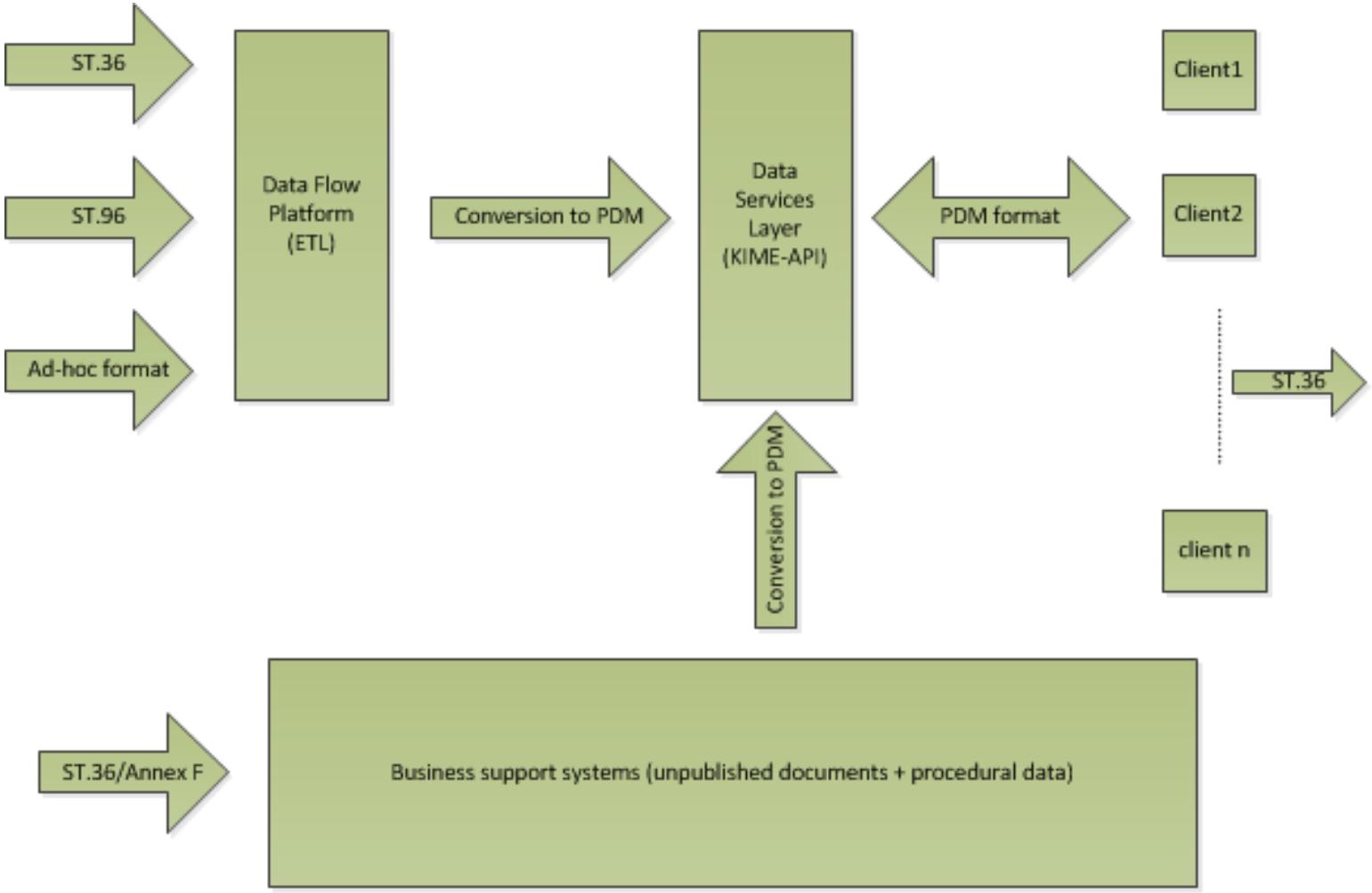
Europäisches Patentamt
European Patent Office
Office européen des brevets

EPO plan for WIPO Standards dealing with XML

Fernando Ferreira
Data Standards Coordinator

Geneva, 22 March 2016

Future data flow



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
  - 1986: SGML becomes an ISO standard: ISO 8879:1986
  - 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

ST36 Documents

publication reference **EP1565318 B1**

family **32324726**
preceding publication date
classification

application reference **EP20030811688**
language-of-filing
invention-title
referenced-cited
applicants
inventors
abstract
....

publication reference **EP1565318 A4**

family **32324726**
...

application reference **EP20030811688**
....

publication reference **DE60335820 D1**

family **32324726**
...

application reference **DE2003635820**
....

PDM Document

```
<teiCorpus>
  <teiHeader type="family">
    <idno type="family-id">32324726</idno>
  </teiHeader>
  <teiCorpus>
    <teiHeader type="application">
      <idno type="docNumber">EP20030811688</idno>
    </teiHeader>
    <TEI>
      <teiHeader type="publication">
        <orgName type="regional">EP</orgName>
        <idno type="docNumber">1565318</idno>
        <classCode scheme="kindCode">B1</classCode>
      </teiHeader>
      <facsimile>...</facsimile>
      <text>...</text>
    </TEI>
    <TEI>
      <teiHeader type="publication">
        <orgName type="regional">EP</orgName>
        <idno type="docNumber">1565318</idno>
        <classCode scheme="kindCode">A4</classCode>
      </teiHeader>
      <facsimile>...</facsimile>
      <text>...</text>
    </TEI>
  </teiCorpus>
  <teiCorpus>
    <teiHeader type="application">
      <idno type="docNumber">DE2003635820</idno>
    </teiHeader>
    <TEI>
      <teiHeader type="publication">
        <orgName type="regional">DE</orgName>
        <idno type="docNumber">60335820</idno>
        <classCode scheme="kindCode">D1</classCode>
      </teiHeader>
      <facsimile>...</facsimile>
      <text>...</text>
    </TEI>
</teiCorpus>
```

Simple Patent Family Layer
family-id
priority-claims
patent classifications (ECLA, ICO)

Patent Application Layer
application-reference
patent-classifications (ECNO)

Patent Publication Layer
publication-reference
invention-title
designation-of-states
dates-of-public-availability
parties (applicants and inventors)
previously-filed-application
patent-classifications
language-of-filing
priority-claims
preceding-publication-date
date-of-coming-into-force
language-of-publication
references-cited
text: abstract, description, claims

The PDM schema is built by reusing standard TEI elements and associating them with a well-defined meaning from the patent information domain

```
<TEI>
  <teiHeader>
    <listChange>
      <change xml:id="version1">...</change> (10)
      <change xml:id="versionI2">...</change>
    </listChange>
  </teiHeader>
  <facsimile> (2)
    <graphic xml:id="id_paglE302"  url="page1.png"/> (9)
    <graphic xml:id="id_pag2E302"  url="page2.png"/>
  </facsimile>
  <text> (1)
    <front>
      <div type="abstract" xml:lang="EN">...</div> (4)
      <div type="abstract" xml:lang="FR">...</div>
    </front>
    <group>
      <text change="version1"> (11)
        <body>
          <div xml:lang="en" type="claims">...</div> (5)
          <div xml:lang="en" type="description"> (6)
            <p facs= "#id_paglE302">The present invention ...</p> (7), (8)
          </div>
        </body>
      </text>
      <text change="version2">...</text>
      <figure> (3)
        <graphic url="figure1.png"/>
      </figure>
      <figure>
        <svg xmlns="http://www.w3.org/2000/svg"> (12)
          <rect id="d002uscom1" x="75" y="75" width="30" height="30"/>
        </svg>
      </figure>
    </group>
  </text>
</TEI>
```

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)

Encoding Text Annotations

```
<TEI>
  <teiHeader type="publication">
    ...
    <note type="standoff_annotation" subtype="word_info">
      <list>
        <item>
          <date when="2012-05-24"/>
          <author type="manual">JP50045</author>
          <term type="synonym">automobile</term>
          <ptr target="#string-range(d1e001, 5, 9)"/>
        </item>
        ...
      </list>
    </note>
    ...
  </teiHeader>
  <text>
    <body>
      <div xml:lang="en" type="description">
        <p xml:id="d1e001" xml:lang="en">The invention relates to a car.</p>
        ...
      </div>
    </text>
</TEI>
```

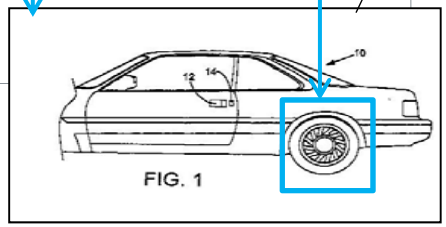
Encoding Figure Annotations

```
<TEI>
  <teiHeader type="publication">
    ...
    <note type="mask" subtype="examiner_annotations">
      <list>
        <item xml:id="mid333x5>
        <date when="2012-05-02"/>
        <author type="manual">CA00000</author>
        <ptr target="#doc1fig1"/>
        <figure>
          <svg xmlns="http://www.w3.org/2000/svg">
            <rect id="doc1mask1" x="75" y="75" width="40" height="30"/>
          </svg>
        </figure>
        </item>
      </list>
    </note>
    ...
    <note type="standoff_annotation" subtype="examiner_comments">
      <list>
        <item>
        <date when="2012-08-24"/>
        <author type="manual">EG50045</author>
        <ptr target="#doc1mask1"/>
        <label>this is the tyre of the car</label>
        </item>
        ...
      </list>
    </note>
    ...
  </teiHeader>
  <text>
    <body>
      <figure
        <graphic xml:id="doc1fig1" url="image-1.tif"/>
      </figure>
    </body>
  </text>
</TEI>
```

Mask

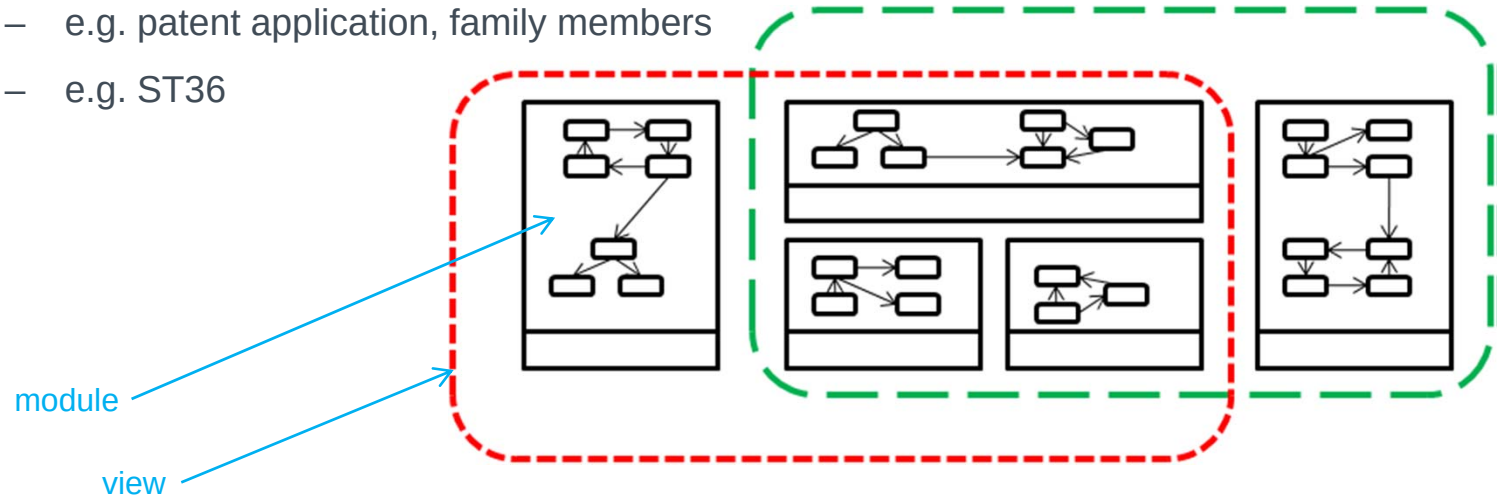
Annotation

Figure



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

Europäisches Patentamt
European Patent Office
Office européen des brevets

Thank you for your attention

Questions?