
cited_appln_id INT(11)
citn_gener_auth VARCHAR(2)
Indexes
tls217_appln_ecla
appln_id INT(10)
epo_class_auth VARCHAR(2)
epo_class_scheme VARCHAR(4)
epo_class_symbol VARCHAR(50)
Indexes
tls218_docdb_fam
appln_id INT(10)
docdb_family_id INT(10)
Indexes
tls211_pat_publn
pat_publn_id INT(10)
publn_auth VARCHAR(2)
publn_nr VARCHAR(15)
publn_kind VARCHAR(2)
appln_id INT(10)
publn_date DATE
publn_lg VARCHAR(2)
publn_first_grant SMALLINT(2)
publn_claims SMALLINT(6)
Indexes
tls205_tech_rel
appln_id INT(10)
tech_rel_appln_id INT(10)
Indexes
tls201_appln
appln_id INT(10)
appln_auth VARCHAR(2)
appln_nr VARCHAR(15)
appln_kind VARCHAR(2)
appln_filing_date DATE
ipr_type VARCHAR(2)
appln_title_lg VARCHAR(2)
appln_abstract_lg VARCHAR(2)
internat_appln_id INT(10)
Indexes
tls221_inpadoc_prs
appln_id INT(11)
prs_event_seq_n INT(11)
prs_gazette_date DATE
prs_code CHAR(4)
IS01ep VARCHAR(2)
IS02ep VARCHAR(4)
IS03ep VARCHAR(20)
IS04ep VARCHAR(2)
IS05ep DATE
IS06ep VARCHAR(2)
IS07ep VARCHAR(300)
IS08ep VARCHAR(2)
IS09ep VARCHAR(50)
IS10ep VARCHAR(700)
IS11ep VARCHAR(20)
IS12ep DATE
IS13ep DATE
IS14ep VARCHAR(2)
IS15ep VARCHAR(50)
IS16ep VARCHAR(50)
IS17ep VARCHAR(50)
IS18ep DATE
IS19ep VARCHAR(50)
IS20ep VARCHAR(10)
IS21ep VARCHAR(30)
IS22ep VARCHAR(50)
IS23ep DATE
IS24ep VARCHAR(100)
IS25ep DATE
IS26ep DATE
IS27ep VARCHAR(1)
Indexes
tls206_ascii
ID INT(10)
prof_person_id INT(11)
prof_doc_sn_id INT(11)
prof_appln_id INT(11)
prof_wk_country CHAR(2)
prof_wk_number TEXT
prof_wk_kind TEXT
prof_source TEXT
prof_a_i_flag TEXT
prof_seq_nr TEXT
prof_country TEXT
prof_nationality TEXT
persons_in_residence TEXT
persons_in_uspto_role TEXT
prof_last_name TEXT
prof_first_name TEXT
prof_middle_names TEXT
prof_street TEXT
prof_city TEXT
prof_state TEXT
prof_zip_code TEXT
prof_name_addr TEXT
Indexes
tls207_pers_appln
person_id INT(10)
appln_id INT(10)
applt_seq_nr SMALLINT(4)
invt_seq_nr SMALLINT(4)
Indexes
tls216_appln_contn
appln_id INT(10)
parent_appln_id INT(10)
contn_type VARCHAR(3)
Indexes
tls206_person
person_id INT(10)
person_ctry_code VARCHAR(3)
doc_std_name_id INT(10)
2 more...
Indexes
tls208_doc_std_nms
doc_std_name_id INT(10)
doc_std_name VARCHAR(150)
Indexes
tls209_appln_ipc
appln_id INT(10)
ipc_class_symbol VARCHAR(15)
ipc_class_level VARCHAR(1)
ipc_version DATE
ipc_value VARCHAR(1)
ipc_position VARCHAR(1)
ipc_gener_auth VARCHAR(2)
Indexes
tls204_appln_prior
appln_id INT(10)
prior_appln_id INT(10)
prior_appln_seq_nr SMALLINT(4)
Indexes
tls202_appln_title
appln_id INT(10)
appln_title VARCHAR(3500)
Indexes
tls203_appln_abstr
appln_id INT(10)
appln_abstract VARCHAR(10000)
Indexes
IFRIS
ipc_cluster_frac_ifris
appln_id INT(10)
ipc_class_symbol VARCHAR(15)
ipc_class_level VARCHAR(1)
ipc_version DATE
ipc_value VARCHAR(1)
ipc_position VARCHAR(1)
ipc_gener_auth VARCHAR(2)
nb_ipc BIGINT(21)
frac_ipc DECIMAL(5,4)
domaines VARCHAR(4)
fields VARCHAR(4)
sfields VARCHAR(6)
cluster_index DECIMAL(2,0)
Indexes
ipc_technology_frac_ifris
appln_id INT(10)
ipc_class_symbol VARCHAR(15)
ipc_class_level VARCHAR(1)
ipc_version DATE
ipc_value VARCHAR(1)
ipc_position VARCHAR(1)
ipc_gener_auth VARCHAR(2)
nb_ipc BIGINT(21)
frac_ipc DECIMAL(5,4)
domaines VARCHAR(4)
fields VARCHAR(4)
sfields VARCHAR(6)
Indexes
ipc_technology_cluster_ifris
domaines VARCHAR(4)
lib_domaines VARCHAR(25)
fields VARCHAR(4)
lib_fields VARCHAR(45)
sfields VARCHAR(6)
lib_sfields_long VARCHAR(85)
lib_sfields_short VARCHAR(85)
cluster_index DECIMAL(2,0)
cluster_label VARCHAR(80)
Indexes
ipc_technology_ifris
domaines VARCHAR(4)
lib_domaines VARCHAR(25)
fields VARCHAR(4)
lib_fields VARCHAR(45)
sfields VARCHAR(6)
lib_sfields_long VARCHAR(85)
lib_sfields_short VARCHAR(85)
ipc VARCHAR(10)
not_ipc VARCHAR(10)
not_appln_id VARCHAR(4)
Indexes
ipc_technology_lib
domaines VARCHAR(4)
lib_domaines VARCHAR(25)
fields VARCHAR(4)
lib_fields VARCHAR(45)
sfields VARCHAR(6)
lib_sfields_long VARCHAR(85)
lib_sfields_short VARCHAR(85)
Indexes
nomen_geo_ifris
lib_ctry_harm VARCHAR(100)
ctry_harm VARCHAR(2)
continent VARCHAR(40)
region VARCHAR(40)
Indexes
ctry_lib_ifris
ctry_final VARCHAR(2)
lib_ctry_harm VARCHAR(100)
ctry_harm VARCHAR(2)
Indexes
tls209_appln_ipc_ifris
appln_id INT(10)
ipc_class_symbol VARCHAR(15)
ipc_class_level VARCHAR(1)
ipc_version DATE
ipc_value VARCHAR(1)
ipc_position VARCHAR(1)
ipc_gener_auth VARCHAR(2)
Indexes
invt_avec_address
inpadoc_family_id INT(10)
key_appln VARCHAR(19)
appln_id INT(10)
appln_auth VARCHAR(2)
appln_filing_date DATE
appln_filing_year INT(4)
appln_first_priority_year INT(4)
person_id INT(10)
key_invt VARCHAR(26)
person_name VARCHAR(300)
person_ctry_code VARCHAR(3)
person_address VARCHAR(500)
name_parsing VARCHAR(300)
Indexes
tls201_appln_ifris
key_appln VARCHAR(19)
appln_id INT(10)
appln_auth VARCHAR(2)
appln_nr VARCHAR(15)
appln_kind VARCHAR(2)
appln_filing_date DATE
ipr_type VARCHAR(2)
appln_title_lg VARCHAR(2)
appln_abstract_lg VARCHAR(2)
internat_appln_id INT(10)
appln_filing_year INT(4)
appln_first_priority_year INT(4)
no_appt_invt INT(1)
no_ipc INT(1)
artificial INT(1)
singleton INT(1)
Indexes
applt_addr_ifris
key_applt VARCHAR(26)
appln_id INT(10)
person_id INT(10)
doc_std_name_id INT(10)
person_name VARCHAR(300)
person_address VARCHAR(500)
person_ctry_code VARCHAR(3)
source VARCHAR(7)
pm_pp_comp VARCHAR(1)
qualif_comp VARCHAR(4)
name_comp VARCHAR(300)
adr_comp VARCHAR(500)
street_comp VARCHAR(40)
zip_code_comp VARCHAR(5)
city_comp VARCHAR(50)
ctry_comp VARCHAR(2)
methode VARCHAR(4)
frac_applt DECIMAL(5,4)
adr_final VARCHAR(500)
ctry_final VARCHAR(2)
Indexes
bridgeSept2009
appln_id_2009 INT(10)
appln_id_2011 INT(10)
Indexes
tls204_appln_prior_ifris
appln_id INT(10)
prior_appln_id INT(10)
prior_appln_seq_nr SMALLINT(4)
appln_priority_year INT(4)
Indexes
tls211_pat_publn_ifris
key_publn VARCHAR(19)
pat_publn_id INT(10)
publn_auth VARCHAR(2)
publn_nr VARCHAR(15)
publn_kind VARCHAR(2)
appln_id INT(10)
publn_date DATE
publn_lg VARCHAR(2)
publn_first_grant SMALLINT(2)
publn_claims SMALLINT(6)
publn_year INT(4)
Indexes
dep_avec_address
inpadoc_family_id INT(10)
key_appln VARCHAR(19)
appln_id INT(10)
appln_auth VARCHAR(2)
appln_filing_date DATE
appln_filing_year INT(4)
appln_first_priority_year INT(4)
person_id INT(10)
key_applt VARCHAR(26)
person_name VARCHAR(300)
person_ctry_code VARCHAR(3)
person_address VARCHAR(500)
name_parsing VARCHAR(300)
Indexes
invt_sans_address
inpadoc_family_id INT(10)
key_appln VARCHAR(19)
appln_id INT(10)
appln_auth VARCHAR(2)
appln_filing_date DATE
appln_filing_year INT(4)
appln_first_priority_year INT(4)
person_id INT(10)
key_invt VARCHAR(26)
person_name VARCHAR(300)
person_ctry_code VARCHAR(3)
person_address VARCHAR(500)
sim FLOAT
address_recup VARCHAR(500)
ctry_code_recup VARCHAR(3)
name_parsing VARCHAR(300)
person_id_recup INT(10)
key_invt_recup VARCHAR(26)
appln_id_recup INT(10)
key_appln_recup VARCHAR(19)
prov VARCHAR(2)
Indexes
invt_addr_ifris
key_invt VARCHAR(26)
appln_id INT(10)
person_id INT(10)
doc_std_name_id INT(10)
person_name VARCHAR(300)
person_address VARCHAR(500)
person_ctry_code VARCHAR(3)
source VARCHAR(7)
qualif_comp VARCHAR(4)
name_comp VARCHAR(300)
adr_comp VARCHAR(500)
street_comp VARCHAR(40)
zip_code_comp VARCHAR(5)
city_comp VARCHAR(50)
ctry_comp VARCHAR(2)
methode VARCHAR(4)
frac_invt DECIMAL(5,4)
adr_final VARCHAR(500)
ctry_final VARCHAR(2)
Indexes
dep_sans_address
inpadoc_family_id INT(10)
key_appln VARCHAR(19)
appln_id INT(10)
appln_auth VARCHAR(2)
appln_filing_date DATE
appln_filing_year INT(4)
appln_first_priority_year INT(4)
person_id INT(10)
key_applt VARCHAR(26)
person_name VARCHAR(300)
person_ctry_code VARCHAR(3)
person_address VARCHAR(500)
sim FLOAT
address_recup VARCHAR(500)
ctry_code_recup VARCHAR(3)
name_parsing VARCHAR(300)
person_id_recup INT(10)
key_applt_recup VARCHAR(26)
appln_id_recup INT(10)
key_appln_recup VARCHAR(19)
prov VARCHAR(2)
Indexes

General introduction

* Data sources

* Data coverage

Technical and conceptual frameworks

* What is a relational database ?

* Architecture server-client

* Data model patstat

* Conceptual model (application)

Attributes and tables

* Main different type of patents

* What are the main analytical dimensions ?

* Main tables and examples

* Focus on specific relations : how to catch inventor locations ?

* Live demo (sql queries) and results

What is Patstat IFRIS ?

* Cleaning country code and adding classification

* Problem of Silent :

  * State of the coverage of the addresses

  * enriching : regpat / inp

  * addresses propagation

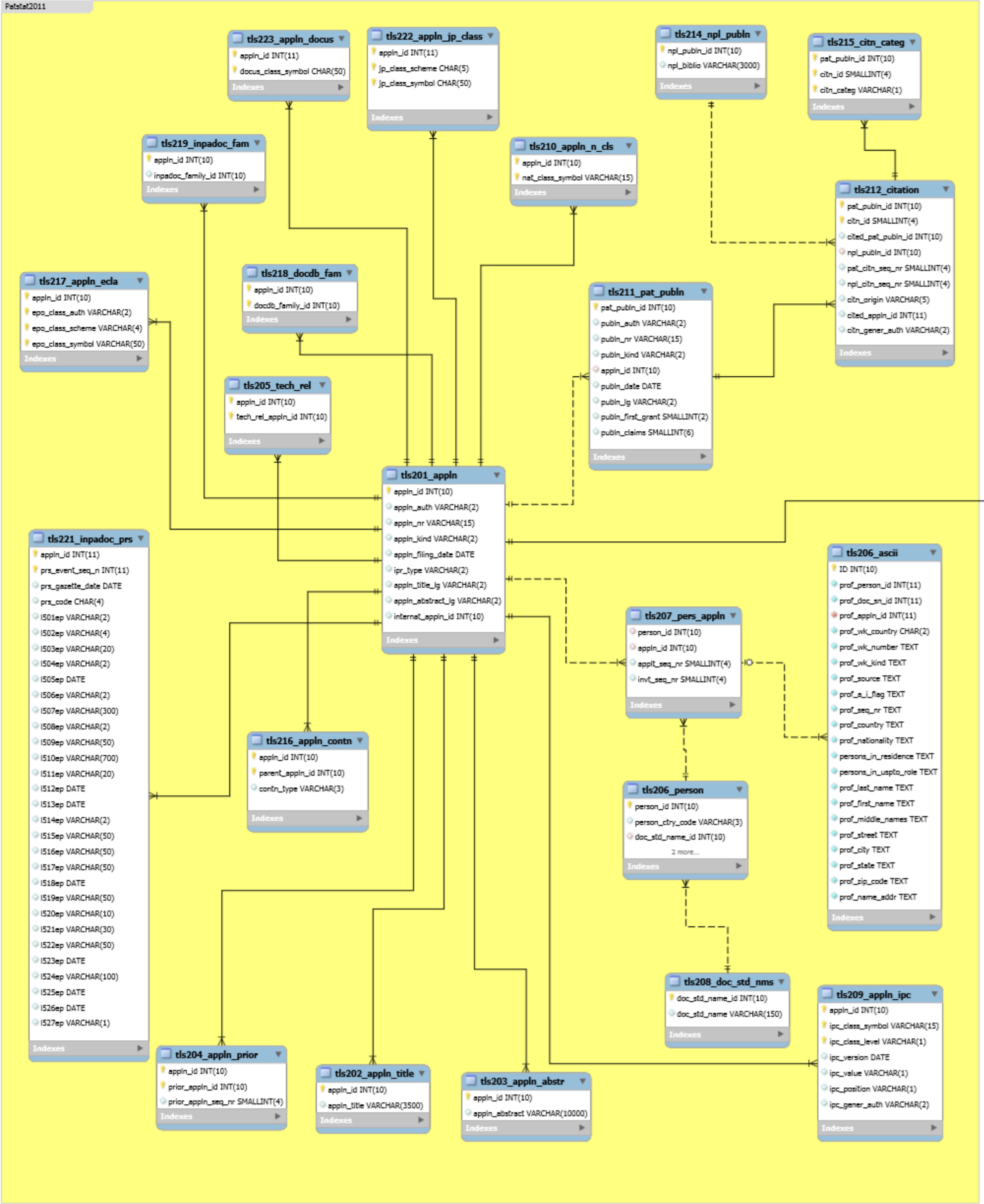
  * artificial : what are they, how do we complete them

* How to characterized technology : IFRIS technology classification

* Some other attributes to facilitate the selection of patents

* Live demos (sql queries) and results

Open discussion and links with Risis datasets



Patstat2011

tls223_appln_docus
appln_id INT(11)
docus_class_symbol CHAR(50)
Indexes

tls222_appln_jp_class
appln_id INT(11)
jp_class_scheme CHAR(5)
jp_class_symbol CHAR(50)
Indexes

tls214_npl_publn
npl_publn_id INT(10)
npl_biblio VARCHAR(3000)
Indexes

tls215_citn_categ
pat_publn_id INT(10)
citn_id SMALLINT(4)
citn_categ VARCHAR(1)
Indexes

tls219_inpadoc_fam
appln_id INT(10)
inpadoc_family_id INT(10)
Indexes

tls210_appln_n_cls
appln_id INT(10)
nat_class_symbol VARCHAR(15)
Indexes

tls212_citation
pat_publn_id INT(10)
citn_id SMALLINT(4)
cited_pat_publn_id INT(10)
npl_publn_id INT(10)
pat_citn_seq_nr SMALLINT(4)
npl_citn_seq_nr SMALLINT(4)
citn_origin VARCHAR(5)
cited_appln_id INT(11)
citn_gener_auth VARCHAR(2)
Indexes

tls217_appln_ecla
appln_id INT(10)
epo_class_auth VARCHAR(2)
epo_class_scheme VARCHAR(4)
epo_class_symbol VARCHAR(50)
Indexes

tls218_docdb_fam
appln_id INT(10)
docdb_family_id INT(10)
Indexes

tls211_pat_publn
pat_publn_id INT(10)
publn_auth VARCHAR(2)
publn_nr VARCHAR(15)
publn_kind VARCHAR(2)
appln_id INT(10)
publn_date DATE
publn_lg VARCHAR(2)
publn_first_grant SMALLINT(2)
publn_claims SMALLINT(6)
Indexes

tls205_tech_rel
appln_id INT(10)
tech_rel_appln_id INT(10)
Indexes

tls201_appln
appln_id INT(10)
appln_auth VARCHAR(2)
appln_nr VARCHAR(15)
appln_kind VARCHAR(2)
appln_filing_date DATE
ipr_type VARCHAR(2)
appln_title_lg VARCHAR(2)
appln_abstract_lg VARCHAR(2)
internat_appln_id INT(10)
Indexes

tls221_inpadoc_prs
appln_id INT(11)
prs_event_seq_n INT(11)
prs_gazette_date DATE
prs_code CHAR(4)
l501ep VARCHAR(2)
l502ep VARCHAR(4)
l503ep VARCHAR(20)
l504ep VARCHAR(2)
l505ep DATE
l506ep VARCHAR(2)
l507ep VARCHAR(300)
l508ep VARCHAR(2)
l509ep VARCHAR(50)
l510ep VARCHAR(700)
l511ep VARCHAR(20)
l512ep DATE
l513ep DATE
l514ep VARCHAR(2)
l515ep VARCHAR(50)
l516ep VARCHAR(50)
l517ep VARCHAR(50)
l518ep DATE
l519ep VARCHAR(50)
l520ep VARCHAR(10)
l521ep VARCHAR(30)
l522ep VARCHAR(50)
l523ep DATE
l524ep VARCHAR(100)
l525ep DATE
l526ep DATE
l527ep VARCHAR(1)
Indexes

tls206_ascii
ID INT(10)
prof_person_id INT(11)
prof_doc_sn_id INT(11)
prof_appln_id INT(11)
prof_wk_country CHAR(2)
prof_wk_number TEXT
prof_wk_kind TEXT
prof_source TEXT
prof_a_i_flag TEXT
prof_seq_nr TEXT
prof_country TEXT
prof_nationality TEXT
persons_in_residence TEXT
persons_in_uspto_role TEXT
prof_last_name TEXT
prof_first_name TEXT
prof_middle_names TEXT
prof_street TEXT
prof_city TEXT
prof_state TEXT
prof_zip_code TEXT
prof_name_addr TEXT
Indexes

tls207_pers_appln
person_id INT(10)
appln_id INT(10)
applt_seq_nr SMALLINT(4)
invt_seq_nr SMALLINT(4)
Indexes

tls216_appln_contn
appln_id INT(10)
parent_appln_id INT(10)
contn_type VARCHAR(3)
Indexes

tls206_person
person_id INT(10)
person_ctry_code VARCHAR(3)
doc_std_name_id INT(10)
2 more...
Indexes

tls208_doc_std_nms
doc_std_name_id INT(10)
doc_std_name VARCHAR(150)
Indexes

tls209_appln_ipc
appln_id INT(10)
ipc_class_symbol VARCHAR(15)
ipc_class_level VARCHAR(1)
ipc_version DATE
ipc_value VARCHAR(1)
ipc_position VARCHAR(1)
ipc_gener_auth VARCHAR(2)
Indexes

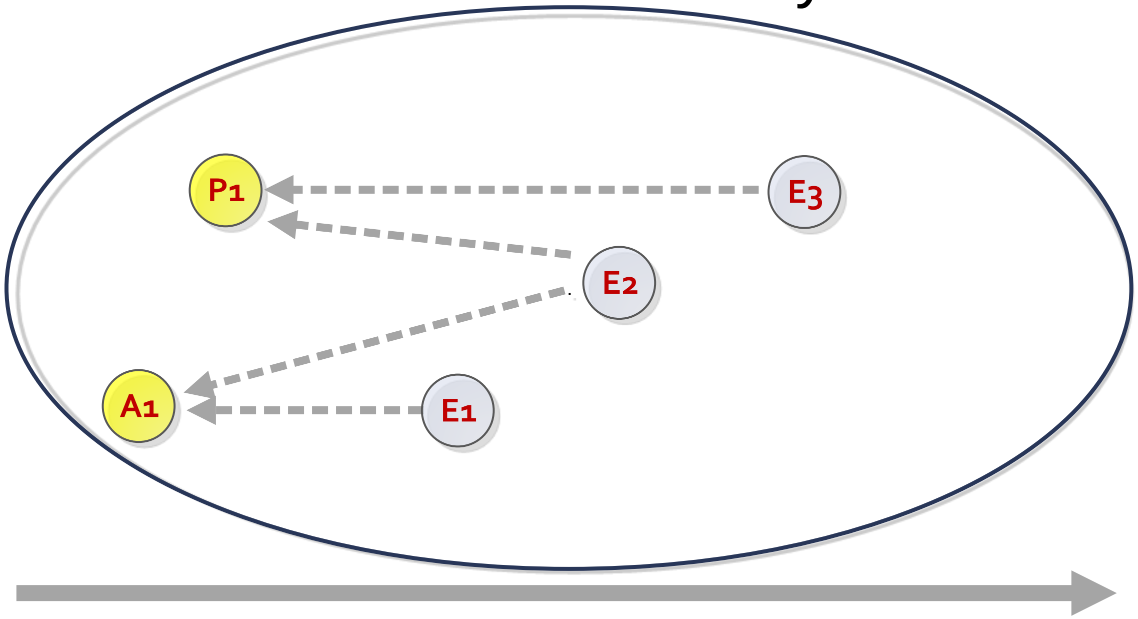
tls204_appln_prior
appln_id INT(10)
prior_appln_id INT(10)
prior_appln_seq_nr SMALLINT(4)
Indexes

tls202_appln_title
appln_id INT(10)
appln_title VARCHAR(3500)
Indexes

tls203_appln_abstr
appln_id INT(10)
appln_abstract VARCHAR(10000)
Indexes

Main types of patents : how to identify priority patents

INPADOC Familly



Filing year

Patent type	Number of patents	%
Priority patents applications non Singleton	27 284 402	39,1%
Priority patents applications singleton	6 617 050	9,5%
Non priority patents (extensions)	23 683 577	34,0%
Artificials	12 108 074	17,4%
Total	**69 693 103**	**100,0%**

A **family** is composed by **first filing patents** (priority patent applications with no priority), and **extensions** applications (with patents mentioned as priorities).

A **singleton** is an application without any family.

First filing (priority patents) applications have the advantage of a date of filing closer to that of the invention (and less redundancy).

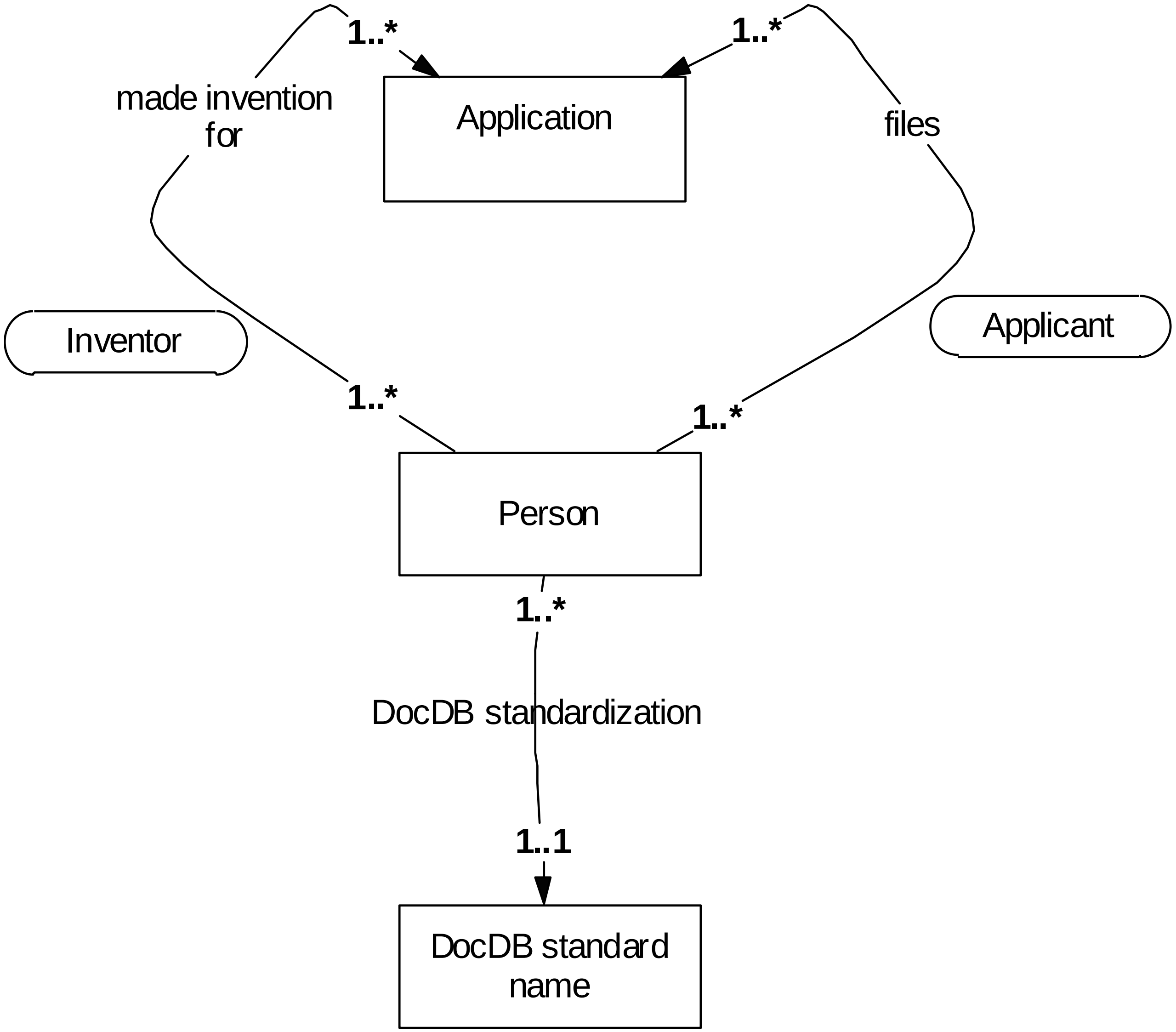
What are the main analytical dimensions

- Geographical : country codes of the applicants or inventors, addresses
- Institutional : patents portfolios of applicant’s names, collaborations (univ - firms)
- Technological through IPC codes

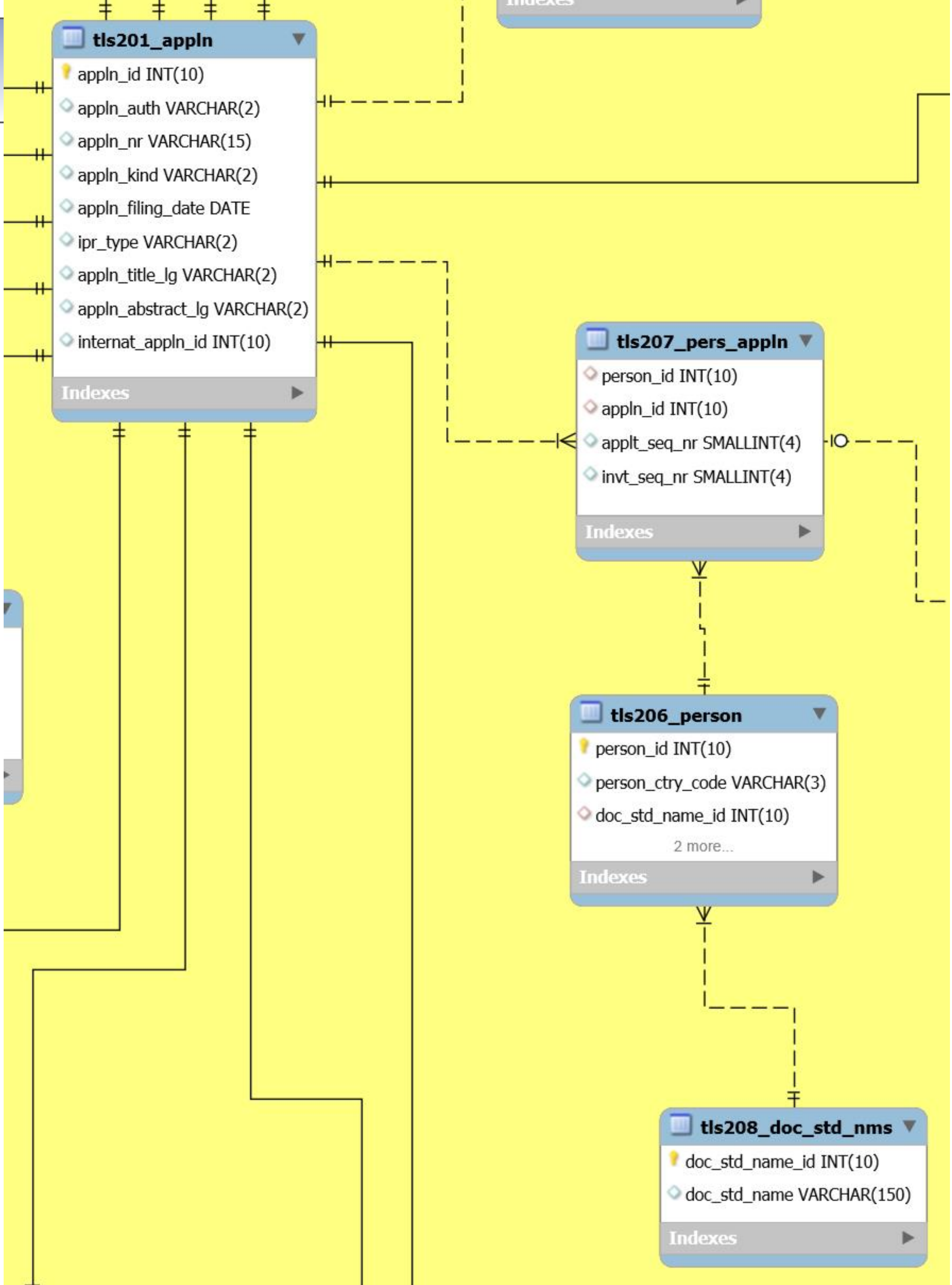
Other possibilties :

- Thematic caracterisation : textual analysis with titles and abstracts
- Intellectual proporty strategies of groups through patent families
- ...

How to identifying inventors and applicants for each application



One to many and one to one relations



tls201_appln
- appln_id INT(10)
- appln_auth VARCHAR(2)
- appln_nr VARCHAR(15)
- appln_kind VARCHAR(2)
- appln_filing_date DATE
- ipr_type VARCHAR(2)
- appln_title_lg VARCHAR(2)
- appln_abstract_lg VARCHAR(2)
- internat_appln_id INT(10)

Indexes

tls207_pers_appln
- person_id INT(10)
- appln_id INT(10)
- applt_seq_nr SMALLINT(4)
- invt_seq_nr SMALLINT(4)

Indexes

tls206_person
- person_id INT(10)
- person_ctry_code VARCHAR(3)
- doc_std_name_id INT(10)

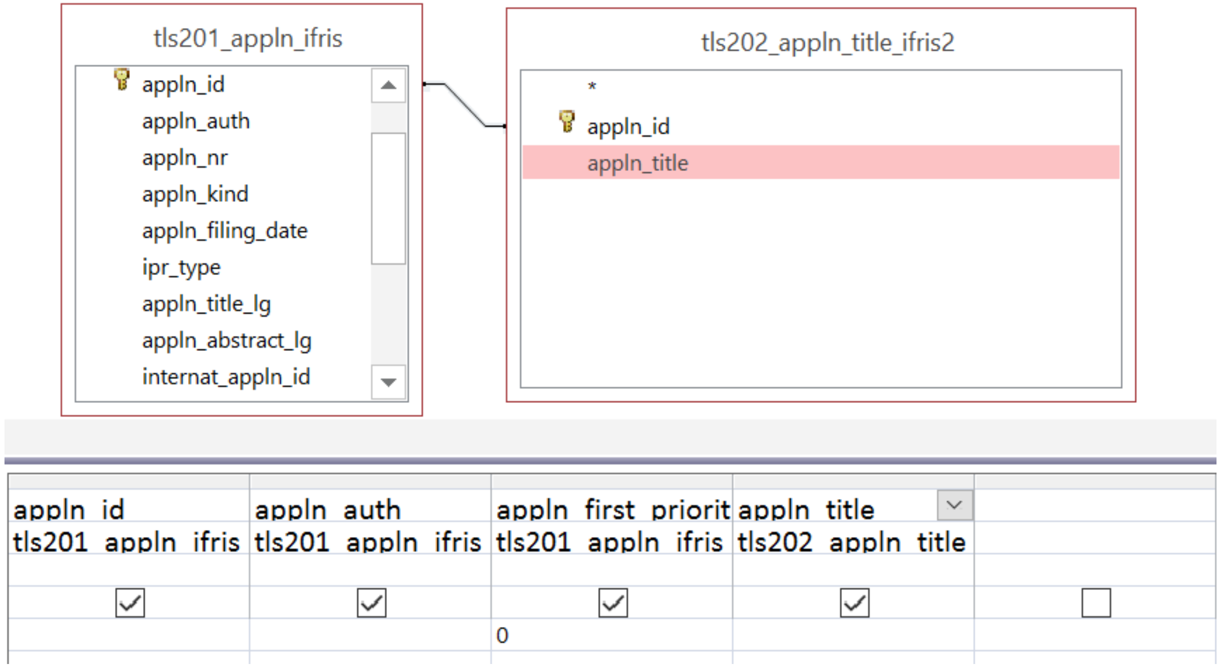
2 more...

Indexes

tls208_doc_std_nms
- doc_std_name_id INT(10)
- doc_std_name VARCHAR(150)

Indexes

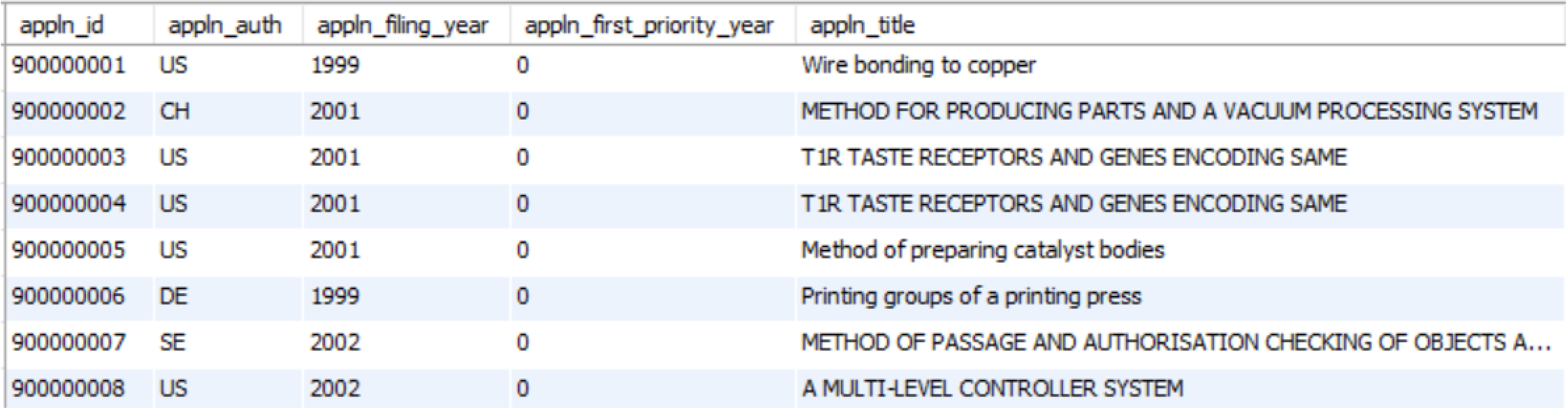
How to list all titles for each priority patents



appln_id	appln_auth	appln_first_priorit	appln_title	
tls201_appln_ifris	tls201_appln_ifris	tls201_appln_ifris	tls202_appln_title	
☑	☑	☑	☑	□
		0		

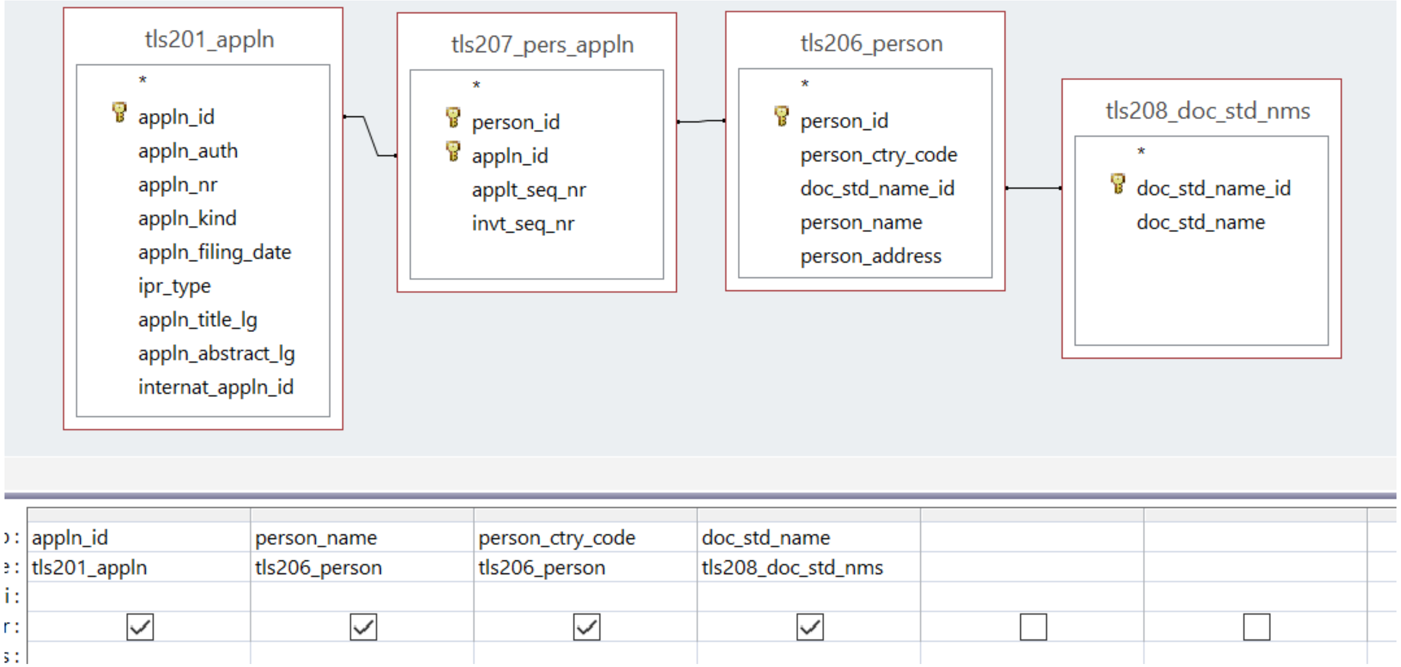
```sql
-- All titles for priority patents (with no first priority year mentioned)
USE patstatSept2011;

SELECT
    a.appln_id,
    a.appln_auth,
    a.appln_filing_year,
    a.appln_first_priority_year,
    b.appln_title
FROM
    tls201_appln_ifris AS a
        INNER JOIN
    tls202_appln_title_ifris2 AS b ON a.appln_id = b.appln_id
WHERE
    a.appln_first_priority_year = 0;
```



appln_id	appln_auth	appln_filing_year	appln_first_priority_year	appln_title
900000001	US	1999	0	Wire bonding to copper
900000002	CH	2001	0	METHOD FOR PRODUCING PARTS AND A VACUUM PROCESSING SYSTEM
900000003	US	2001	0	T1R TASTE RECEPTORS AND GENES ENCODING SAME
900000004	US	2001	0	T1R TASTE RECEPTORS AND GENES ENCODING SAME
900000005	US	2001	0	Method of preparing catalyst bodies
900000006	DE	1999	0	Printing groups of a printing press
900000007	SE	2002	0	METHOD OF PASSAGE AND AUTHORISATION CHECKING OF OBJECTS A...
900000008	US	2002	0	A MULTI-LEVEL CONTROLLER SYSTEM

How to identifying inventors and applicants for each application



tls201_appln
- *
- appln_id
- appln_auth
- appln_nr
- appln_kind
- appln_filing_date
- ipr_type
- appln_title_lg
- appln_abstract_lg
- internat_appln_id

tls207_pers_appln
- *
- person_id
- appln_id
- applt_seq_nr
- invt_seq_nr

tls206_person
- *
- person_id
- person_ctry_code
- doc_std_name_id
- person_name
- person_address

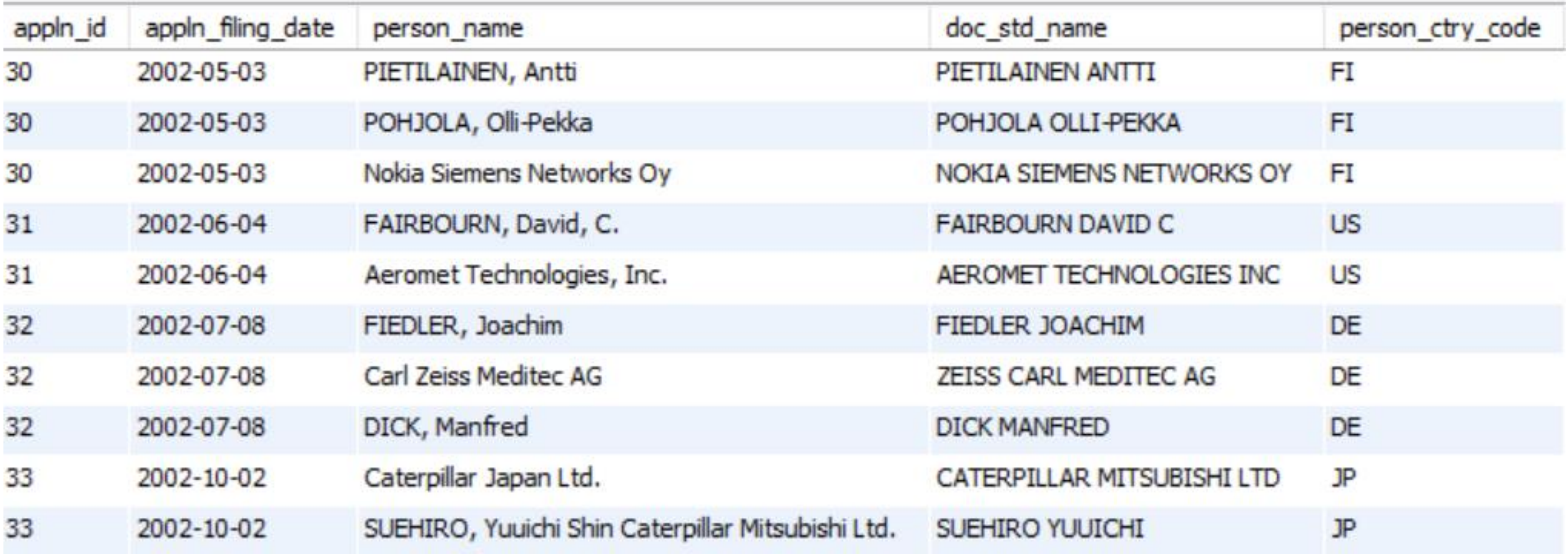
tls208_doc_std_nms
- *
- doc_std_name_id
- doc_std_name

:	appln_id	person_name	person_ctry_code	doc_std_name		
:	tls201_appln	tls206_person	tls206_person	tls208_doc_std_nms		
i:						
r:	☑	☑	☑	☑	□	□
:						

```
-- MySQL Query : all applicants and all inventors for each application

USE patstatSept2011;

SELECT
    a.appln_id, a.appln_filing_date, c.person_name, d.doc_std_name, c.person_ctry_code
FROM
    tls201_appln AS a
        INNER JOIN
    tls207_pers_appln AS b ON a.appln_id = b.appln_id
        INNER JOIN
    tls206_person AS c ON c.person_id = b.person_id
        INNER JOIN
    tls208_doc_std_nms AS d ON d.doc_std_name_id = c.doc_std_name_id
ORDER BY a.appln_id ASC;
```



appln_id	appln_filing_date	person_name	doc_std_name	person_ctry_code
30	2002-05-03	PIETILAINEN, Antti	PIETILAINEN ANTTI	FI
30	2002-05-03	POHJOLA, Olli-Pekka	POHJOLA OLLI-PEKKA	FI
30	2002-05-03	Nokia Siemens Networks Oy	NOKIA SIEMENS NETWORKS OY	FI
31	2002-06-04	FAIRBOURN, David, C.	FAIRBOURN DAVID C	US
31	2002-06-04	Aeromet Technologies, Inc.	AEROMET TECHNOLOGIES INC	US
32	2002-07-08	FIEDLER, Joachim	FIEDLER JOACHIM	DE
32	2002-07-08	Carl Zeiss Meditec AG	ZEISS CARL MEDITEC AG	DE
32	2002-07-08	DICK, Manfred	DICK MANFRED	DE
33	2002-10-02	Caterpillar Japan Ltd.	CATERPILLAR MITSUBISHI LTD	JP
33	2002-10-02	SUEHIRO, Yuuichi Shin Caterpillar Mitsubishi Ltd.	SUEHIRO YUUICHI	JP

General introduction

* Data sources
* Data coverage

Technical and conceptual frameworks

* What is a relational database ?
* Architecture server-client
* Data model patstat
* Conceptual model (application)

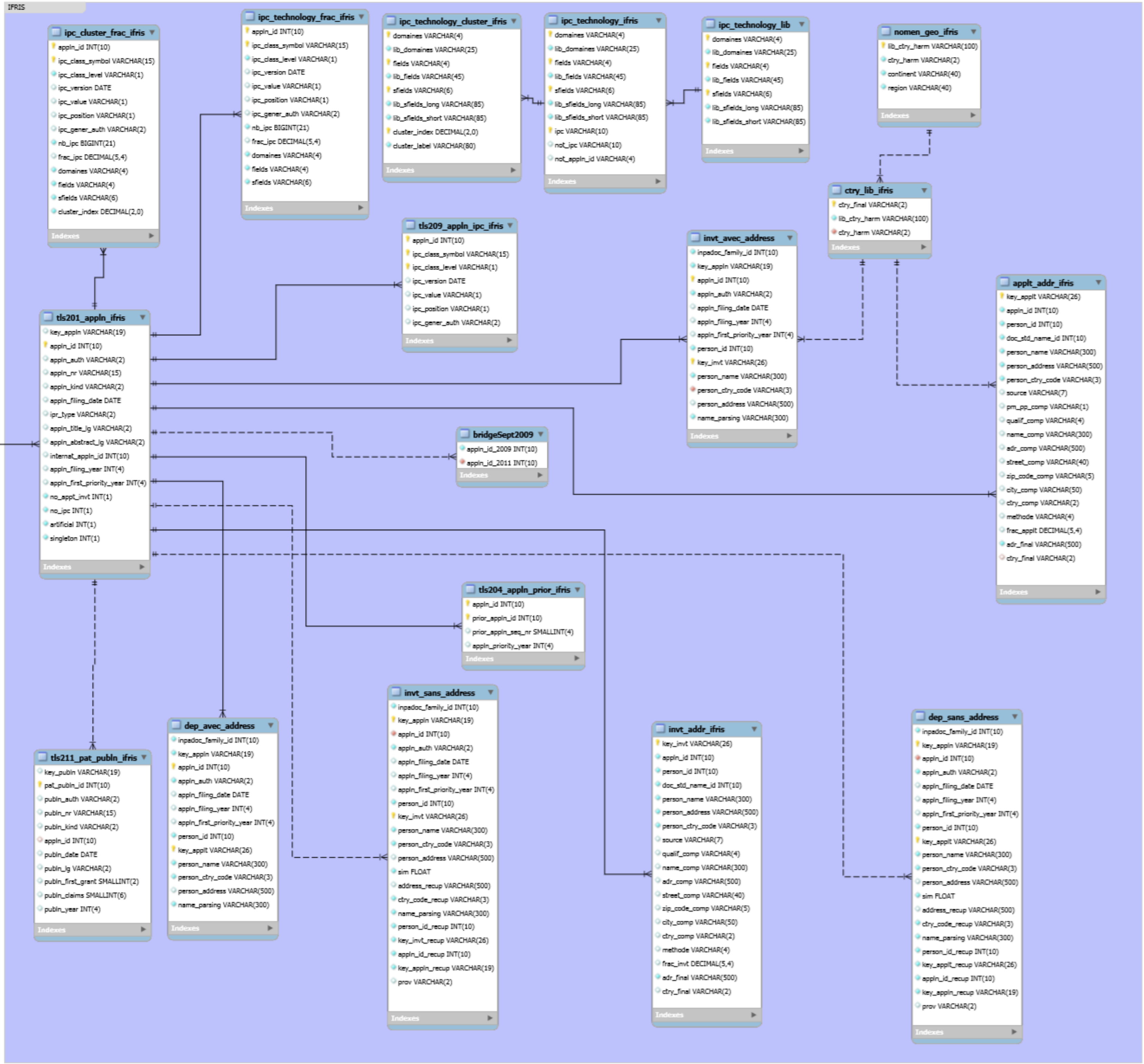
Attributes and tables

* Main different type of patents
* What are the main analytical dimensions ?
* Main tables and examples
* Focus on specific relations : how to catch inventor locations ?
* Live demo (sql queries) and results

What is Patstat IFRIS ?

* Cleaning country code and adding classification
* Problem of Silent :
    * State of the coverage of the addresses
    * enriching : regpat / inp
    * addresses propagation
    * artificial : what are they, how do we complete them
* How to characterized technology : IFRIS technology classification
* Some other attributes to facilitate the selection of patents
* Live demos (sql queries) and results

Open discussion and links with Risis datasets



Three types a new information:

- New tables (technology...)
- New attributes
- New values from external sources or propagations

Missing information : examples for inventors and applicants addresses

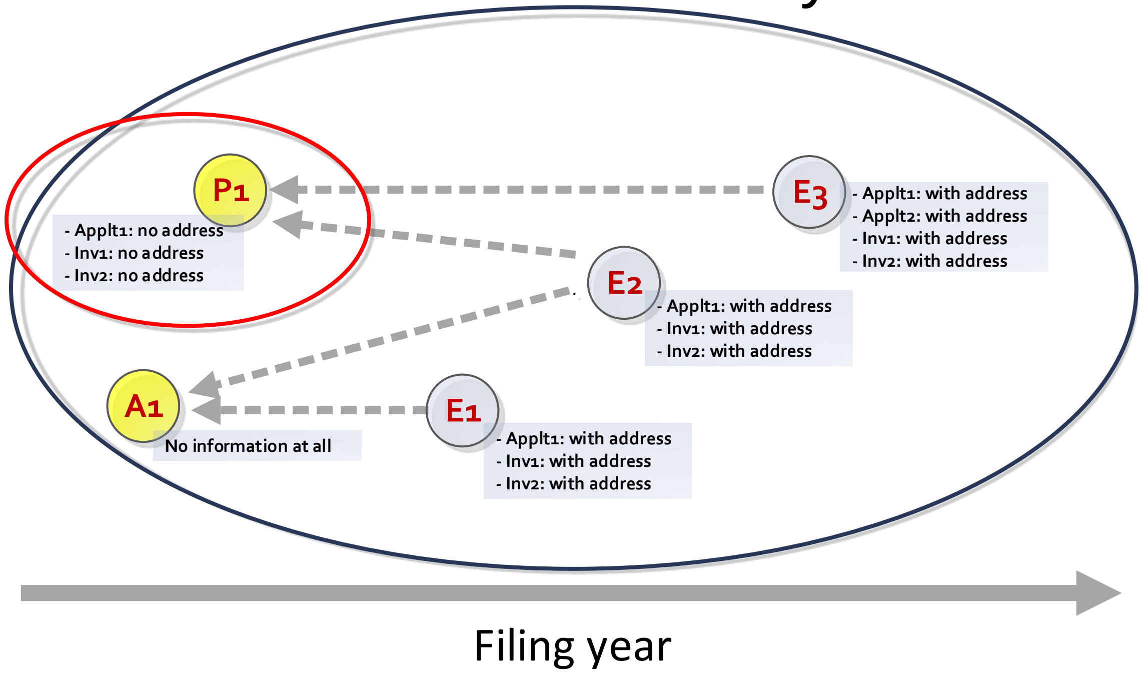
Problem: depending of the patent authority of the addresses are missing.

We had developed :

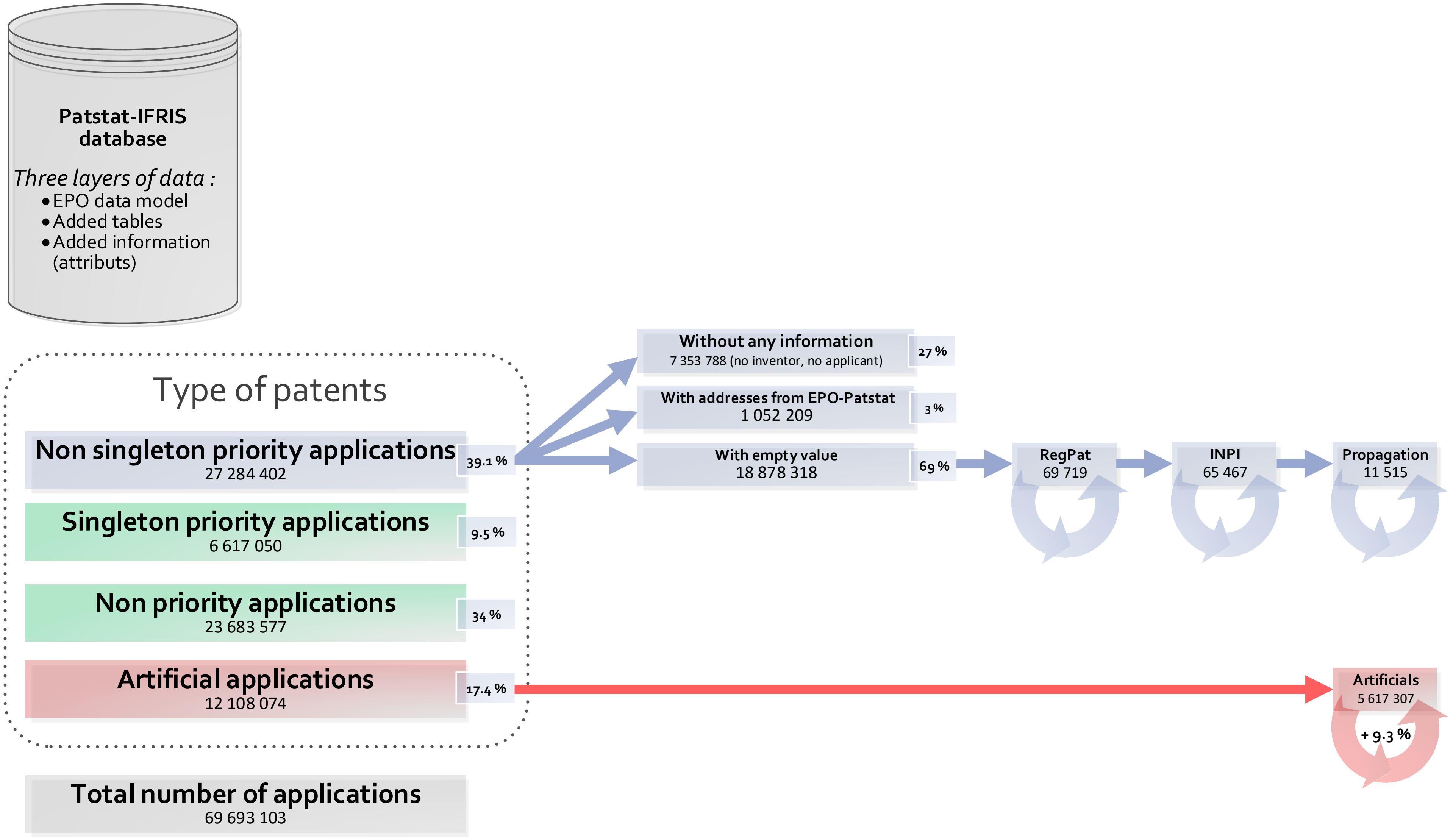
- a method to fill the missing addresses based on a string comparison of all the inventor and applicant names within INPADOC families
- Focus on priority patents

Propagation of the addresses through INPADOC families

INPADOC Familly



Filling of the missing addresses for the priority patent P1 base on the comparison of his inventor and applicant names with the other names accessible through the INPADOC family.



Patstat-IFRIS database

Three layers of data :

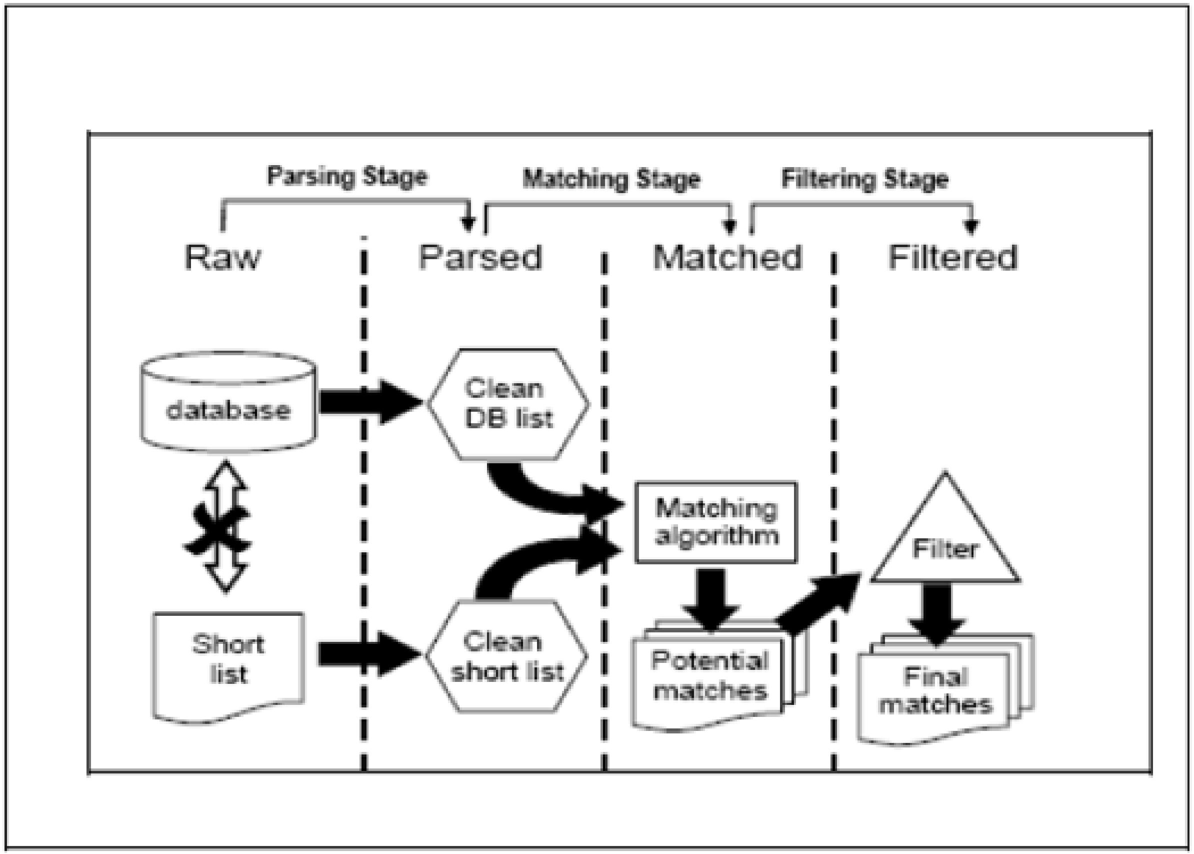
- EPO data model
- Added tables
- Added information (attributs)

Type of patents

Non singleton priority applications	27 284 402	39.1 %
Singleton priority applications	6 617 050	9.5 %
Non priority applications	23 683 577	34 %
Artificial applications	12 108 074	17.4 %
Total number of applications	69 693 103	

Without any information	7 353 788 (no inventor, no applicant)	27 %
With addresses from EPO-Patstat	1 052 209	3 %
With empty value	18 878 318	69 %

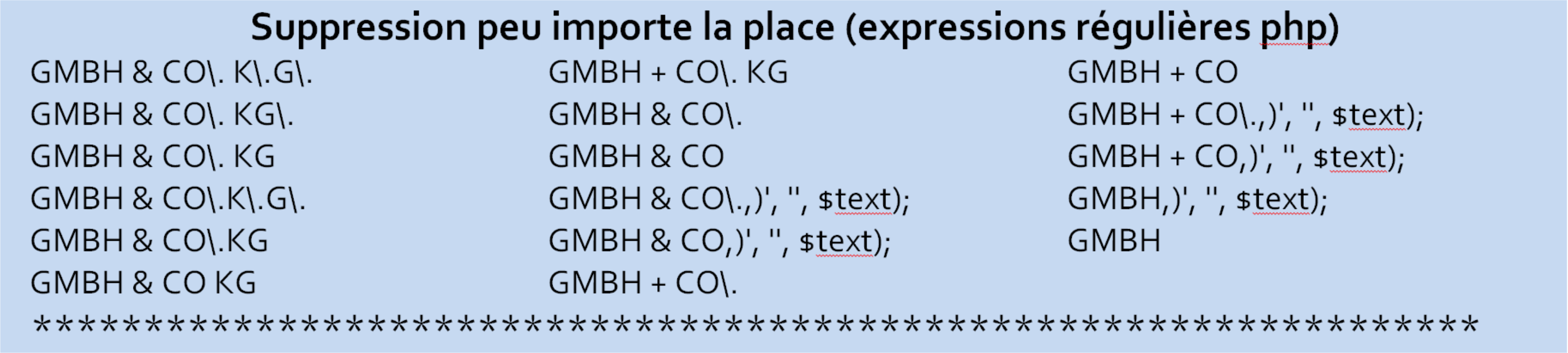
Nearly complete coverage for EPO, USPTO and FR patent authority



How to play the “Names Game”: Patent retrieval comparing different heuristics (Raffo et al., RP, 2009)

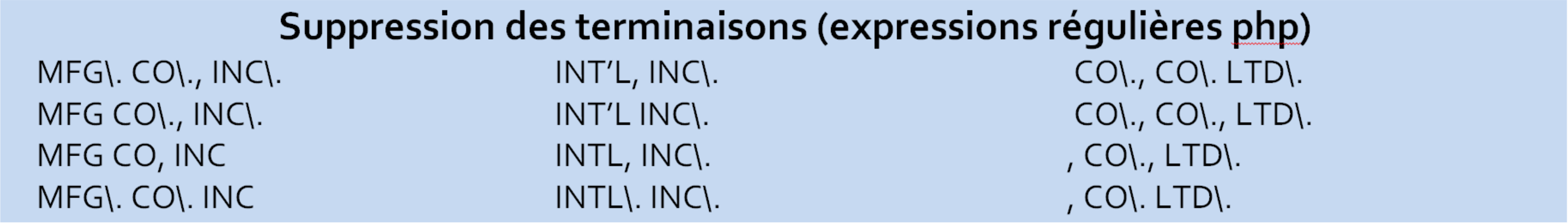
To be sure to have the best proximity score we are doing some pre-processing cleaning during the parsing step.

Parsing step (exemples of cleaning)
Magerman (2006)



Suppression peu importe la place (expressions régulières php)

```
GMBH & CO\. K\.G\.
GMBH & CO\. KG\.
GMBH & CO\. KG
GMBH & CO\.K\.G\.
GMBH & CO\.KG
GMBH & CO KG
GMBH + CO\. KG
GMBH & CO\.
GMBH & CO
GMBH & CO\.,)', '', $text);
GMBH & CO,)', '', $text);
GMBH + CO\.
GMBH + CO
GMBH + CO\.,)', '', $text);
GMBH + CO,)', '', $text);
GMBH,)', '', $text);
GMBH
*******************************************************************
```

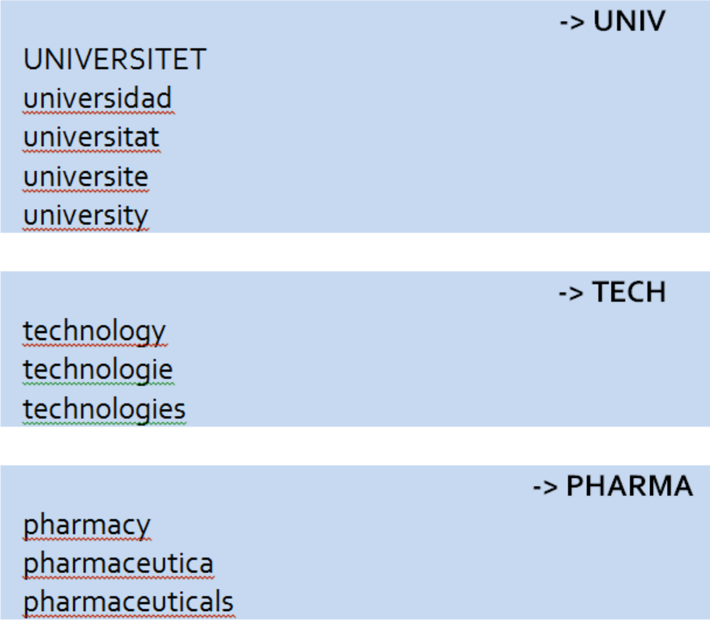


Suppression des terminaisons (expressions régulières php)

```
MFG\. CO\., INC\.
MFG CO\., INC\.
MFG CO, INC
MFG\. CO\. INC
INT’L, INC\.
INT’L INC\.
INTL, INC\.
INTL\. INC\.
CO\., CO\. LTD\.
CO\., CO\., LTD\.
, CO\., LTD\.
, CO\. LTD\.
```

More than 1 200 different rules for applicants

Parsing step (harmonising names)



-> UNIV

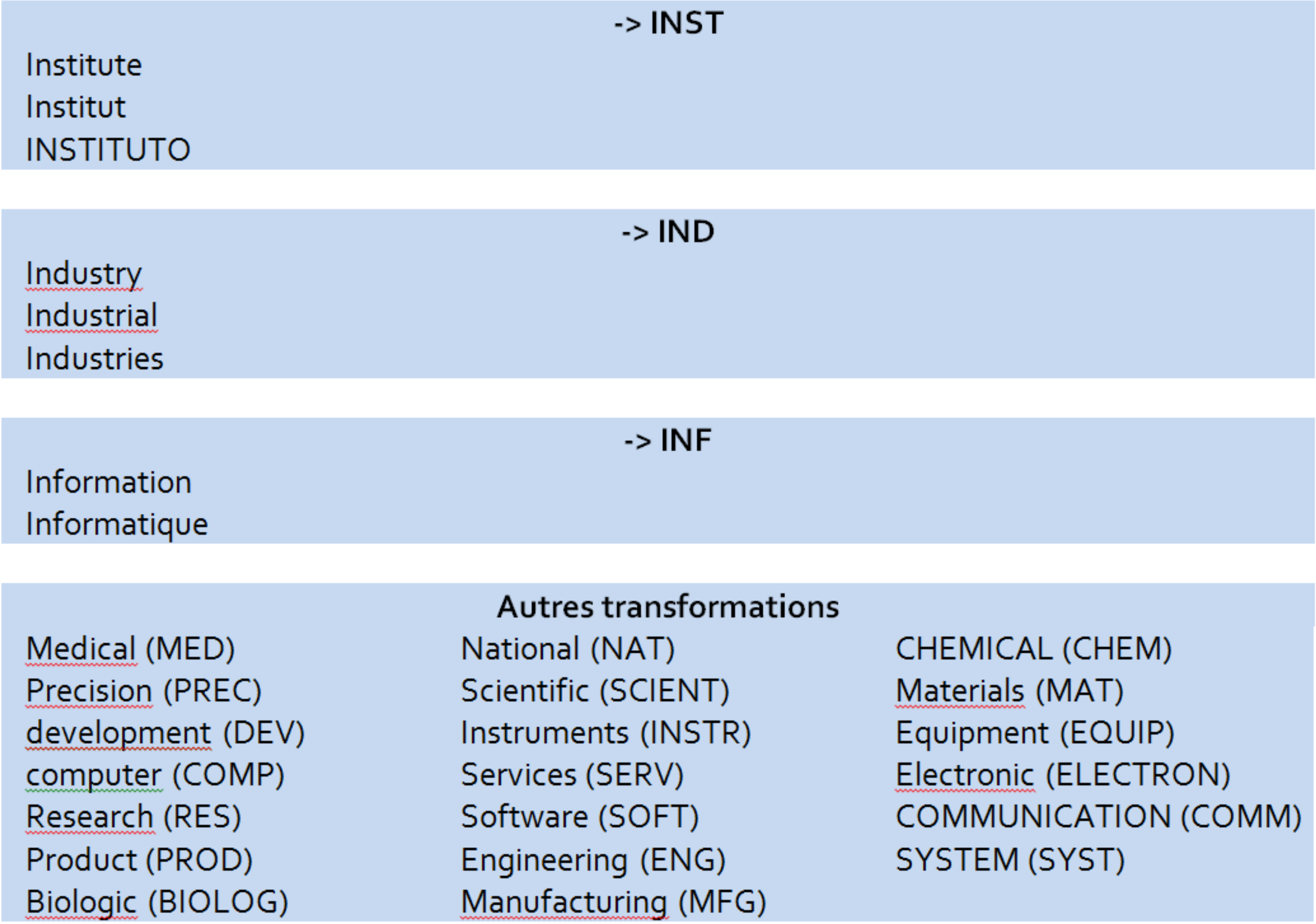
UNIVERSITET
universidad
universitat
universite
university

-> TECH

technology
technologie
technologies

-> PHARMA

pharmacy
pharmaceutica
pharmaceuticals



-> INST

Institute
Institut
INSTITUTO

-> IND

Industry
Industrial
Industries

-> INF

Information
Informatique

Autres transformations

Medical (MED)
Precision (PREC)
development (DEV)
computer (COMP)
Research (RES)
Product (PROD)
Biologic (BIOLOG)

National (NAT)
Scientific (SCIENT)
Instruments (INSTR)
Services (SERV)
Software (SOFT)
Engineering (ENG)
Manufacturing (MFG)

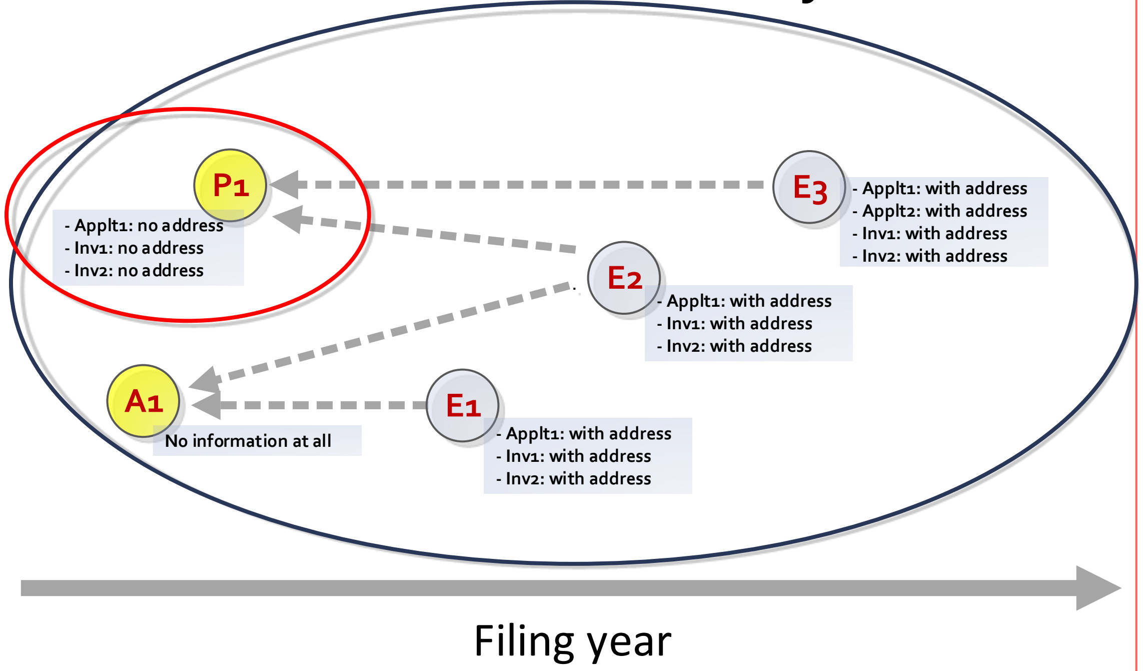
CHEMICAL (CHEM)
Materials (MAT)
Equipment (EQUIP)
Electronic (ELECTRON)
COMMUNICATION (COMM)
SYSTEM (SYST)

45 rules for applicants

665 geographical informations removed at the end of the string (country & continent)

Matching step

INPADOC Familly

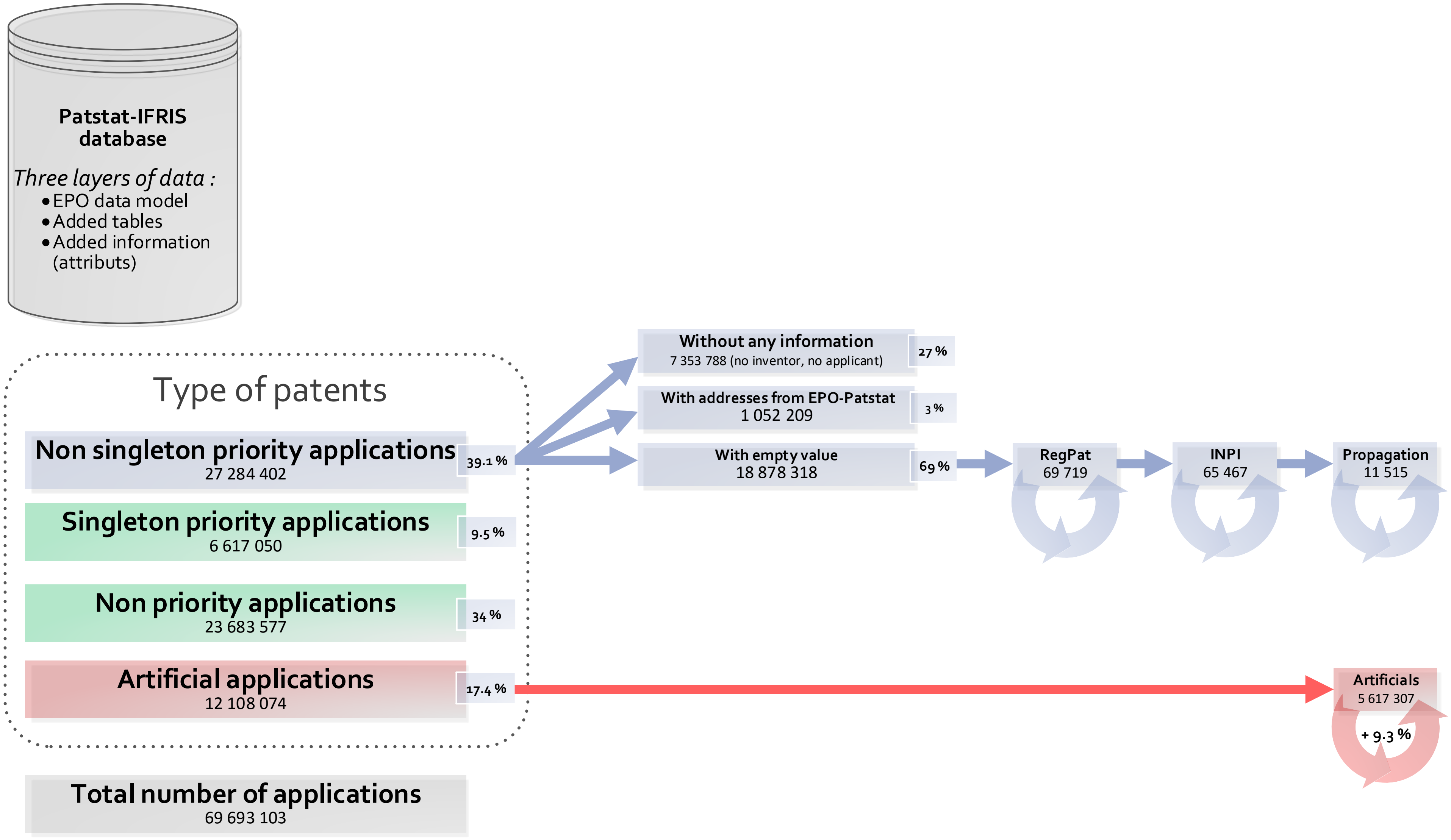


String similarity by using weighted Jaccard on bi-gram $$1-\frac{A \cap B}{A \cup B}$$

As there is also **inventor addresses on the applicant addresses list** (a physical person that has the intellectual property for its patent, is in the applicant list for this patent), we are comparing in a second time inventor name and applicant name.

Filtering step

Comparison of the year of filing to select the closest candidate patents with information to fill the priority patent, with a threshold (+- 5 years).



Patstat-IFRIS database

Three layers of data :

- EPO data model
- Added tables
- Added information (attributs)

Type of patents

Non singleton priority applications
27 284 402 — 39.1 %

Singleton priority applications
6 617 050 — 9.5 %

Non priority applications
23 683 577 — 34 %

Artificial applications
12 108 074 — 17.4 %

Total number of applications
69 693 103

Without any information
7 353 788 (no inventor, no applicant) — 27 %

With addresses from EPO-Patstat
1 052 209 — 3 %

With empty value
18 878 318 — 69 %

RegPat
69 719

INPI
65 467

Propagation
11 515

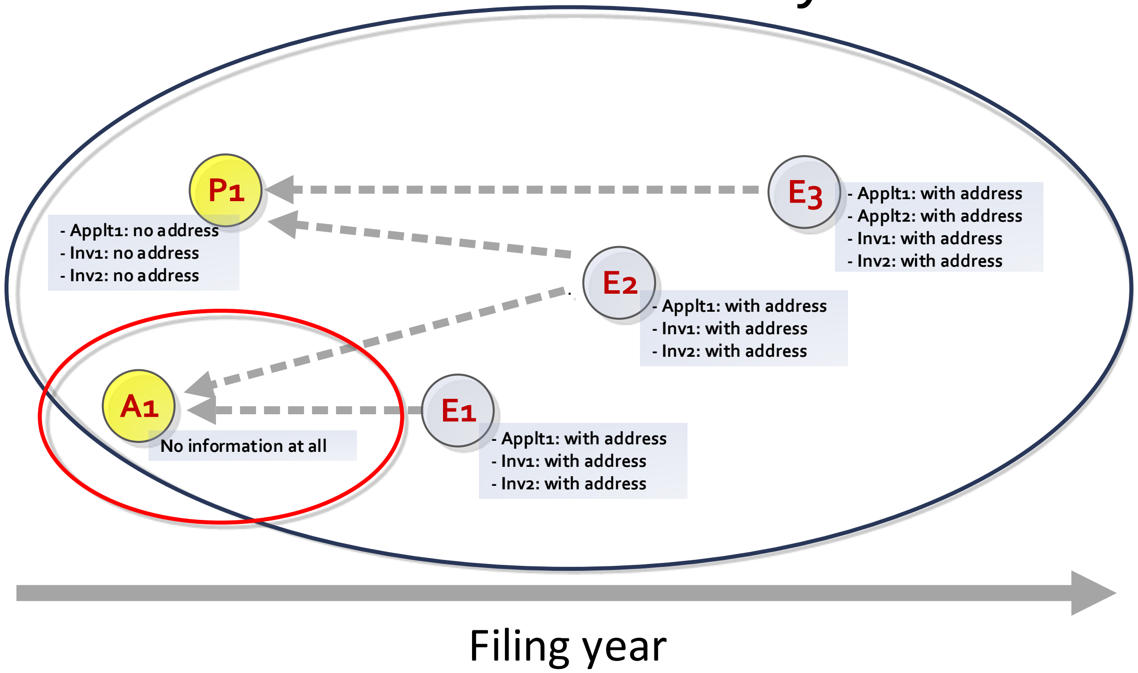
Artificials
5 617 307
+ 9.3 %

Why and how to use artificial patents ?

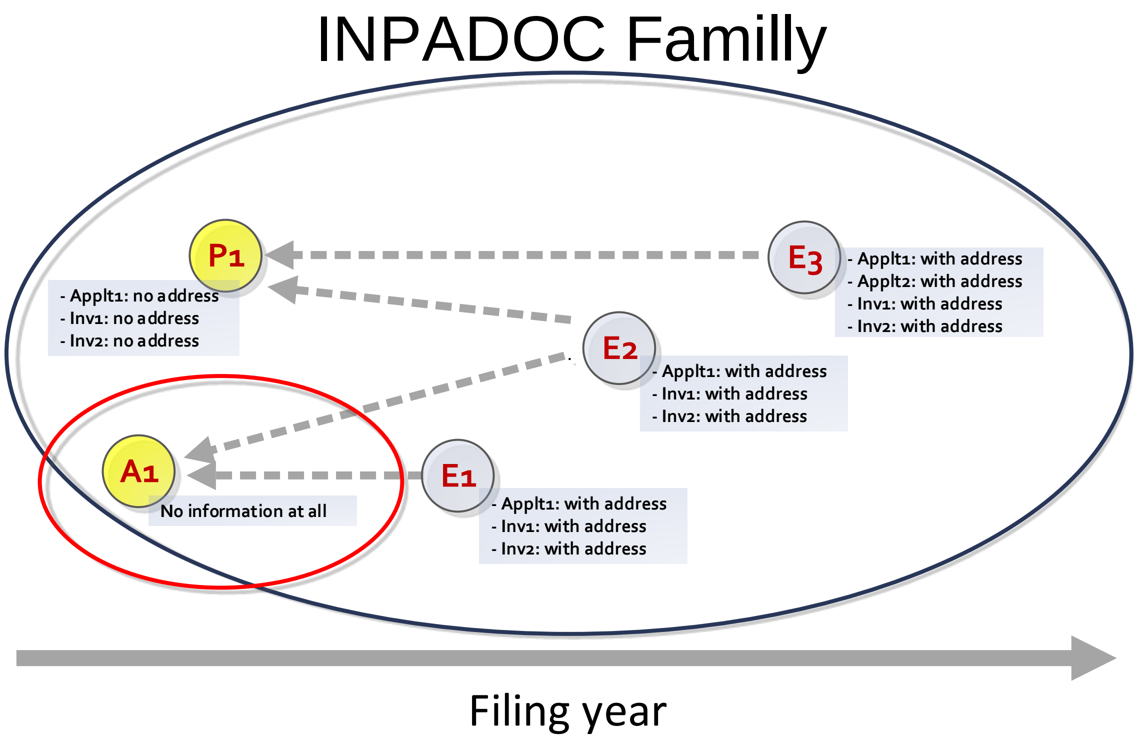
	Nb Applications	Code	Description
82,6%	57 585 029	**0**	non artificial
9,4%	6 571 987	**1**	patent mentioned by a priority with no corresponding filing in DocBD (provisionnal in US)
2,2%	1 550 321	**2**	patent mentioned in a citation in an other patent
5,7%	3 985 766	**3**	patent mentioned with a wrong filing date
	69 693 103		

Large amount of artificial : patents mentioned by other patents but not present in DocBD database.

INPADOC Familly



Filling of the missing addresses for the priority patent A1 based on the comparison of its inventor and applicant names with the other names accessibles trought the INPADOC family.

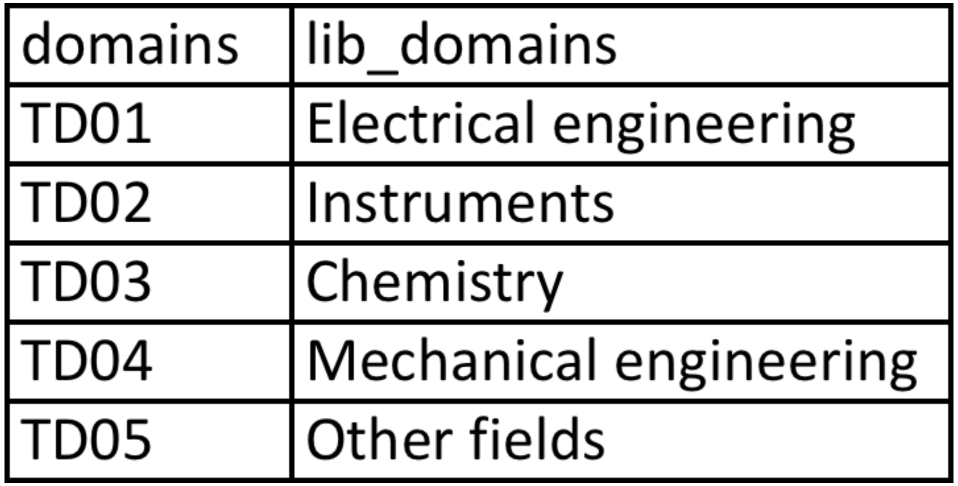


Selection of the candidates : prioritisation of direct links inside family (against indirect), and threshold on the filing date (the closest patent is chosen)

Building technological categories and subfields

- A systematic technology classification based on the codes of the International Patent Classification (IPC codes)

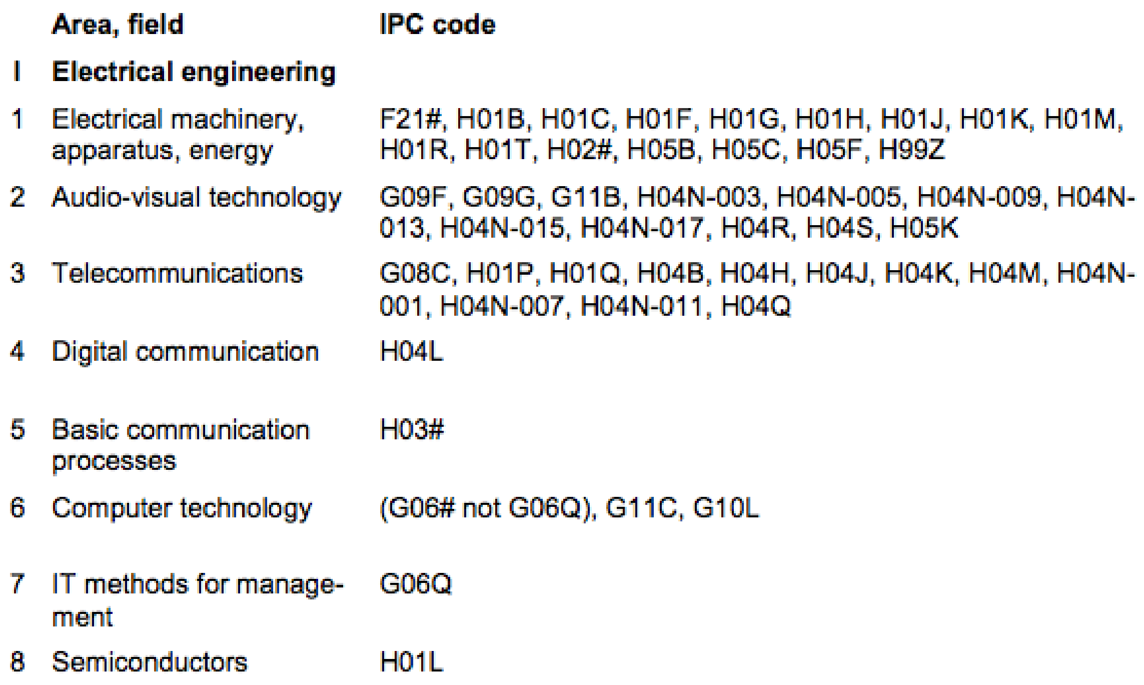
- Fractional (or integer) counting
  - 5 domains
  - 35 technological fields
  - 401 technological subfields



domains	lib_domains
TD01	Electrical engineering
TD02	Instruments
TD03	Chemistry
TD04	Mechanical engineering
TD05	Other fields

Patent Technology Classification

Ex: Domain TD01 Electrical engineering and its fields



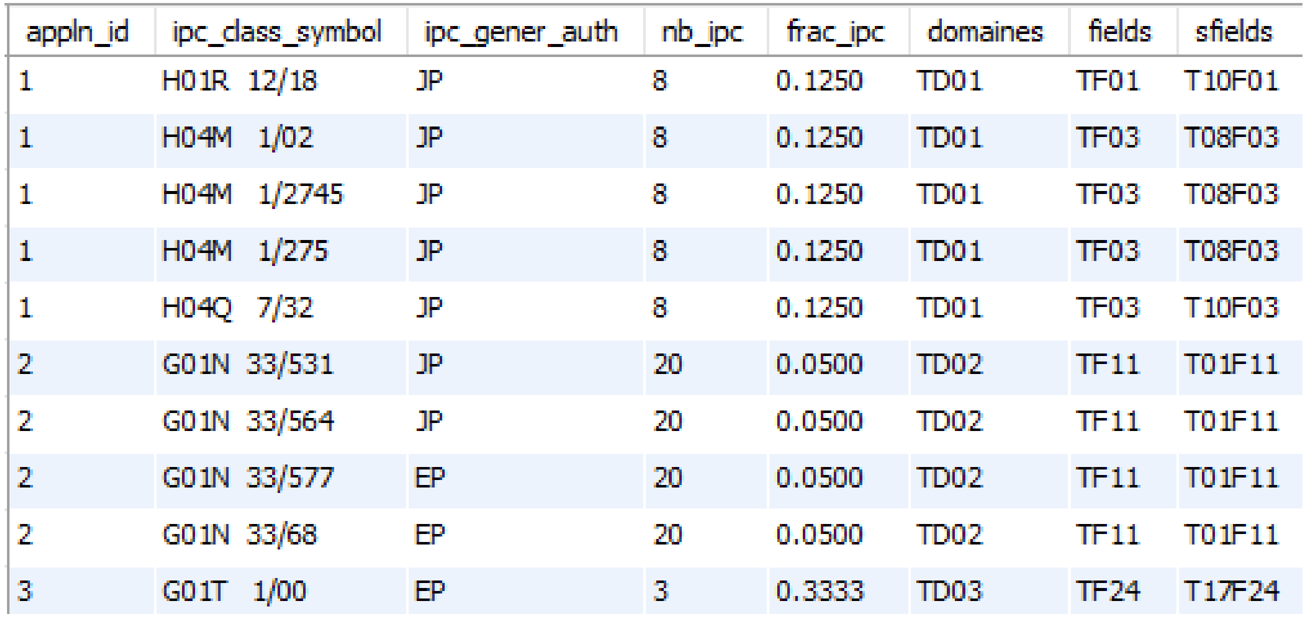
	Area, field	IPC code
I	**Electrical engineering**	
1	Electrical machinery, apparatus, energy	F21#, H01B, H01C, H01F, H01G, H01H, H01J, H01K, H01M, H01R, H01T, H02#, H05B, H05C, H05F, H99Z
2	Audio-visual technology	G09F, G09G, G11B, H04N-003, H04N-005, H04N-009, H04N-013, H04N-015, H04N-017, H04R, H04S, H05K
3	Telecommunications	G08C, H01P, H01Q, H04B, H04H, H04J, H04K, H04M, H04N-001, H04N-007, H04N-011, H04Q
4	Digital communication	H04L
5	Basic communication processes	H03#
6	Computer technology	(G06# not G06Q), G11C, G10L
7	IT methods for management	G06Q
8	Semiconductors	H01L

www.wipo.int/ipstats/.../pdf/wipo_ipc_technology.pdf

Breakdown of patents by fields (fract. Counting), top 10

Rank	fields	lib_fields	Priority patents	Total
1	TF29	Other special machines	56575470	111549810
2	TF01	Electrical machinery, apparatus, energy	54944640	101025296
3	TF02	Audio-visual technology	43646700	84425790
4	TF10	Measurement	42331880	81792380
5	TF23	Chemical engineering	35797524	88527124
6	TF19	Basic materials chemistry	35771970	110793202
7	TF34	Other consumer goods	35582400	59575050
8	TF35	Civil engineering	35185376	56135870
9	TF28	Textile and paper machines	32724956	69821356
10	TF26	Machine tools	23257770	42442260

```sql
-- Fractional counting and
-- technological fields
USE patstatSept2011;
SELECT
    appln_id,
    ipc_class_symbol,
    ipc_gener_auth,
    nb_ipc,
    frac_ipc,
    domaines,
    fields,
    sfields
FROM
    patstatSept2011.ipc_technology_frac_ifris;
```

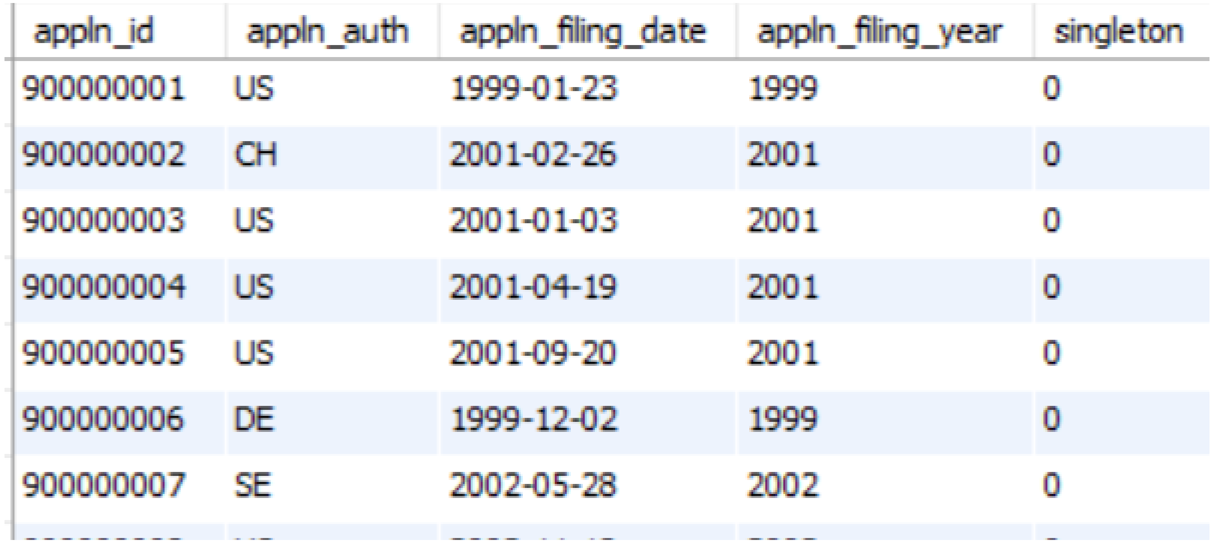


appln_id	ipc_class_symbol	ipc_gener_auth	nb_ipc	frac_ipc	domaines	fields	sfields
1	H01R 12/18	JP	8	0.1250	TD01	TF01	T10F01
1	H04M 1/02	JP	8	0.1250	TD01	TF03	T08F03
1	H04M 1/2745	JP	8	0.1250	TD01	TF03	T08F03
1	H04M 1/275	JP	8	0.1250	TD01	TF03	T08F03
1	H04Q 7/32	JP	8	0.1250	TD01	TF03	T10F03
2	G01N 33/531	JP	20	0.0500	TD02	TF11	T01F11
2	G01N 33/564	JP	20	0.0500	TD02	TF11	T01F11
2	G01N 33/577	EP	20	0.0500	TD02	TF11	T01F11
2	G01N 33/68	EP	20	0.0500	TD02	TF11	T01F11
3	G01T 1/00	EP	3	0.3333	TD03	TF24	T17F24

Examples useful attributes you can get in Patstat-IFRIS

Singleton : 0 to identifying directly the non singleton applications (demo)

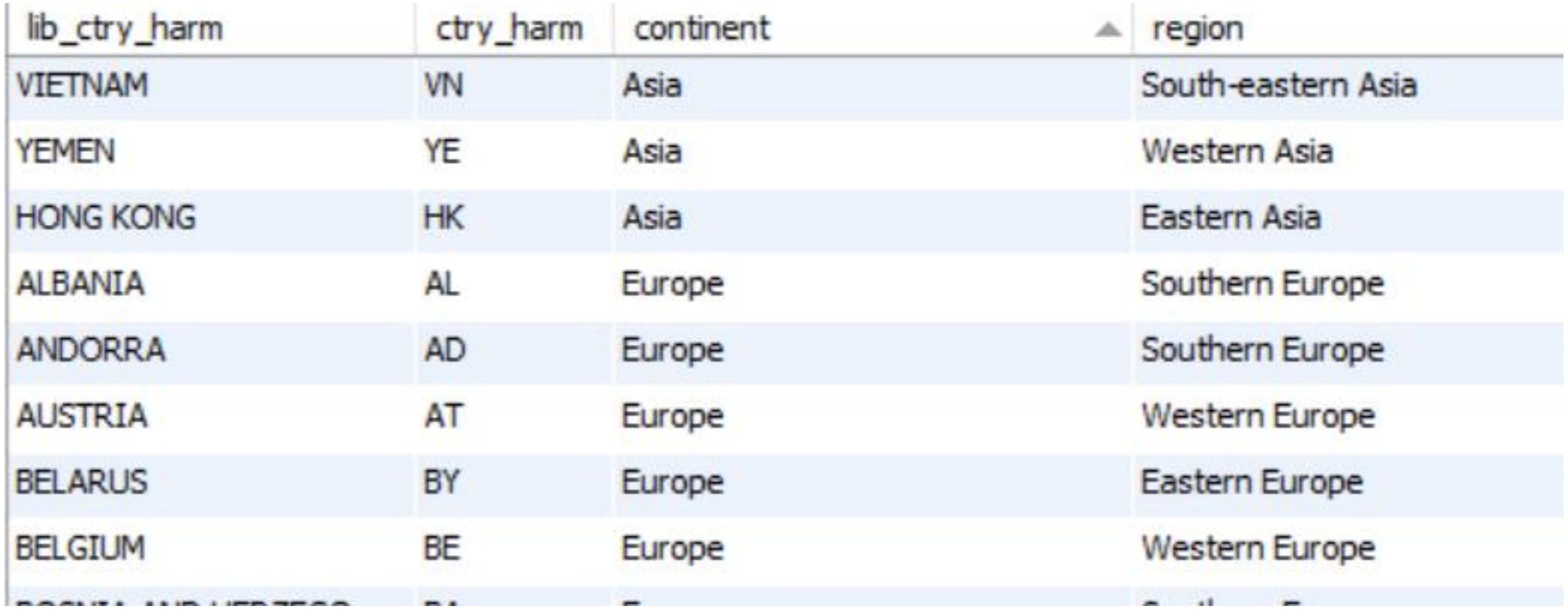
```
-- List of non singleton priority patents (filing date -> filing year)
USE patstatSept2011;
SELECT
    appln_id,
    appln_auth,
    appln_filing_date,
    appln_filing_year,
    singleton
FROM
    tls201_appln_ifris
WHERE
    singleton = 0
    AND appln_first_priority_year = 0;
```



appln_id	appln_auth	appln_filing_date	appln_filing_year	singleton
900000001	US	1999-01-23	1999	0
900000002	CH	2001-02-26	2001	0
900000003	US	2001-01-03	2001	0
900000004	US	2001-04-19	2001	0
900000005	US	2001-09-20	2001	0
900000006	DE	1999-12-02	1999	0
900000007	SE	2002-05-28	2002	0

Cleaned and harmonized patstat country codes (ISO Norme 3166_2) with the CIA Factbook continents and subcontinents classification (demo).

```
-- List of harmonized country codes and continent names
USE patstatSept2011;
SELECT
    *
FROM
    nomen_geo_ifris;
```



lib_ctry_harm	ctry_harm	continent	region
VIETNAM	VN	Asia	South-eastern Asia
YEMEN	YE	Asia	Western Asia
HONG KONG	HK	Asia	Eastern Asia
ALBANIA	AL	Europe	Southern Europe
ANDORRA	AD	Europe	Southern Europe
AUSTRIA	AT	Europe	Western Europe
BELARUS	BY	Europe	Eastern Europe
BELGIUM	BE	Europe	Western Europe

```
-- List of non singleton priority patents (filing date -> filing year) in France
USE patstatSept2011;
SELECT
    ctry_final, person_name, COUNT(frac_applt) AS NbPatents
FROM
    applt_addr_ifris_with_artif
WHERE ctry_final = 'FR'
GROUP BY ctry_final , person_name;
```