Authority Files & Data Quality

The Data Consumer Perspective

Committee on WIPO Standards (CWS) : Tenth Session

Magdalena Zelenkovska, Senior Patent Data Manager
Patent Database Section, Global Databases Division
Infrastructure and Platforms Sector

Geneva, November 23, 2022

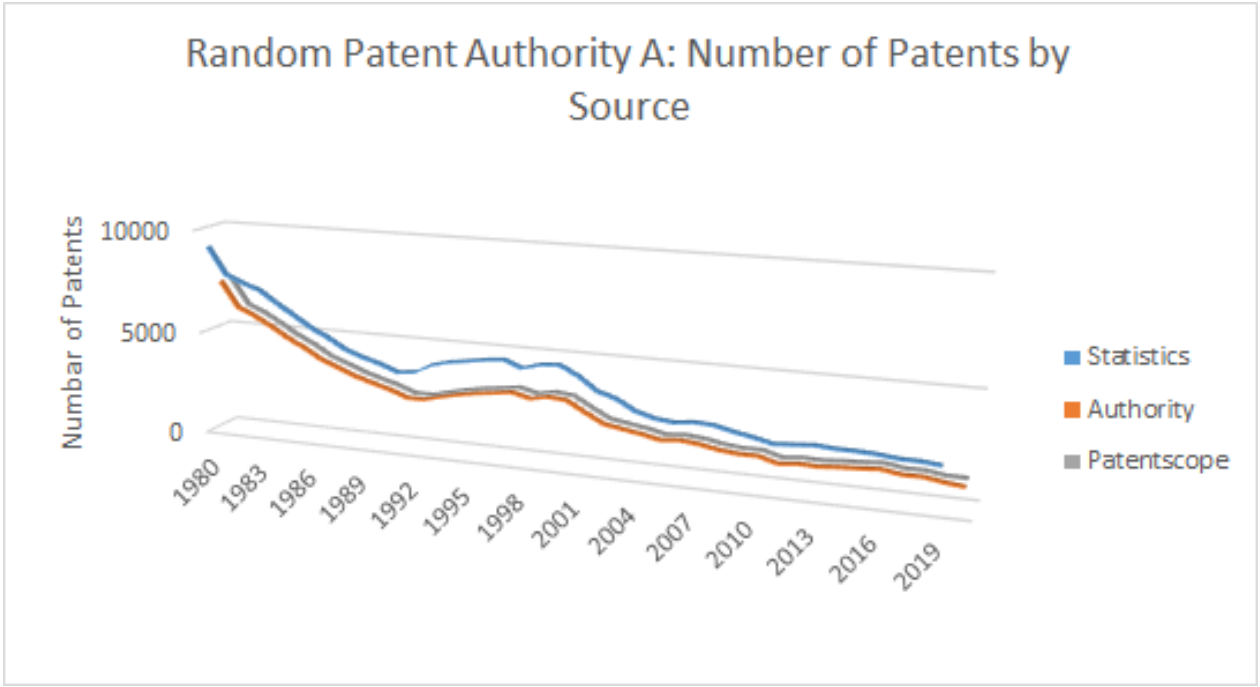
Authority Files as a Data Quality Tool

- Authority File Use Cases
  - Online Search Tools like PATENTSCOPE
  - IP Office Digitization Projects
  - PCT Minimum Documentation
- Quality Data
  - Data Fit for Purpose
  - Context dependent
  - Meets the needs of data consumer

Authority Files and The Single Version of Truth

- WIPO Patent Data Sources (raw and aggregate)
  - WIPO Statistics Data Center: https://www3.wipo.int/ipstats/
  - Authority File Portal: https://www.wipo.int/standards/en/authority_file.html
  - PATENTSCOPE: https://patentscope.wipo.int/
- Aggregate and Raw Data as received by offices

Authority Files and The Single Version of Truth



Authority Files – a Tool to Improve Accuracy

- ***Accuracy is a measure of the degree to which the data matches the real-life entity.***

- The Authority File can be used to measure and improve the accuracy of our patent collections such as publication and application numbers or kind codes

- ***The Definition File*** – optional, but crucial in ensuring accuracy (paragraph 36 of WIPO ST.37)

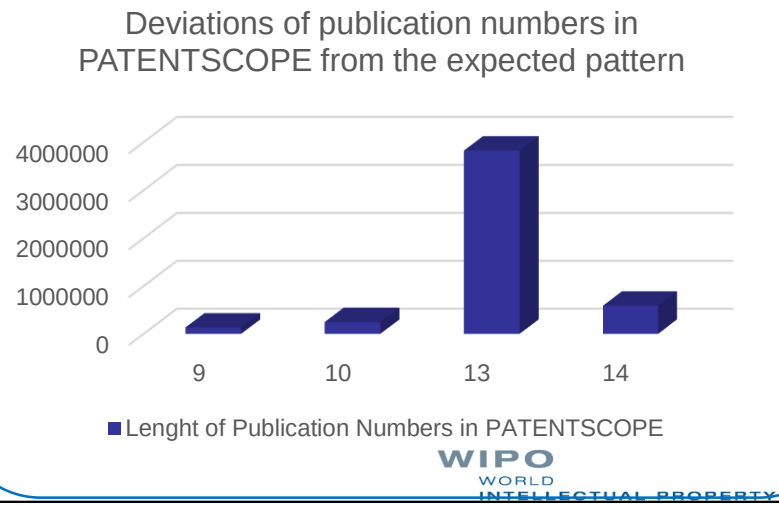
< (e.g.) Numbering System for Unexamined Publication & Publication of Application >

IP Right (2digits) Patent (10), Utility Model (20)		Year (4digits)				Sequence Number (7digits)						
1	0	1	9	8	3	0	0	0	0	0	0	1

< (e.g.) Numbering System for Examined Publication >

IP Right (2digits) Patent (10), Utility Model (20)		Sequence Number (11digits)										
1	0	0	1	1	7	6	9	0	0	0	0	0

Description of KIPO's Authority File: Numbering System



Authority Files – A Tool to Enforce Integrity

- ***Data Integrity or Coherence is associated to completeness, accuracy and consistency.***

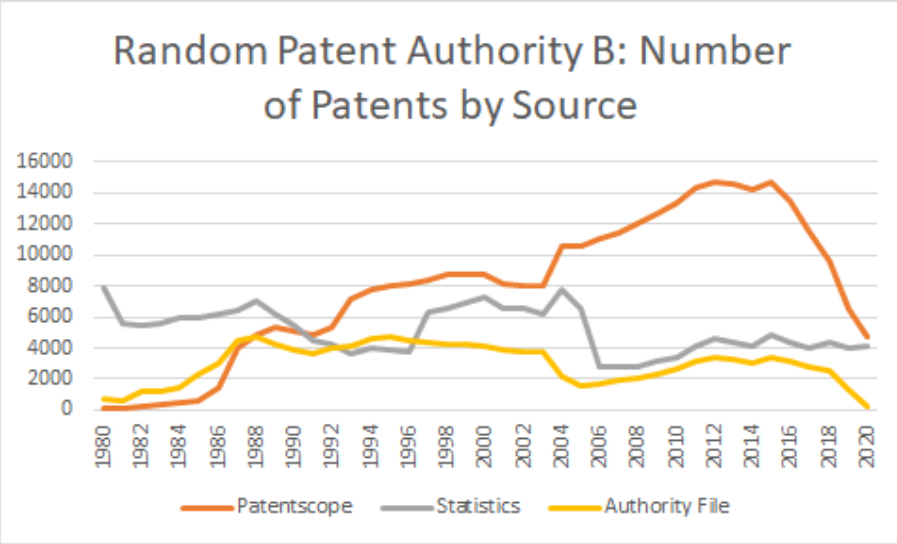
- Data Integrity can refer both to referential integrity and to internal consistency (no holes in the data)

- Putting authority files in the center and mapping different sources to them enforces integrity automatically

Authority Files – A Tool to Measure Completeness

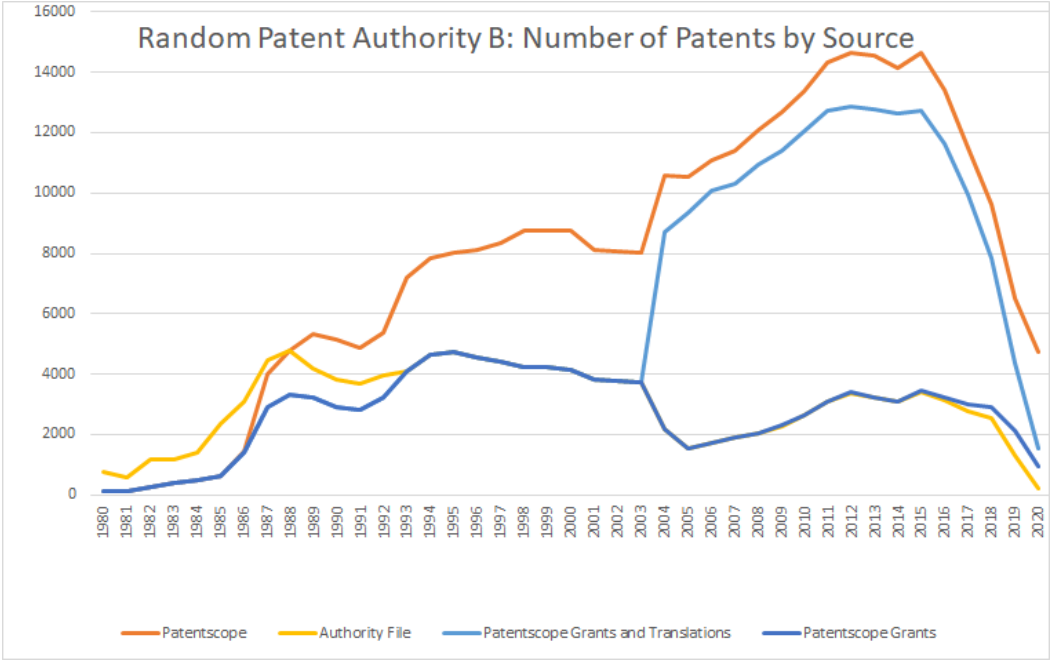
- ***A measure for the presence (or absence) of all mandatory data***

- Measured at data set, record and column level
- Authority files
  - designed to be the ultimate source of benchmark data
  - expected to match gazette data
  - should be a superset of any other source of the same data set



Authority Files – A Tool to Measure Completeness

- Dataset Level Completeness
- PATENTSCOPE Kind Codes: A,B,T,U,Y
- Authority File Kind Codes B,T,Y
- No Definition File available
- Conclusion: unusual and sudden increase of numbers due to translations; authority file incomplete as it contains only a subset of all publications
- ST.37 Paragraph 3: «all publication numbers assigned by the IP office"



Complete Authority Files – record and column level

- ST. 37 Paragraph 8 defines the minimum data elements to uniquely identify a patent document: publication information

- ST. 37 Paragraph 9 defines the optional elements – record level completeness
  - Exception codes – very useful to make the line graphs above match perfectly
  - Priority Applications – data enrichment
  - Application identification – extremely useful for identifying priorities – a must for building reliable patent families
  - Text Searchable Information – added for the purposes of PCT Minimum documentation

Authority File – The need for Consistency

- ***Ensures that data values are represented consistently within a data set, between and across data sets.***

- Consistency can be on record level or a cross-record level

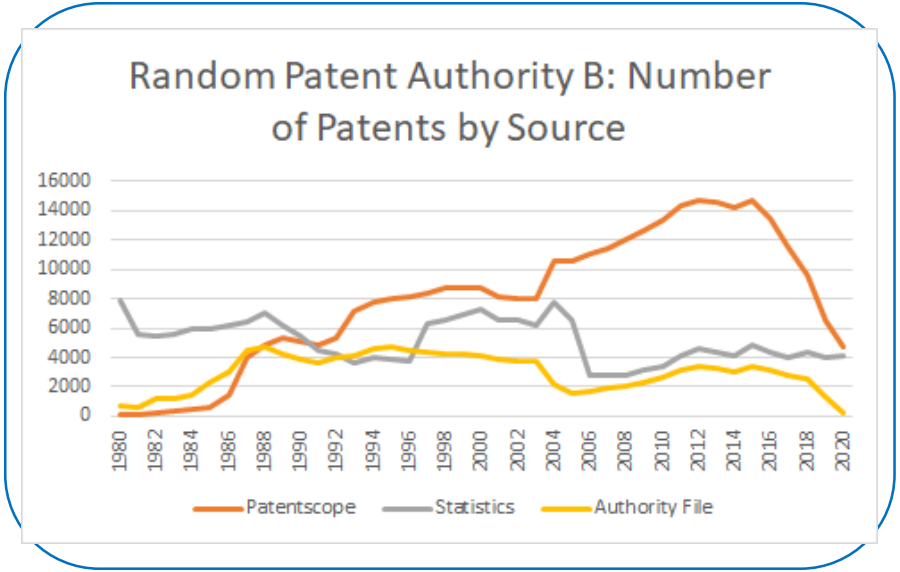
- Example : An authority file where exception Codes are only provided for a portion of the authority file (random date ranges)

Authority File - Reasonability

- ***Measure of how much data pattern meets expectations***

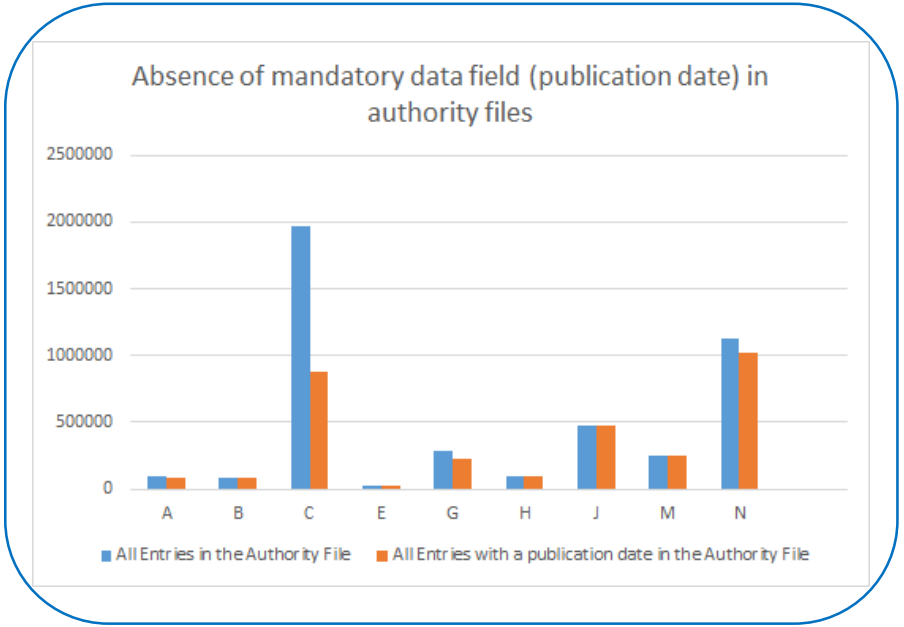
- Could be based on comparison of benchmark data or past instances of similar data

Example :



Authority File – A Tool to Enforce Uniqueness

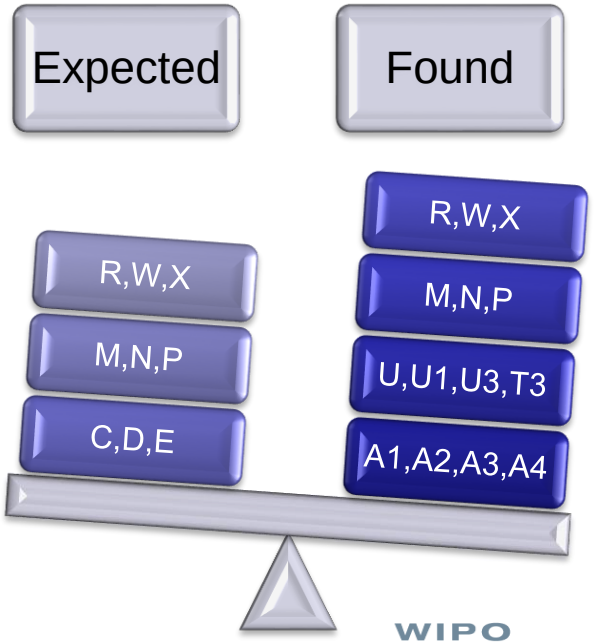
- ***No entity exists more than once in a specific data set***
- A key value relates to one and only one entity.
- Example : Only complete publication including an office code, a publication number, a kind code and a publication date can ensure uniqueness of a record in an authority file



Authority File - Validity

- ***Refers to whether data values are consistent with a defined domain of values***
- A domain of values could be a set of values, a range of values or rule-based.

Example : Invalid values for Exception Codes



Authority File - Timeliness

IP Office	Authority file	Definition file	Coverage	Remark
AT	XML		1990-01-01 to 2022-03-01	Updated monthly; published here biannually.
AU	HTML		1905-12-04 onwards	Updated quarterly
CA	TXT	PDF	Until 2021-12-31	Comprehensive; CIPO intends to produce an authority file for published patent documents once each year.
CN	HTML	HTML	1985-09-10 onwards	Every six months
CZ	XML (ZIP)	XML (ZIP)*	1903-01-01 onwards	Updated annually
DE	TXT	TXT	Date range up until 2022-03-07; for DE patents and utility models the start of the date range is 1978-01-01. For other types, all data available.	Comprehensive; Updated at a yearly interval
EA	HTML	HTML	1996-07-01 and updated monthly	Authority file is updated on monthly basis https://www.eapo.org/en/?publs=authfile

- ***Measures how up-to-date is the data and how frequently it is likely to change.***
- ST. 37 recommends at least one annual update of the WIPO Authority File portal in March of each year

Authority Files – User Perspective Summary

- From a user perspective the following should be encouraged
  - Sharing of all publications by the office
  - Sharing of the optional information such as definition file and additional data elements
  - Timely sharing of data
  - Validation of the content of data elements to avoid empty or invalid data
  - Promote collaboration between the authors of statistics and authors of the authority files for a better mutual usage as benchmarks.

- As a community we also need to share any experiences with the digitization of gazettes as the ultimate source for complete and consistent authority files and reflect on ways to reuse and improve those experiences

CONTACT

- WIPO’s PATENTSCOPE https://patentscope.wipo.int/
- patentscope-data@wipo.int for data-related issues
- patentscope@wipo.int for feature-related issues
- magdalena.zelenkovska@wipo.int