The TEI as a modeling infrastructure: TEI beyond the TEI realms

Laurent Romary

The TEI as a modelling infrastructure: TEI beyond the TEI waters

Laurent Romary, Inria

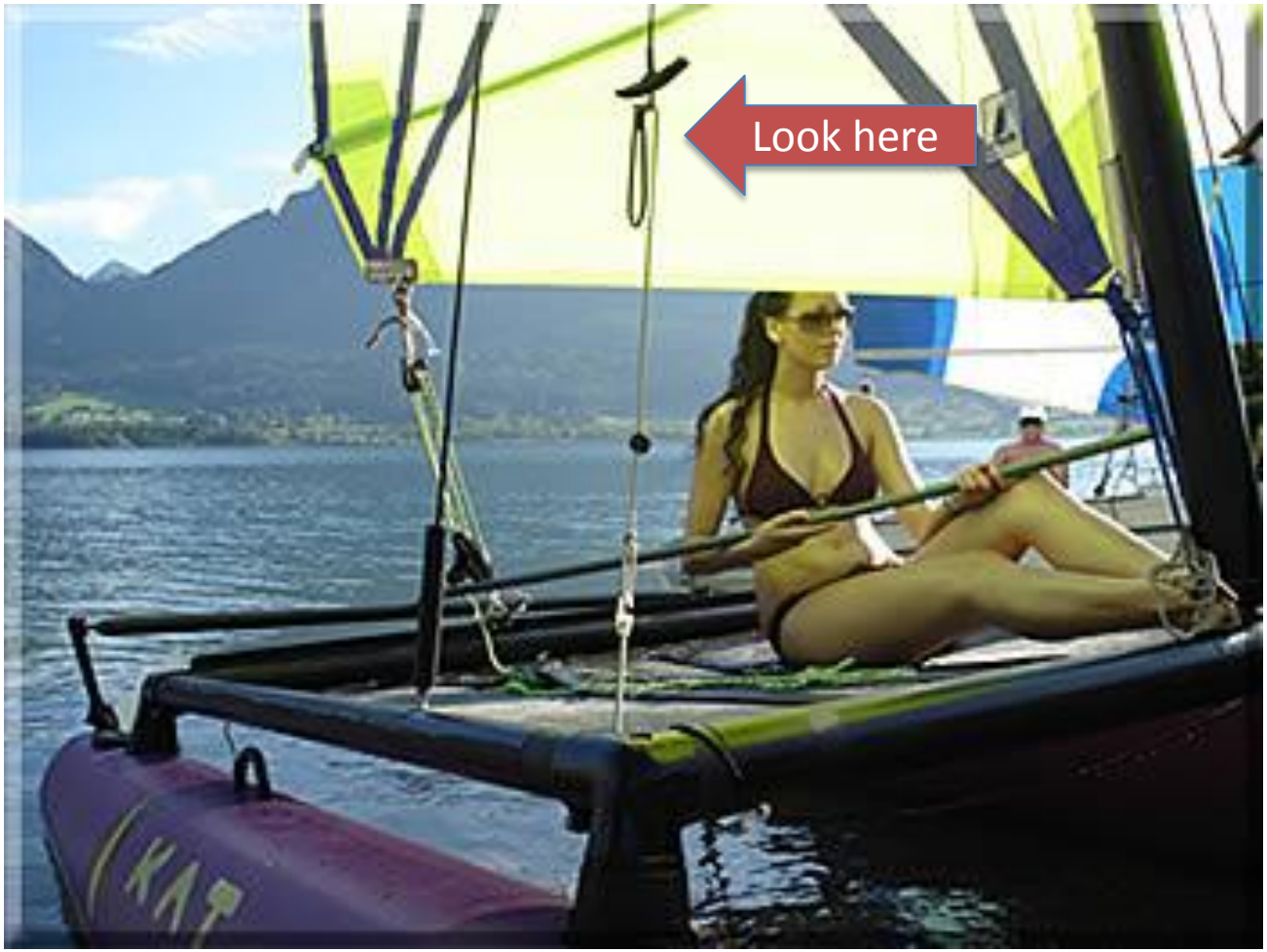
team ALMAnaCH

Overview of the cruise

- The TEI as a standard: what does it mean?
- TEI resilience in a variety of contexts
  - Modelling complex families of document at EPO
  - TEI Lex 0: Tightening the dictionary chapter
  - Welcoming foreign vocabularies: EAD in ODD
- Whither TEI?

STANDARDS: UNDERSTANDING WHAT THEY ARE

Sailing is relaxing



Sailing is fun

Really fun…

Up to a point…

Technical problem

- Trapeze
  - harness + wire attached to the upper part of the mast
  - Allows sailors to hang outside the boat
    - Increase control of speed and craft
- Incidents by sailing involving trapeze wires
  - Risk of being trapped underwater
  - Sailors must be able to detach themselves at any time

ISO 10862

- *Small craft – Quick release system for trapeze harness* (ISO/TC 188: Small craft)
- An ISO standard to prevent death and injury to sailors attached to sailing trapezes on small craft, by ensuring that they can release themselves from the wire hooking them to the boat in emergency
- *Requirements and test methods* to ensure the correct operation of safety release devices
  - Release time shorter than 5 sec.
  - Device operable with only one hand (and with full finger neoprene gloves)
  - Safety mechanism should not be released inadvertently

Note: In Europe, harnesses complying with the standard will bear the CE mark

RWO Trapeze Bar

Standards

- What they are
  - Reference background for the management of a technical process
  - The three pillars of a standard: **consensus**, **publicity**, **maintenance**
  - A small constraint for each, a large benefit for all!
    - Reading it, understanding it, implementing it…
- What they aren't
  - Something coming from nowhere… participating is always an option
  - Mandatory documents: common ground for a transaction
  - Regulations: unless a state or an international organisation takes it up
- The need for standards developing organizations (SDO)
  - Ensure that the three pillars of standardisation are in place

CAN THE TEI BE CONSIDERED AS PART OF THE STANDARDS FLEET?

The Text Encoding Initiative Consortium

- Mission and organisation
  - Develops guidelines for the representation of text-based documents
  - Community based: institutional and individual membership
- Standardisation process
  - 2 releases per year of the TEI guidelines
  - Community feedback on Github
  - Triage and decision made by TEI technical council
  - All documents are freely available (CC-BY+BSD 2-clause license)
- Vertical standardisation
  - Anything that has to do with text
  - Worked on infrastructural aspects by necessity (ODD)
  - The TEI community had a seminal role in setting up XML
- Business model
  - Contributions of the members to the consortium

In the beginning

Text archives
Humanities
Standards
SGML

Not intended (immediately) for individual scholars

1. Novembre 1987: Vassar College, Poughkeepsie

A quick historical overview

- 1960's — GML (Generalized Markup Language) by IBM
- 1970's & 1980's — ANSI initiates project to develop a Standard text-description language based on GML
- 1983 — SGML becomes an industry standard
- 1986 — SGML (Standard Generalized Markup Language) becomes an ISO standard: ISO 8879:1986
- 1987 — TEI (Text Encoding Initiative)
- 1990 — HTML 1.0 (HyperText Markup Language)
- 1992 — TEI edition P3 (Michael Sperberg-McQueen and Lou Burnard, eds)
- 1997/1998 — XML 1.0 (eXtensible Markup Language) (Tim Bray, Jean Paoli and Michael Sperberg-McQueen, eds)

TEI in a nutshell

- TEI namespace:
  - xmlns="http://www.tei-c.org/ns/1.0"
- TEI documentation:
  - http://www.tei-c.org/release/doc/tei-p5-doc/en/html/
- TEI processor, Roma:
  - http://www.tei-c.org/Roma/
- TEI document model
  - Read: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DS.html
- TEI architecture: modules, classes
- TEI vocabulary: more than 500 elements…
  - Read: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/CO.html

TEI –core principles (1)

- The TEI document as a digital surrogate of a physical source
  - A TEI document is always part of a digital library workflow
    - Source – surrogate – enrichment – publication
    - Documented in the header; encoded in the content
  - Born digital documents may as well encounter a succession of changes/versions
- The TEI document as an autonomous object in a DL workflow
  - Embedded meta-data + content
  - Multiple “hands”: annotations

TEI –core principles (2)

- Favoring the semantics rather than the layout
  - (quasi) No presentational construct
  - Publication requires a transformation stage (XSLT; ePub, pdf, HTML, etc.)
- Document structure (model of a “text”)
  - Macro-structure: front-body-back
  - Meso-structure: divisions, paragraphs/lists/figures/etc.
  - Micro-structure: in-line annotation mechanisms
    - Dates, names, notes, references, foreign expressions, etc.

The TEI guidelines

- Online documentation
  - Prose description organized in chapters
  - Specific documentation for each element
  - Access to all examples from the guidelines
- Schema(s)
  - RelaxNG, W3C (, DTD)
  - Available online from the *Roma* interface
  - Delivered as packages (Ubuntu, Oxygen)
- The TEI guidelines as specifications
  - Documentation and schemas are generated from one single specification file
    - Expressed in a TEI sub-language: ODD (One Document Does it all)

The central role of customization

- Each TEI project starts with the definition of a customization
  - Module selection
  - Sub-setting elements
  - Reducing possible values or content models
  - Adding, when necessary, new descriptive object
- ODD as the technical platform for customization

Consequences

- Family of formats
  - Comparison of two TEI-based projects through their ODDs
- Support for third-party projects
  - In-house maintenance of customization and documentation
    - E.g. DTAbF at the Berlin Brandenburg Academy of Sciences
  - Even non TEI application!
    - E.g. EAD in ODD
- Customization as the seed for the guidelines evolution
  - Changes introduced for a specific project may be doomed to be useful for a wider community

EXPLORING NEW WATERS: TEI FOR 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
- Ad hoc representation formats exist, e.g. JATS
  - But they lack resilience…

The European Patent Office

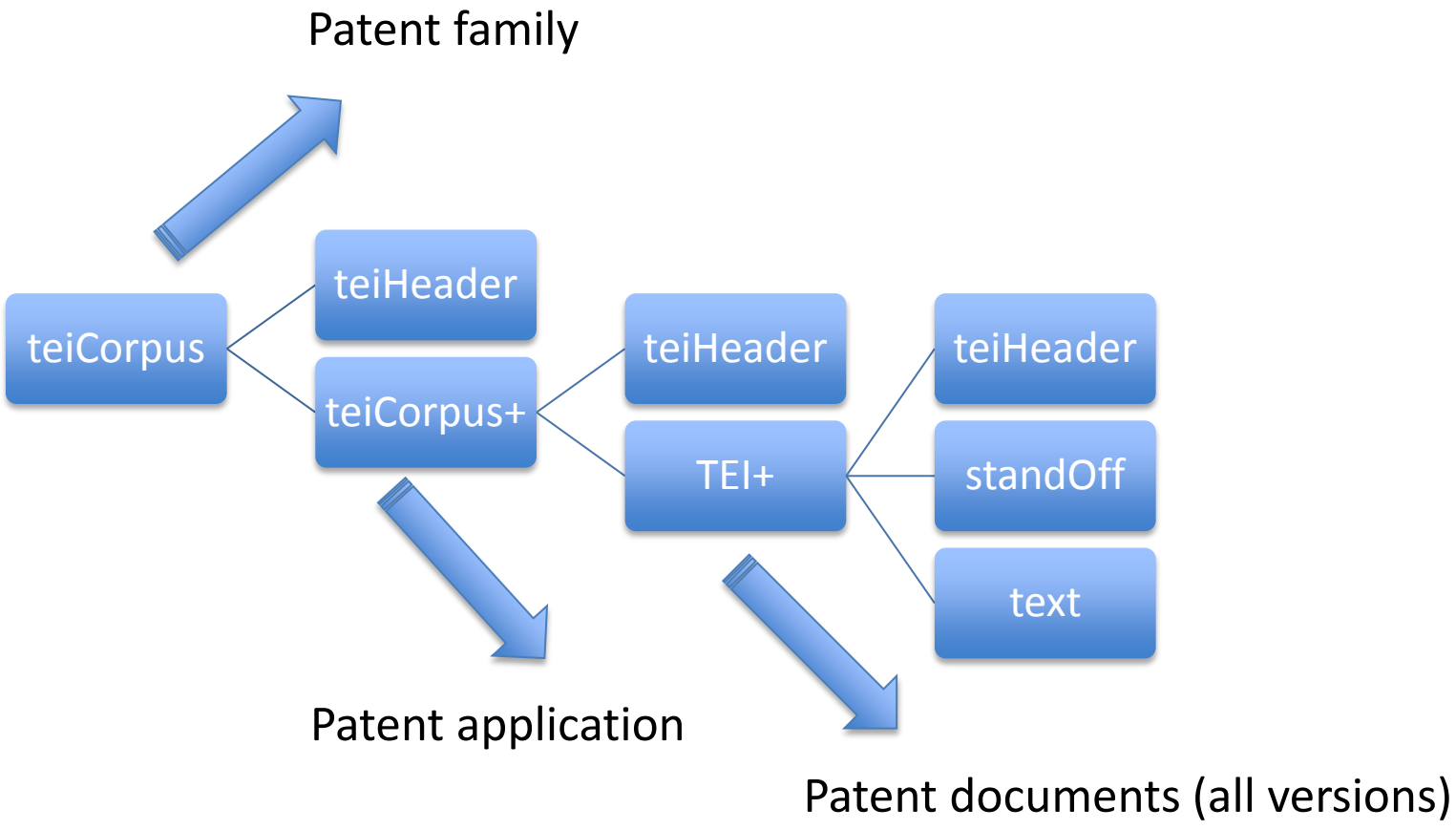
- The European one-stop shop for patent applications
- Examination of each application by experts from the field (examiners)
  - Based on existing patents as well as scholarly publications (aka *Non Patent Literature*)
- Some figures
  - Several thousands of examiners
  - 200 million documents
  - 2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

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”)
  - Forms 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

- The TEI appeared to be the optimal choice

The Patent Document Model



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> (code name: bePatient!)
- 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>)
  - Alternative components: e.g. CALS table model
- Next steps
  - All scholarly publications (NPL)
  - All official communications

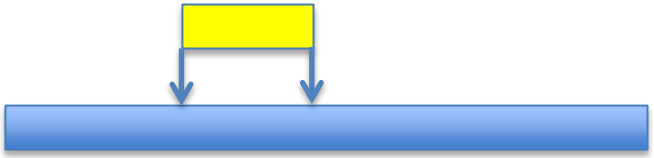
With thanks to the stand-off gang…

INTEGRATION OF THE STAND-OFF PROPOSAL IN THE PDM MODEL

The simple picture

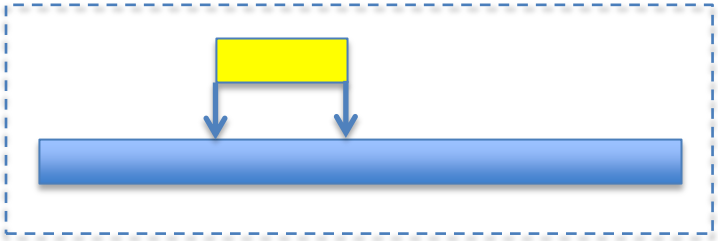
Inline annotation:

Intertwined with the source text



Stand off annotation:

Source text is referenced from outside



Embedded stand off annotation:

Stand off annotations attached to the same document as the source

Why embedded stand-off annotation?

- In line (!) with the TEI philosophy
- Each time the source document is seen as the reference organisational unit
  - Corpus management
  - Transmission workflow
  - Multiple annotation layers
  - Competing annotations
    - E.g. Manual vs. automatic annotation

Standoff: A long-standing issue

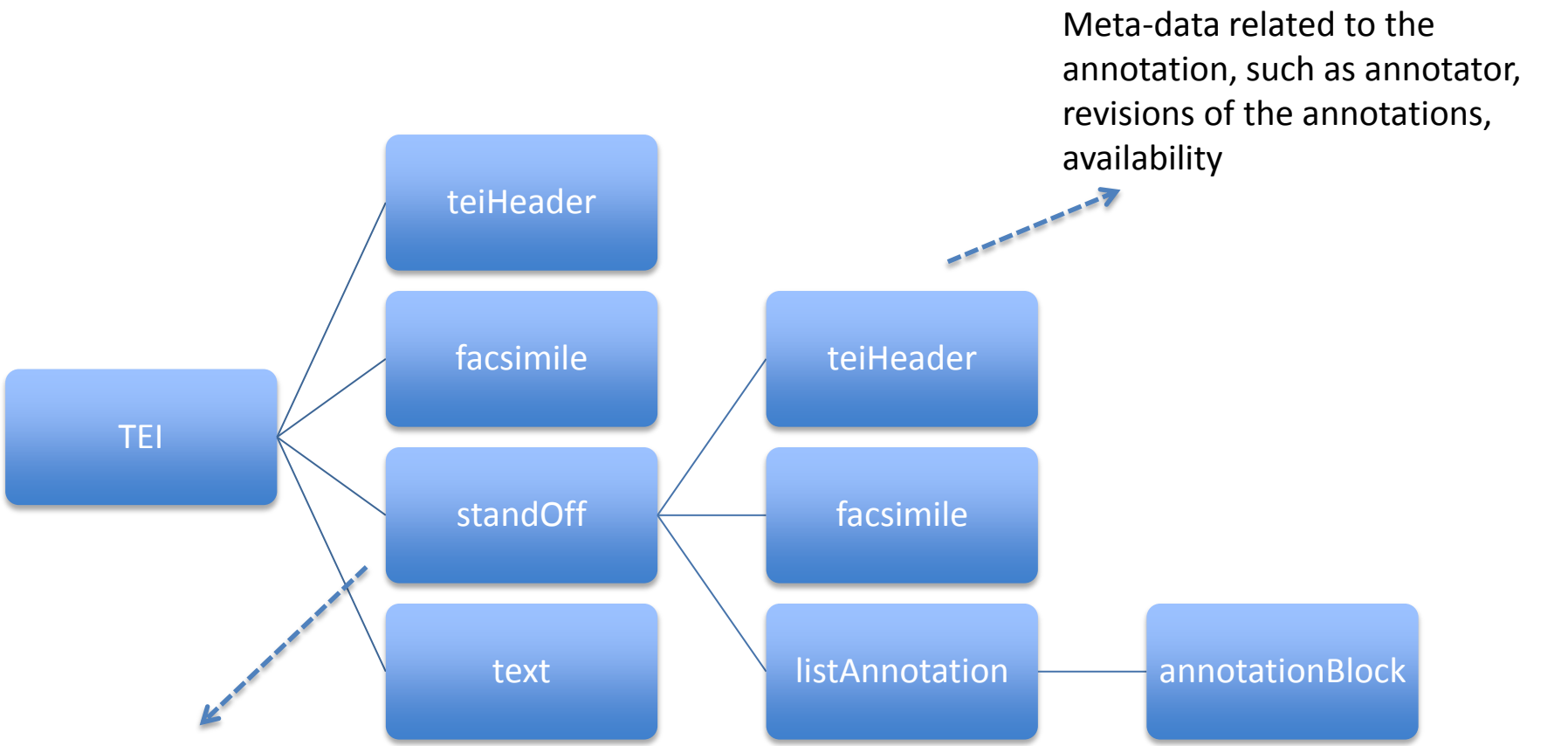
- The idea of standoff annotation is not new in general
  - Thompson & McKelvie, 1997
- Standoff annotation has been a core concept in the TEI guidelines since the beginning
  - Cf. Chapter: Linking, Segmentation, and Alignment
  - Availability of <anchor>, , <interp>, <link>, @ana
- But: not integrated in the TEI architecture
  - Stand-off elements can appear anywhere in a TEI document
  - Usual trade-off between on-site vs. grouping (<back>)
- The NLP community has also developed its own means
  - GraF (Ide & Suderman 2007) , Paula (Zeldes et al. 2009), etc.

- Need for a proper, and inclusive, treatment of standoff annotations in the TEI
  - Better integration, more guidance

Embedded standoff: Basic concept

- Building up an autonomous document containing primary source and additional annotations
  - Annotations are conveyed with their specific meta-data
  - Annotations have their specific place in the TEI document architecture
  - Standoff annotations may be recursively organized
  - Standoff annotations may point to textual as well as facsimile content
  - Well-defined elementary annotation units
  - Coherence with existing models (Open Annotation, ISO TC 37) should be ensured
- Typical use-cases
  - Annotated corpora
    - Treebanks
  - Text mining
    - Named entity recognition, keyword/terms extraction
  - Human annotations on a document
    - critical editions, patent examination, peer review…
- Strong relation with interlinear annotation

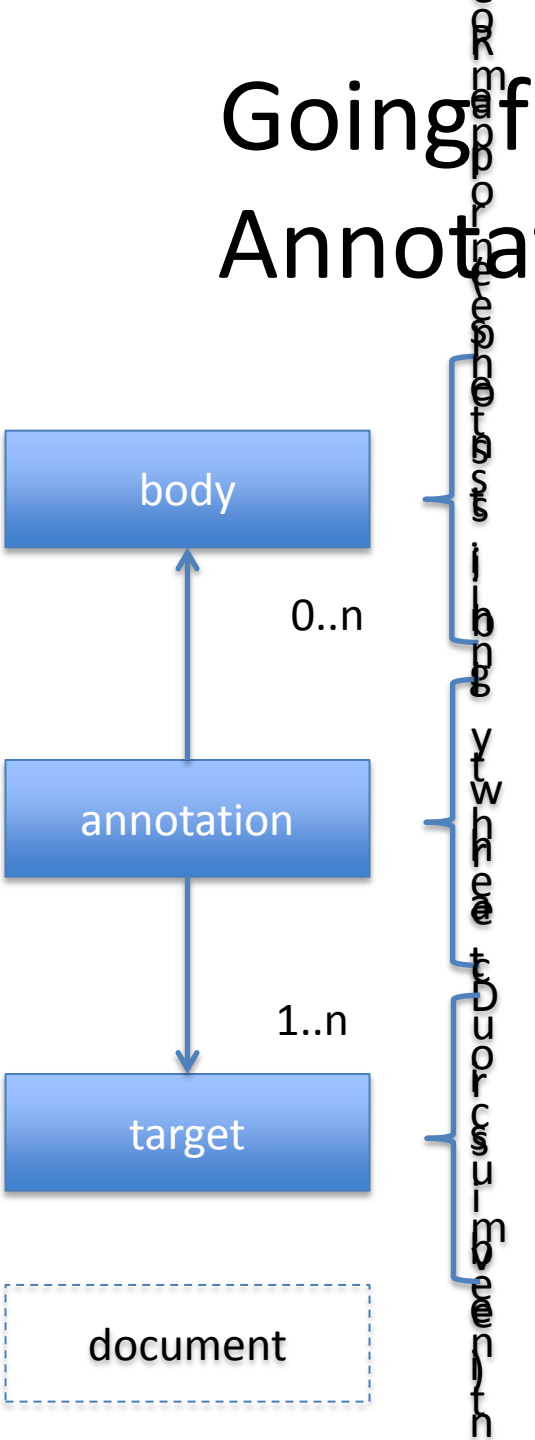
Annotations in TEI: <standOff>



Timeline

- 2011: Paper by Thomas Schmidt in jTEI (https://jtei.revues.org/142)
- August 2012: new tickets by Javier Pose (EPO)
- January 2014: Workshop in Berlin
  - Draft of a first proposal
  - Setting-up a github environment
- 2012-2016: ISO 24624 project (Editor: Thomas Schmidt)
  - Need for a annotation grouping component (<annotationBlock>)
- May 2015: Council meeting in Ann Arbor
  - Several updates to the proposal
  - “Stabilisation” of element names, with memory leaks
- March 2016: TEI release 6.0.0
  - New element <annotationBlock> for interlinear annotation
- August 2016: publication of ISO 24624 Transcription of Spoken Language

Going further: mapping the Web Annotation Data Model (WADM)



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

<interp type=“” inst=”” ana=“”>

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

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

Prototypical example

Dates in a named entity recognition context

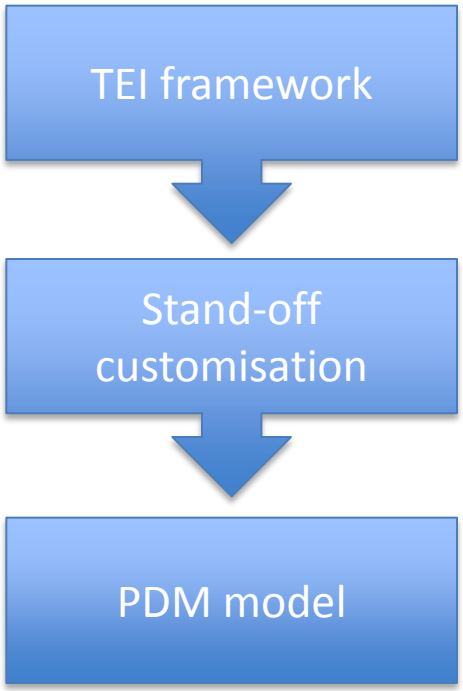
```
<annotationBlock>
   <date xml:id="E4N1" from=“1944-08-17“ to=“1944-08-25”>
      17 - 25 août 1944</date>
   <interp ana="#E4N1" inst="#d1e173"/>
   <span xml:id="d1e173" from="#E4T6" to="#E4T10" />
</annotationBlock>
```

Great advantage on readiness and programmatic treatment

standOff: A resilient model for the patent document model

- Covers a whole range of annotation types
  - Manuel annotations by examiners (features, clarity, claim trees)
  - Commentaries by examiner (preliminay to so-called “search report”)
  - Automatically created annotations
    - Bibliography
    - Various technical entities (chemistry, biology, physics)
    - Argumentative objects

Chaining ODDs in PDM




```
<schemaSpec ident="standoff-proposal"
start="TEI teiCorpus">
```

```
<schemaSpec ident="tei_docdb" prefix="tei_" start="TEI teiCorpus"
source="https://raw.githubusercontent.com/laurentromary/stdfSpec/A
nArbor/Specification/standoff-proposal_doc.xml">
```

PDM: A real life demonstration of the TEI power to be used in digital library projects

- Complex re-combination of TEI components with more exploratory developments
  - Numerous feedback to the TEI guidelines on GitHub
- Wide document type coverage:
  - E.g. representation of the patent legal corpus (European Patent Convention, Guidelines, Case law)
- The <teiCorpus> element is a useful object ;-) for large-scale applications

With thanks to Toma Tasovac, Belgrade Center for Digital Humanities and the DARIAH WG on lexical resources

CHANNELLING TUMULTUOUS WATERS: THE TEI LEX-0 INITIATIVE

Where we are coming from..

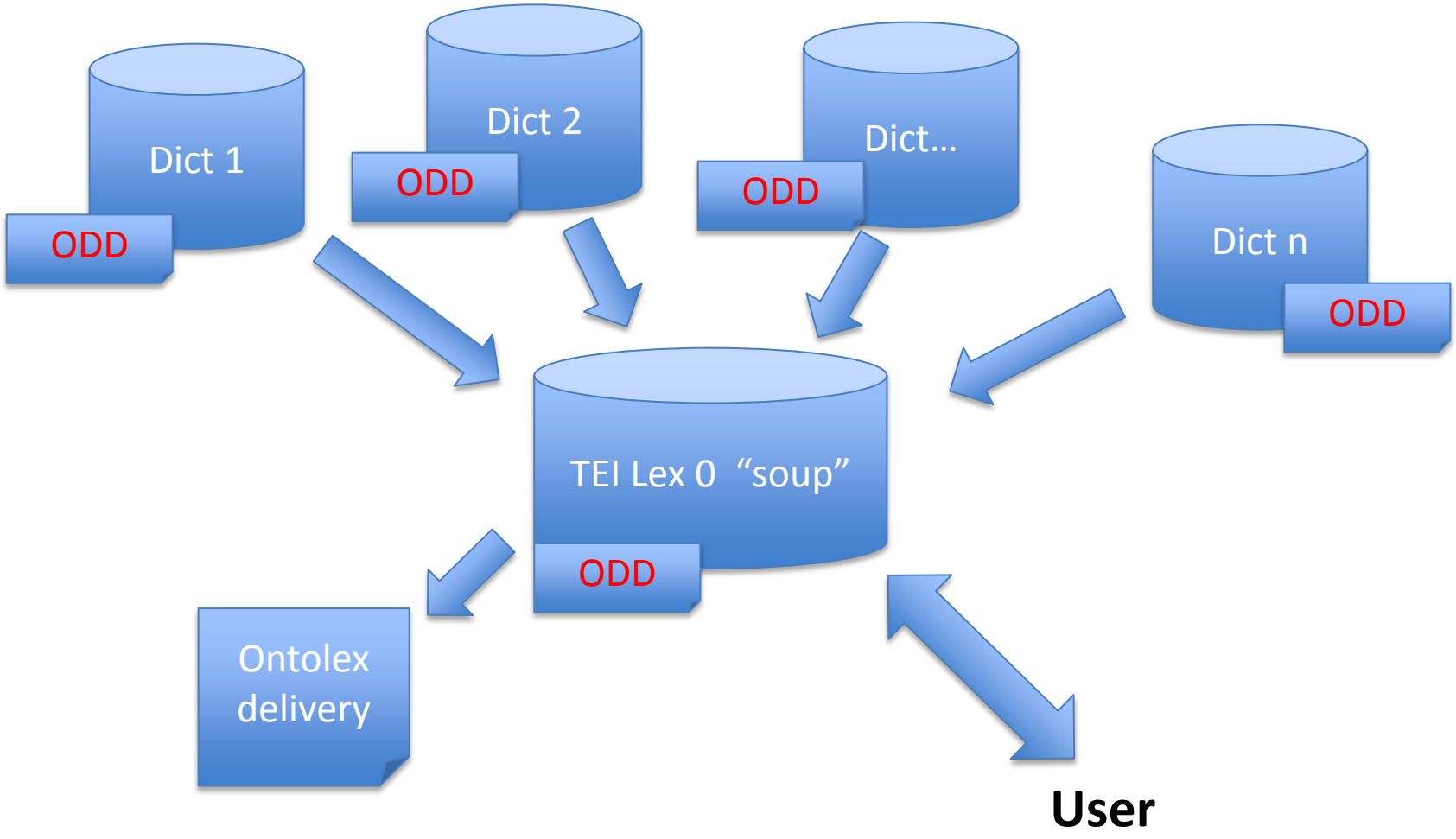
```
<entry>
    <def>Un animal sans queue ni tête</def>
    <hom>
        <form>I don't remember why but I need a variant here...</form>
    </hom>
    <gramGrp>
    <note>Oups, forgot to mention some grammatical constraints</note>
    </gramGrp>
    <form>
        <orth>maybe I could put the lemma here</orth>
    </form>
    <usg type="equiv">rabbit</usg>
    <xr type="translation"><ref>rabbit</ref></xr>
</entry>
```

Tightening the TEI dictionary chapter

- Objective of TEI Lex-0
  - Improving consistent encoding of lexical entries across lexicographic projects
- Various use cases in mind
  - Target format (analysing and comparing)
    - Cf. TEI Analytics (MONK project)
  - Generic dictionary tools
  - Education (discussions – arguments are as important as schema)
- Position wrt current chapter
  - Provide further constraints; at times departing from the guidelines…
  - Not necessarily an editing/publishing format

TEI Lex-0 should be primarily seen as a format that existing TEI dictionaries can be univocally transformed to in order to be queried, visualised, or mined in a uniform way

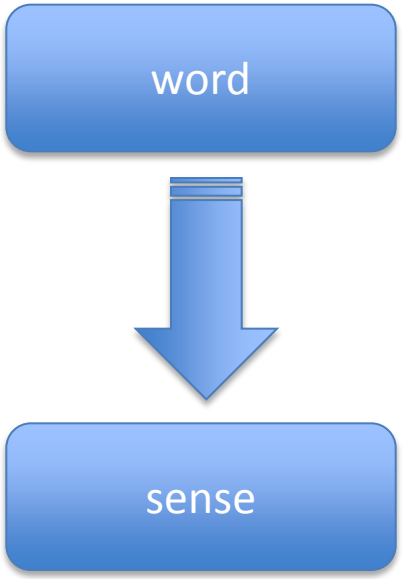
The ELEXIS centralized hub



Institutional setting

- Initial work
  - COST Action European Network of e-Lexicography (ENeL)
    - Working Group "Retrodigitised Dictionaries" (Toma Tasovac and Vera Hildenbrandt)
- Current
  - DARIAH
    - Working Group "Lexical Resources" (Laurent Romary and Toma Tasovac)
  - Support from H2020-funded European Lexicographic Infrastructure (ELEXIS)
- Further alignment with ISO 24613 (LMF), currently under revision

Enforcing the semasiological model




```
<entry>
    <form type="lemma">
        ...
    </form>
    <sense>
        <def>...</def>
    </sense>
</entry>
```

Overview of requirements

- General organisation of a dictionary entry
- Constraints on form and grammatical information
- Cross-references
- Embedded entries
- Usage
- Etymology

Simplifying the dictionary micro-structure

- Current situation
  - Containing vs. contained entries
    - <superEntry> – <entry> – <re>
  - Structured vs. unstructured entries
    - <entry> – <entryFree>
- The TEI Lex-0 vision
  - Representing all entry-like objects as <entry>
    - Note: cf. ticket on <lbl> and <pc> in <entry>
    - Making more use of <dictScrap>
    - Making <entry> recursive

Recursive entry - example

```
<entry type="wordFamily">
  <form type="base">
    <orth>Haus-</orth>
  </form>
  <pc>,</pc>
  <form type="base">
    <orth>haus-</orth>
  </form>
  <pc>:</pc>
  <!-- possibly some shared usg information -->
  <entry type="wordForm">
    <form type="lemma">
      <orth expand="Hausaltar">-altar</orth>
      <pc>,</pc>
      <gramGrp>
        <gen value="masculine">der</gen>
      </gramGrp>
    </form>
    <sense>...</sense>
  </entry>
  <entry type="wordForm">
    <form type="lemma">
      <orth expand="Hausandacht">-andacht</orth>
      <pc>,</pc>
    </form>
        <!-- ... -->
   </entry>
   <!-- ... -->
</entry>
```

Reducing the content model of <entry>

- Allowed in <entry>
  - <form>, <sense>, <entry>, <etym>, <gramGrp>, <usg>, <xr>, <pc>, and <dictScrap>
- Not allowed in <entry>
  - <def>, <hom>, and <cit>
- Additional features
  - Mandatory @xml:id
  - Mandatory @xml:lang to indicate the object language
  - Encouraging using a @type on <entry>
    - Ongoing discussion to determine a coherent set of values

Usage information

- a label which can be attached at various points in the entry hierarchy in order to signal e. g. restrictions in terms of geographic regions, domains of specialized language or stylistic properties for the particular lexical item that it is attached to
  - label-like descriptors (often abbreviated) and as fuller narrative expressions
    - E.g. <usg type="style" norm="expletive">Schimpfwort</usg>
    - <usg type="hint">(рекла сељанка на њиви за време врућине)</usg> (“(said by a peasant woman in the field in hot weather)”)

Providing more coherence to usg/@type

- usg/@type is made mandatory
- usg/@norm is encouraged
- Reference works of Svendsen (2009) and Atkins and Rundell (2008)
  - Cf. usual linguistic notions of diachronic, diatopic, diastratic etc.
- Dropping values that are superfluous in the current guidelines given other TEI lex choices
  - lang, gram, syn, hyper, colloc, comp, obj, subj, verb
- New recommended values
  - temporal, domain, sociocultural, meaningType, frequency, attitude, normativity, hint

Implementation

- ODD specification available on
  - https://github.com/DARIAH-ERIC/lexicalresources
- Change documentation via GitHub tickets
- Pushing request to the TEI guidelines when we think what we do is of general interest

We've got a ticket to ride…

- Numerous evolutions that the group has initiated or to which members of the group contributed
  - Historical reminder: generalisation of <gramGrp> as container of grammatical information (2011-2012)
  - model.entryPart.top for <pc> and <lbl>
    - Reconciling the lexical and the editorial view
  - Generalising the use of @notation (orth, pron, hyph, stress, syll, pRef, oRef)
  - Improving <etym>: att.typed and recursivity
  - <entry> in <def> (chr-emil)
  - Deprecating oVar and oRef; before a deeper reform is set in place…
  - Towards a model.sensePart class for <sense>
  - And recursive entries!

Next steps

- Ensuring the best possible convergence between initiatives
  - Constantly fighting against silos
  - Adapting existing endeavours rather than re-inventing
- Disseminating knowledge and competence on lexical standards
  - Training: Lexical master class, #DARIAHteach
  - Documenting: blogs, papers and … the SSK
- Standardisation as a never-ending activity: improving things further
  - Integrating standards in concrete usage scenarios to identify usability, constraints etc.
  - Standardising as knowledge transfer

With thanks to Veerle Vanden Daelen and Charles Riondet

COMBINING STREAMS OF HETEROGENEOUS ARCHIVAL DATA

EHRI: an infrastructure for historical research about the holocaust

- EHRI's second phase (2015-2019) as an EU financed project with a total budget of almost 8 mio €
- 24 partner institutions from 17 countries: Research institutions, archives and e-science specialists
- EHRI's goal: Support research into the Holocaust and help networking of Holocaust researchers and archives

EHRI: Partner institutions

- NIOD, Institute for War, Holocaust and Genocide Studies (Amsterdam): Overall project coordination
- Yad Vashem (Jerusalem)
- CEGESOMA (Brussels)
- King's College (London)
- Institute for Contemporary History (Munich)
- Jewish Museum in Prague
- DANS (Den Haag)
- Wiener Library (London)
- Vienna Wiesenthal Institute for Holocaust Studies
- ŻIH (Warsaw)
- Mémorial de la Shoah (Paris)
- International Tracing Service (Arolsen)
- USHMM (Washington D.C.)
- Bundesarchiv (Berlin / Koblenz)
- Elie Wiesel National Institute for the Study of the Holocaust in Romania (Bucharest)
- Hungarian Jewish Archives (Budapest)
- Vilna Gaon State Jewish Museum
- Dokumentačné stredisko holokaustu (Bratislava)
- Foundation Jewish Contemporary Documentation Center CDEC (Milan)
- The Jewish Museum of Greece (GR)
- Ontotext (Sofia)
- INRIA (Le Chesnay)
- Stowarzyszenie Centrum Badań nad Zagładą Żydów (Warsaw)
- Kazerne Dossin: Memorial, Museum and Documentation Centre on Holocaust and Human Rights (Mechelen)

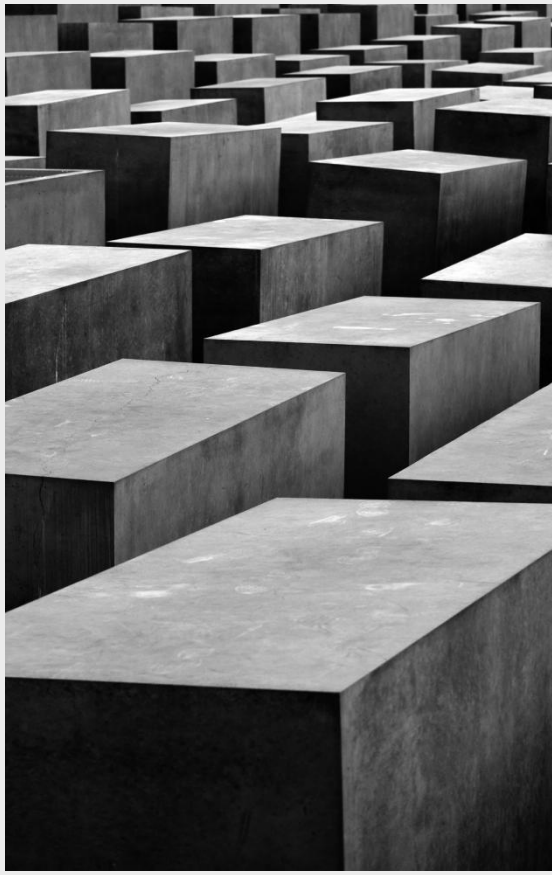
Why EHRI?

- **Fragmentation and dispersal of archival sources**
  - Geographical scope Holocaust
  - Attempts to destroy the evidence
  - Migration of Holocaust survivors
  - Multiple documentation projects after the war
- **Internationalization Holocaust research**
  - Holocaust in Eastern Europe
  - New levels of collaborative research
- **New opportunities for digital research**

EHRI Aims

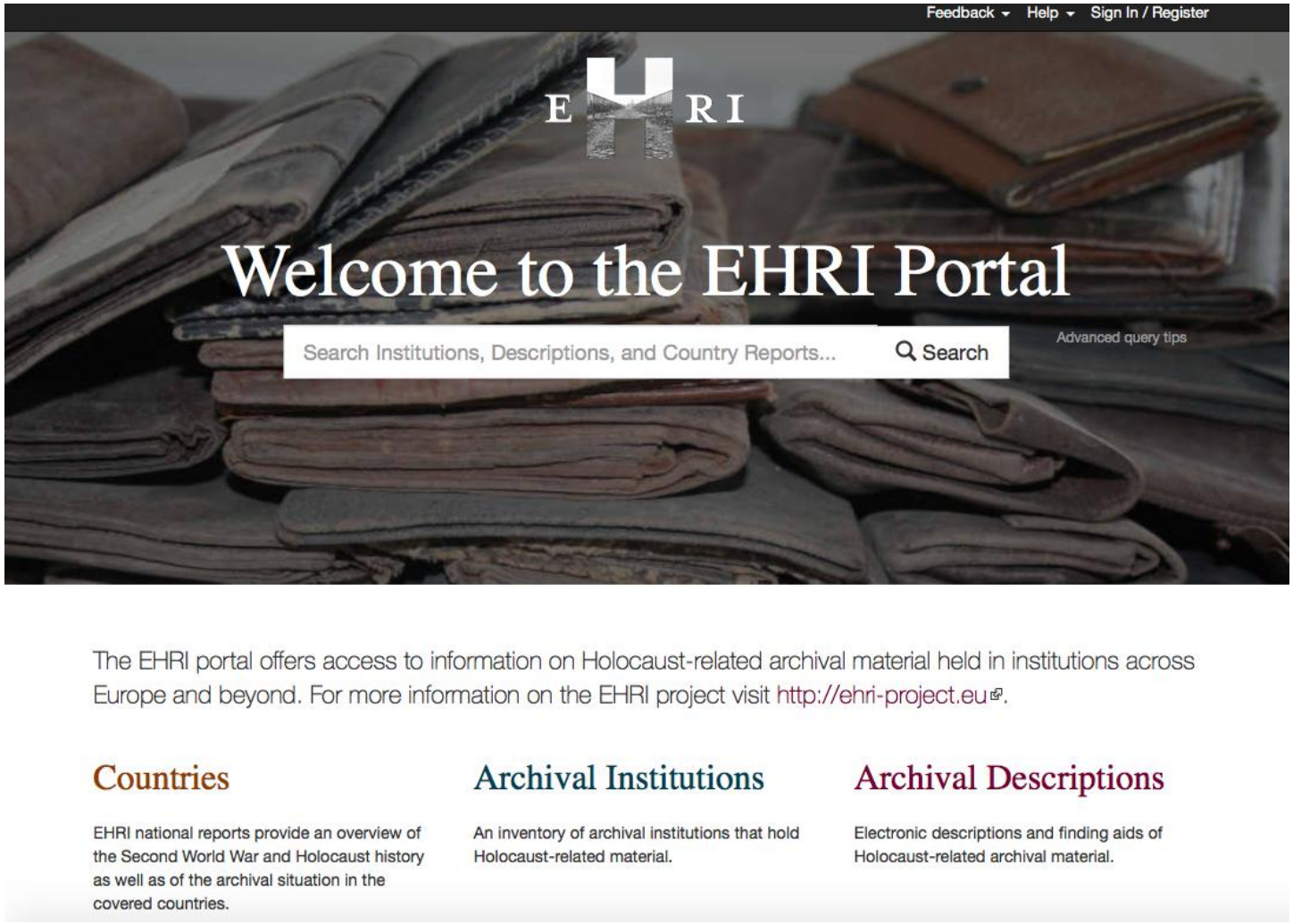
The main objective of EHRI is to support the Holocaust research community by

1. **integrating** information on key archival **collections** and **institutions** into an online portal
2. **encouraging** collaborative Holocaust **research** and investigating new methodologies



www.ehri-project.eu

EHRI Portal - https://portal.ehri-project.eu/

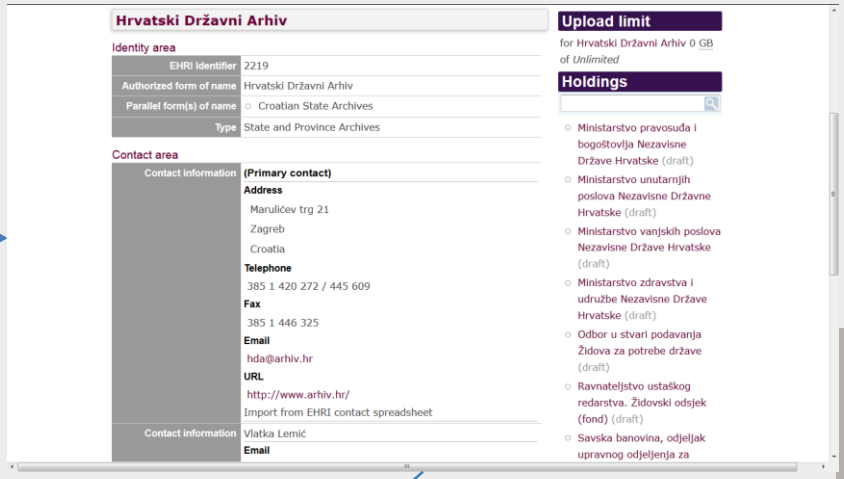


57 country reports, >1,900 descriptions of institutions, >230,000 archival descriptions

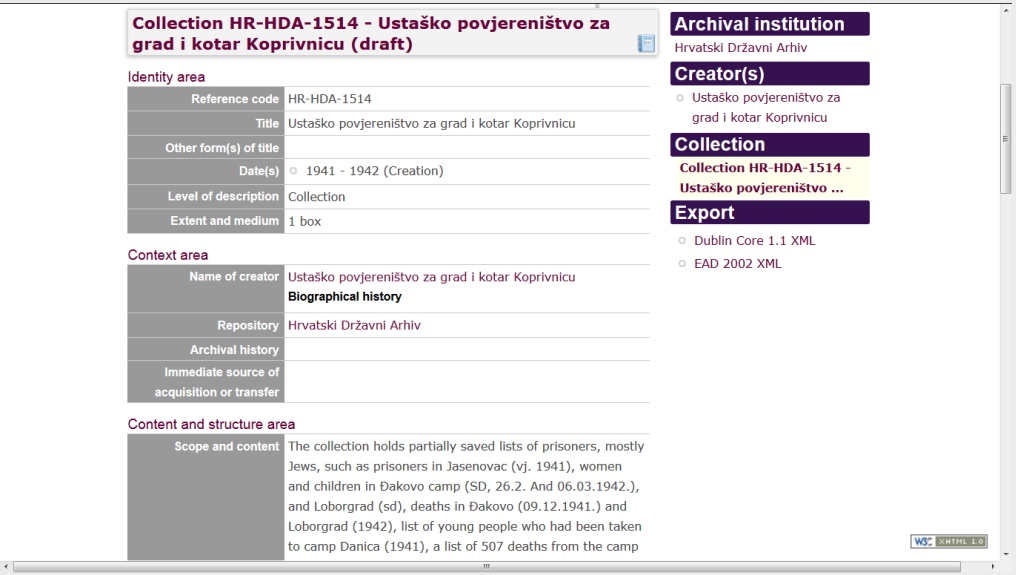
EHRI Database

Country reports (57 countries)

Entry on the individual archive (over 1,900)

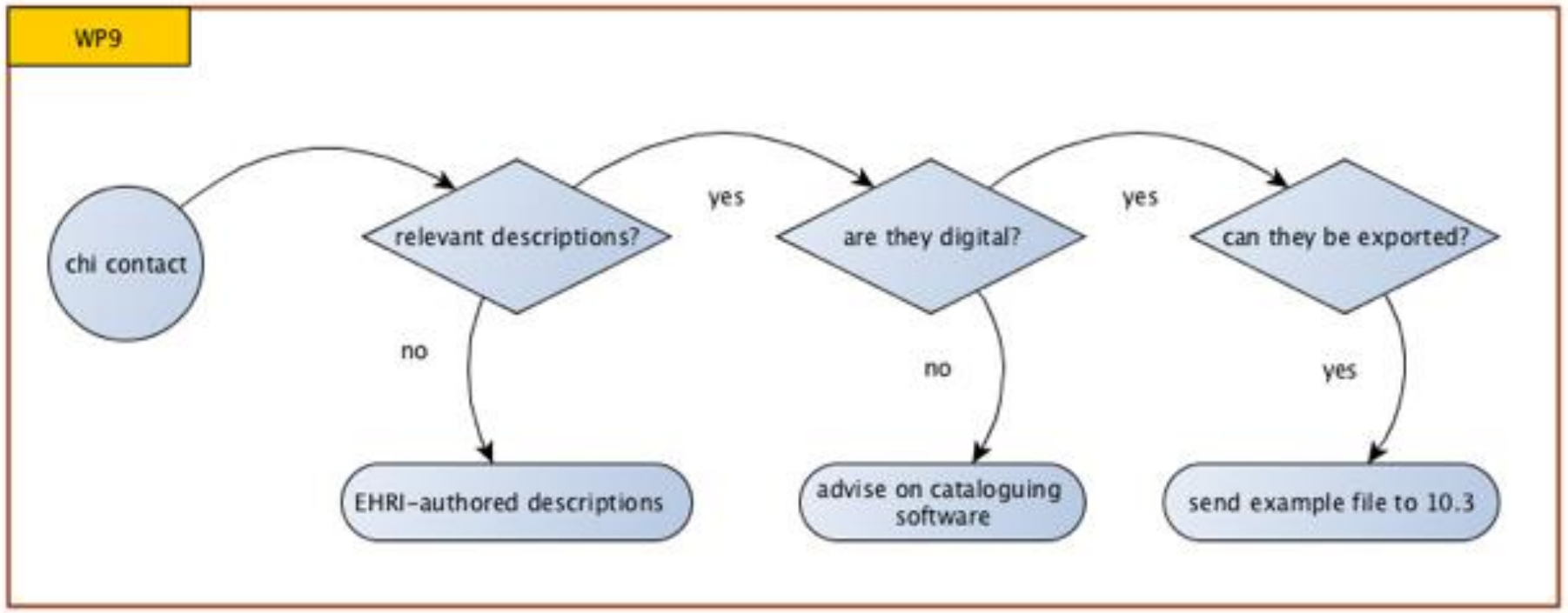


Individual entries (collections / units) (tens of thousands)

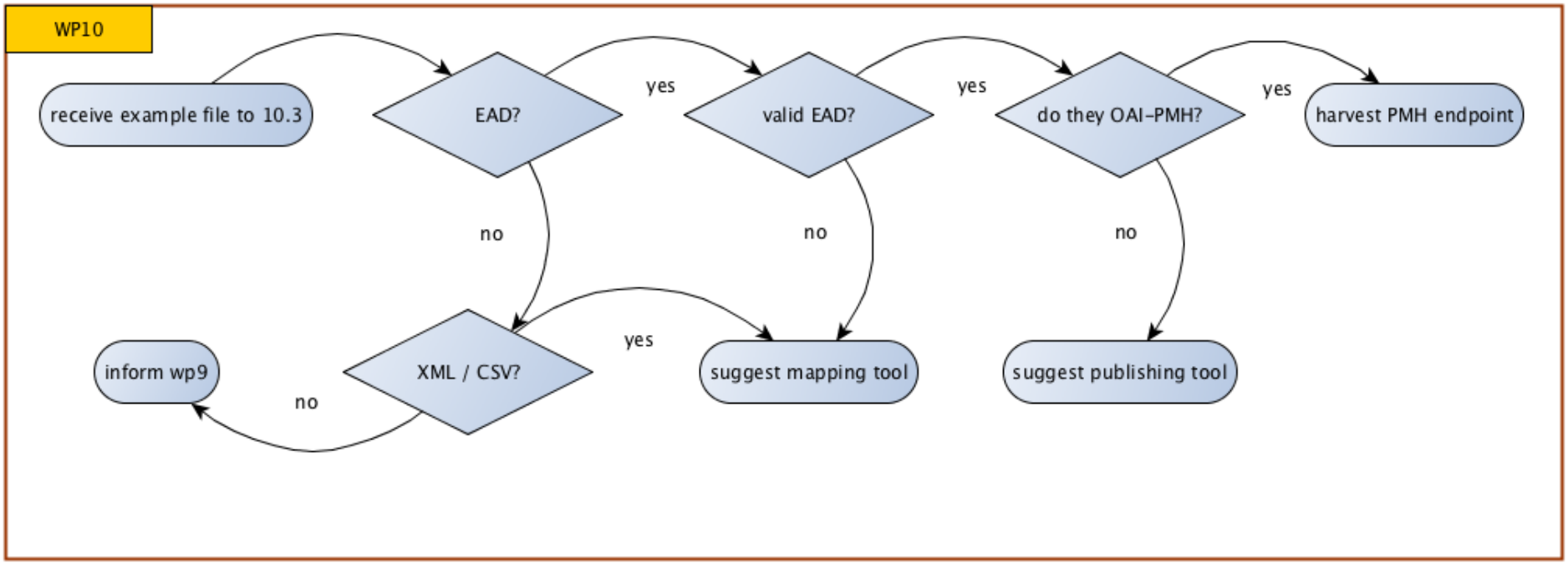


[EHRI Portal]

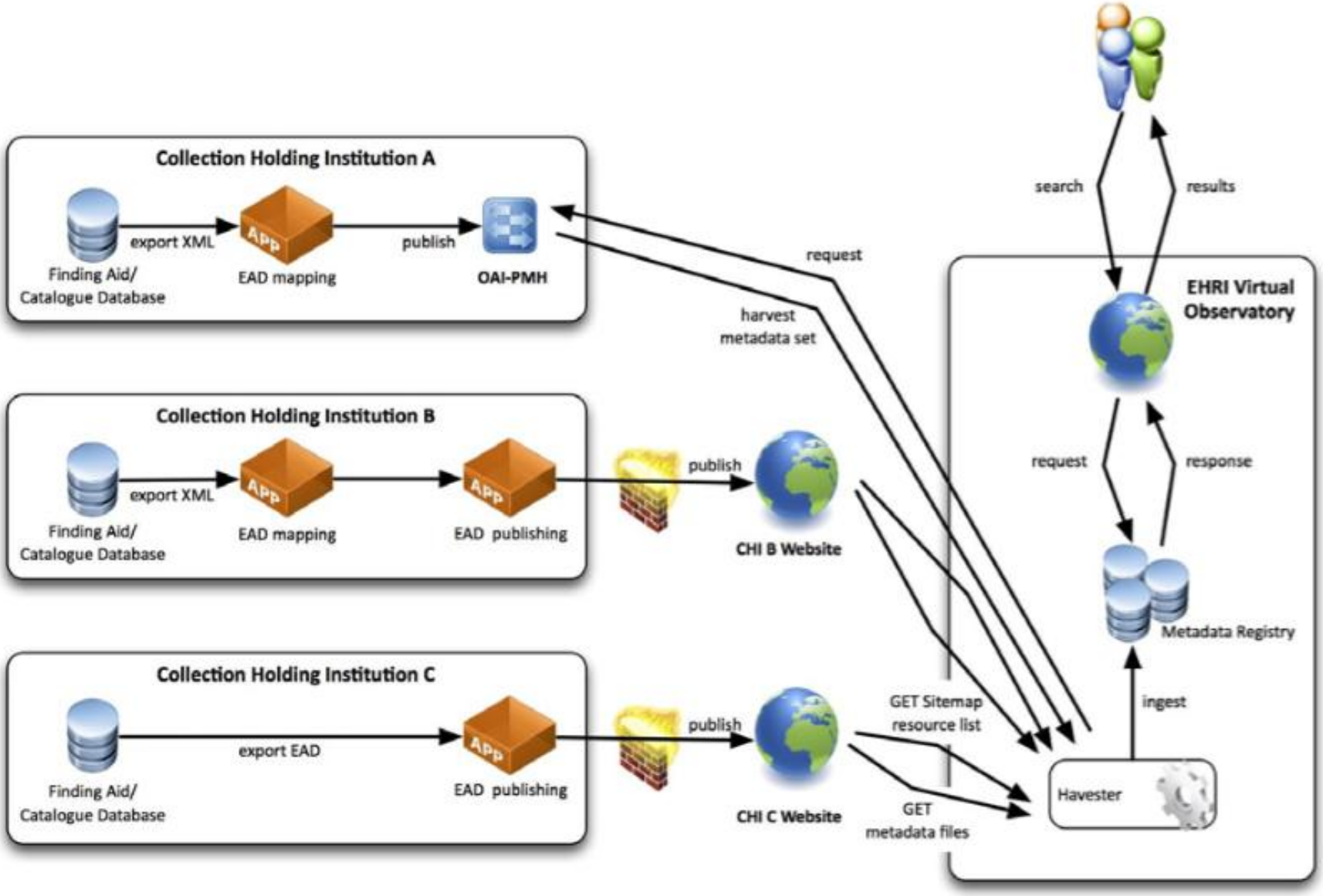
Integration of collection descriptions (I)



Integration of collection descriptions (II)



Data Integration via mapping & publishing tool



EAD - Encoded Archival Description

- Scope:
  - Machine processing of finding aids (XML)
  - Collection descriptions
- Initiated in 1993, EAD 1.0 (1998) … EAD3 (2015)
- Inspired by TEI (cf. D. Pitti)
- Maintained by the *Society of American Archivists* and the *Library of Congress*
- Allows one to align onto the ISAD(G) archival standard
- EAD is too permissive (Shaw 2001; Bunn 2013)
  - Tension between information exchange and archival description
- Future: RiC (Records in Context) ➔ New paradigm? Definitive solution?

EAD in EHRI: strategic view

- EAD2002 is the pivot format for automatic ingestion of archival descriptions.
- Ingestion of data in many formats
  - EAD1, Dublin Core, home-made formats
  - EAD2002 with very different encoding guidelines
- EHRI must have its own specific description rules
  - Narrowing EAD encoding possibilities
  - Adding quality checks
  - Content-oriented rules
  - Not modifying the EAD2002 reference schema
- A strategy centered on the archival network
  - Content-oriented rules based both on EHRI and CHI input data models
  - Integrating the human readable documentation in the validation process

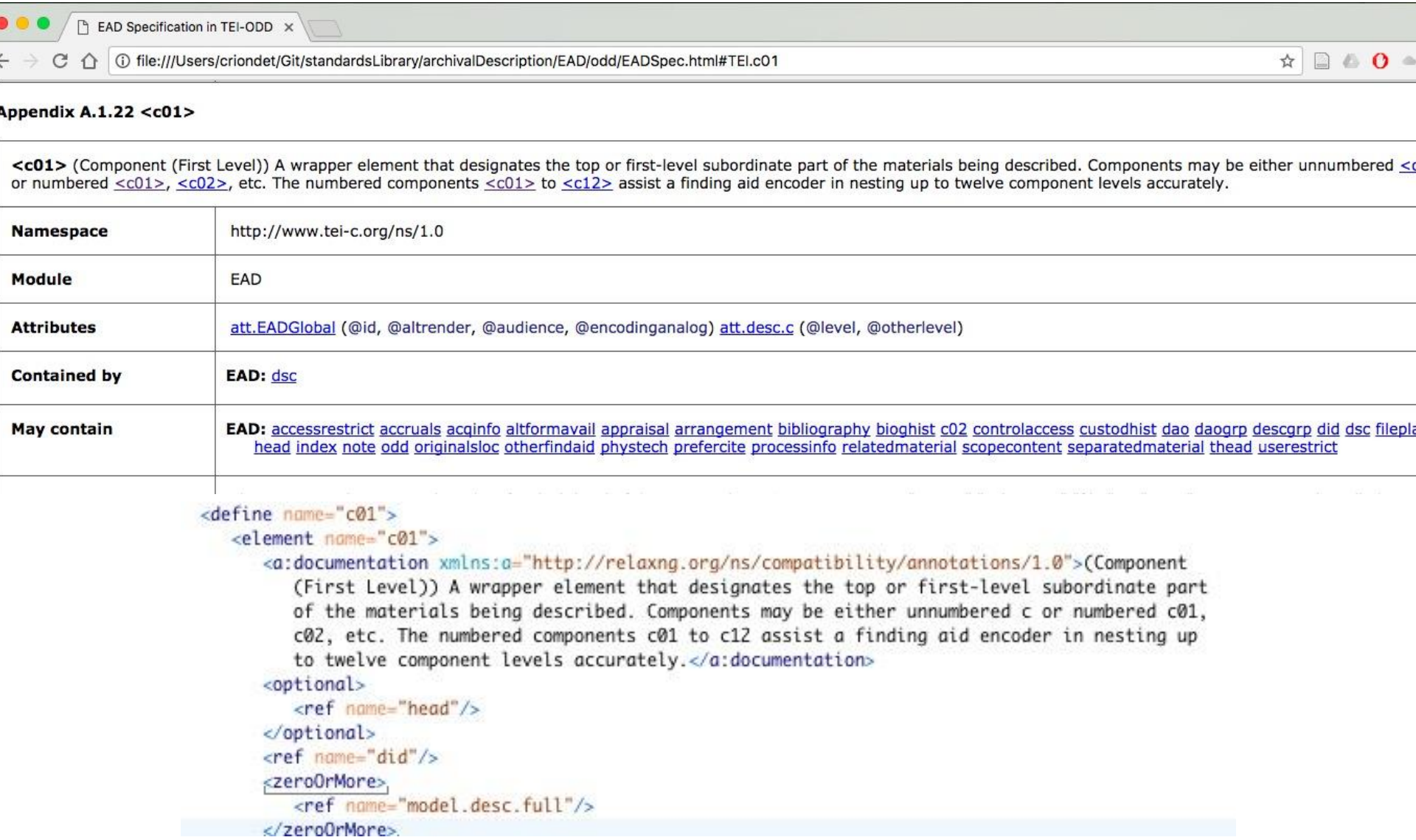
Using ODD for external vocabulary

- Is this an exotic idea?
  - The ODD specification language is not tied to TEI
  - The ODD processor can generate documentation and schema independantly of the TEI architecture
  - Still, it is possible to take up bits of the TEI architecture
    - E.g. some attributes, specific crystals (e.g. bibliography)
- Other known applications: XHTML (M. Holmes), Springer DTD, etc.

Defining an XML element with ODD

```
<elementSpec ident="c01" module="EAD">
    <gloss>Component (First Level)</gloss>
    <desc>A wrapper element that designates the top or first-level subordinate
        part of the materials being described. Components may be either
        unnumbered <gi>c</gi> or numbered <gi>c01</gi>, <gi>c02</gi>, etc. The
        numbered components <gi>c01</gi> to <gi>c12</gi> assist a finding aid
        encoder in nesting up to twelve component levels accurately.</desc>
    <classes>
        <memberOf key="att.EADGlobal"/>
        <memberOf key="att.desc.c"/>
    </classes>
    <content>
        <rng:optional>
            <rng:ref name="head"/>
        </rng:optional>
        <rng:ref name="did"/>
        <rng:zeroOrMore>
            <rng:ref name="model.desc.full"/>
```

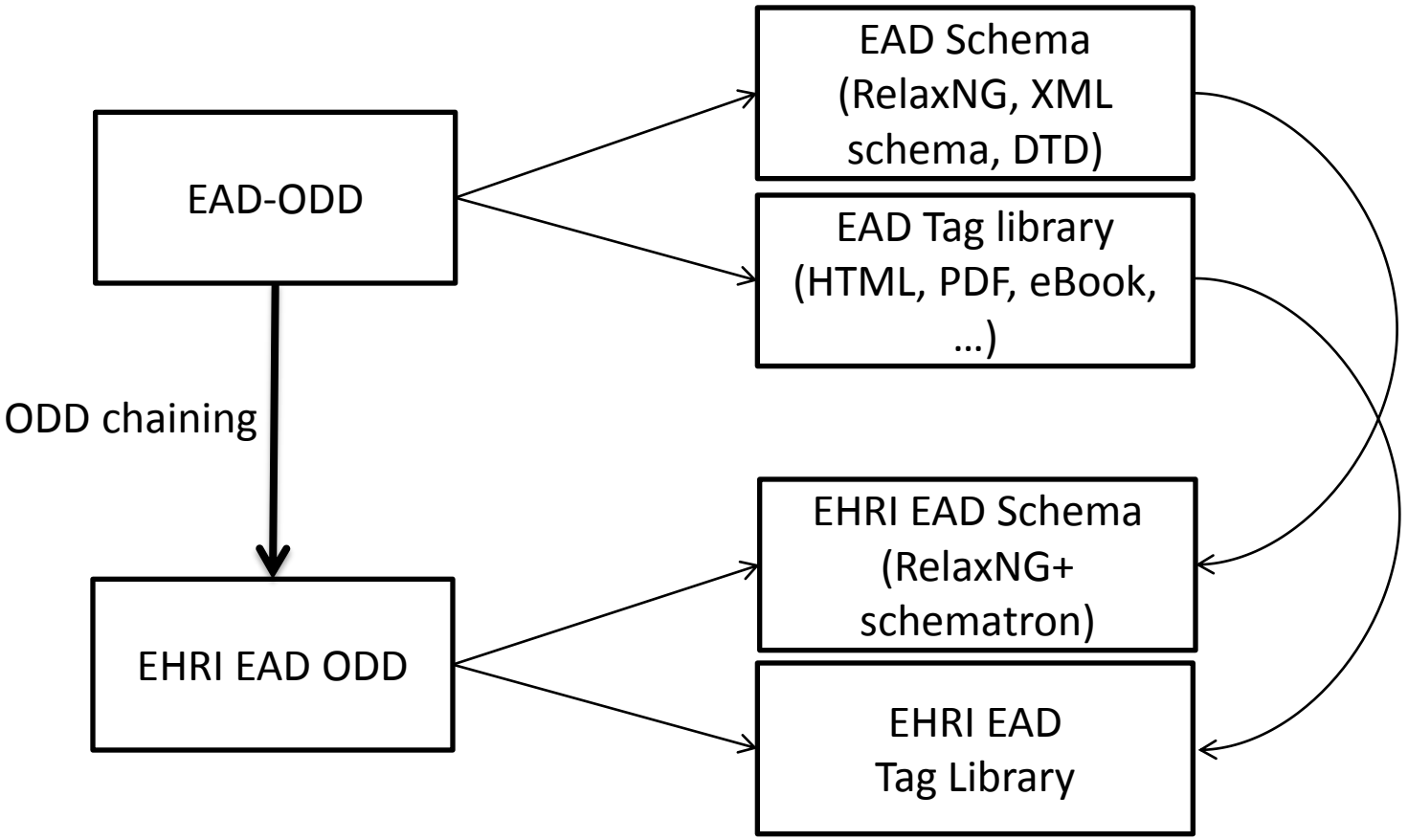
Generating documentation and schema



EAD in ODD within EHRI in a nutshell

- Full coverage of EAD 2002
  - Official EAD schema (RelaxNG) : www.loc.gov/ead/ead.rng
  - Guidelines provided by the Library of Congress: http://loc.gov/ead
- Using the ODD chaining possibilities
  - One master model + derived specific customisations
- Maintained in Github (Parthenos project)
  - http://github.com/ParthenosWP4/standardsLibrary/

Flexible and customizable methodology



Quality checking with Schematron

- Emphasize EAD validation errors
- Align the descriptions with EHRI constraints
  - Required elements in EHRI (but not in EAD)
    - E.g. scopecontent (description of the content of the documents)
  - Content normalisation (dates, codes, …)
- Highlight some description elements that could be improved
  - E.g. content related elements (existence of copies of the material, bibliographic references, …)

- Sorted in categories (roles)
  - MUST: mandatory for import process
  - SHOULD: mandatory for description process, i.e. In terms if archival description. Not technically mandatory, but may cause comprehension issues
  - COULD: non mandatory rules. Enhance the general quality of the description, without any obligation. Pointing that informational element.

Schematron in EHRI: example

```
<constraintSpec ident="labelDesirable" scheme="isoschematron" type="EHRI" mode="add">
  <desc><gi>unitdates</gi> COULD have a <att>label</att> attribute or an
      <att>encodinganalog</att> attribute, describing the type of date</desc>
  <constraint>
    <rule xmlns="http://purl.oclc.org/dsdl/schematron" context="ead:unitdate"
      see="&path;#EAD.unitdate"><assert xmlns="http://purl.oclc.org/dsdl/schematron"
        role="COULD" test="normalize-space(@label) or normalize-space(@encodinganalog)"
        >unitdates COULD have a label attribute or an encodinganalog attribute,
        describing the type of date</assert></rule>
  </constraint>
</constraintSpec>
```

Schematron in EHRI (cont.)

```
<constraintSpec ident="authfilenumberPossible" scheme="isoschematron" type="EHRI">
  <desc>Access points COULD be chosen in authority lists. The list is declared with a
      <att>source</att> attribute. The related id of this authority should be declared
    in an <att>authfilenumber</att> attribute. Note that EHRI provides URLs for
    vocabularies and authorities. Check the <ref target="http://ehri-project.eu">EHRI
      website</ref> for more information</desc>
  <constraint>
    <rule xmlns="http://purl.oclc.org/dsdl/schematron" context="ead:controlaccess"
      see="&path;#EAD.controlaccess">
      <assert xmlns="http://purl.oclc.org/dsdl/schematron" role="COULD"
        test=".[@authfilenumber and @source]">Access points COULD be chosen in an
        authority list. This list should be declared in a @source attribute. The related
        id of this authority should be declared in an @authfilenumber attribute.
      </assert>
    </rule>
  </constraint>
</constraintSpec>
```

Connect validation and mapping process to ad hoc documentation

2.1.11. <archdesc>

<archdesc> (Archival Description) A wrapper element for the bulk of an EAD document instance, which describes the content, context, and extent of a body of archival materials, including administrative and supplemental information that facilitates use of the materials. Information is organized in unfolding, hierarchical levels that allow for a descriptive overview of the whole to be followed by more detailed views of the parts, designated by the element Description of Subordinate Components <dsc> . Data elements available at the <archdesc> level are repeated at the various component levels within <dsc>, and information is inherited from one hierarchical level to the next.

Namespace	http://www.tei-c.org/ns/1.0
Module	EAD
Attributes	att.EADGlobal (@id, @altrender, @audience, @encodinganalog) att.relatedencoding (@relatedencoding) att.desc.c (~~@level~~, @otherlevel)

@level The hierarchical level of the materials being described by the element.

Derived from att.desc.c

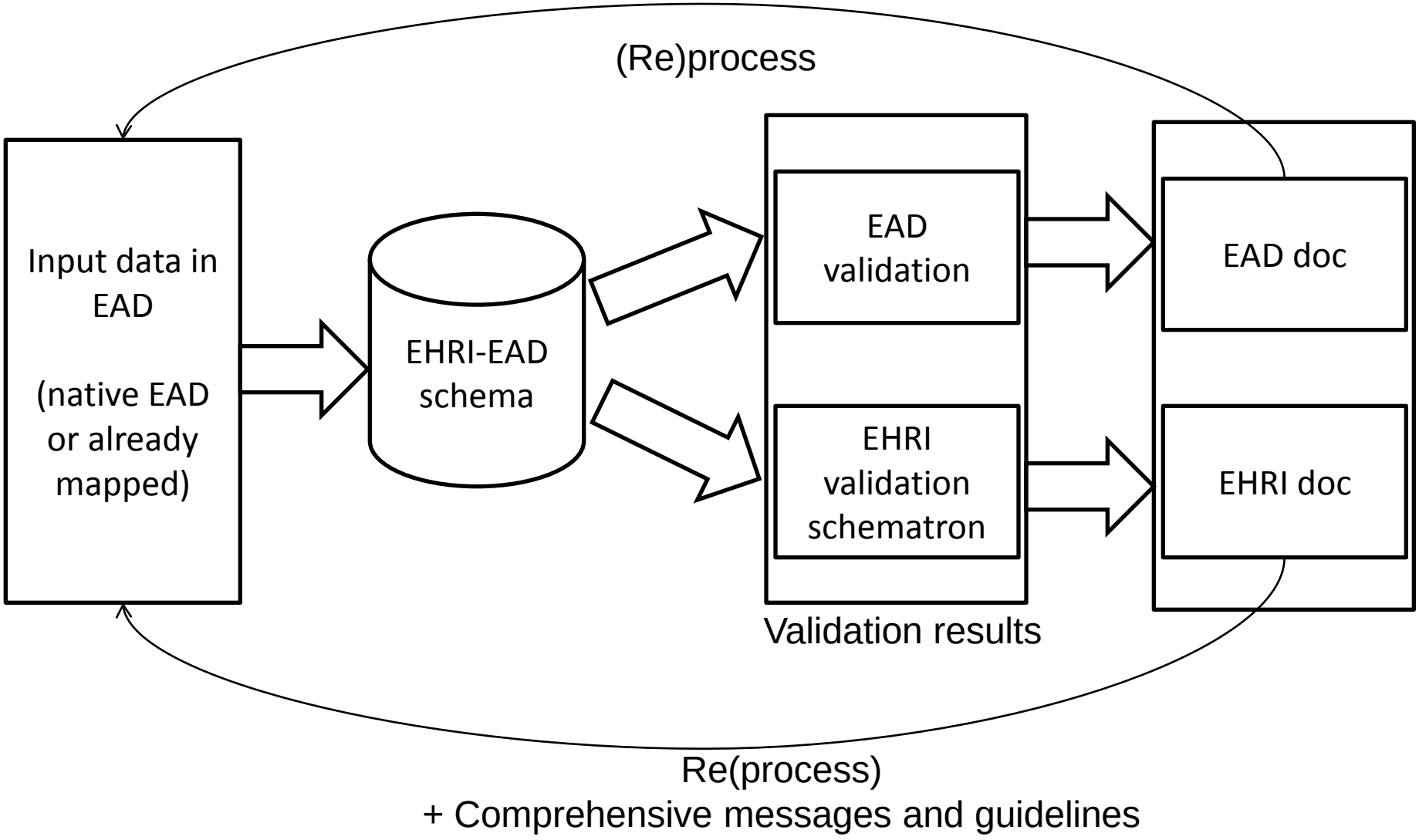
Status **Required**

Schematron If the attribute @level has the value 'otherlevel', an attribute @otherlevel MUST be added

```
<s:rule context="ead:ead"
 see="https://cdn.rawgit.com/EHRI/data-validations/92c8e39f/ODD-RelaxNG/EAD/EHRI_EAD_doc.html#EAD.att.desc.c">
<s:assert role="MUST"
 test="not(@level = 'otherlevel') or (@otherlevel and not(@otherlevel = ''))">If the attribute level has the value 'otherle
 MUST be added</s:assert>
</s:rule>
```

Schematron The <archdesc> element can have for @level the value 'fonds', not the subcomponents <c01> to <c06>

Workflow


(Re)process
Input data in EAD
(native EAD or already mapped)
EHRI-EAD schema
EAD validation
EHRI validation schematron
Validation results
EAD doc
EHRI doc
Re(process)
+ Comprehensive messages and guidelines

Where do we go from this?

- A generic modelling workflow for heterogeneous multi-source data
  - Demonstrates the power of the TEI architecture
- Contribution to the maintenance of archival standards
  - Bridge between EAD2002 and EAD3
  - TEI as a possible framework for taking care of EAD in the long run and facilitating the integration of EAC (and EAG)
- TEI – EAD conversions
  - E.g. interoperability with <msDesc> components
- Dreamy…
  - Adding stand-off (yes <standOff>!) annotations extracted from plain text EAD content

WHICH COURSE FOR THE TEI?

The TEI is doing well – the hidden TEI

- Antonio Zampolli price by ADHO
  - Reflects that the TEI is pervading all fields in the (digital?) humanities
- TEI has become a natural component of a humanities project based on textual sources
  - Many small editions are flourishing everywhere
  - Now recommended or requested by funding organisations (DFG!)
  - Numerous training events (cf. DiXiT, DARIAH master classes)
- Taken up by larger organisations
  - Academies, Dictionary projects, EPO… especially in Europe

Consolidating our conceptual model

- TEI as a rich space of elementary constructs
  - Attributes (classes), “entities”, bibliographical and dictionary entries, etc.
- Multifarious document types for various communities
  - From scholarly editions to dictionaries, including computer mediated communication, scientific information, etc.
  - More precise guidelines for specific applications
    - Collaboration with ISO (standards), DARIAH (recommendations)
    - Reducing syntactic freedom in specific application domains, not in TEI as a whole
  - Complementing our stock: onomasiological constructs, standOff
- Strong conceptual basis with pure ODD
  - For TEI and non TEI based application
  - Starting point for offering support to other dissemination formats (JSON, LOD) – Interfacing the trends
  - XML is likely to remain central for a long time for sustainable back-office content

Focusing, enlarging?

- Enlarging our expert basis (e.g. dictionaries, stand-off)
  - Stronger role for SIGs
  - Close coordination with council
  - Bringing in more technical experts from outside
- Institutional partnership
  - Archives, Clarin, DARIAH, MEI, Europeana
    - Further enforcement of the TEI guidelines
    - Sharing our technical platform
      - E.g. EAD maintenance
  - Thinking together the sustainability of TEI material
    - Repositories (Tapas?)
    - The TEI already offers a strong basis for sustainability
- Need for a stable communication framework
  - Lively conference and journal (jTEI)
  - Investing in the web site and the wiki

Should we/you be afraid of standards?

```
<cit>
    <quote>Yes you should be afraid, but you should be more afraid of
    not having them</quote>
    <author>Wendell Piez</author>
</cit>
```

For the long winter evenings

Laurent Romary. TBX goes TEI -- Implementing a TBX basic extension for the Text Encoding Initiative guidelines. *Terminology and Knowledge Engineering 2014*, Jun 2014, Berlin, Germany. 2014, Terminology and Knowledge Engineering, TKE 2014. <hal-00950862v2>

Laurent Romary, Andreas Witt. Méthodes pour la représentation informatisée de données lexicales/Methoden der Speicherung lexikalischer Daten. *Lexicographica*, de gruyter Mouton, 2014, 30. <hal-00991745>

Laurent Romary. TEI and LMF crosswalks. *JLCL - Journal for Language Technology and Computational Linguistics*, 2015, 30 (1), <http://www.jlcl.org>. <hal-00762664v4>

Laurent Romary. An abstract model for the representation of multilingual terminological data: TMF - Terminological Markup Framework. *TAMA 2001*, Feb 2001, Antwerp, Belgium. 2001. <inria-00100405>

Laurent Romary. Standards for language resources in ISO – Looking back at 13 fruitful years. *edition - die Terminologiefachzeitschrift*, Deutscher Terminologie-Tag e.V. (DTT), 2015. <hal-01220925>

...

https://cv.archives-ouvertes.fr/laurentromary