TEI challenges in an accelerating digital world

Laurent Romary

TEI challenges in an accelerating digital world

Laurent Romary

Inria & DARIAH

Argument

- Assessing the impact of increasing interest for the TEI in “other” communities
- Benefits: stabilizing the TEI as reference standard for text documents
- Challenges: further requirements on the TEI model
- The case of scientific and technical information

SCIENTIFIC INFORMATION?

Characterising scientific documents

- Expert documents describing a specific scientific and technical progress with respect to the state of the art
- Three main domains
  - Scholarly publications
  - Standardisation documents
  - Patents
- Some common characteristics
  - Authorship: the basis of scientific attribution
  - Structure: usually a formal internal organisation
  - Vocabulary: technical terms are essential to convey (or hide) meaning
  - Network of references: relating to the state of the art
  - Certification: workflow, responsibilities, metadata

Authorship

Publications - *The essence of publishing*

- Importance of attribution
- Reflects the context and time of the research (project, affiliation, biography)
- The hidden hand of reviewers

Standards - *Priority to the institution*

- Consensus building => large expert group
- ISO: no authors but project leaders
- W3C: editors

Patents - *A variety of roles*

- Applicant/inventor/representative
- Opponents
- … and *examiners*

Structure

Publications - *Semi-formal*

- Title/authors/affiliations/abstract
- Loosely structured content
- Formulas, Tables, figures, graphics
- References

Standards - *Very formal*

- Introduction/scope/terms and definitions/description/references/annexes
- Formulas, Tables, figures, graphics

Patents – *Very formal*

- Title/inventors/claims/abstract/ description
- Multilingualism (EPO)
- Formulas, Tables, figures, graphics

Language

Publications - *Semi-formal*

- Loose keywords, when any
- Community of practices
- Creativity is part of the publication process…

Standards - *Very formal*

- Central role of the *terms and definitions* section
- Based on the principles of terminology

Patents - *Obfuscating*

- Achieving widest coverage and preventing retrieval

Network

Publications - *Semi-formal*

- References pointing to previous publications in the same domain
- Citation is an essential aspect of scholarly fame…

Standards - *Very formal*

- Section 2: normative references
- Possible additional bibliographic section at the end

Patents - *Very formal*

- Citations in the application description
- Citations as annotations from the examiner
- Impact on acceptance or refusal

Workflow

Publications - *Semi-formal*

- Traditional (vestigial?) concept of peer-review
- From author’s initial manuscript to publisher’s version
- Evolution in the role of each version (e.g. prior art)

Standards - *Very formal*

- Decision process reflecting membership structure
- ISO: WD, CD, DIS, FDIS, IS
- One single reference document

Patents - *Very formal*

- Review by patent examiners
- Coordination of multiple submissions: national, US, Europe, etc.
- Importance of initial submission date

THREE EXISTING SCENARIOS

SCENARIO 1 – INTERFACING A PUBLICATION REPOSITORY

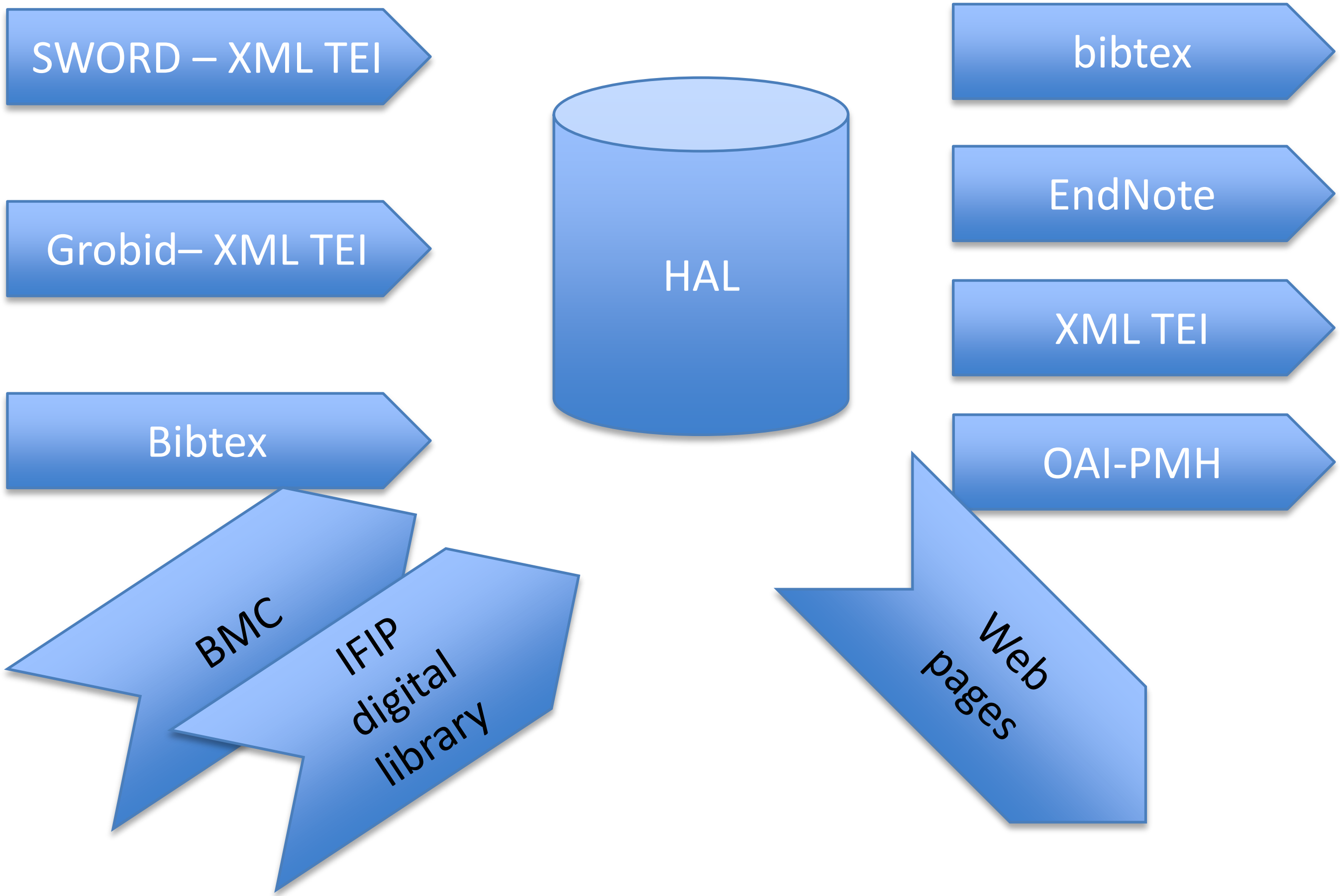
Publication repositories

- An infrastructure for scholarship
  - Open dissemination of scholarly papers
  - Metadata and documents
- An attempt to counter-balance the hegemony of private publishers
  - Cf. Green open access
- An essential tool for assessment and strategic planning
  - E.g. H2020 open access policy

HAL – the French national repository

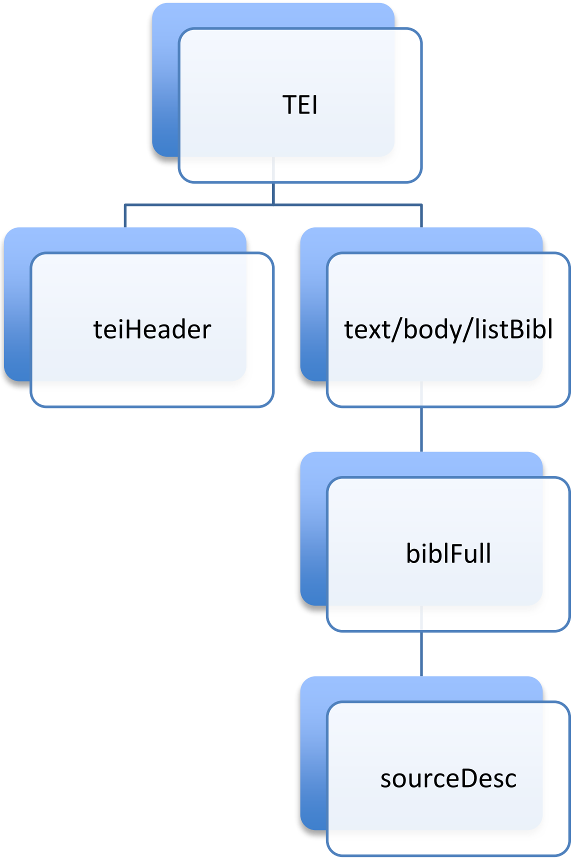
- Developed and maintained at CCSD, Lyon
- Offers portals to various higher education and research institutions
  - E.g. HAL-Inria portal
- Accompanied by incentive measures
  - E.g. Inria’s deposit mandate
- 356 640 documents and 1 039 779 as of 2015-09-16

HAL information ecology



TEI: back-office production format for all exports and imports

The three-tier HAL model



- Source information
- [illegible]
- [illegible]
- [illegible]

Perspective

- The TEI as an in-depth metadata format
  - Going beyond the standard Dublin Core representations available in OAI-PMH interfaces
  - Recording precise data for scientific information (affiliations, licences, etc.)
- Preparing for an extension to full-text management
  - PDF 2 TEI (Grobid)
  - Re-publication framework (HTML, ePub)

SCENARIO 2 – MASSIVE INGESTION OF SCHOLARLY PAPERS

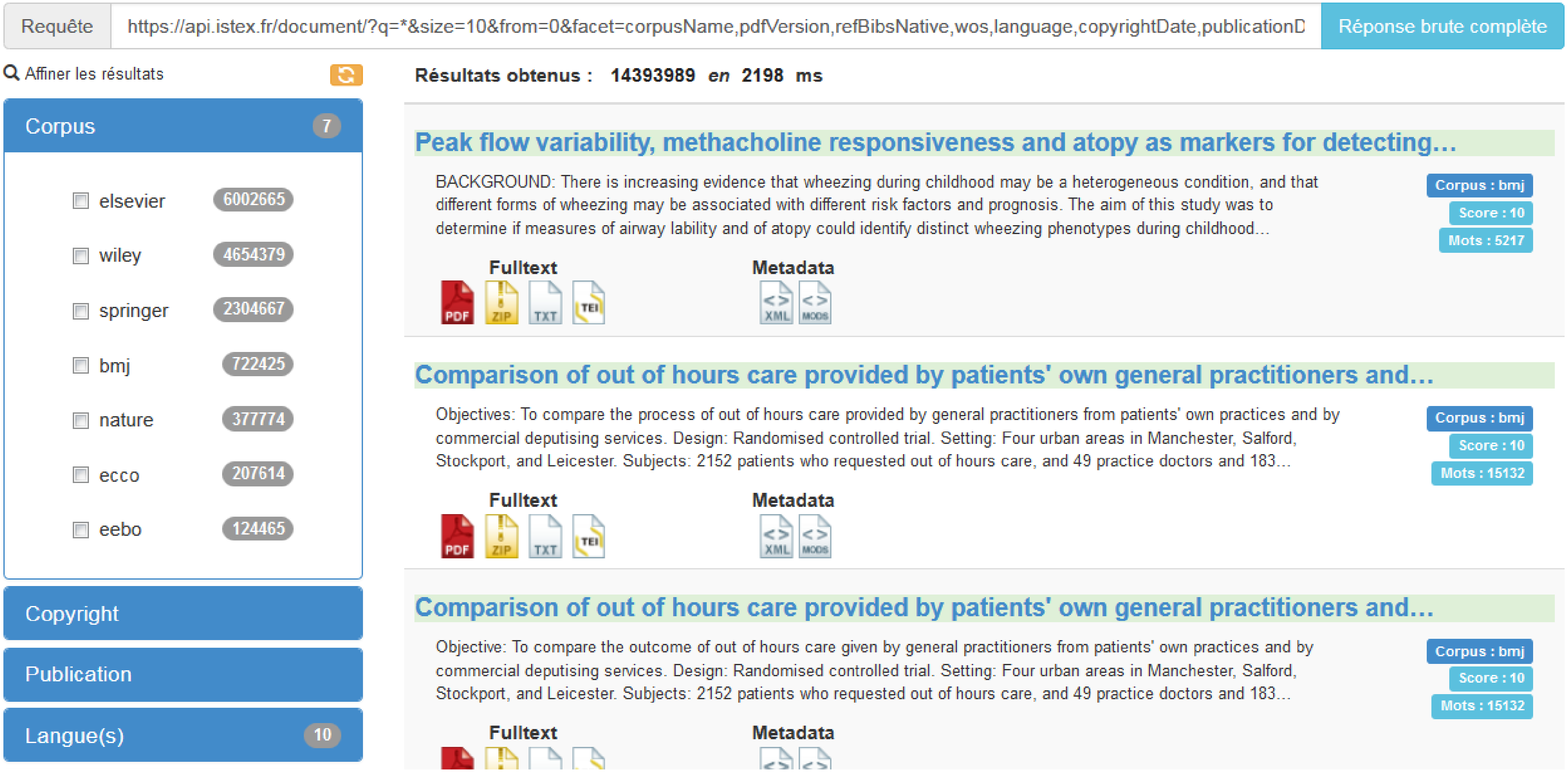
Context

- Big deals and national licences
  - Global negotiation frameworks for journal subscriptions
  - Additional requirements concerning archival and re-use (text and data mining)
- Necessity to put together an infrastructure for ingesting and delivering massive amount of scholarly papers
  - Heterogeneous formats from publishers
  - Homogeneous delivery platform

The national Istex project

- ANR funded project (*Investissement d’avenir*)
- Currently 15 million objects (target: 20M)
- TEI as the target format for all contents
  - Ensuring a continuity from simple meta-data to full-text
- Production lines
  - Conversion of publishers’ formats
    - Metadata, unstructured full-text, structured full-text
  - Automatic meta-data extraction from PDF (Grobid)

The Istex document repository



Requête https://api.istex.fr/document/?q=*&size=10&from=0&facet=corpusName,pdfVersion,refBibsNative,wos,language,copyrightDate,publicationD Réponse brute complète

Affiner les résultats

Corpus 7

□	elsevier	6002665
□	wiley	4654379
□	springer	2304667
□	bmj	722425
□	nature	377774
□	ecco	207614
□	eebo	124465

Copyright

Publication

Langue(s) 10

Résultats obtenus : **14393989** *en* **2198** ms

Peak flow variability, methacholine responsiveness and atopy as markers for detecting…

BACKGROUND: There is increasing evidence that wheezing during childhood may be a heterogeneous condition, and that different forms of wheezing may be associated with different risk factors and prognosis. The aim of this study was to determine if measures of airway lability and of atopy could identify distinct wheezing phenotypes during childhood…

Corpus : bmj
Score : 10
Mots : 5217

Fulltext: PDF ZIP TXT TEI
Metadata: XML MODS

Comparison of out of hours care provided by patients' own general practitioners and…

Objectives: To compare the process of out of hours care provided by general practitioners from patients' own practices and by commercial deputising services. Design: Randomised controlled trial. Setting: Four urban areas in Manchester, Salford, Stockport, and Leicester. Subjects: 2152 patients who requested out of hours care, and 49 practice doctors and 183…

Corpus : bmj
Score : 10
Mots : 15132

Fulltext: PDF ZIP TXT TEI
Metadata: XML MODS

Comparison of out of hours care provided by patients' own general practitioners and…

Objective: To compare the outcome of out of hours care given by general practitioners from patients' own practices and by commercial deputising services. Design: Randomised controlled trial. Setting: Four urban areas in Manchester, Salford, Stockport, and Leicester. Subjects: 2152 patients who requested out of hours care, and 49 practice doctors and 183…

Corpus : bmj
Score : 10
Mots : 15132

Fulltext
Metadata

TEI as a pivot format for interchange

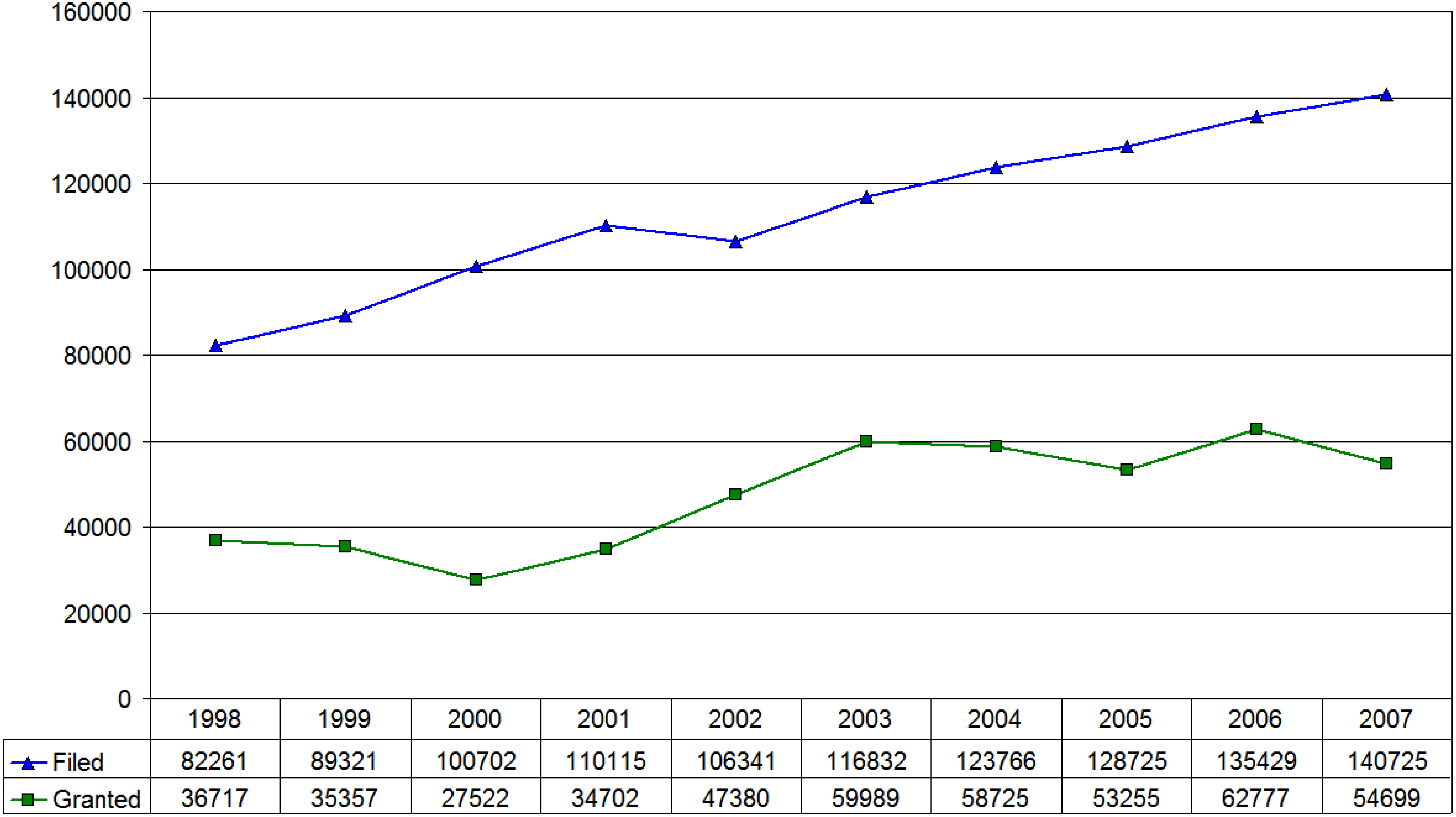
- General strategy: *no information should be lost*
  - Nearly everything in sourceDesc
  - + Keywords, Summary, Copyright
- Strict author description
  - Deep encoding of names
  - Deep encoding of affiliations (Web of Science - 3-level)
  - Deep encoding of addresses – getting the country right
- Precise publishing information
  - Pagination, DOIs, volume, issue, journals name(s)
  - Yes, biblStruct is cool!
- Necessity of constantly adapting the target model
  - (That’s why JATS sucks, if you ask)

SCENARIO 3 – MANAGING THE BACK-OFFICE OF THE EPO

The European Patent Office

- The European one-stop shop for patent applications
- Examination of each application by experts from the field (examiners)
  - Based on existing patents as well as scholarly publications (aka *Non Patent Literature*)
- Some figures
  - Several thousands of examiners
  - 200 million documents
  - 2 billion annotations…

European patent applications and granted patents

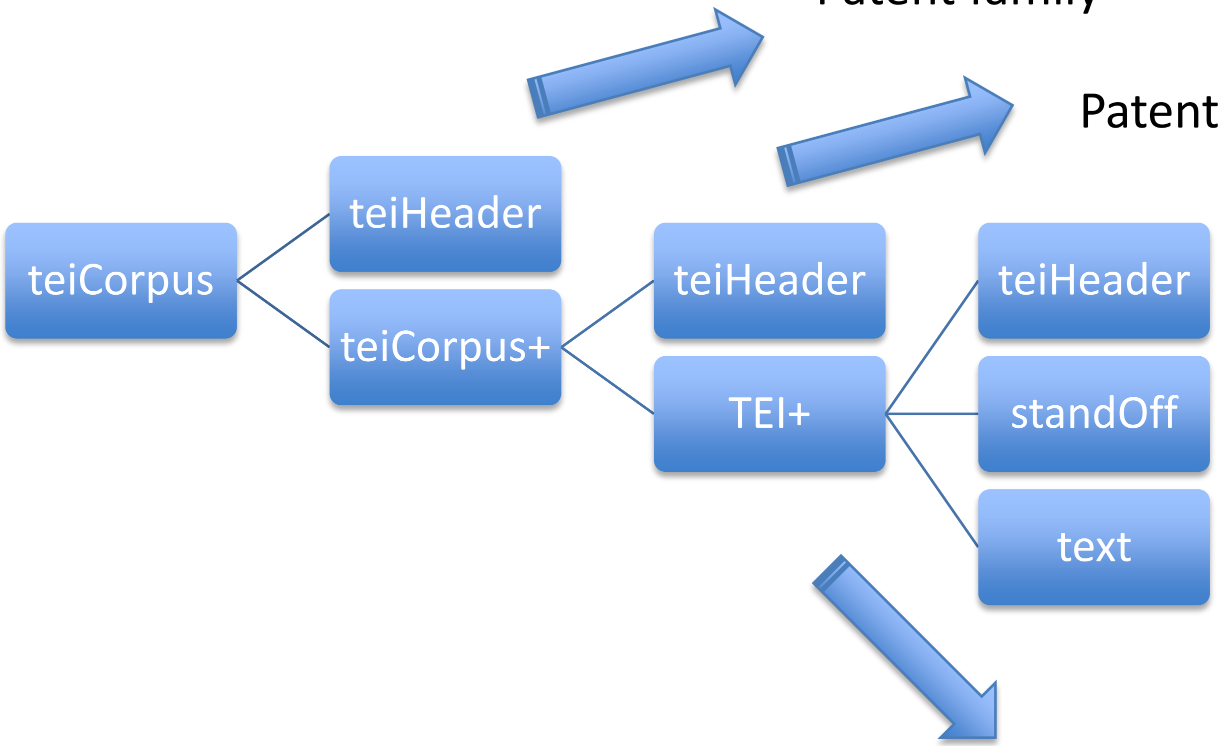


	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
▲ Filed	82261	89321	100702	110115	106341	116832	123766	128725	135429	140725
■ Granted	36717	35357	27522	34702	47380	59989	58725	53255	62777	54699

The (simplified) patent life-cycle

- Patent application in one or several patent offices
  - USPTO, Japan, EPO (directly or initiated in a specific country)
  - First application: reference date for the patent (“coming into force”)
  - Form a “Patent family”
- Examination process for one application
  - Search report, communications, decision, appeal, opposition
  - Patent documents may be revised at each stage
- Necessity to have a single model for dealing with all stages and versions
- Again, the TEI appeared to be the optimal choice

The Patent Document Model



Patent family

Patent application

Patent documents (all versions)

The situation so far

- Complete implementation in the back-office system
  - Integration of several so-far dispersed data-bases
  - First large-scale implementation of <standOff>
- Quite a few customisations – maintained in a reference ODD specification
  - Re-use of TEI attributes at various places
    - @type, @cert, @sortKey
  - Bibliographic references to patents
  - Complex classification mechanism (<classCodeGroup>)
  - <body> in <interp> …
  - Let’s <party>!
- All in all a large scale demonstration of the TEI possibilities

WHY THE TEI?

Why the TEI, indeed?

- Two essential features
  - The very rich TEI vocabulary
    - covers (nearly) all the features needed in a scientific information scenario
  - The huge customisation capabilities offered by the TEI architecture (modules, classes, ODD)
    - Allows one to fine-tune models for highly constrained environments
- One more feature…
  - The short decision cycle of the TEI standardisation process
    - “external” maintenance of most of the schema components and documentation
      - Towards “thin” specifications and documentations

A vision for scientific information

- A family of formats within the TEI information space
  - Sharing common components in a stable environment
  - Comparison of information coverage thanks to the ODDs
  - Avoiding fragmentation
    - The JATS, BITS syndrome, with the argument that “one size does not fit all”
    - There is no one size in the TEI, but a full-blown online catalogue…
- Strategy
  - Pushing compliance as far as one can
  - Strong re-use of TEI components (attribute, elements, classes)
  - Push proposals to the TEI council as early as possible

But life is not always easy…

- Such large-scale applications are a stress-test for the TEI
  - Do we consider these use cases as valid TEI ones
  - How do we process requests for change outside our core TEI culture?
- Illustrating complexity
  - “Simple” TEI evolutions: authors and abstracts
  - Complementing the TEI with missing components
    - Defining a new crystal: <standOff>
    - Blending with external vocabularies: TBX and terminologies

“SIMPLE” CASES — AUTHORS AND ABSTRACTS

Multifarious authors

- The <author> element may be implemented in quite a variety of ways:
  - `<author>Peter Stadler</author>`

  -
    ```
    <author>
      <persName>Peter Stadler</persName>
    </author>
    ```

  -
    ```
    <author>
      <persName>Stadler, Peter</persName>
    </author>
    ```

  -
    ```
    <author>
      <persName>
        <forename>Peter</forename>
        <surname>Stadler</surname>
      </persName>
    </author>
    ```

But an author is more than this

```xml
<author>
 <persName>
   <forename type="first">Laurent</forename> <surname>Romary</surname>
 </persName>
 <email>laurent.romary@inria.fr</email>
 <idno type="halauthor">49567</idno>
 <idno type="ORCID">http://orcid.org/0000-0002-0756-0508</idno>
 <idno type="arXiv">http://arxiv.org/a/Romary_L</idno>
 <affiliation>
   <org type="laboratory”>
     <orgName>Institut National de Recherche en Informatique et en Automatique</orgName>
     <address>
        <addrLine>Domaine de Voluceau-Rocquencourt BP 105 78153 Le Chesnay Cedex</addrLine>
       <country key="FR"/>
     </address>
     <ref type="url">http://www.inria.fr/index.fr.html</ref></org></affiliation></author>
```

Simplified author description attached to a publication in HAL

How do we deal with this?

- Making full use of macro.phraseSeq
- Email, address
  - Already there
- Idno
  - Required a series of ticket (since 2008)
    - Make it an identifier, not a number
    - Allow it in <author> (and aligning <person> :-})
- Affiliation
  - Already there, and you can put quite a lot in it…

Going further — Allowing biographies

- Biographies are a standard component of author descriptions in scholarly publication
  - They may change heavily in size and content within a corpus of publication
  - They may range from a simple text to a structured content
- Two tickets in place
  - 543: make the content of <occupation> more structured
    - Easy…
  - 542: make <occupation> part of <author>
    - Ambered!

Expected example

```
<author>
  <persName>
        <forename type="first">Olivier</forename>
        <surname>Le Deuff</surname>
  </persName>
   <email>oledeuff@gmail.com</email>
   <occupation>
      <p>Docteur en sciences de l’information et de la communication, <hi rend="bold">Olivier
Le Deuff</hi> a soutenu en 2009 une thèse sur <hi rend="italic">La culture de l’information en
reformation</hi>. Il est chercheur au laboratoire PREFics (Plurilinguismes, Représentations,
Expressions Francophones - information, communication, sociolinguistique), une composante du
PRES Université européenne de Bretagne, Université Rennes 2. Il est également webmaster du
Guide des égarés (<ref target="#www.guidedesegares.info"/>).</p>
   </occupation>
</author>
```

Is an author a person or a name?

- The content model of <person> would be an ideal candidate for <author>
  - E.g. <occupation> can be a child of <person>
- Still <author> does not have a clear-cut semantics
  - “***contains*** the name(s) of an author”
    - “TEI Council subgroup unanimously believes that it is incorrect to put information inside <author> other than the author’s name.”
  - So why a wrapper? Only to qualify the person as an author?
  - Why such an elaborate content model?
  - Where to put information attached to authors in a scientific information se case?
    - “(Other information, like e-mail address and occupation, should be inside the <bibl> … that is, <author> is not a substitute for <bibl>, it contains only the name(s).)”
- <author> has been partially conceived in the perspective of the poor <respStmt>

The situation so far

- A huge tension within model.respLike
  - Specific and potentially rich constructs
    - Author, editor, funder, meeting, principal, sponsor
  - A very limited respStmt element (name it or leave it…)
- Reflect a cultural tension
  - "Any more prosopographical information should be stored in the *standard* TEI manner"
- What if we have other types of intervening parties?
  - inventor, applicant, examiner etc.

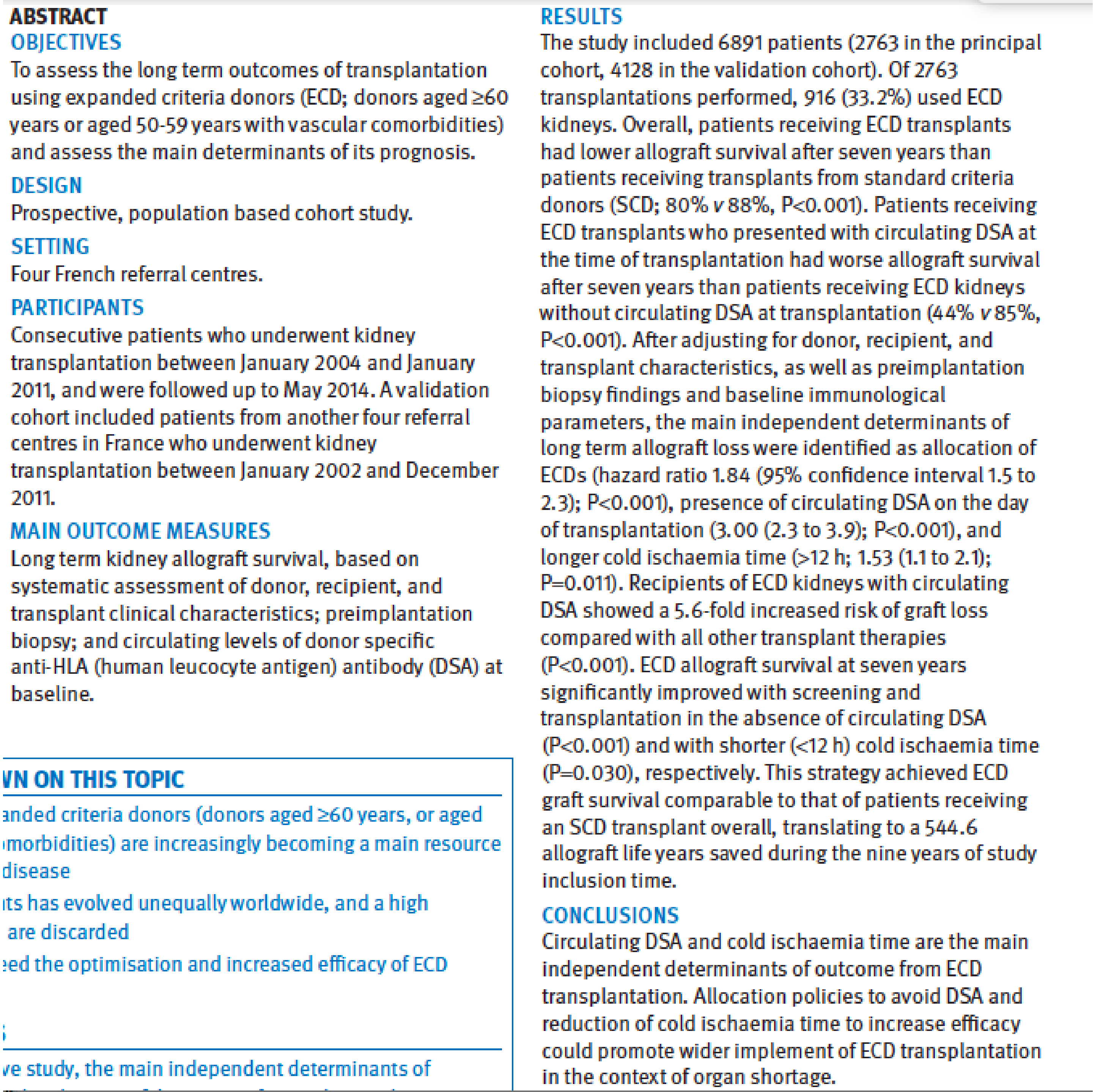
Towards a new element?

- <party> to replace <respStmt> with a more person-like component
- Modelled upon <author>
  - Member of model.respLike
  - Content model: macro.phraseSeq?
- Seen as a generic element for other respLike objects
  - <party role=“author”> = <author>
  - [rem. We need inheritance…]
- Already implemented in the EPO PDM model…
  - <party type=“inventor”></party>

A similar tension with <abstract>

- An abstract is a standard component of the metadata for a scholarly paper
- Initial ticket to get an abstract in the TEI header (471)
  - – Easy achievement in profileDesc
  - – (model.pLike | model.listLike)+
- But soon, real life knocked at the door…

A standard case in the bio-medical domain



ABSTRACT

OBJECTIVES

To assess the long term outcomes of transplantation using expanded criteria donors (ECD; donors aged ≥60 years or aged 50-59 years with vascular comorbidities) and assess the main determinants of its prognosis.

DESIGN

Prospective, population based cohort study.

SETTING

Four French referral centres.

PARTICIPANTS

Consecutive patients who underwent kidney transplantation between January 2004 and January 2011, and were followed up to May 2014. A validation cohort included patients from another four referral centres in France who underwent kidney transplantation between January 2002 and December 2011.

MAIN OUTCOME MEASURES

Long term kidney allograft survival, based on systematic assessment of donor, recipient, and transplant clinical characteristics; preimplantation biopsy; and circulating levels of donor specific anti-HLA (human leucocyte antigen) antibody (DSA) at baseline.

RESULTS

The study included 6891 patients (2763 in the principal cohort, 4128 in the validation cohort). Of 2763 transplantations performed, 916 (33.2%) used ECD kidneys. Overall, patients receiving ECD transplants had lower allograft survival after seven years than patients receiving transplants from standard criteria donors (SCD; 80% *v* 88%, $P<0.001$). Patients receiving ECD transplants who presented with circulating DSA at the time of transplantation had worse allograft survival after seven years than patients receiving ECD kidneys without circulating DSA at transplantation (44% *v* 85%, $P<0.001$). After adjusting for donor, recipient, and transplant characteristics, as well as preimplantation biopsy findings and baseline immunological parameters, the main independent determinants of long term allograft loss were identified as allocation of ECDs (hazard ratio 1.84 (95% confidence interval 1.5 to 2.3); $P<0.001$), presence of circulating DSA on the day of transplantation (3.00 (2.3 to 3.9); $P<0.001$), and longer cold ischaemia time (>12 h; 1.53 (1.1 to 2.1); $P=0.011$). Recipients of ECD kidneys with circulating DSA showed a 5.6-fold increased risk of graft loss compared with all other transplant therapies ($P<0.001$). ECD allograft survival at seven years significantly improved with screening and transplantation in the absence of circulating DSA ($P<0.001$) and with shorter (<12 h) cold ischaemia time ($P=0.030$), respectively. This strategy achieved ECD graft survival comparable to that of patients receiving an SCD transplant overall, translating to a 544.6 allograft life years saved during the nine years of study inclusion time.

CONCLUSIONS

Circulating DSA and cold ischaemia time are the main independent determinants of outcome from ECD transplantation. Allocation policies to avoid DSA and reduction of cold ischaemia time to increase efficacy could promote wider implement of ECD transplantation in the context of organ shortage.

WN ON THIS TOPIC

anded criteria donors (donors aged ≥60 years, or aged
morbidities) are increasingly becoming a main resource
disease

its has evolved unequally worldwide, and a high
are discarded

eed the optimisation and increased efficacy of ECD

;

ve study, the main independent determinants of

BMJ 2015;351:h35 57 | doi: 10 .113 6/bmj.h3557

Another cultural clash

- Ticket 548: request to have divisions in <abstract>
- No strong enthusiasm…
  - “quite a major change”/”transgressive”/”very undesirable change”/”put it in the front”
    - Hack with list, typed ab, etc.?
  - Having <div> in the header is a total protonic reversal!
- Project: designing a div-like element specific to the header

THE RETURN OF TBX

TBX?

- Term Base eXchange
- Implement an onomasiological model for lexical data
  - Concept to term
- Published as an ISO standard (ISO 30042, building upon ISO 16642)
- A long time ago in another galaxy…
  - Was initiated within the TEI guidelines
- Bringing it back to the TEI is more than a cool thing to do
  - But a nice ODD exercise and a nightmare for the council…

TEI and onomasiological representations

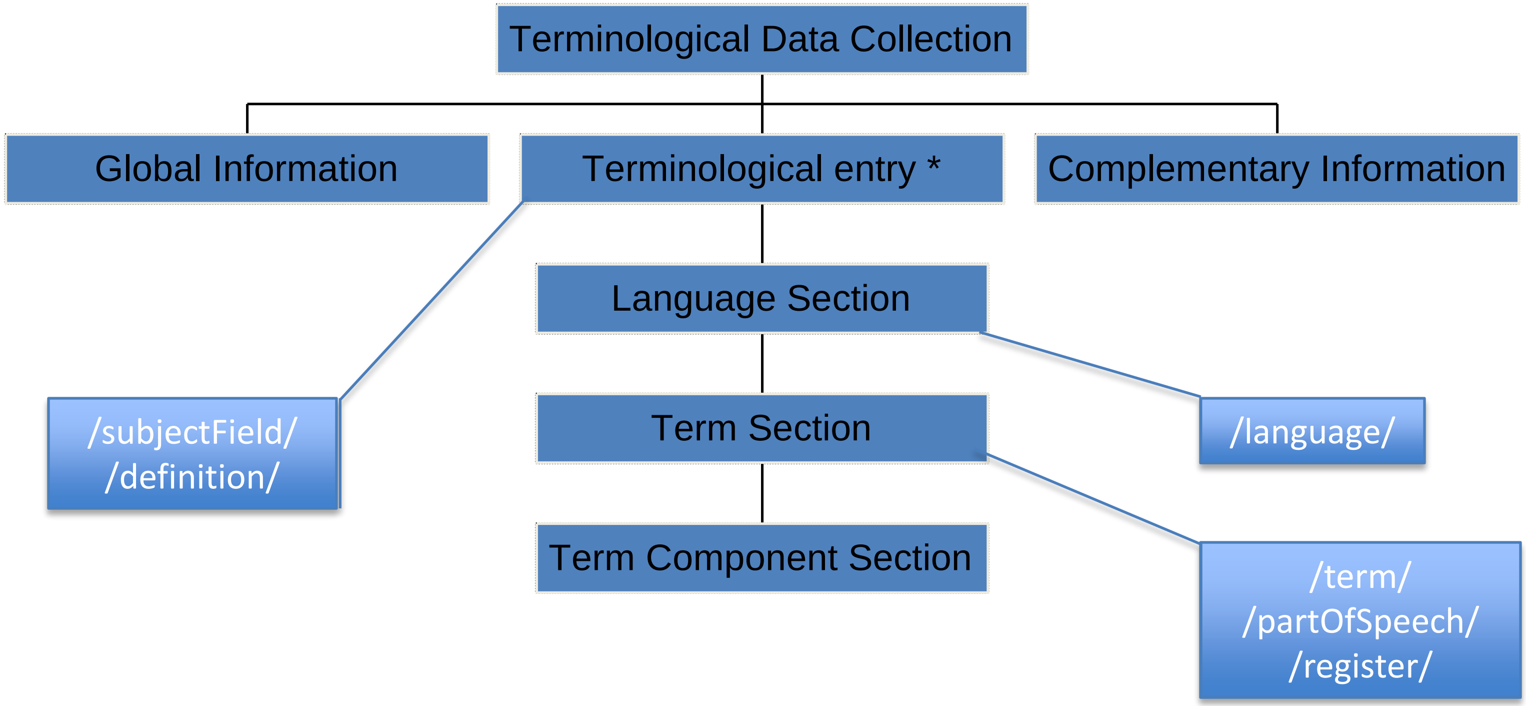
- ISO 6156:1987 (Mater)
- 1989: Setting up the TEI
  - Specific chapter of the TEI guidelines dedicated to the representation of terminological data
  - The SGML-based representation integrated in the TEI framework remained there until the P4 edition
- ISO 12200 (Martif): 1999, improves the TEI proposal (bracketing), but breaks the link to the TEI by going ISO
  - document structure strongly inspired from the TEI (e.g. the header-text organisation; entries embedded within a <text> and <body> hierarchy);
  - reaching out to the translation and localisation industry
- 1999-2003: Abstracting away
  - ISO 12620:1999: data categories and ISO 16642 (TMF): 2003, meta-model
  - Basic for the specification of a variety of terminological formats
- ISO 30042:2008 TBX (TermBase eXchange) , after work carried out in LISA
- Current: TBX-Basic, TBX-Min …

Conceptual Principles

- Concept orientation
  - All terminological information pertaining to one concept including all terms (designing this concept) in all languages and all descriptive and administrative data must be handled as one terminological unit
- Term autonomy
  - All terms belonging to one concept should be managed (in one terminological entry) as autonomous (repeatable) blocks of data categories without any preference for a specific term
    - ≠ thesaurus

ISO 704:2009 Terminology work -- Principles and methods

ISO 16642 (TMF) meta-model



Serializing TMF in TBX (ISO 30042)

```
<termEntry>
  <descrip type=“subjectField”>Medicine</descrip>
  <descrip type=“definition”>An acute viral infection involving the respiratory tract. It is marked by inflammation of the NASAL MUCOSA; the PHARYNX; and conjunctiva, and by headache and severe, often generalized, myalgia.(MeSH)</descrip>
  <langSet>
    <tig xml:lang='fr'>
      <term>grippe</term>
      <descrip type=“partOfSpeech”>Noun</descrip>
     <descrip type=“register”>vernacular</descrip>
    </tig>
  </langSet>
  <langSet>
    <tig xml:lang='en'>
      <term>influenza</term>
      <descrip type=“partOfSpeech”>Noun</descrip>
      <descrip type=“register”>all</descrip>
    </tig>
  </langSet>
</termEntry>
```

Implementing a TBX based extension for the TEI guidelines

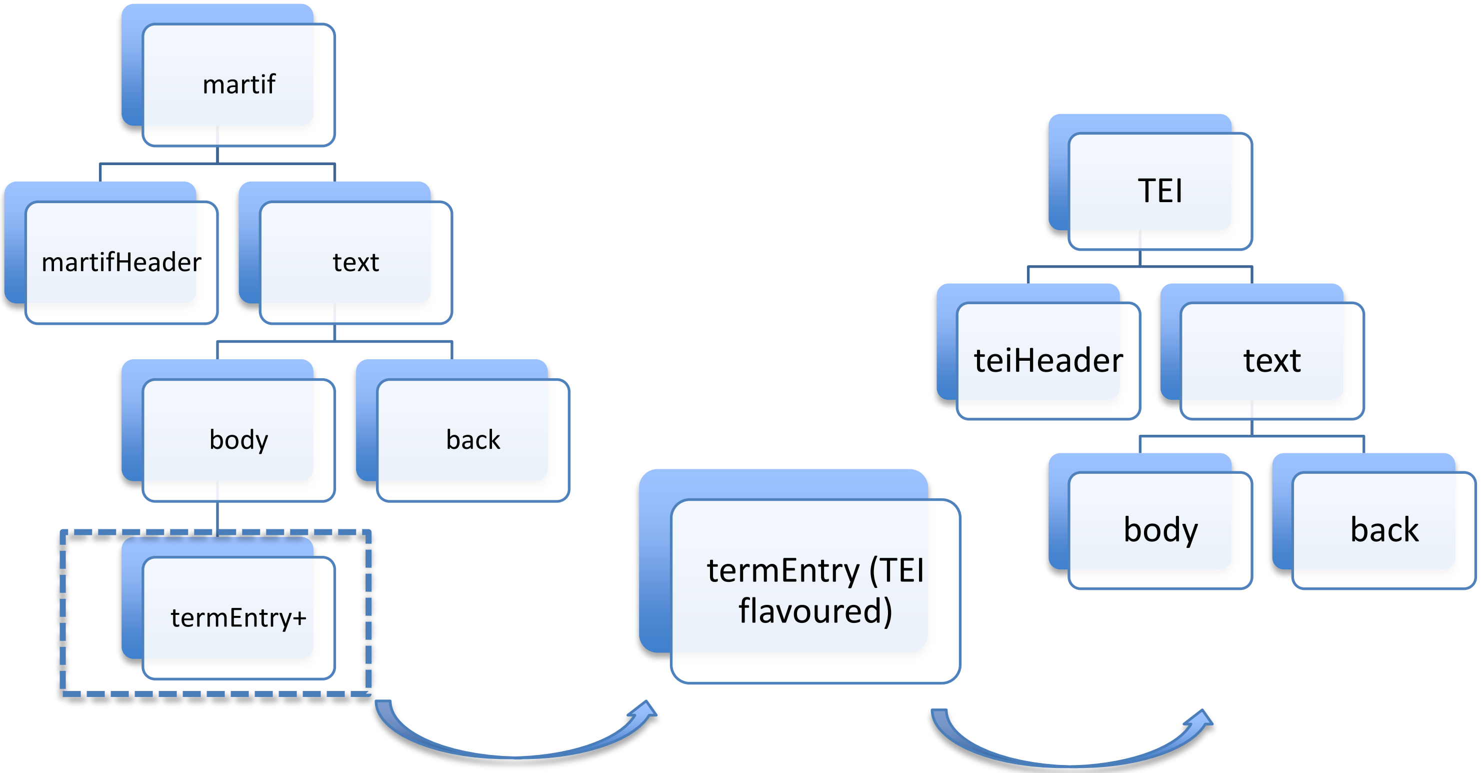
- Addressing a new (?) user community: digital humanists
  - Importance of primary sources and construction of secondary digital objects (annotations, indexes, etc.)
  - Onomasiological sources in field linguistics
- Taking the best from the onomasiological work in the last 40 years
  - Avoiding simplistic representations such as SKOS and thesaurus standards
- Fostering more convergence in standardization
  - Favoring reuse of components from various standardization worlds

- But do we have any kind of histocompatibility?

Tissue typing

- Host: the TEI document structure
  - Terminological entries can occur at many places
    - Specific section, inline, between other TEI elements
  - As far as building up a terminological database in TEI
    - E.g. recording bilingual philosophical vocabulary from Wittgenstein's works
  - Keeping all inline annotation facilities for textual fields
    - Names, dates, foreign expressions, notes, pointers, feature structures…
  - Improved documentation with the rich TEI header
    - Important from a scholarly perspective
- Graft: a TBX-like terminological entry
  - Structural skeleton
    - Inspired from TBX-Basic (DCA style)
    - Note that TBX already has an ODD spec!
  - Data categories
    - Initially reduced to a very small number of meaningful categories for a DH scenario
    - subjectField, definition, source, partOfSpeech, grammaticalGender, etc.
    - In particular: no project management data categories

The transplant process



martif

martifHeader

text

body

back

termEntry+

termEntry (TEI flavoured)

TEI

teiHeader

text

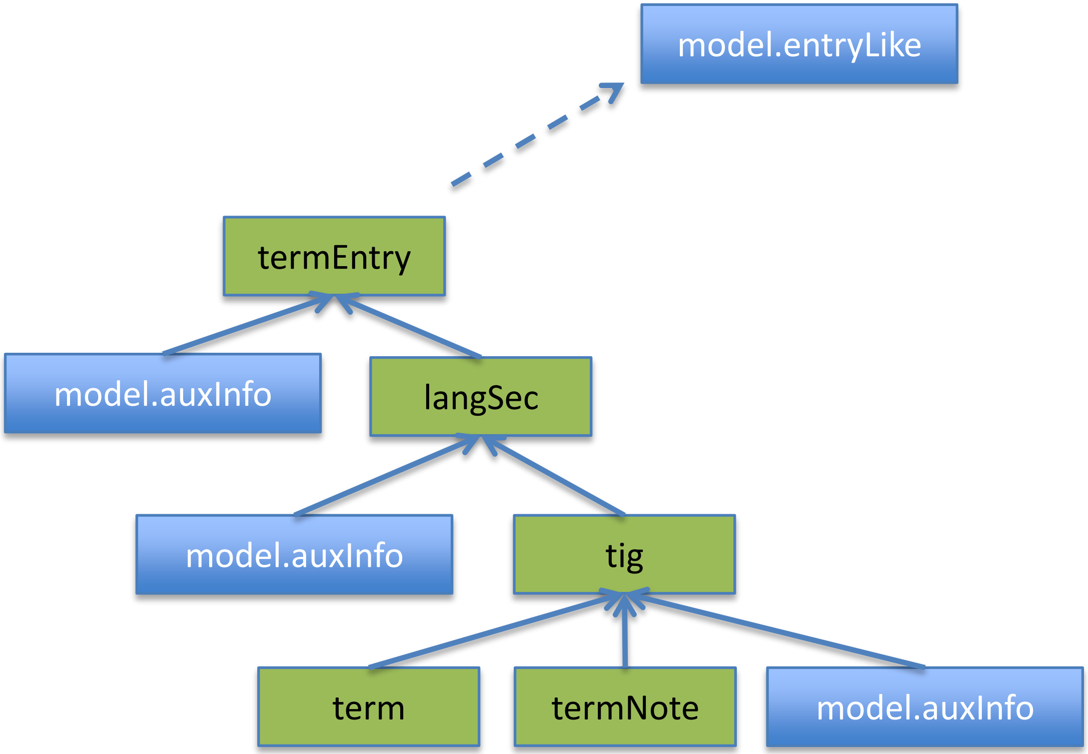
body

back

Harvest and adapt

Insert wherever <entry> can occur

TBX in the ODD architecture



Surgery report

- Ensuring the graft by means of namespaces
- Incompatible tissues
  - Attributes
    - att.global attribute class: @xml:id, @xml:lang, @xml:base, @xml:space
    - @target => att.pointing: making ID/IDREF be URI
  - Outdated element
    - <tbx:xref> (cf. URI mechanism)
  - TEI elements in their own namespace
    - <tei:term>
    - <tei:hi>: bringing the semantic back on tracks
    - <tei:ref>, <tei:ptr>, <tei:note>
- Second life
  - Rich textual content model

The patient after surgery

```
<termEntry xmlns="http://www.tbx.org">
 <descrip type="subjectField" xml:lang="fr">Industrie mécanique</descrip>
 <langSet xml:lang="de">
  <descripGrp>
   <descrip type="definition">endloser Riemen …</descrip>
   <admin type="source">De Coster, Wörterbuch, …</admin>
  </descripGrp>
  <tei:note>wird zum Antrieb der Lichtmaschine, des Ventilators …</tei:note>
  <tig>
    <tei:term>Keilriemen</tei:term>
    <admin type="source">De Coster, …</admin>
  </tig>
 </langSet>
 <langSet xml:lang="fr”>… </langSet>
</termEntry>
```

A chimera?

- TBX@TEI is not a proper subset of TBX
  - Document structure
  - Changes in the content model of <termEntry>
- Still:
  - Can be used to generate TBX compatible data univocally
  - Is probably the most optimal way the get DH colleagues to be acquainted with good terminological practices
- And the customization can be customized!
  - Getting rid of unwanted TEI objects for TBX afficionados
  - Providing description of more complex data categories

Towards a new chapter in the TEI guidelines? Or maintenance of an official customization?

A NEW MODEL FOR EMBEDDED STAND-OFF ANNOTATIONS

A long-standing issue

- Stand-off annotation has been a core concept in the TEI guidelines since the beginning
  - Cf. Chapter: Linking, Segmentation, and Alignment
- But: not integrated in the TEI architecture
  - Stand-off elements can appear anywhere in a TEI document
  - Trade-off between on-site vs. grouping (<back>)
- Plus: An old conflict with the NLP community
- Need for a proper, and inclusive, treatment of stand-off annotations

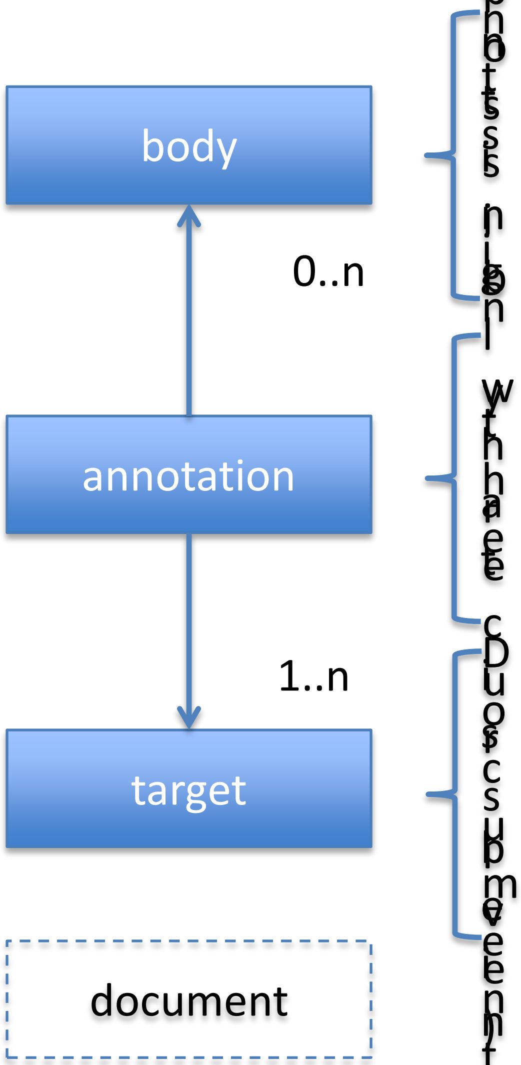
Basic concept

- Building up an autonomous document containing primary source and additional annotations
  - Annotations are conveyed with their specific meta-data
  - Stand-off annotations may be recursively organized
  - Stand-off annotations may point to textual as well as facsimile content
  - Coherence with existing models (Open Annotation, ISO TC 37) should be ensured
- Typical use-cases
  - Human annotations on a document
    - critical editions, patent examination, open peer review
    - Internal prosopography, entities at large
  - Text mining
    - Named entity recognition, keyword/terms extraction
  - Annotated corpora
    - Treebanks
- Strong relation with interlinear annotation

Timeline

- August 2012: new tickets by Javier Pose (EPO)
- January 2014: Workshop in Berlin
  - Draft of a first proposal
  - Setting-up a github environment
- May 2015: Council meeting in Ann Arbor
  - Several updates to the proposal

The Open Annotation model – TEI (possible) implementation


body

0..n

annotation

1..n

target

document

<bibl>, <person>, <place>, <fs>, <note>, <body>, MAF, SynAF

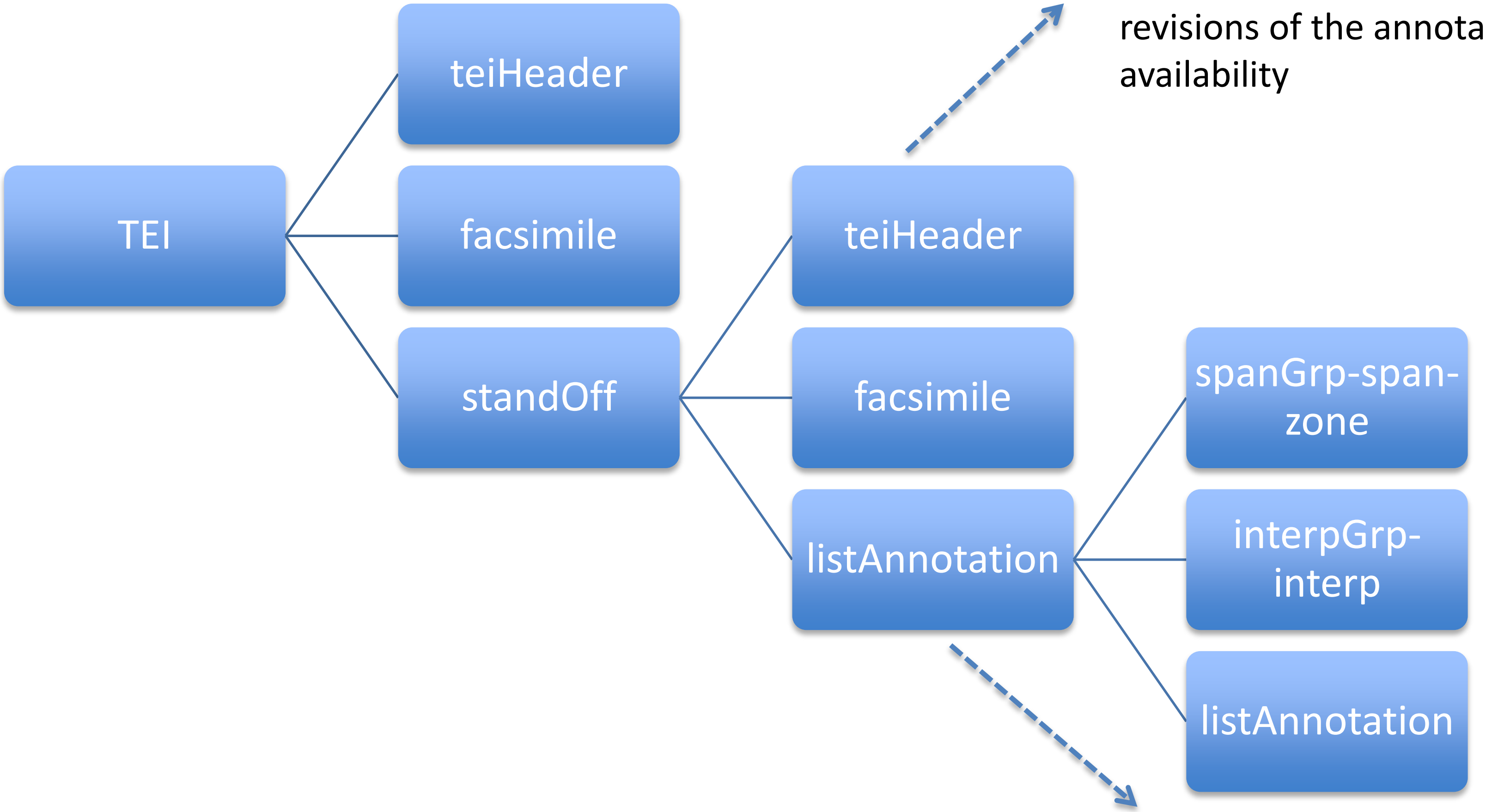
<interp type=“” inst=”” ana=“”>

<span type=“” from=“” to=“”>

<zone type="" corresp="#_theSurface"
 ulx="1253" uly="802” lrx="22" lry="29"/>

Any TEI object (with @xml:id) or <surface>

Annotations in TEI: <standOff>



Meta-data related to the annotation, such as annotator, revisions of the annotations, availability

Recursive construct: allows the organisation of annotations par method, annotator, campaign

standOff – main components

- listAnnotation — a double semantics
  - Equivalent to <text> in TEI document
  - Groups elementary annotation chunks
- span — business as usual
  - Identifies a markable within the full-text of the document
  - Need to improve guidance about the use of pointers
- interp — extended usage
  - Attributes
    - @type: provides the type of the annotation
      - Cf. @type on the parent standOff element
    - @resp: the entity who is responsible for this annotation
    - @inst: lists the components (span or surface) to be annotated
    - @ana: points to annotation content (body in OA speak)

Application: interlinear annotation

- Encoding interlinear annotation as inline content (in <text>)

```
<listAnnotation who="#SPK0" start="#T9" end="#T12" xml:id="au1“>
  <u xml:id="u1">
    <seg xml:id="seg45" type="utterance" subtype="declarative">
      <w xml:id="w43">Nee</w> <pc xml:id="pc3">,</pc> <w xml:id="w44">hab</w> <w
xml:id="w45">kein</w> <w xml:id="w46">Führerschein</w>
    </seg>
  </u>
  <spanGrp type="en">
    <span from="#T9" to="#T12">No, I don't have a driver's license.</span>
  </spanGrp>
  <spanGrp type="pos“>
    <span from="#w43" to="#w43">NE</span>
    <span from="#pc3" to="#pc3">$,</span>
    <span from="#w44" to="#w44">VAIMP</span>
    <span from="#w45" to="#w45">PIAT</span>
    <span from="#w46" to="#w46">NN</span>
  </spanGrp>
</listAnnotation>
```

Application: interlinear annotation

- Encoding interlinear annotation as stand-off markup
  - In <standOff>

```
<listAnnotation corresp="#u1">
        <spanGrp xmlns="http://www.tei-c.org/ns/1.0" type="en">
          <span from="#T9" to="#T12">No, I don't have a driver's license.</span>
        </spanGrp>
        <spanGrp xmlns="http://www.tei-c.org/ns/1.0" type="pos">
          <span from="#w43" to="#w43">NE</span>
          <span from="#pc3" to="#pc3">$,</span>
          <span from="#w44" to="#w44">VAIMP</span>
          <span from="#w45" to="#w45">PIAT</span>
          <span from="#w46" to="#w46">NN</span>
        </spanGrp>
</listAnnotation>
```

  - In <body>

```
<u xml:id="u1" who="#SPK0" start="#T9" end="#T12">
        <seg xml:id="seg45" type="utterance" subtype="declarative">
          <w xml:id="w43">Nee</w><pc xml:id="pc3">,</pc>
          <w xml:id="w44">hab</w> <w xml:id="w45">kein</w> <w xml:id="w46">Führerschein</w>
        </seg>
</u>
```

Issues (many)

- Which header do we need?
  - Standoff annotation requires very specific meta-data
  - If we adopt the TEI header, we need to make it more flexible…
    - Should we have a convergence with biblFull (where profileDesc is missed, BTW, SF:533, deeply ambered)
  - Stand-off annotations may be generated by humans and machines
    - how to put <author> (editionStmt) and <appInfo> (encodingDesc) at the same place?
- How do we provide guidance concerning annotations?
  - Mapping the OA model to precise TEI constructs?
  - Allowing a wide variety of possible vocabularies depending on the use case?
    - TBX entries, MathML, full-text annotation (<body>?)

WHERE DO WE GO FROM HERE?

Reflecting on our practices

- Cf. missed opportunities
  - ISO could have gone TEI, we could have prevented JATS to occur, and yes the Computational linguistic community should have been in since ages…
- Going for more inclusiveness
  - Avoiding fragmentation – the TEI should not fork too much
  - Be ready to adapt (note: none of the proposed changes are backward incompatible…)
- Evolutions at the benefit of everyone
  - Convergence between primary and secondary sources
- Organisational point of view
  - Shall we acknowledge external vocabularies as part of the TEI architecture?
  - Shall we host external document types in the TEI framework?
  - Shall we go towards sub-committee in the TEI technical council?
  - How do we maintain global coherence?
- A strategy of anticipation?

- [Gabriel Bodard] Just to note, Carrie, that this has already been implemented in the latest EpiDoc release (try validating against http://www.stoa.org/epidoc/schema/latest/tei-epidoc.rng and see if it works for you), in anticipation of forthcoming TEI compliance...

MERCI!