Europäisches Patentamt
European Patent Office
Office européen des brevets

Quality measures for bulk data products

Martin Kracker

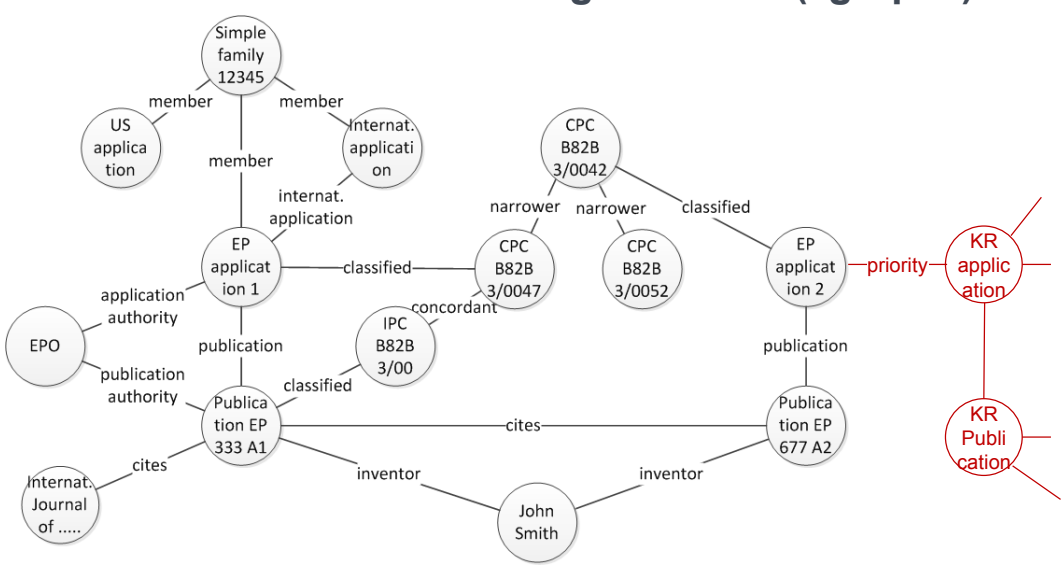
Agenda

- Data quality control for bulk data
  - Linked open EP data
  - PATSTAT data
- Completeness of data:
  Coverage map

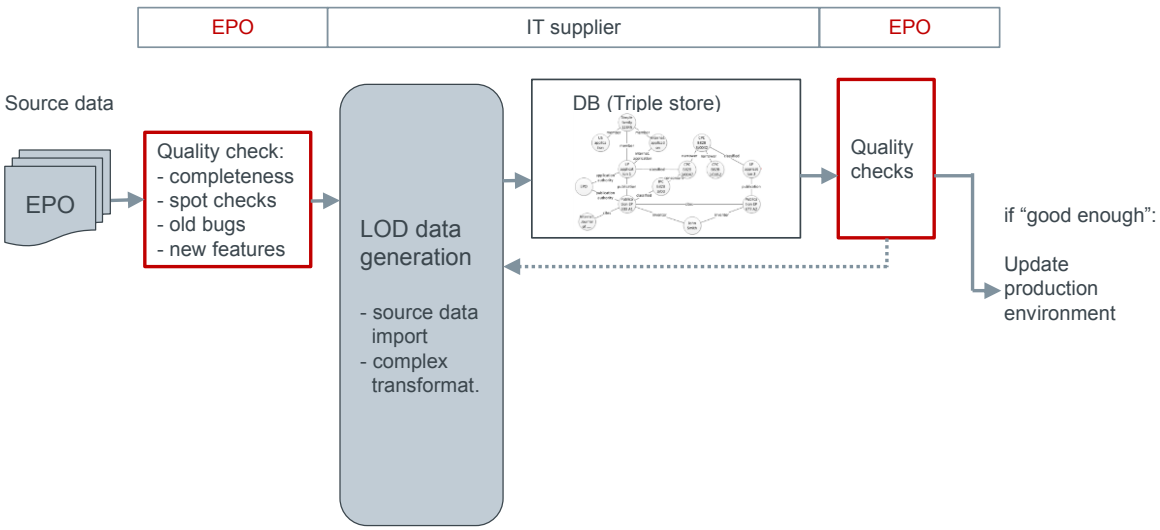
Linked open EP data

- Data product containing **EP bibliographic data** and **CPC scheme**
- New format: Linked data (aka Semantic Web) (RDF)
- Free-of-charge, open license, updated weekly
- Target user group:
  Patent non-experts, web developers, data scientists
- Launched: April 2018:
  epo.org/linked-data

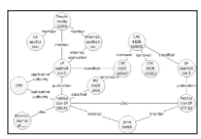
Linked data can be seen as a huge network ("graph")



Overview of *Linked open EP data* production process

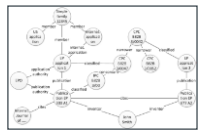


Testing each new version



- After changes of transformation scripts:
  Run ≈ 130 SPARQL test scripts (each testing one or more constraints) to identify deviations between data and specification
  - Plausibility of number of resources by class
  - Existence of mandatory properties
  - Absence of non-defined properties
  - Single-values properties must not occur multiple times
  - Domain and range of relation properties
  - Data formats and business rules like "All granted EP publications must have a B1 publication"

Weekly testing

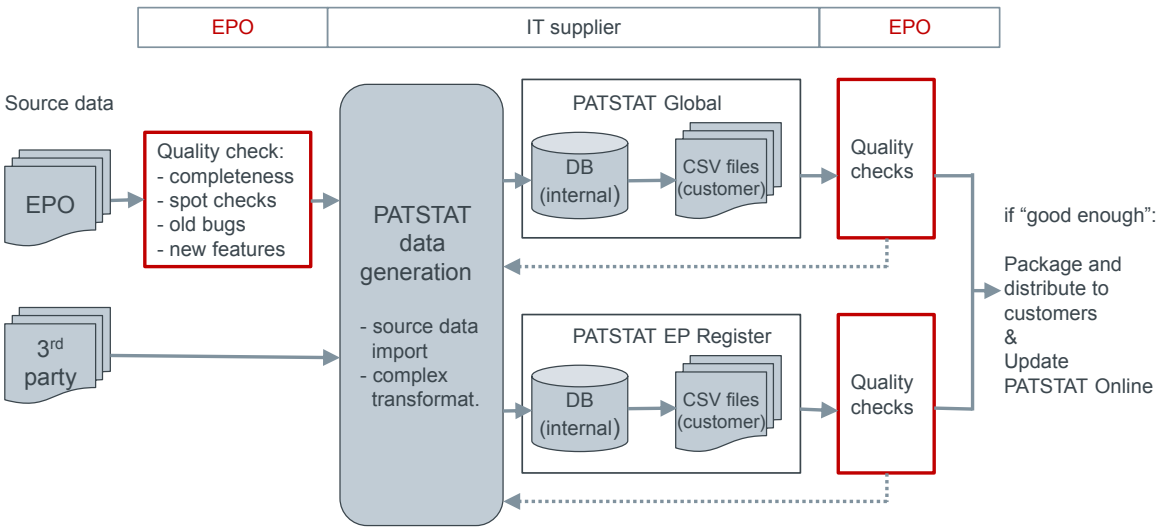


- Weekly tests after each publication day:
  Manual check whether number of
  - Compare number of newly published of A1, A2, ..B9 publications with official number in official European Publication Server
  - Detailed spot checks for selected publications

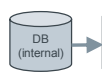
Agenda

- Data quality control for bulk data
  - Linked open EP data
  - PATSTAT data
- Completeness of data:
  Coverage map

Overview of *PATSTAT* production process



Testing data in PATSTAT DB

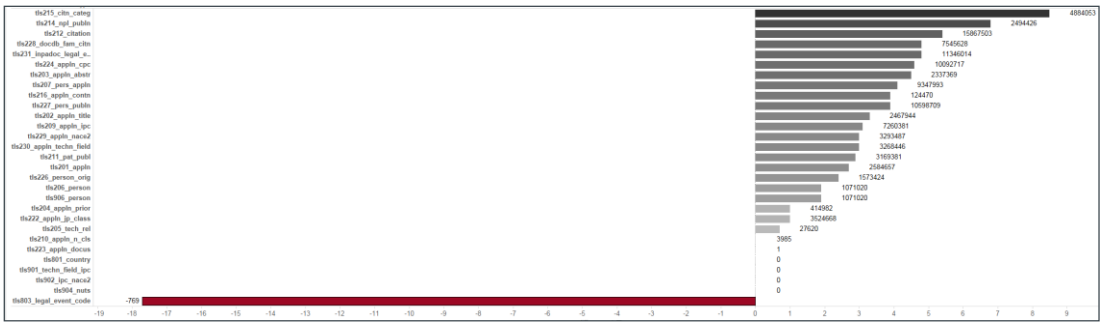


- Automated Test:
  Run ≈ 1 550 SQL test scripts to identify deviations between DB content and specification (Data Catalog)
  - referential integrity, cardinalities, ...
  - domains of attributes, e.g. country codes
  - re-compute and compare derived attributes, e.g. INPADOC family
  - plausibility of data content, business rules
  - monitor attributes which are empty or have constant values

- Manual tests
  - whether bugs of the previous versions have been fixed
  - analyse logged errors & warnings delivered by IT supplier

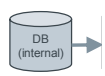
Testing completeness of PATSTAT data (1)

- Visually checking growth of each table



Difference (%) of number of rows of each table, comparing current and previous version

Testing completeness of PATSTAT data (2)



- Visually comparing the number of instances of key data items (application, publications, IPC, CPC, applicants, legal events, ..) from the current to the previous database using Tableau charts.

auth	9999	2912	2019	2018	2017	2016	2015	2014	2013	2012	2011	2010	2009	2008	2007
DD	0														
DE	0			368000	80	41	1	0	0	0	0	0	0	0	
DK	0			1250	77	56	26	25	15	10	7	5	4	3	
DL	0														
DM	0														
DO	0			4500	37	3	0	0	0	0	0	0	0	0	
DZ	0					100	0	0	0	0	0	0	0	0	
EA	6				157	135	4	0	1	0	0	0	0	0	
EC	20					100	0	0	0	0	0	0	0	0	
EE	4				250	25	-11	-9	-4	0	-2	-1	-1	0	
EG	11					77	5	0	0	0	0	0	0	0	
EM	44			100	816	57	14	2	1				0	0	
EP	-2			89900	107	83	4	1	1				0	0	
ER	0														
ES	0			6340	16	29	30	27	17				4	3	
ET	0														
FI	0			9700	62	33	1	0	0				0	0	
FR	0			2800	187	66	0	0	0				0	0	
GA	0												-100		
GB	0			753300	2	9	2	0	0	0	0	0	0	0	

auth: **EE**
year: **2013**
diff_percent: **-4**
diff: **-4**
db_old: **92**
db_new: **88**

Difference (%) of number of selected entities, comparing current and previous version

Testing PATSTAT's CSV files

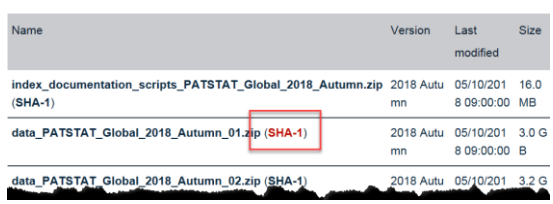
- In total: ≈ 270 GB in ≈ 200 files

Applied checks:

- Are the file names as specified?
- Is the header of the CSV file as specified?
  (spelling, missing fields, order of sequence)
- Character level checks
  - UTF8 conformance
  - invalid XML characters (control characters)
  - consistent number of string delimiters (double quotes)
  - no <back slash><double quote> \"

Test recommended for subscribers

- To detect data corruption during data transmission from Download Area:
  →**Compare SHA-1 values**



Name	Version	Last modified	Size
index_documentation_scripts_PATSTAT_Global_2018_Autumn.zip (SHA-1)	2018 Autumn	05/10/2018 09:00:00	16.0 MB
data_PATSTAT_Global_2018_Autumn_01.zip (SHA-1)	2018 Autumn	05/10/2018 09:00:00	3.0 GB
data_PATSTAT_Global_2018_Autumn_02.zip (SHA-1)	2018 Autu	05/10/201	3.2 G

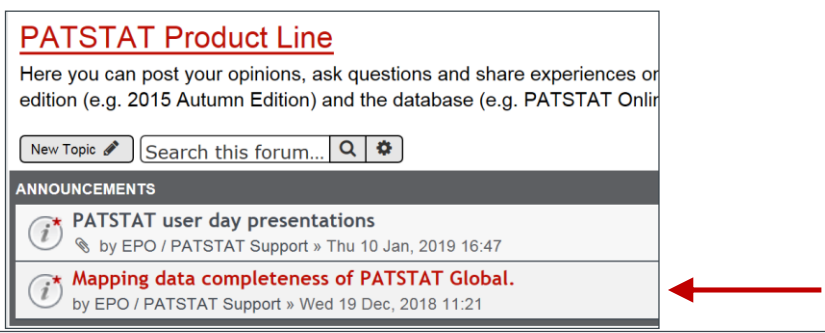
- Check completeness of your data loading
  →**Check logs of your ETL tool**
  →**Apply row-counting script** and compare to the reference figures (see documentation folder in the delivery package)

Agenda

- Data quality control for bulk data
  - Linked open EP data
  - PATSTAT data
- Completeness of data:
  Coverage map

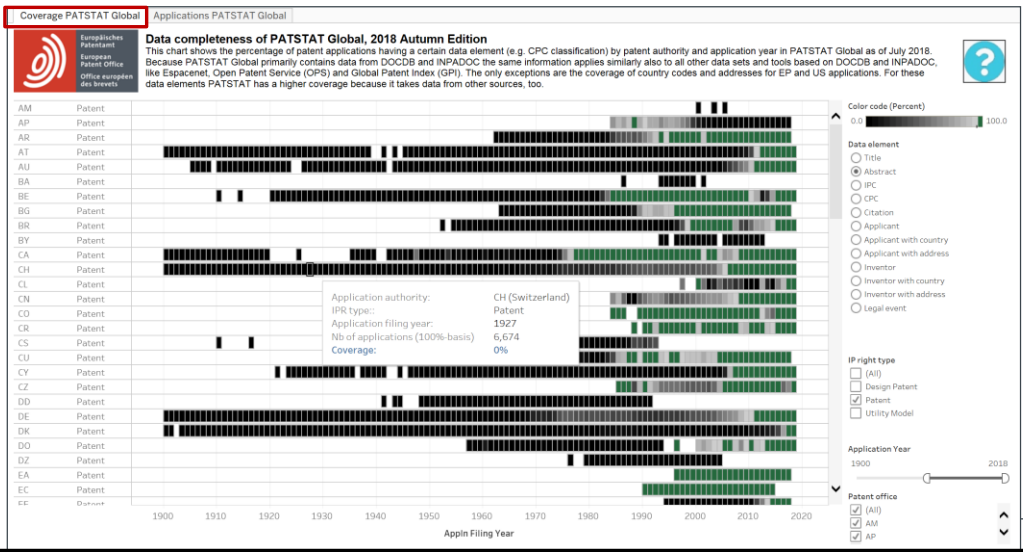
Coverage Map for PATSTAT Global / DOCDB (1)

- Provides insight into amount and relative completeness of existing data
- Interactive chart
- Accessible from PATSTAT Forum;
  https://forums.epo.org/mapping-data-completeness-of-patstat-global-7984



Coverage Map for PATSTAT Global / DOCDB (2)

https://public.tableau.com/profile/patstat.support#!/vizhome/CoverageofPATSTAT2018AutumnEdition/CoveragePATSTATGlobal



20

Thank your for your attention

Martin Kracker
European Patent Office
Vienna

mkracker@epo.org