WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

IC-SDV 2018

Automatic Text categorization in the International Patent Classification

IPCCAT-Neural

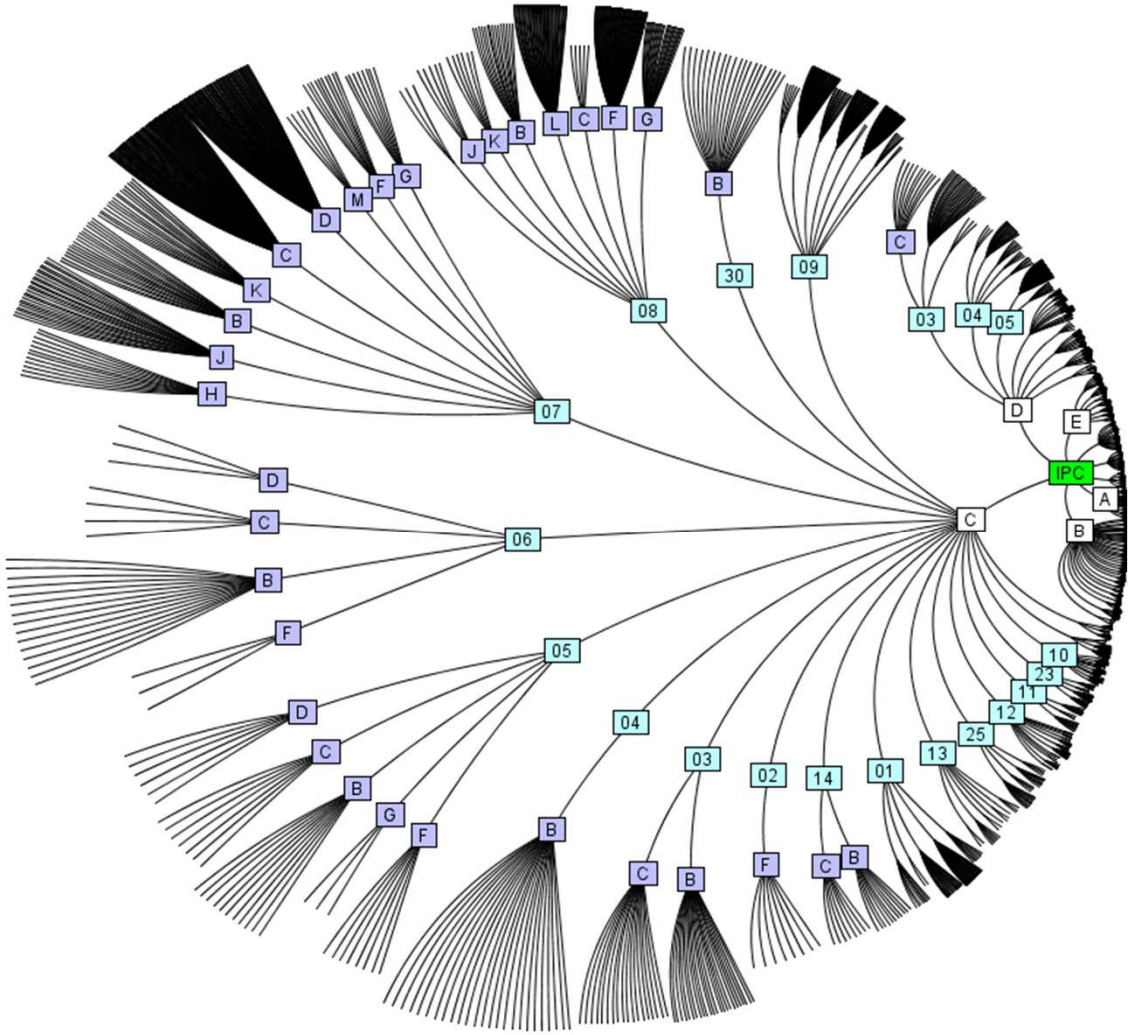
Nice
April 23, 2018

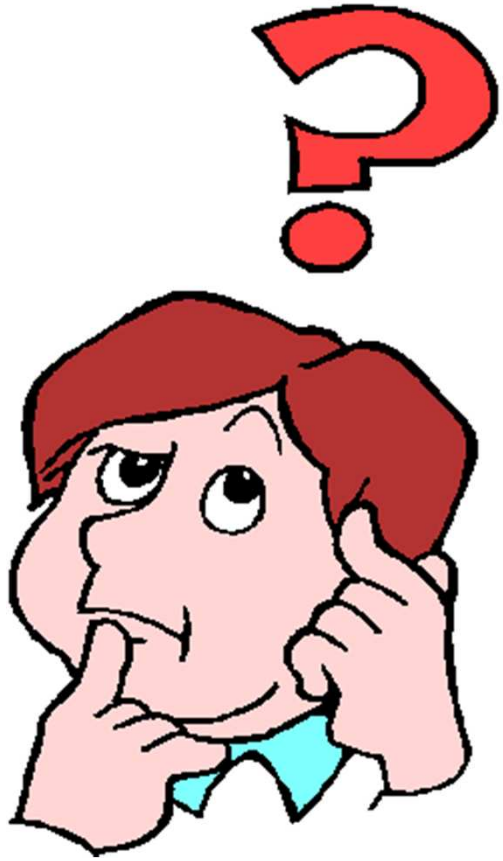
Patrick FIÉVET & Jacques GUYOT

IPCCAT-neural : automatic text categorization in the IPC

- **What is it about?**
  - Patent Classifications : **IPC** (and CPC)
  - Automatic text **CAT**egorization in the specific context of patent documents
  - Artificial Intelligence (AI) to imitate previous practices (already classified patent documents)

IPCCAT-neural : automatic text categorization in the IPC





WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

IPCCAT-neural : automatic text categorization in the IPC

- **Initial problems to be solved:**
  - IPCs allotment in small Patent Offices
  - Automatic routing of patent/technical documents according to their technical domains

IPCCAT-neural : automatic text categorization in the IPC

- **Is it a new?**
  - No, see IPCCAT presentation to ICIC 2003
  - IPCCAT at Main Group level exists since 2004 with a yearly retraining take into account:
    - new vocabulary in patent documents
    - Evolution of the IPC system

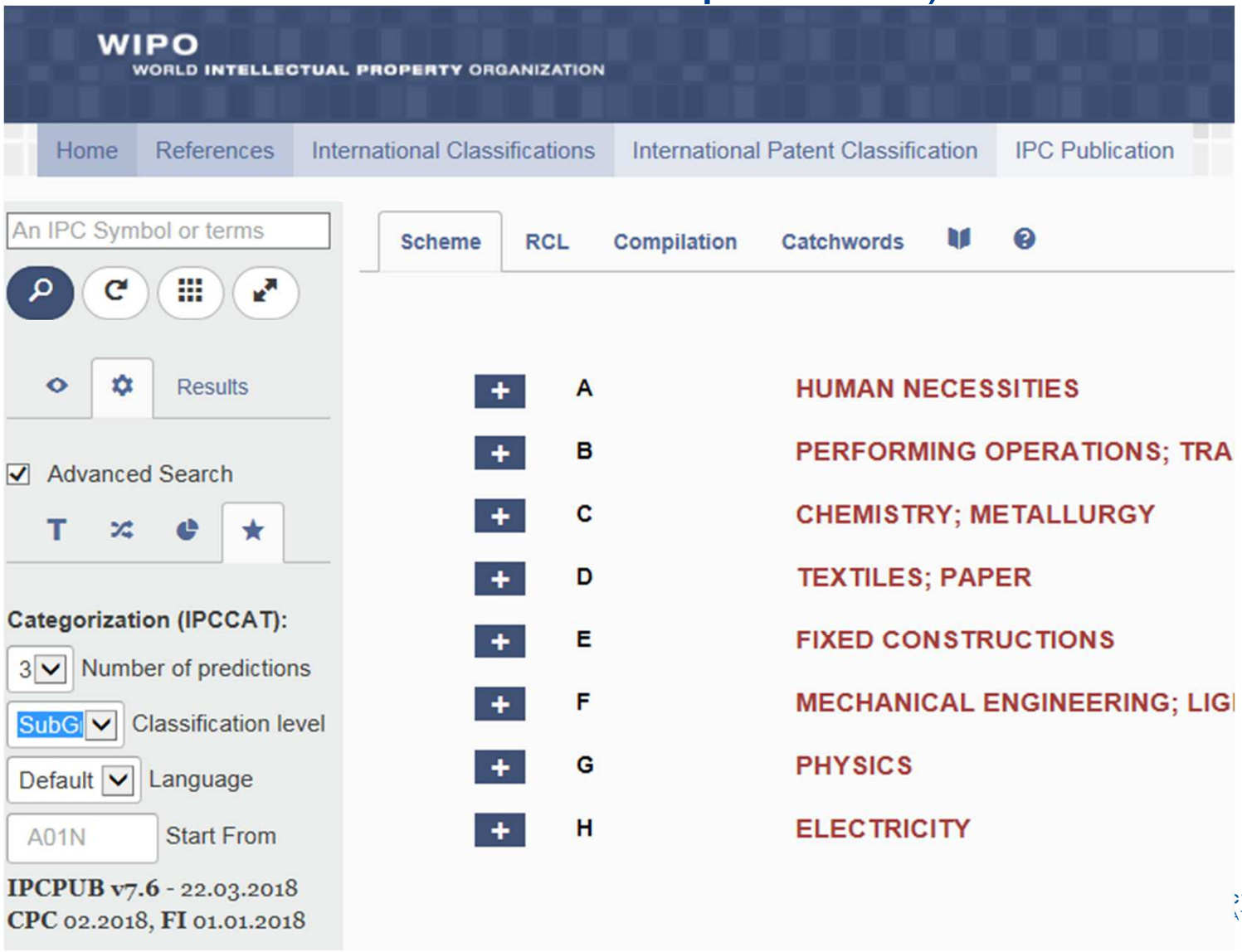
IPCCAT-neural : Principles

- Large collection (millions) of patent documents already classified, preferably with good-reputation practices
- Minimum number of documents for each IPC symbol
  - Use of the CPC (converted into IPC through concordance)
  - Complement with IPC symbols
- **Training /Testing phase : 80% / 20%**
  - Precision measure: Three guess evaluation on millions of test cases

IPCCAT-neural : Principles

- **Production use: 100% of the collection**
  - Web service + 5 confidence levels
  - User interface through IPC publication platform (IPCPUB)

IPCCAT-neural : user interface (through IPC Publication platform)



IPCCAT-neural : automatic text categorization in the IPC

- **Challenges?**
  - IPC coverage Vs. availability of training collections
  - Precision Vs. Recall (e.g. for Prior art Search)
  - Absolute Vs. relative quality of IPCCAT-Neural

IPCCAT-neural : recall challenges

- **Recall: "First" IPC does not mean "Main" IPC**
  - One IPC is usually not enough for patent classification but is enough for routing
  - **Automatic Prediction** (guess) of the **most appropriate IPC** symbol**s** on the basis on a text input (e.g. patent abstract) **with an associated level of confidence in each of these predictions**
  - **Confidence levels can be used to decide upon the number of IPC**s** to be allotted**

IPCCAT-neural : precision challenges

- **Precision: "Mycat" based on classic neural networks**
  - **Trained system** based on (many) **neural networks**
  - No evidence that more recent technology e.g. Deep Learning would perform better than "classic" neural networks (because the classification is known in the training collection)
  - Cascaded predictions improve precision

IPCCAT-neural : quality challenges

- **IPCCAT quality is relative to IPC quality in its training collection:**
  - Quality of guessed IPC vs. IPC theory. IPCCAT Imitates human practices (good and bad ones)
  - Limited by documents fragments used during its training
- **IPCCAT offers consistent and repeatable predictions**
  - That human beings are usually not be able to achieve

IPCCAT-neural 2017 performances

	English Set	French Set
Data Source	DOCDB	DOCDB
Number of Training Patents	27'731'470	4'460'815
Number of Testing Patents	1'386'574	223'041
Total Number of Example Patents	26'344'897	4'237'774
Total Number of Classes in IPC 2017.01	130	130
Number of Trained Classes	130	124
Coverage at Class Level	**100%**	**95.38%**
Precision of Classification at Class Level ("Three Guesses")	**96.04%**	**93.58%**
Total Number of Sub-Classes in IPC 2017.01	639	639
Number of Trained Sub-Classes	639	631
Coverage at Sub-Class Level	**100%**	96.87%
Precision of Classification at Sub-Class Level ("Three Guesses")	**93.99%**	**89.97%**
Total Number of Main Groups in IPC 2017.01	7'420	7'420
Number of Trained Main Groups	7'374	7'170
Coverage at Main Group Level	99.38%	96.63%
Precision of Classification at Main Group Level ("Three Guesses")	**89.27%**	**82.91%**
Total Number of Sub Groups in IPC 2017.01	72'981	72'981
Number of Trained Sub Groups	72'137	66'430
Coverage at Sub Group Level	98.84%	91.02%
Precision of Classification at Group Level ("Three Guesses")	**82.50%**	**72.09%**

IPCCAT-neural 2018

- Where are we today?

IPCCAT-neural 2018: text categorization in the IPC at subgroup level !

- Automatic prediction in 99% of the IPC i.e. among **72,137 categories**
- Top-three guesses with **80% precision**

IPCCAT-neural 2018: text categorization in the IPC at subgroup level

Training collection:

- **27.7 million in EN**
- 4.4 million in FR

IPC coverage(using IPC and CPC through concordance):

- **99%** at subgroup level (**EN**)
- 91% at subgroup level (FR)

IPCCAT-neural 2018: text categorization in the IPC at subgroup level

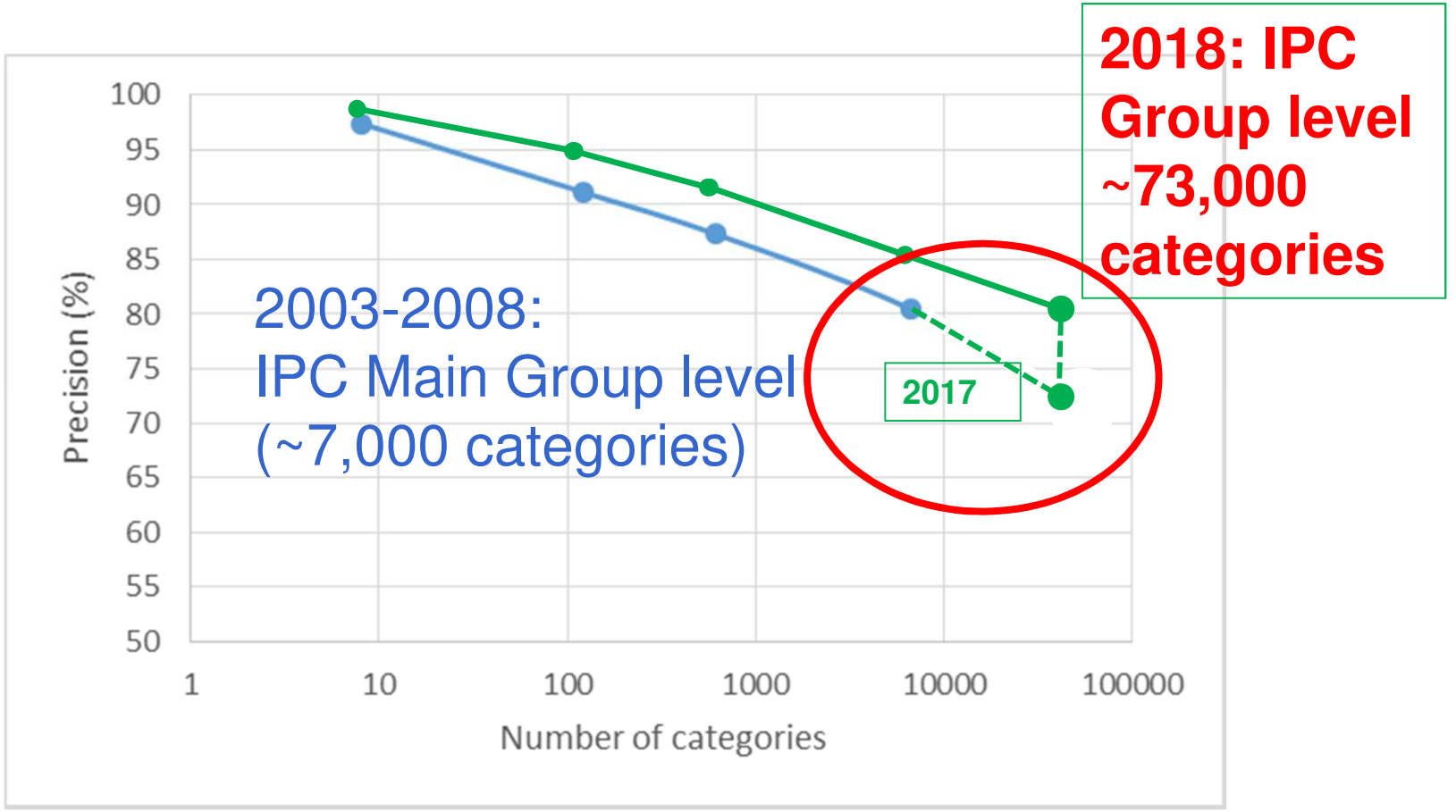
Precision:

- based on **1.5 million of test cases: 82% (EN)**
- three-guesses evaluation:
  - **82.5 % at subgroup level (EN)**
  - **72% at subgroup level (FR)**

IPCCAT-neural text categorization in the IPC at subgroup level

- **Why was It actually doable?**
  - Recent evolution of the IPCCAT classifier available on-demand as open source by the **Olanto foundation see** http://olanto.org/foundation
  - **Added value in data processing:**
    - Training based on patent documents computed from DOCDB XML excerpts (title + abstract)
    - Computation of both IPC and CPC classifications
    - Progress in computing power opens new R&D horizons e.g. GPU, text processing,…

Evolution of IPCCAT R&D over years



IPCCAT-neural 2018

- Potential use of IPCCAT technology

IPCCAT-neural practical use

- **What it could be for?**
  - Improvement of the consistency in patent classification
  - Reduction of the backlog of IPC reclassification through automation of the residual IPC reclassification of patent documents after some years: **Potential alternative to IPC reclassification Default transfer**

IPCCAT-neural for IPC reclassification

- **Additional Challenges:**
  - **non-EN language:**
    - Large training collection, with good IPC coverage
    - Consistency in classification practices
    - Preserve possibility of dedicated solution
  - **Costs containment for WIPO**
    - Streamline production of training collection(s)
    - Streamline IPCCAT retraining

IPCCAT-neural: **on the way to assist IPC reclassification**

- **Chronology:**
  1. **Evidence** that text categorization works at IPC subgroup level with **acceptable precision**: **Done**

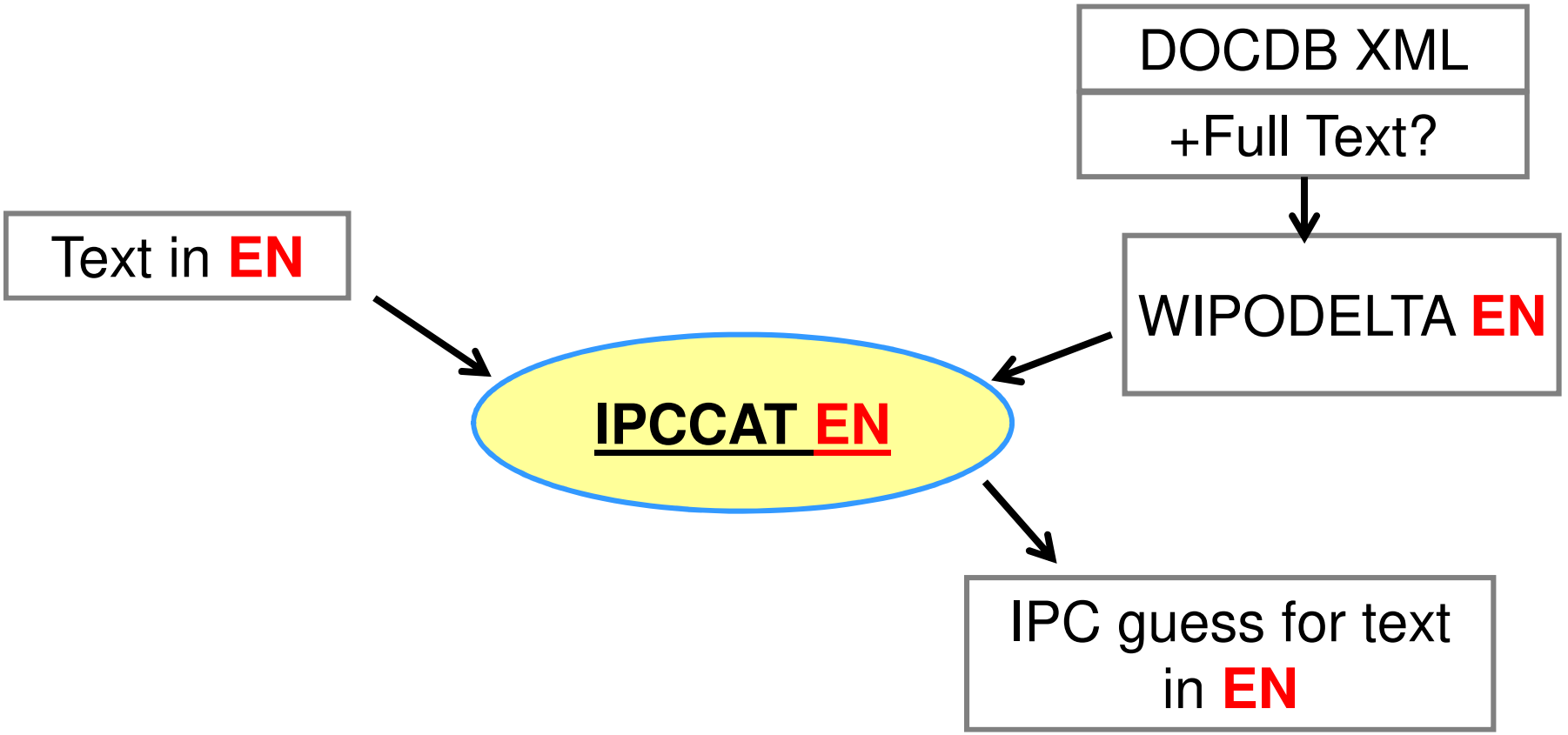
  2. Integration of IPCCAT neural at sub-group level into **IPCPUB v 7.6** **Done**

  3. Confirmation that **Cross-lingual text categorization** can assist in other languages than EN, even in absence of large training collections: **Done**

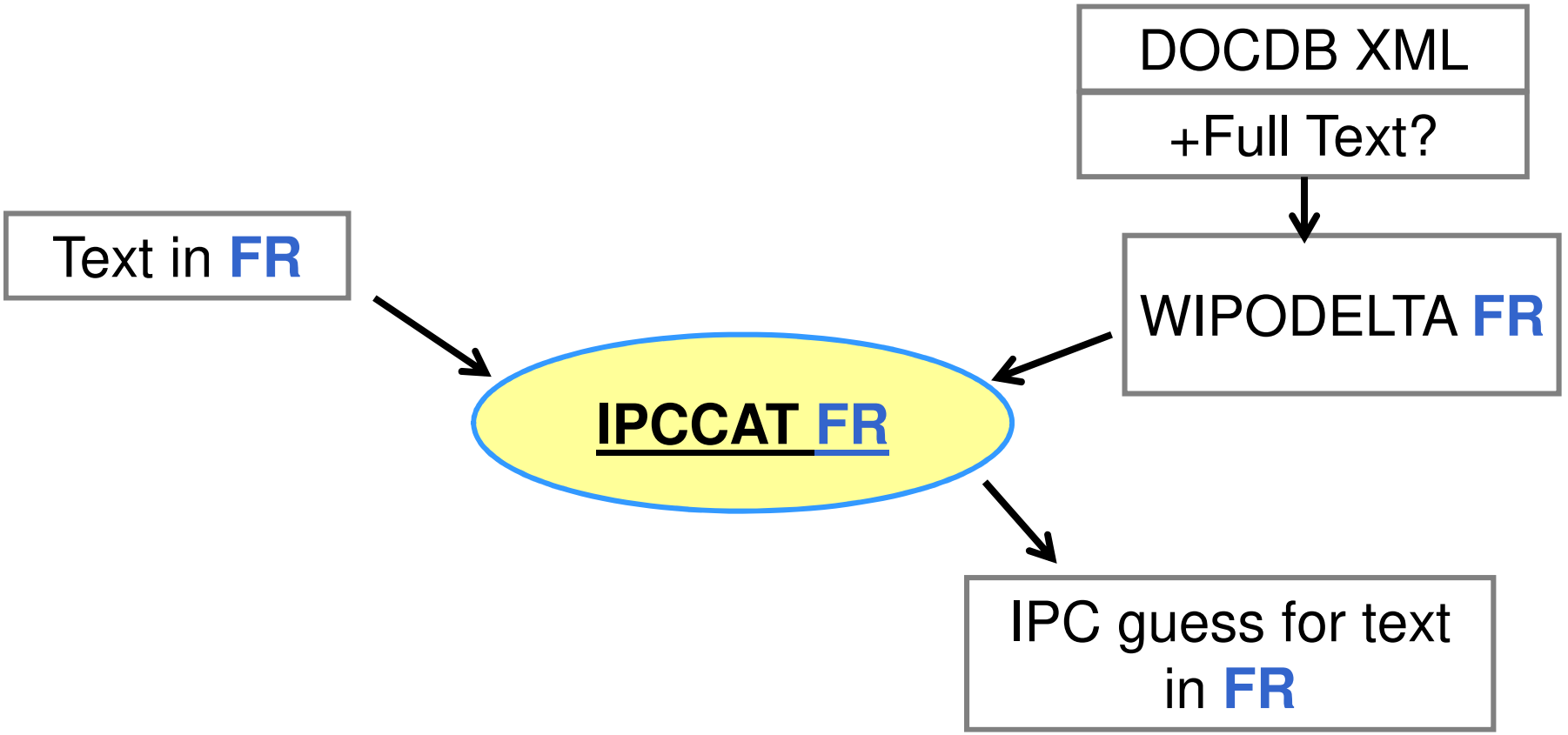
IPCCAT-neural: **on the way to assist IPC reclassification**

- **Chronology: (Still a long way to go)**
  4. Incentives for R&D in automated text categorization: **WIPO DELTA** training collection Q2 2018 **in progress**
  5. Propose alternatives to Default Transfer e.g. **more than one symbol** based on IPCCAT guesses **and confidence levels**(**2019**)
  6. CE Decisions, WIPO resource planning, etc… (**2019**)
  7. **Development of** the production-scale solution integrating **neural cross-lingual text categorization** (based on **IPCCAT neural and** WIPO translate) (**202x**)
  8. **Integration into IPCWLMS for Stage 3 reclassification** (**202x**)

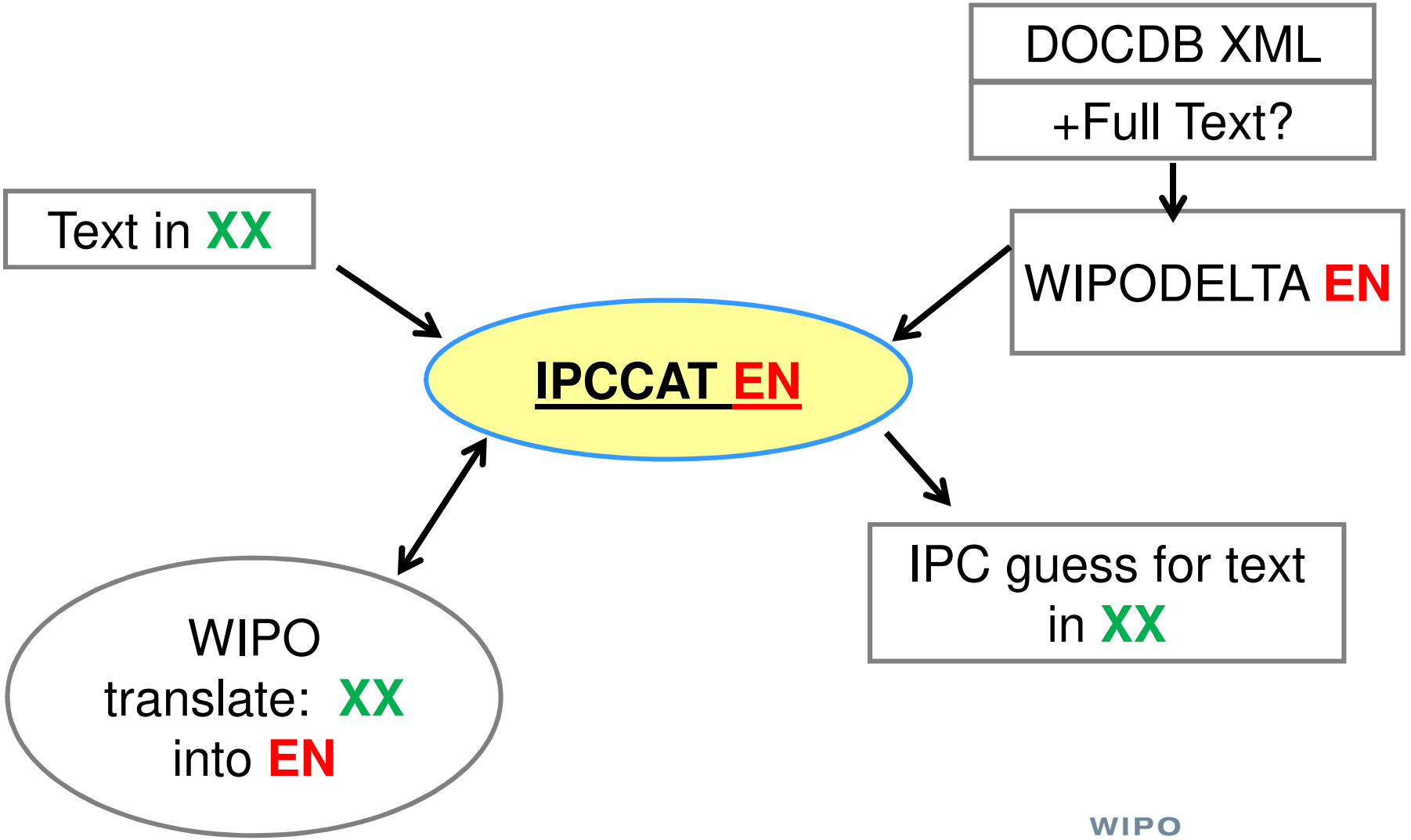
IPCCAT-neural cross lingual principle



IPCCAT-neural cross lingual principle



IPCCAT-neural cross lingual



Text categorization in the IPC

- **Other 2018 perspectives:**
  - **Cross lingual text categorization in the IPC at subgroup level**
    - **Confirmation of expectations through p**rototyping of ES, FR, EN, DE, RU support through use of automatic translation by commercial product (bound by budget limitations)

    e.g. DE text translated text into EN and submitted to IPCCAT neural trained with EN documents
    - Available through IPCPUB interface or web service (Q2 2018)
    - IPCCAT retraining based on IPC 2018.01 (Q3 2018)

Incentive to R&D in text categorization: **WIPO-Delta training collection**

- **Incentives for research and development institutes interested in automatic text categorization :**
  - WIPO DELTA 2018 EN collection available upon request
  - Fully specified XML format
  - Xx million documents classified in the IPC
  - **Complement the public WIPO-ALPHA training collection**
  - **http://www.wipo.int/classifications/ipc/en/ITsupport/Categorization/dataset/index.html**

Thank you for your attention!

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION