Europäisches Patentamt
European Patent Office
Office européen des brevets

A ROADMAP FOR PATSTAT

Davide Lingua

D. 5.1.1 Publication

EPO Worldwide Patent Statistical Database (PATSTAT)

... or what can happen when we listen to our customers

So what is PATSTAT ?

EPO Worldwide Patent Statistical Database (PATSTAT)

For the first time, the EPO has provided an *off-line* worldwide patent database:

- we provide the data (tables) & database model
- we show how to store the data
- researchers can work with the entire database on a standard laptop PC

Thanks to the OECD (Science, Technology, Industry Section) in Paris, France, for their earlier database work

PATSTAT: the story

- Early prototype developed by OECD (Science, Technology, Industry Section)
- EPO took over responsibility for production
- First distributed in April 2006
- Restricted distribution in development period
- Since October 2007 publicly available
- It has established itself as a database of choice for
  - statisticians
  - academics
  - policy advisors

PATSTAT: some facts

- Produced twice yearly as a snapshot of the whole database
- Designed for integration into a relational database
- Distributed on DVDs
- A collection of tables in csv format
- Approx. zipped size: 12 Gb
- Approx. database (indexed) size: 100 Gb
- It is not a "plug & play" product! You need:
  - SQL knowledge to make queries
  - support from an IT specialist for getting it up & running
  - appropriate HW/SW

Data sources for PATSTAT

EPO sources:

- DOCDB, the EPO master bibliographic database
- EPO ESPACE Bulletin database, for name & address data

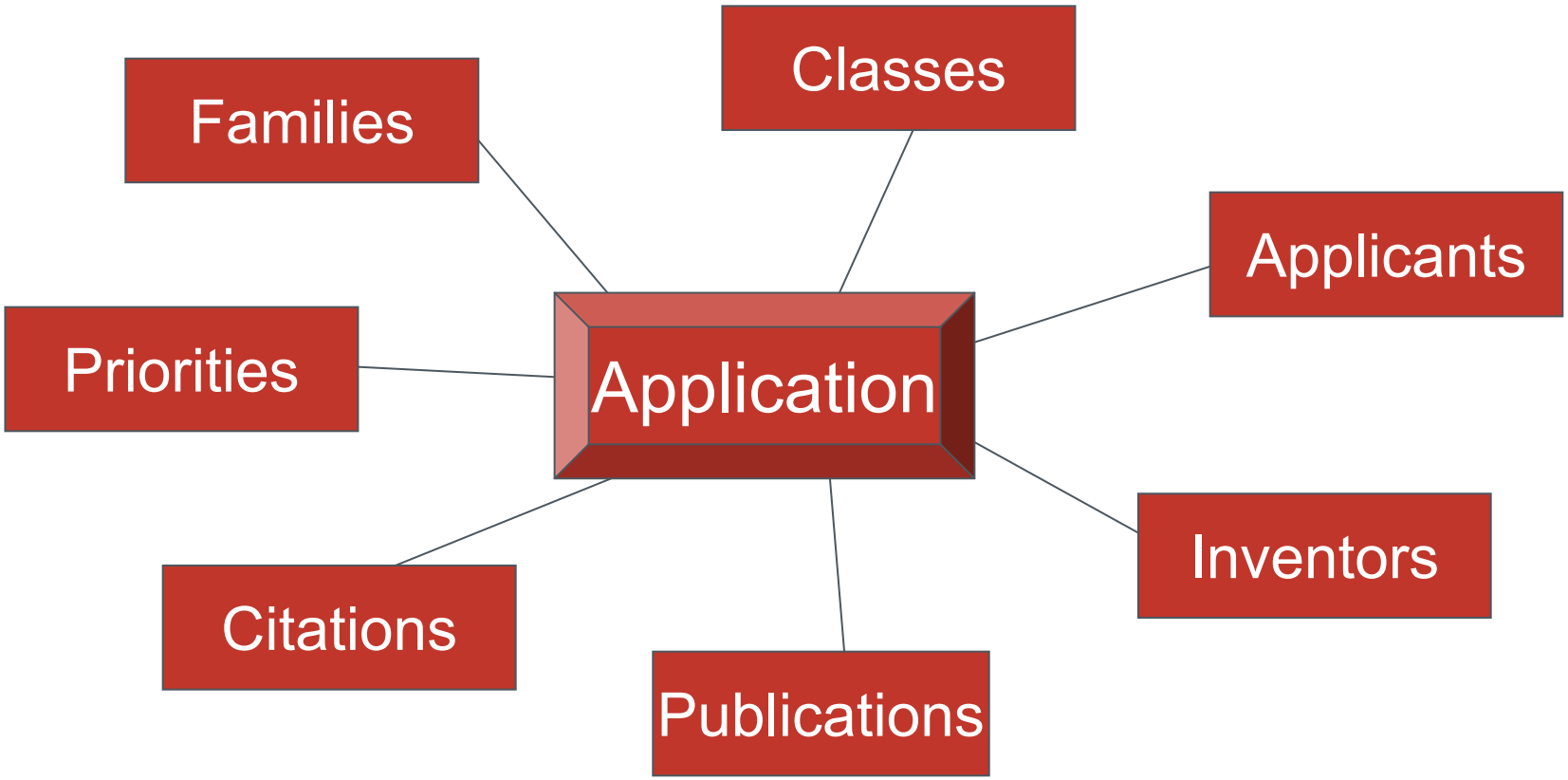
Other sources:

- US publications data from public ftp site for name & address data

So how is PATSTAT different from DOCDB?

- PATSTAT comes with a suggested database structure
- For DOCDB a structure must be designed by the user
- DOCDB and PATSTAT have the same publication, application, priority, citation data, ...
- DOCDB has more abstracts
- PATSTAT has more address data for EP and US patents
- PATSTAT contains pre-calculated INPADOC families and simple "DOCDB" family

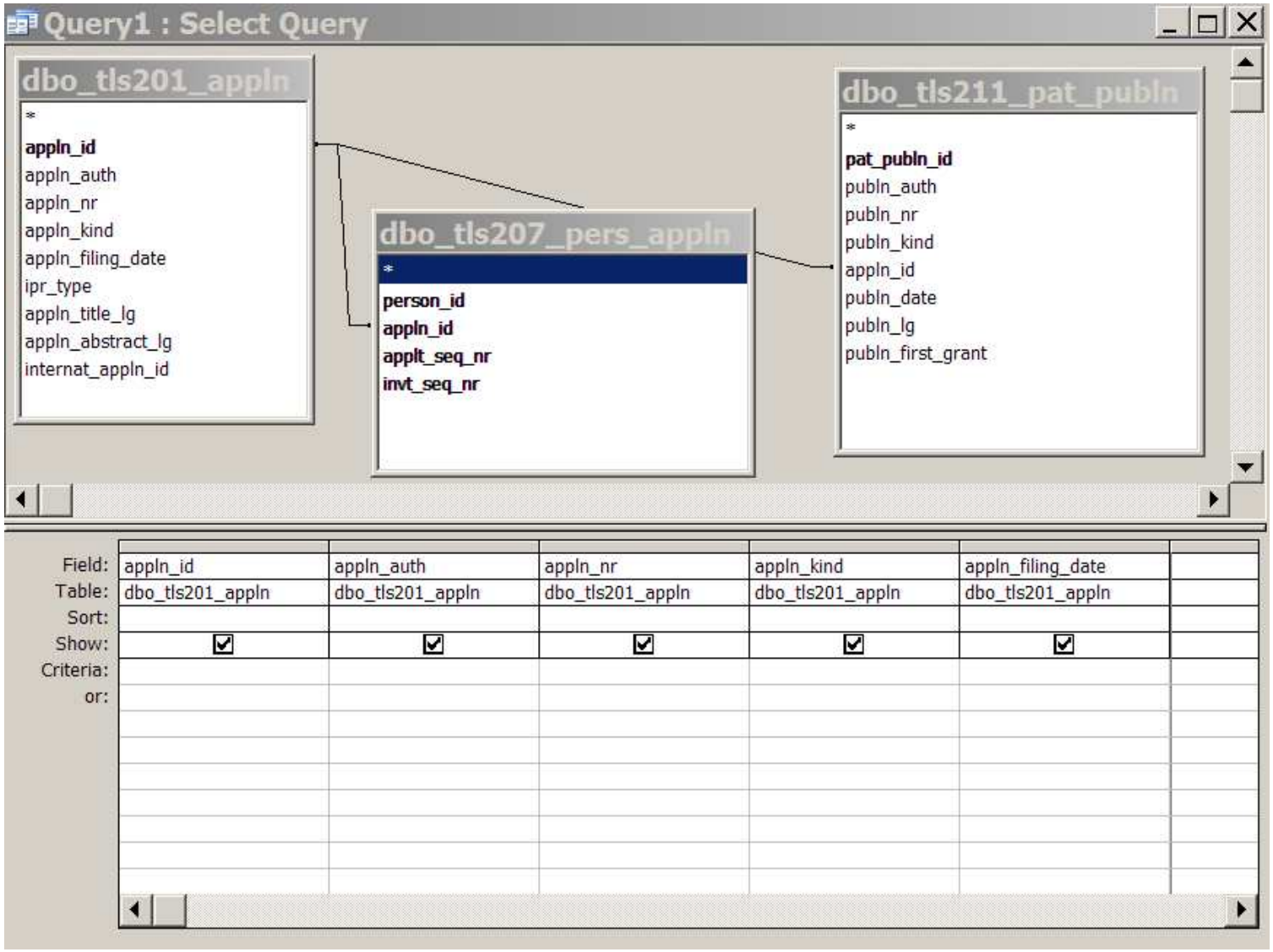
PATSTAT Central Database



Physical tables in the April 2010 edition

Table name	Number of rows in SQL Server in April 2010
TLS201_APPLN	64,741,334
TLS202_APPLN_TITLE	47,367,201
TLS203_APPLN_ABSTR	approx 18 million
TLS204_APPLN_PRIOR	27,990,886
TLS205_TECH_REL	2,113,937
TLS206_PERSON	36,705,580
TLS207_PERS_APPLN	131,735,282
TLS208_DOC_STD_NMS	16,633,097
TLS209_APPLN_IPC	296,015,742
TLS210_APPLN_N_CLS	24,262,953
TLS211_PAT_PUBLN	72,470,099
TLS212_CITATION	92,900,289
TLS214_NPL_PUBLN	13,942,281
TLS215_CITN_CATEG	17,400,424
TLS216_APPLN_CONTN	1,700,695
TLS217_APPLN_ECLA	97,833,161
TLS218_DOCDB_FAM	57,505,125
TLS219_INPADOC_FAM	64,741,334

Application_ID is the central key



TLS201_APPLN

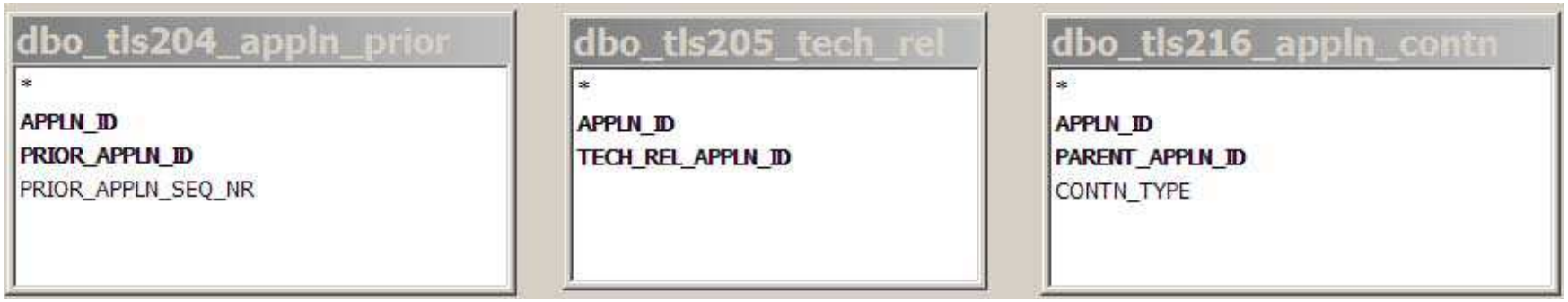
- "True" applications
  if appl_ID =< 57.505.125
- Unpublished priorities if appl_ID
  between 58.000.000 and 63.983.731
- "dummy" D2 applications from
  citations if appl_ID between 64.000.000 and 65.252.476
- appl_kind = W for PCT applications
- PCT origin is given in internat_appl_ID
- ipr_type: UM, DP, PI

Relationships between applications

- Priorities under the Paris Convention (INID 30) → Table 204
- References to other related domestic patent documents (INID 60) → Table 216
  - addition
  - continuation in part
  - division
  - reissue
  - substitute
- Data related to other International Conventions (INID 80) eg. PCT → Table 201
- Technical relations → Table 205

Priorities

- TLS205 and TLS216: meaningful information only from the 1990ies onwards
- Earlier technical relations and continuations were attributed "Paris Convention priority" status



Persons: applicants (assignees, grantees, proprietors) and inventors

- Person is physical or legal entity
- Person can be applicant AND inventor
- doc_std_name: max. 30 char.
- Names and addresses from most recent publication
- "docdba" elements are used (as received by EPO)
- USPTO data: 1/3 of sequence data missing ➔ person_name and doc_std_name_id might not match
- special txt files for US data with individual elements

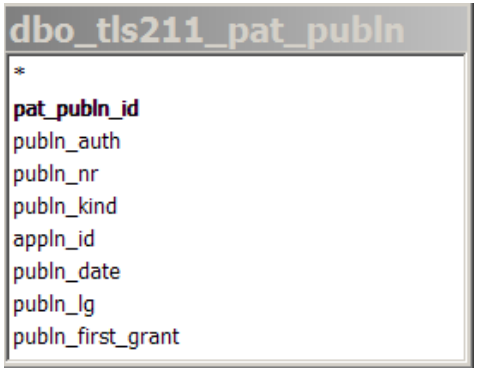
Persons and address data

Authority	**US**		**EP**	**Other (GB, IE,...)**
Publication type	Published applications	Published patents	All	All
Range Jan. 1976- Nov. 2005	DOCDB XML	OECD patent database	ESPACE BULLETIN	DOCDB XML
Range Nov. 2005-today (*)	USPTO website	USPTO website	ESPACE BULLETIN	DOCDB XML

(*) Sept. 2005 to today for published applications

- person_ctry_code from data format "docdb" (standardised), except US, EP
- person_ctry_code: coverage is 50%, ***not JP***
- Special EP cases: "*data withheld*", "*the designation of the inventor has not yet been filed*"
- EP data reflect the last changes (from Bulletin database)

Patent publications



- For publications having more than one occurrence (example EP or WO A9), only the last occurrence is loaded
- Invalid or empty dates: 9999-12-31
- Abstracts and titles are attributed to the application, in case of multiple occurrences:
  - English has priority
  - the most recent is selected
- publn_first_grant: no guarantee, INID codes 450 and 470

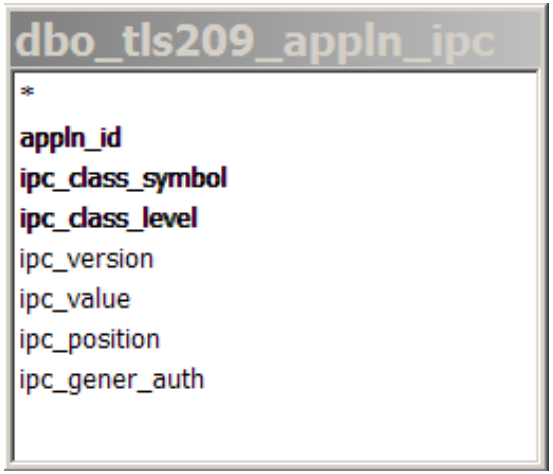
Citations

- Reliable data for: AP, AU, BE, CZ, DE, EP, ES FR, NL, SG, US, WO
- Batches only for: JP, DK, LU, GR, TR
- Euro-PCT: npl_biblio = "See references of WO 0046271A1"
- Patent citations hidden in NPL citations:
  - npl_publn_id > 0
  - npl_citn_seq_nr > 0 , and
  - *cited_pat_publn_id > 0*

TLS209: IPC-8 classification

- IPC 1-7 symbols are NOT given
- 570.000 documents have no IPC8 but a IPC1-7 symbol
  - 10% published after 2006!
  - 10% have an ECLA symbol
- IPC8 classes are aggregated and de-duplicated at application and simple family level
- ***Core*** and ***advanced*** symbols are given
- Jan. 2011: revision of IPC-8 -

  core symbols will be deleted!



TLS217: ECLA classification

- Nanotech codes:
  - epo_class_scheme = ICO
  - epo_class_symbol = Y01N6:00
- Environmentally Sound Technologies (in September 2010)
  - epo_class_scheme = ICO
  - epo_class_symbol = Y02B10:00, Y02C10:00,...
- Schemes covered:
  - EC
  - ICO
  - ECNO
  - IDT

DOCDB Simple Patent Family versus INPADOC Extended Family

DOCDB Simple versus INPADOC Extended

- INPADOC Extended Family
  - is covering a technology
  - might be slight differences in technical content
  - members do not have to share more than one priority with at least one other member, directly or indirectly

- DOCDB Simple Patent Family
  - is covering one invention
  - technical content covered is identical
  - members have to share identical priority pictures

DOCDB Simple versus INPADOC Extended

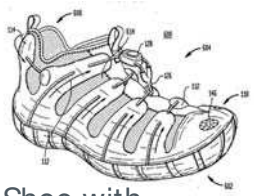
- INPADOC Extended Family
  - broadest definition of a patent family
  - supports identification technological trends
  - supports definition of geographical coverage

- DOCDB Simple Patent Family
  - subset of INPADOC Extended Family
  - particularly suited for prior art search
  - tailored to the needs of EPO examiners

Family list
9 application(s) for: **US2007011914 (A1)**

1. **Shoe with lacing**
   Publication info: **EP1743540 (A1)** — 2007-01-17
2. **Shoe with anatomical protection**
   Publication info: **EP1864584 (A1)** — 2007-12-12
3. **Shoe with lacing**
   Publication info: **US2007011910 (A1)** — 2007-01-18
4. **Shoe with lacing**
   Publication info: **US2007011911 (A1)** — 2007-01-18
5. **Shoe with lacing**
   Publication info: **US2007011912 (A1)** — 2007-01-18
6. **Shoe with wraparound lacing**
   Publication info: **US2008047165 (A1)** — 2008-02-28
7. **Shoe with anatomical protection**
   Publication info: **US2007011914 (A1)** — 2007-01-18
8. **SHOE WITH LACING**
   Publication info: **WO2007011737 (A2)** — 2007-01-25
9. **SHOE WITH ANATOMICAL PROTECTION**
   Publication info: **WO2007143228 (A2)** — 2007-12-13

one INPADOC extended =
four DOCDB simple families



Shoe with
anatomical protection

Extended family for **US 2007011914A1** VIEW Application - Priority Picture

PUBLICATION					APPLICATION				INACTIVE PRIORITY				
Dos	CC	ID	KC	DATE	CC	ID	KC	DATE	CC	ID	KC	DATE	LMI
	US	2007011914	A1	18-01-2007 EN 1 *	US	44896706	A	07-06-2006	US	32859306	A	10-01-2006	2
									US	19521405	A	02-08-2005	2
									US	18297005	A	15-07-2005	2
	EP	1864584	A1	12-12-2007 EN 1 **	EP	07010961	A	04-06-2007					
	WO	2007143228	A2	13-12-2007 EN 1 **	US	2007013782	W	07-06-2007					

continuation in part

PRIORITY PICTURE

CC	ID	KC	DATE
US	44896706	A	07-06-2006

PUBLICATION					APPLICATION				INACTIVE PRIORITY				
Dos	CC	ID	KC	DATE	CC	ID	KC	DATE	CC	ID	KC	DATE	LMI
	EP	1743540	A1	17-01-2007 EN 1 *	EP	06014723	A	14-07-2006					
	US	2007011912	A1	18-01-2007 EN 1 **	US	32859306	A	10-01-2006					
	WO	2007011737	A2	25-01-2007 EN 1 **	US	2006027388	W	13-07-2006					

continuation in part

PRIORITY PICTURE

CC	ID	KC	DATE
US	18297005	A	15-07-2005
US	19521405	A	02-08-2005
US	32859306	A	10-01-2006

PUBLICATION					APPLICATION				INACTIVE PRIORITY				
Dos	CC	ID	KC	DATE	CC	ID	KC	DATE	CC	ID	KC	DATE	LMI
	US	2007011910	A1	18-01-2007 EN 1 *	US	18297005	A	15-07-2005					
	US	2008047165	A1	28-02-2008 EN 1 **	US	90114707	A	14-09-2007	US	90114707	A	14-09-2007	

first filing & division of

PRIORITY PICTURE

CC	ID	KC	DATE
US	18297005	A	15-07-2005

PUBLICATION					APPLICATION				INACTIVE PRIORITY				
Dos	CC	ID	KC	DATE	CC	ID	KC	DATE	CC	ID	KC	DATE	LMI
	US	2007011911	A1	18-01-2007 EN 1 *	US	19521405	A	02-08-2005					

continuation in part

PRIORITY PICTURE

CC	ID	KC	DATE
US	19521405	A	02-08-2005
US	18297005	A	15-07-2005

Patent families: TLS218 and TLS219

- TLS218: simple family
  - Family-identifier is created in DOCDB
  - Family-identifier is a surrogate key, unique but change is not excluded (application might change family)
  - Surrogate keys remain the same through PATSTAT editions

- TLS219 INPADOC family
  - Family-identifier is created in PATSTAT
  - inpadoc_family_id changes with every edition of PATSTAT

Changes in April 2010 edition

- ECLA: table TLS217 contains complete ECLA/ICO instead of only Y01N nanotech codes
- Outsourcing of PATSTAT production
  - DOCDB XML instead of mainframe extraction
  - EP person data: ESPACE BULLETIN instead of mainframe extraction
  - improved quality process
  - systematic user acceptance testing
  - "dummy" D2 application reduced
  - improved US applicant name coverage

Planned developments

- EST to become part of ECLA/ICO codes (Y02)
- Extension with worldwide legal data: PRS in CSV format with PATSTAT "ApplicationID"
- Unique application ID ➔ stable appl_ID
- publn_first_grant from routine to DOCDB XML feed
- Number of claims for US data
- PCT address data
- Name standardisation
- PATSTAT visualisation project

Questions?

...you have the floor!

Thank you for your attention

Davide Lingua
dlingua@epo.org

PS: See you in Vienna!