EPO plan for WIPO Standards dealing with XML

Fernando Ferreira
Data Standards Coordinator

Geneva, 22 March 2016
Future data flow

Data Flow Platform (ETL) → Conversion to PDM → Data Services Layer (KIME-API) → PDM format → Client 1 → Client 2 → Clients n

ST.36 → Ad-hoc format → ST.96

ST.36/Annex F → Business support systems (unpublished documents + procedural data)
Patent Document Model: Drivers

• Improve Patent Data: standardization, integration, enrichment

• Strengthen Security: protect data from unauthorized access and corruption

• Improve IM agility to deliver IT services by providing simple and efficient access to reusable data services at enterprise level
Why PDM?

- **Common data model** for patent related documentation supporting use of both patent and non patent literature in search, examination, translation and dissemination

- Encode **semantic annotations** and descriptions of the primary content (data enrichment)

- Ability to **manage changes** to the data in a consistent way and **evolve** the model over time to meet new requirements
The Patent Document Model

- The Patent Document Model (PDM) is a common model for patent related documentation based on Text Encoding Initiative (TEI)

- **Principles:**
  - Encode simple families with corresponding applications and publications
  - Encode annotations (manual and automatic)
  - Compatible with DocDB XML (based on ST.36)
  - Independent of the physical implementation

- **Applied at EPO today:**
  - KIME: reusable and high performance data service for provisioning of published patent documents in PDM format (45M families, 90M patents)
  - ANSERA: figure and non-boolean search application
  - TAPAS: benchmark platform for future search applications
Text Encoding Initiative - Background

• XML Standard for the representation of texts in digital form

• History:
  – 1960: GML by IBM
  – 1987: TEI
  – 1992: TEI edition P3 (Michael Sperberg-McQueen and Lou Burnard, eds)
  – 1997/1998: XML 1.0 (Tim Bray, Jean Paoli and Michael Sperberg-McQueen, eds)

• Features:
  – A wide range of standardized, general purpose modelling elements for describing
digital sources (meta-data), understanding and representing the content (text, tables,
formulas, figures), enriching (annotations, links), versioning and disseminating
  – Customizable and extensible in a standard way
  – Open and supported by tools

• Users: from individual scholars to large digitisations
The PDM schema is built by reusing standard TEI elements and associating them with a well-defined meaning from the patent information domain.
Standard elements for encoding text, figures and facsimile

- Text (1), figure (2) and facsimile (3)
- Aligning text and facsimile (8), (9)
- Divisions: abstract (4), claims (5) and description (6)
- Headers, paragraphs (7), formulas, tables, lists and enumerations
- Versioning (10), (11)
- Including external schemas (12)
The invention relates to a car.
Modular PDM

- **Logical layers** dividing the model into different modules
- **Views** providing visibility to part(s) of the model
- **Modules:**
  - **core-PDM:** all the patent-related universal information independently of specific intended uses or applications;
  - **add-on modules:** additional functional oriented data added to the core-PDM (e.g. OCR, versioning, annotations, legal status, confidential and public information, ST36 compatible information, etc.)
- **Views / Filters:**
  - e.g. patent application, family members
  - e.g. ST36
PDM roadmap

• Extend PDM modular design
• Support for public as well as confidential content
• Support for different types of annotations (manual or automatic)
• Support additional data:
  – legal status,
  – procedural state,
  – non-patent literature,
  – examiner communications
• Support for document- changes and versioning
• Support for translations
Thank you for your attention

Questions?