# The Text Encoding Initiative as an Infrastructure

Laurent Romary

DARIAH-EU

Digital Research Infrastructure
for the Arts and Humanities

# THE TEXT ENCODING INITIATIVE AS AN INFRASTRUCTURE

LAURENT ROMARY, INRIA & DARIAH

28 FEBRUARY 2017

www.dariah.eu

# A brief history of time

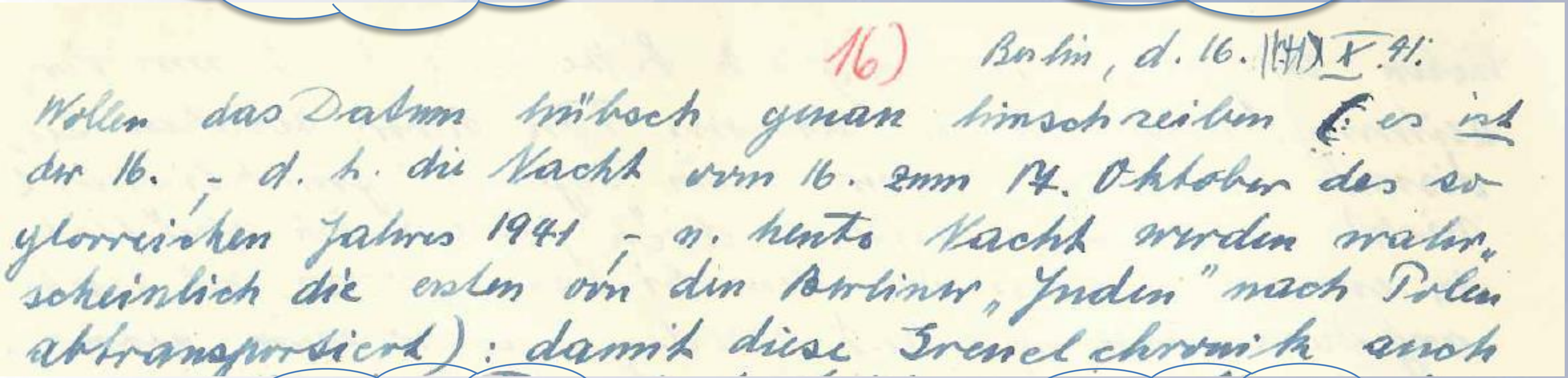
- DARIAH: a 10+ year story
  - The central role of data, expertise and standards
- TEI: 30 years of building standards and gathering competence
  - Integrated in the DH landscape
- Picture for the future
  - Complementarity: sustaining and reaching out

DARIAH-EU

# Sources: evidence based research in the humanities

Documenting:
origin, date,
material

Qualifying:
authorship, research
value, authenticity



16) Berlin, d. 16. (14) X. 41.

Wollen das Datum hübsch genau hinschreiben (es ist der 16. – d. h. die Nacht vom 16. zum 17. Oktober des so glorreichen Jahres 1941 – u. heute Nacht werden wahrscheinlich die ersten von den Berliner „Juden" nach Polen abtransportiert): damit diese Greuelchronik auch

Analyzing:
layout, transcription,
names, dates

Communicating:
corpus, rights,
contextualization

DARIAH-EU

Source: *L. Alt's diary*

# Sources in the digital world

- Increased number of digitized and born-digital sources
- Dealing with the digital life cycle
  - Creation, curation, enrichment, communication, archiving… and certification
- Consequence for DARIAH
  - Building on the **Long-established tradition of digital arts and humanities** research in Europe
  - Contributing to sustaining **data**, **technical services** and above all **expertise**
  - The central role of researchers: DARIAH as **collective intelligence**
  - **Sharing with outsiders:** DARIAH as a platform

DARIAH-EU

# Building up an infrastructure

- **2006:** DARIAH @ ESFRI Roadmap

Digital Research Infrastructure for the Arts and Humanities DARIAH

- **2008 – 2011:** Preparatory Phase project — *Preparing DARIAH*

DARIAH Digital Research Infrastructure for the Arts and Humanities

- **2011 – 2013:** Transition Phase establishing the DARIAH-ERIC

DARIAH-EU

- **2014:** DARIAH-ERIC

DARIAH-ERIC

DARIAH-EU

DARIAH-EU

Digital Research Infrastructure
for the Arts and Humanities

# DARIAH as an ERIC

| Founding Members | |
|---|---|
| Austria | 🇦🇹 |
| Belgium | 🇧🇪 |
| Croatia | 🇭🇷 |
| Cyprus | 🇨🇾 |
| Denmark | 🇩🇰 |
| France | 🇫🇷 |
| Germany | 🇩🇪 |
| Greece | 🇬🇷 |

| Founding Members | |
|---|---|
| Ireland | 🇮🇪 |
| Italy | 🇮🇹 |
| Luxembourg | 🇱🇺 |
| Malta | 🇲🇹 |
| Netherlands | 🇳🇱 |
| Serbia | 🇷🇸 |
| Slovenia | 🇸🇮 |

| Recent members | |
|---|---|
| Poland | 🇵🇱 |
| Portugal | 🇵🇹 |

Cooperating partners in:

- Switzerland
- Sweden
- UK

www.dariah.eu

# Virtual Competence Centers

**VCC 1 – e-Infrastructure**

1. A+H Infrastructure Services
2. A+H Research Environment demonstrators
3. A+H Service Environment
4. Data Federation and interoperability
5. Developer community
6. Preservation Infrastructure
7. Reference Software Packages

**VCC 2 - Research & Education**

1. Community engagement
2. Training and Education Programme
3. Understanding research practices
4. Virtual Research Environment

**VCC 3 – Scholarly Content Management**

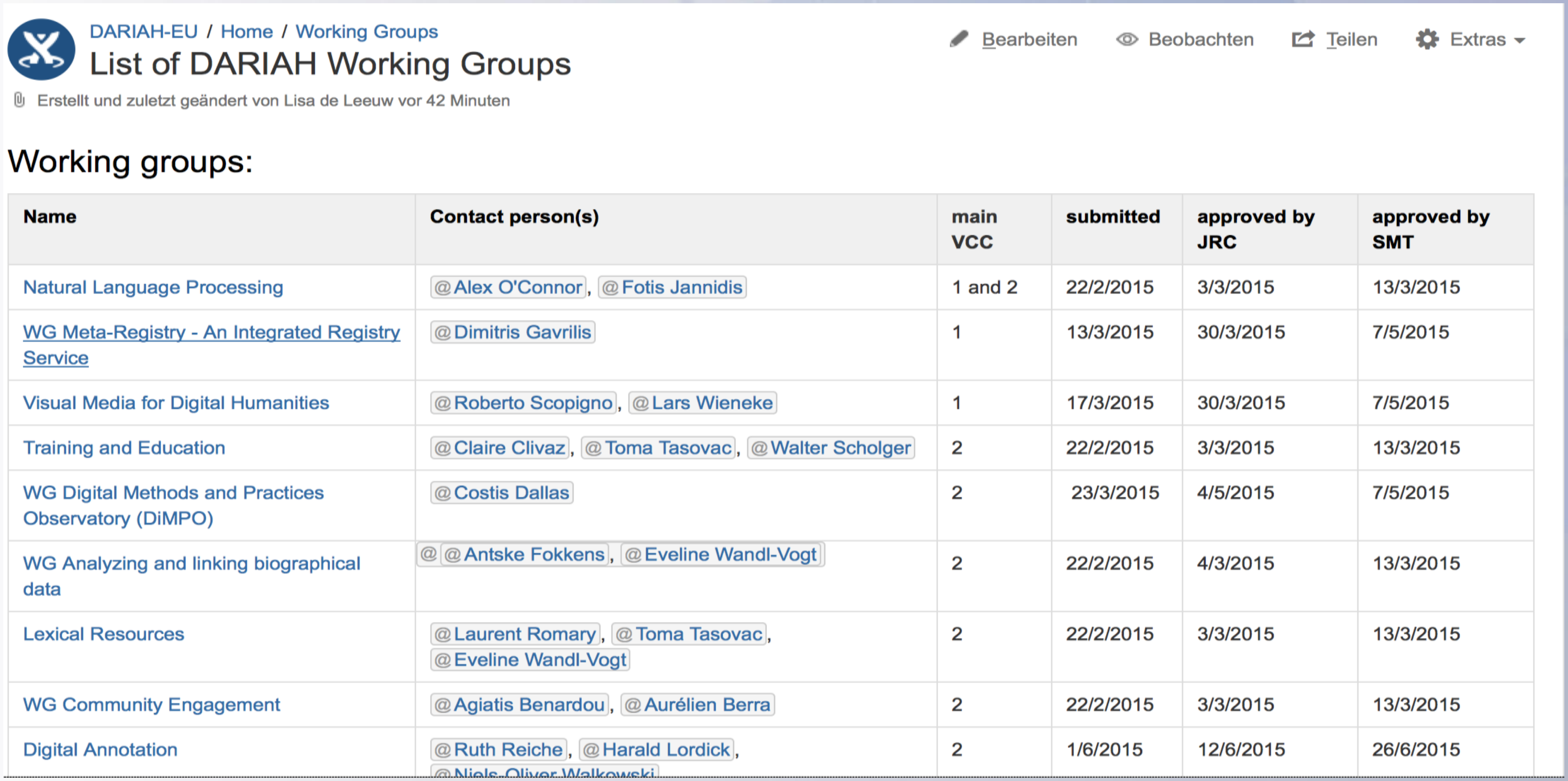
1. Best Practices and Open Access
2. Curation
3. Dissemination and Digital Publishing
4. Enrich digital scholarly content
5. Reference Data Registries

**VCC 4 – Advocacy, Impact & Outreach**

1. Ensuring capacity in DARIAH
2. Ensuring Participation in DARIAH
3. High-level Advocacy
4. Impact and value
5. Outreach

DARIAH-EU

# Working Groups – going bottom-up



DARIAH-EU / Home / Working Groups

## List of DARIAH Working Groups

Erstellt und zuletzt geändert von Lisa de Leeuw vor 42 Minuten

### Working groups:

| Name | Contact person(s) | main VCC | submitted | approved by JRC | approved by SMT |
|---|---|---|---|---|---|
| Natural Language Processing | @Alex O'Connor, @Fotis Jannidis | 1 and 2 | 22/2/2015 | 3/3/2015 | 13/3/2015 |
| WG Meta-Registry - An Integrated Registry Service | @Dimitris Gavrilis | 1 | 13/3/2015 | 30/3/2015 | 7/5/2015 |
| Visual Media for Digital Humanities | @Roberto Scopigno, @Lars Wieneke | 1 | 17/3/2015 | 30/3/2015 | 7/5/2015 |
| Training and Education | @Claire Clivaz, @Toma Tasovac, @Walter Scholger | 2 | 22/2/2015 | 3/3/2015 | 13/3/2015 |
| WG Digital Methods and Practices Observatory (DiMPO) | @Costis Dallas | 2 | 23/3/2015 | 4/5/2015 | 7/5/2015 |
| WG Analyzing and linking biographical data | @ @Antske Fokkens, @Eveline Wandl-Vogt | 2 | 22/2/2015 | 4/3/2015 | 13/3/2015 |
| Lexical Resources | @Laurent Romary, @Toma Tasovac, @Eveline Wandl-Vogt | 2 | 22/2/2015 | 3/3/2015 | 13/3/2015 |
| WG Community Engagement | @Agiatis Benardou, @Aurélien Berra | 2 | 22/2/2015 | 3/3/2015 | 13/3/2015 |
| Digital Annotation | @Ruth Reiche, @Harald Lordick, @Niels-Oliver Walkowski | 2 | 1/6/2015 | 12/6/2015 | 26/6/2015 |

DARIAH-EU

# Services

Access

Expertise

Interoperability

Hosting content

Tools and Software

Training

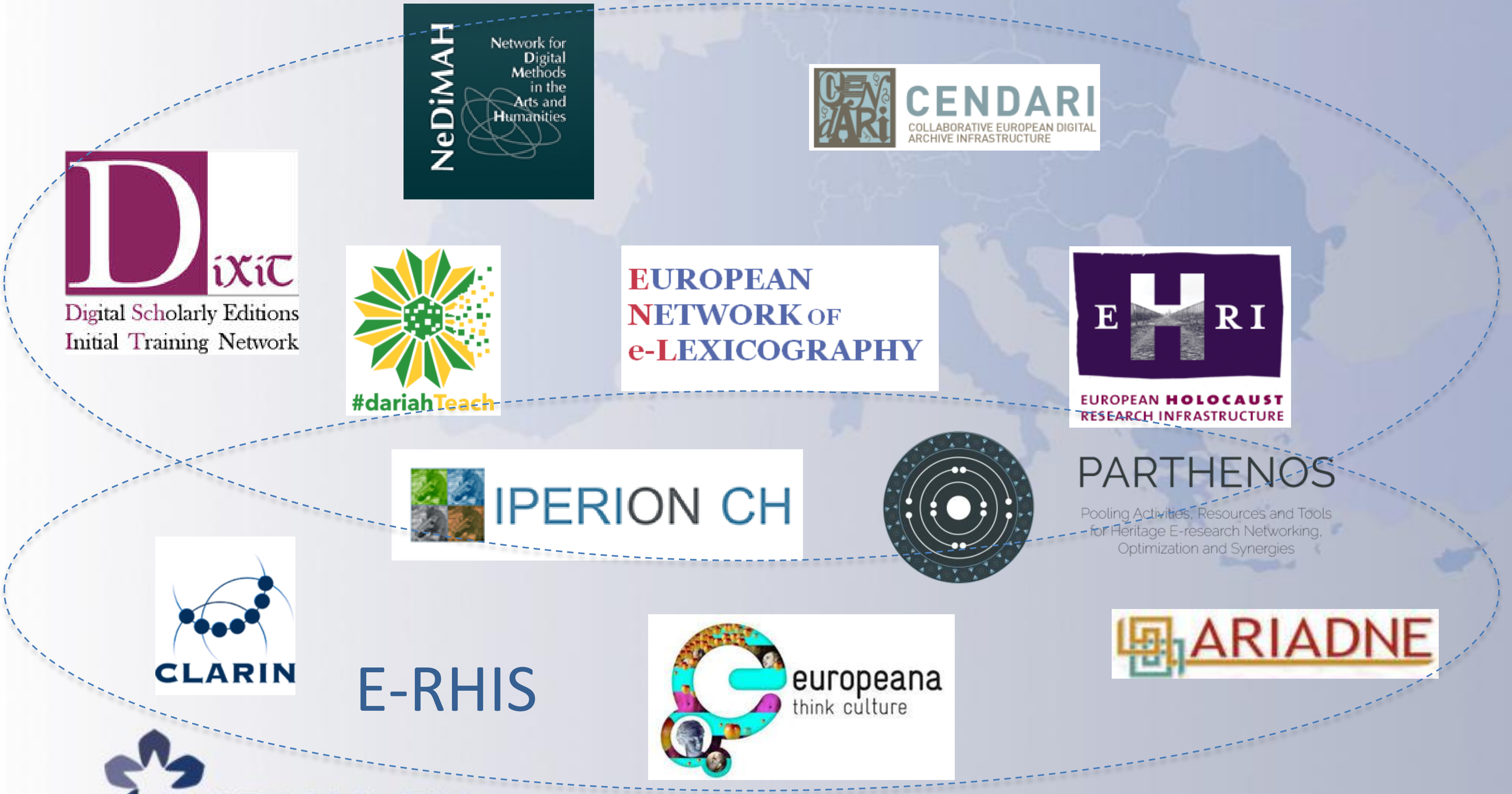
Summer school

Event

Coordination

DARIAH-EU

# Communities within and around DARIAH



NeDiMAH Network for Digital Methods in the Arts and Humanities

CENDARI COLLABORATIVE EUROPEAN DIGITAL ARCHIVE INFRASTRUCTURE

DiXiT Digital Scholarly Editions Initial Training Network

#dariahTeach

EUROPEAN NETWORK OF e-LEXICOGRAPHY

EHRI EUROPEAN HOLOCAUST RESEARCH INFRASTRUCTURE

IPERION CH

PARTHENOS Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies

CLARIN

E-RHIS

europeana think culture

ARIADNE

DARIAH-EU

DESIR
a DARIAH-EU project

- Horizon 2020 funded project
- Call: H2020-INFRADEV-2016-2017

(Development and long-term sustainability of new pan-European research infrastructures)

- Starting date: January 2017
- Duration: 36 months
- Consortium: 15 partners from all over Europe

## WP3: GROWTH

- New Countries: UK, Czech Republic, Spain, Switzerland, Finland, Israel.
- Prepare DH RI Country Reports and Develop Specific Accession Strategy and Action Plans
- Coordinate, Monitor and Support Enlargement

DARIAH-EU

# Data fluidity

- Progress so far in several directions
  - Improvement of digital competence: #DARIATeach, DARIAH course registry
  - Development and maintenance of reference standards: TEI, ISO TC 37, archival standards
  - Progress in tooling and data hosting: Nakala, Ortolang
- One main issue: using and re-using content

DARIAH-EU

# The loneliness of the researcher

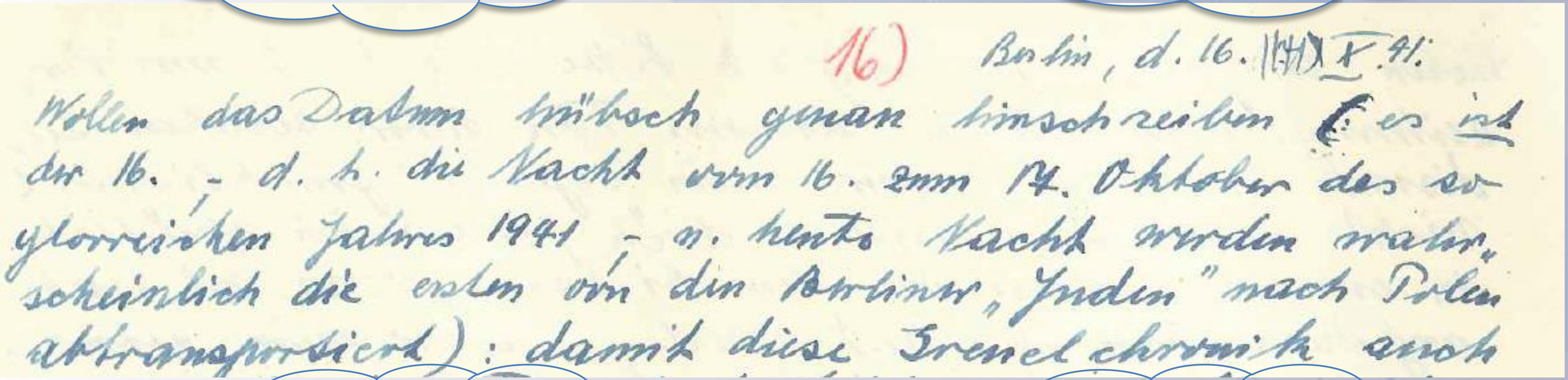
- Conditions of access, re-use and communication of primary sources
  - How much can I take and re-use from a Cultural Heritage Institution ?
    - Illustration, citation, scans
  - How much am I allowed to disseminate?
    - Transcriptions, annotations, collations, mash-ups
  - Which recognition will I gain from this work?
    - From traditional publishing to online digital editions

DARIAH-EU

# What am I allowed to do?

Documenting:
origin, date,
material

Qualifying:
authorship, research
value, authenticity



16) Berlin, d. 16. (17.) X. 41.
Wollen das Datum hübsch genau hinschreiben (es ist der 16. – d. h. die Nacht vom 16. zum 17. Oktober des so glorreichen Jahres 1941 – u. heute Nacht werden wahrscheinlich die ersten von den Berliner „Juden" nach Polen abtransportiert): damit diese Greuelchronik auch

Analyzing:
layout, transcription,
names, dates

Communicating:
corpus, rights,
contextualization

DARIAH-EU

Source: *L. Alt's diary*

# Data Re-use Charter

- Framing the conditions of collaboration between Cultural Heritage Institutions and Researchers

- Defines the key elements each of the parties commits to in terms of data access, use and re-use

- Online commitment to guaranty immediate reciprocal awareness, and thus create a trusted network of stakeholders

DARIAH-EU

# Stakeholders

- Cultural Heritage Institutions (big or small)
  - Main source of primary information for research in the humanities (physical and/or digital)
- Equipments (big and small...)
  - Data production by researchers on CH objects
- Data hosting institutions
  - Warrant the stability, the visibility and the long time availability of the primary data
- Researchers
  - Compile, analyse, enrich and disseminate CH content
  - Engaging in person or via their HER institution

# Common commitments

- Access
  - Metadata, primary surrogates, transcriptions
- Licensing
  - May depend on types of data and specific collections
- Dissemination
  - Technical requirements, citation rules, associated publications
- Enrichments
  - Re-use, hosting and visibility of scholarly work
  - Aiming jointly at improving quality of digital resources

DARIAH-EU

# Perspectives

- Timeline
  - Charter outline
  - Setting up a round table with representatives of all types of stakeholders
  - Online prototype with first participants
  - Official kick-off in conjunction with the IFLA conference in 2017
- Making the charter a reference communication tool for CHI‘s and scholars
  - Strong collaboration needed with: Europeana, Clarin and E-RIHS
  - Support from the Parthenos and Iperion project

DARIAH-EU

Digital Research Infrastructure
for the Arts and Humanities

# STANDARDISATION IN DARIAH AND PARTHENOS

www.dariah.eu

# Standards, standards, everywhere…

- Standards: non legally binding documents produced by organisation ensuring
  - International **consensus** building (not a one person's/one group's work)
  - **Communication** (standards cannot be internal to an organisation)
  - **Maintenance** (evolution according to users' needs, technology change etc.)
- ISO, W3C, TEI comply to this principles
- Important distinction:
  - Horizontal standards (cross-domain: ex. XML)
  - vs. Vertical standards (specific to one domain: ex. ISO 24611 for morpho-syntactic annotations, EAD for archival collections)

DARIAH-EU

# Standards in the Arts and Humanities

- Well established practices (examples)
  - Text Encoding Initiative
    - Comprehensive XML vocabulary for textual documents
  - ISO TC 37/SC 4 portfolio
    - Mature set of standards for linguistic annotation
- Niches to be secured
  - MEI (Music Encoding Initiatives)
- Lacunae (or fragmentations) to overcome
  - E.g. Descriptive metadata for audio-visual data
- Library and archival standards to interoperate with
  - METS, MODS, FRBR descriptive framework
  - EAD, EAG, EAC

DARIAH-EU

# Standards and scholarly communities

- Informing
  - Minimal knowledge to express requirements towards implementers
  - Being able to design community data models in terms of existing standards
- Participating
  - Keeping standards as close as possible to the need of communities
  - Training of standards experts within scholarly communities
- Looking ahead
  - Standards should anticipate on future needs and use cases
    - E.g. Provision of generic and customisation mechanisms
      - E.g. Unicode, TEI

DARIAH-EU

# Striving for impact

- The variety of document types in the humanities provide a wealth of rich data models
  - Potential impact on other communities with similar needs
- A typical example: the European Patent Office
  - Some figures
    - Several thousands of examiners
    - 200 million documents
    - 2 billion annotations…
  - A TEI based model for their back-office document platform
    - Families, Applications, Documents
- Importance of a real standardisation strategy and high-quality data models

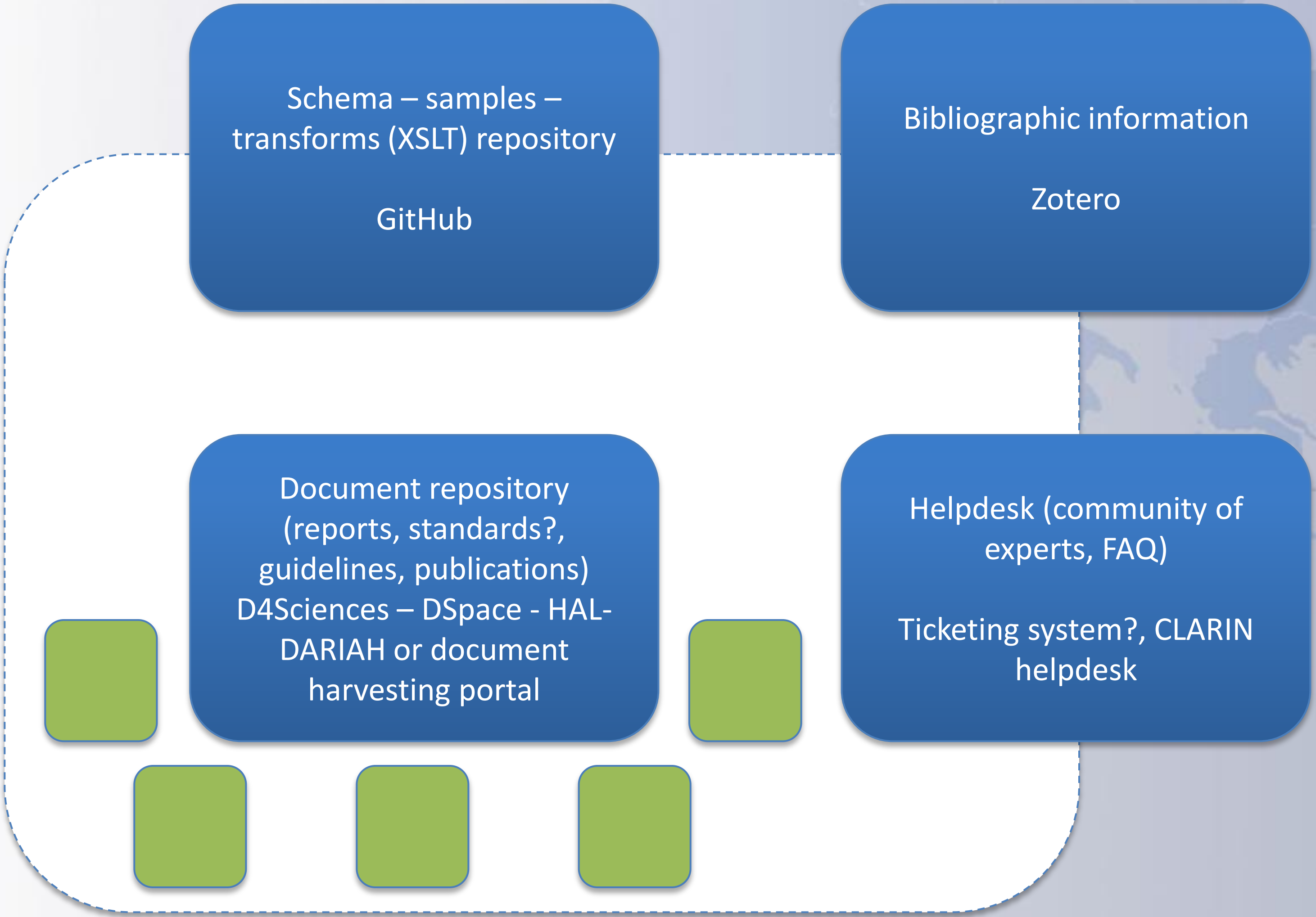
DARIAH-EU

# Implementing the vision through Parthenos

- H2020 Infrastructure project
  - WP4: dedicated to standards
- How to proceed?
  - Documenting, recommending and disseminating information about well-established standards (SSK)
  - Fostering the stabilisation of stable community initiatives
  - Launching standardisation think-tanks for under-covered domains

DARIAH-EU

# Organising knowledge about standards



Schema – samples – transforms (XSLT) repository

GitHub

Bibliographic information

Zotero

Document repository (reports, standards?, guidelines, publications) D4Sciences – DSpace - HAL-DARIAH or document harvesting portal

Helpdesk (community of experts, FAQ)

Ticketing system?, CLARIN helpdesk

# Issues

- Limiting new developments
  - Existing infrastructures
- Ensuring genericity
  - Expert network
- Agile deployment
  - No big plan
- Thinking sustainable right from the onset
  - Our role as infrastructure

DARIAH-EU

Digital Research Infrastructure
for the Arts and Humanities

# THE TEXT ENCODING INITIATIVE: 30 YEARS OF ACCUMULATED WISDOM AND ITS POTENTIAL FOR A BRIGHT FUTURE

www.dariah.eu

# 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 became 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)

DARIAH-EU

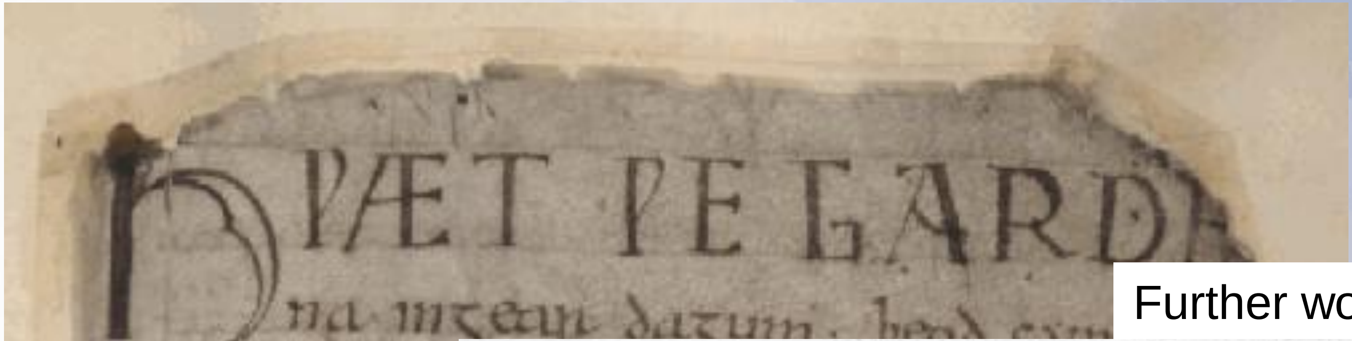
# TEI for digital scholarly work

- A trend towards digital curatorship
  - Describing digital sources: meta-data
  - Understanding and representing the structure of digital sources: content
  - Enriching (annotations, links), versioning, disseminating
- A wide user community
  - From individual scholars to large digitization projects

DARIAH-EU

# The standard scenario?

Digitizing source documents



HƿÆT ƿE GARDE

na ingear dagum. þeod cyn

þrym ge fr

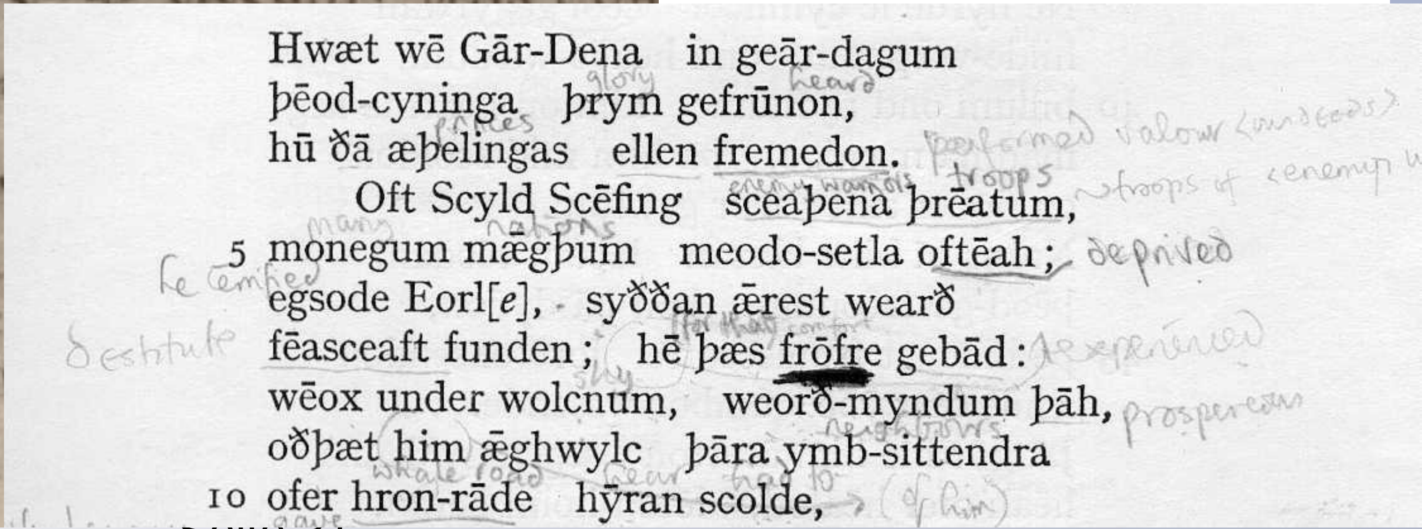
fre medon

þreatum m

of teah e

fea sceaft

Further work on documents



Hwæt wē Gār-Dena   in geār-dagum
þēod-cyninga   þrym gefrūnon,
hū ðā æþelingas   ellen fremedon.
Oft Scyld Scēfing   sceaþena þrēatum,
5 monegum mǣgþum   meodo-setla oftēah;
egsode Eorl[*e*], syððan ǣrest wearð
fēasceaft funden;   hē þæs frōfre gebād:
wēox under wolcnum,   weorð-myndum þāh,
oðþæt him ǣghwylc   þāra ymb-sittendra
10 ofer hron-rāde   hȳran scolde,

glory
heard
princes
performed valour (wondrous)
enemy warriors
troops
~troops of (enemy w
many
nations
deprived
he terrified
Destitute
for that comfort
experienced
sky
prospered
neighbours
whale road
hear
had to
(of him)
gave

# 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
    - Recorded 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”: annotation

DARIAH-EU

# TEI –core principles (2)

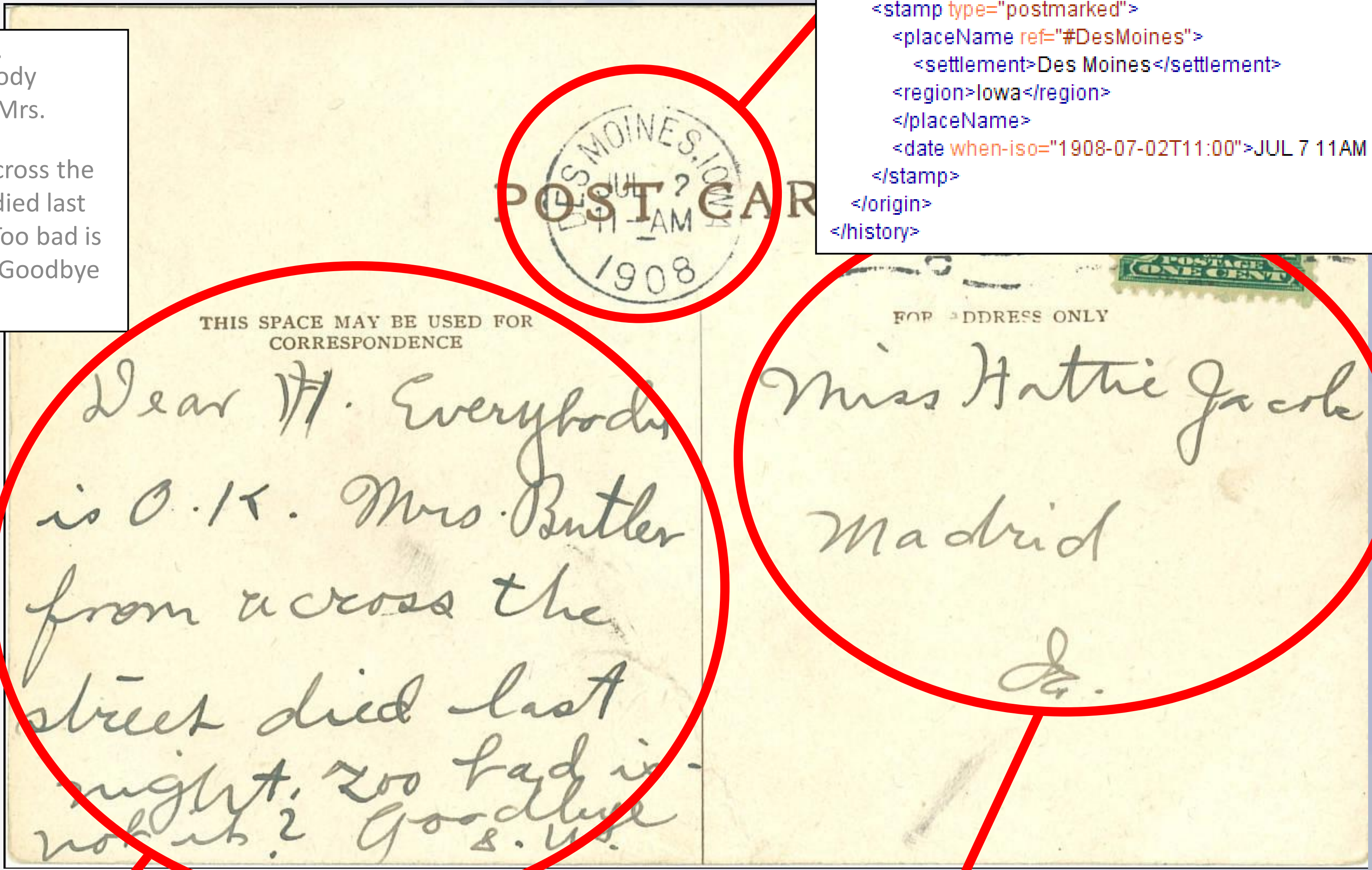
- Favoring the semantics rather than the layout
  - (quasi) No presentational construct
  - Publication requires a transformation stage (XSLT; ePub, pdf, HTML, etc.)
- Document structure
  - Macro-structure: front-body-back
  - Meso-structure: divisions, paragraphs/lists/figures/etc.
  - Micro-structure: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

# ALL YOU CAN ENCODE…

# Examples

- Simple encoded text
  - The Little Riding Hood
- Scholarly paper
  - Towards Higher Ground
- Dictionaries
  - Larousse

- Dear H. Everybody
- is O.K. Mrs. Butler
- from across the
- street died last
- night. Too bad is
- not it? Goodbye
- S. W.



DES MOINES, IOWA
JUL 7
11 AM
1908

POST CARD

POSTAGE ONE CENT

THIS SPACE MAY BE USED FOR CORRESPONDENCE

Dear H. Everybody
is O.K. Mrs. Butler
from across the
street died last
night. Too bad is
not it? Goodbye
S. W.

FOR ADDRESS ONLY

Miss Hattie Jacobs
Madrid
Ia.

```
<history>
  <origin>
    <stamp type="postmarked">
      <placeName ref="#DesMoines">
        <settlement>Des Moines</settlement>
      <region>Iowa</region>
      </placeName>
      <date when-iso="1908-07-02T11:00">JUL 7 11AM 1908</date>
    </stamp>
  </origin>
</history>
```

```
<div type="back" facs="#noble0337b">
  <div type="left">
    <salute>Dear <persName ref="#HJ">H</persName>.</salute>
    <p>Everybody <lb/>is O.K. Mrs. Butler <lb/>from across the <lb/>street died
      last <lb/>night. Too bad is <lb/>not it?</p>
    <signed>Goodbye <lb/><persName>S. W.</persName></signed>
  </div>
```

```
<div type="right">
  <p>
    <address>
      <addrLine>Miss <persName ref="#HJ">Hattie Jacobs</persName></addrLine><lb/>
      <placeName ref="#Madrid"><settlement>Madrid</settlement><lb/>
      <region>Ia</region></placeName>
    </address>.
  </p>
</div>
```

# HOW DO YOU MANAGE THIS?

# 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 as a standardization body (1)

- Consensus building
  - Community based decision process
- Maintenance
  - Two releases per year
- Publication
  - All TEI contents are available under the double CC-BY+BSD 2 clause license.

# TEI as a standardization body (2)

- Organization
  - Consortium of institutional and individual members
  - Conference, journal (jTEI)
- The TEI at work
  - Board: administrative aspects
  - Technical council: coordinates the evolution of the TEI guidelines

DARIAH-EU

# Standardization work

- Community based workflow
  - Mailing list
  - GitHub – bugs and features
    - Recording all issues and decisions
  - Cf. ODD as a specification platform
- Deliverables
  - Documentation — TEI guidelines (more than 500 elements)
  - Schemas — DTD, RelaxNG, W3C
- Additional resources
  - Tools
    - Online customization: Roma
    - Online processing: OxGarage
    - Stylesheets (included in Oxygen)
  - Examples — TEI by Example

DARIAH-EU

# Special Interest Groups (SIGs)

- Computer-Mediated Communication (Michael Beißwenger)
- Correspondence. (Peter Stadler and Joachim Veit)
- Education (TBA)
- Libraries (Stefanie Gehrke and Kevin Hawkins)
- Manuscripts (Dot Porter and Gerrit Brüning)
- Music (Raffaele Viglianti)
- Ontologies (Oyvind Eide and Christian-Emil Ore)
- Scholarly Publishing (Daniel O'Donnell)
- TEI for Linguists (Piotr Bański and Andreas Witt)
- Text and Graphics (John Walsh and Martin de la Iglesia)
- Tools (Serge Heiden)

DARIAH-EU

# Varieties of TEI Conformance

- Pure *TEI-all* subset
  - Most TEI projects
- TEI subset with extensions
  - E.g. adding TBX terminologies in TEI
- Non TEI document with TEI constructs (defined as an ODD)
  - EAG extensions in the EU Cendari project
- Non TEI document defined by means of an ODD document
  - E.g. ISO 24616:2012 Language resources management -- Multilingual information framework

# The central role of customization

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

DARIAH-EU

# 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 n ODD
- Does not prevent one from knowing the TEI components…
  - Most projects can live with just a subset of the TEI ontology
    - With the strong possibility to impact on the guidelines themselves
      - E.g. <abstract>

DARIAH-EU

Digital Research Infrastructure
for the Arts and Humanities

# TEI: RETURN TO THE FUTURE

www.dariah.eu

# TEI: you're not alone…

- The hidden TEI: scientific information at the European Patent Office
- New components in the TEI: <standOff>
- Working with others: ISO LMF

DARIAH-EU

DARIAH-EU

Digital Research Infrastructure
for the Arts and Humanities

# SCIENTIFIC INFORMATION?

www.dariah.eu

# 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

DARIAH-EU

# 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/representati ve
- Opponents
- … and *examiners*

DARIAH-EU

# 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

DARIAH-EU

# 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…

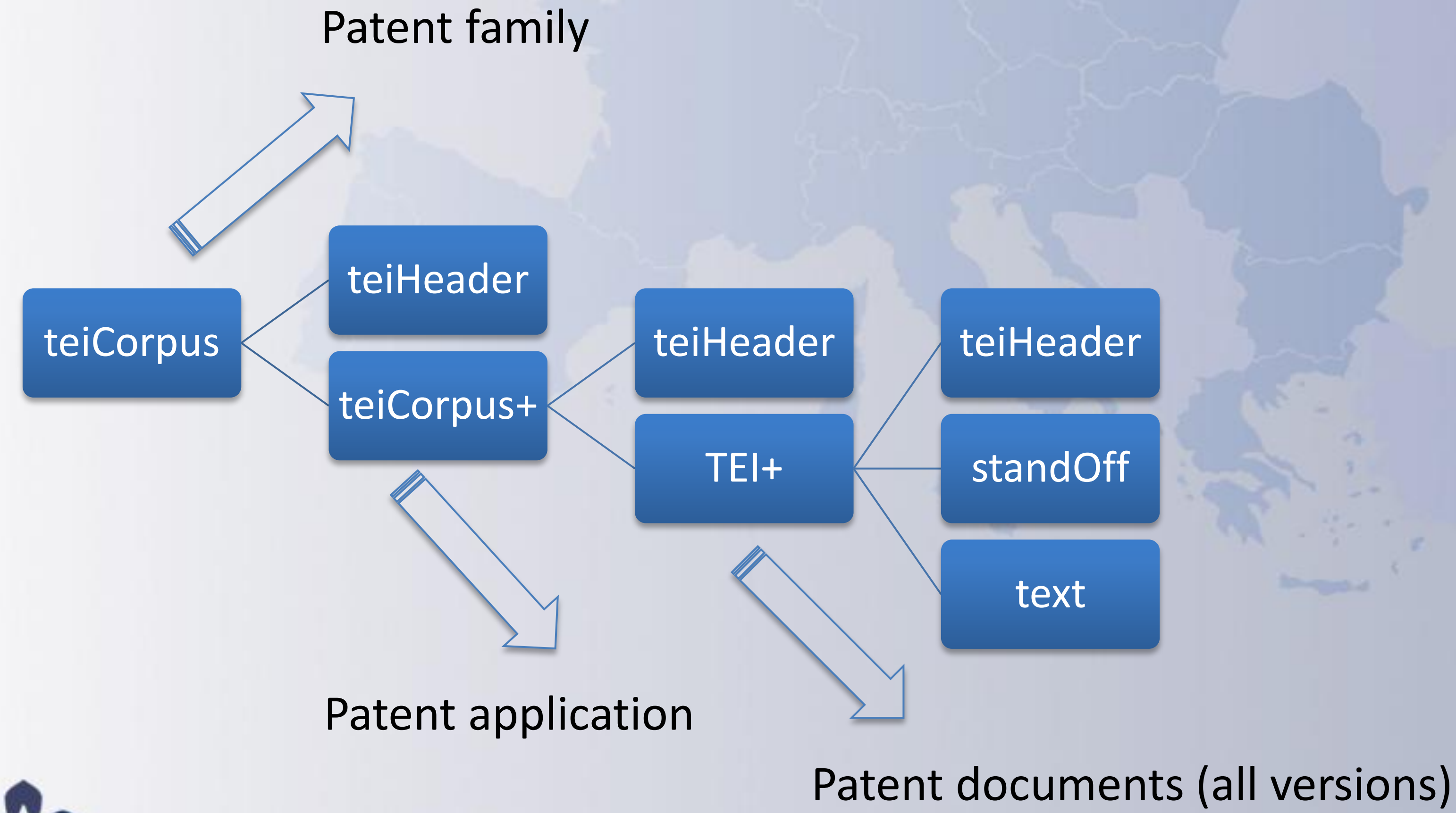
# The (simplified) patent life-cycle

- 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”)
  - 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

- The TEI appeared to be the optimal choice

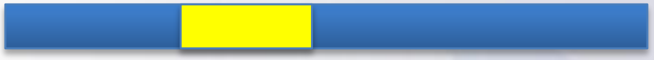
DARIAH-EU

# The Patent Document Model



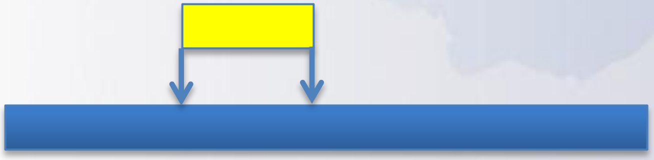
# WAKE UP STAND-OFF!

# The simple picture



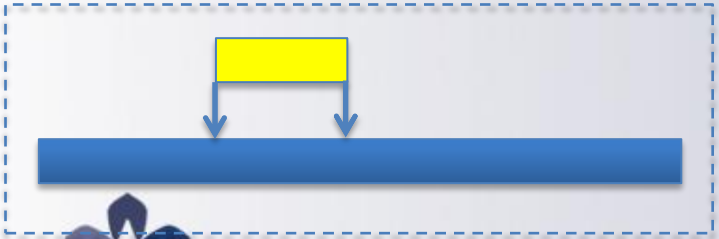
*Inline annotation:*

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

DARIAH-EU

# 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>, <span>, <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

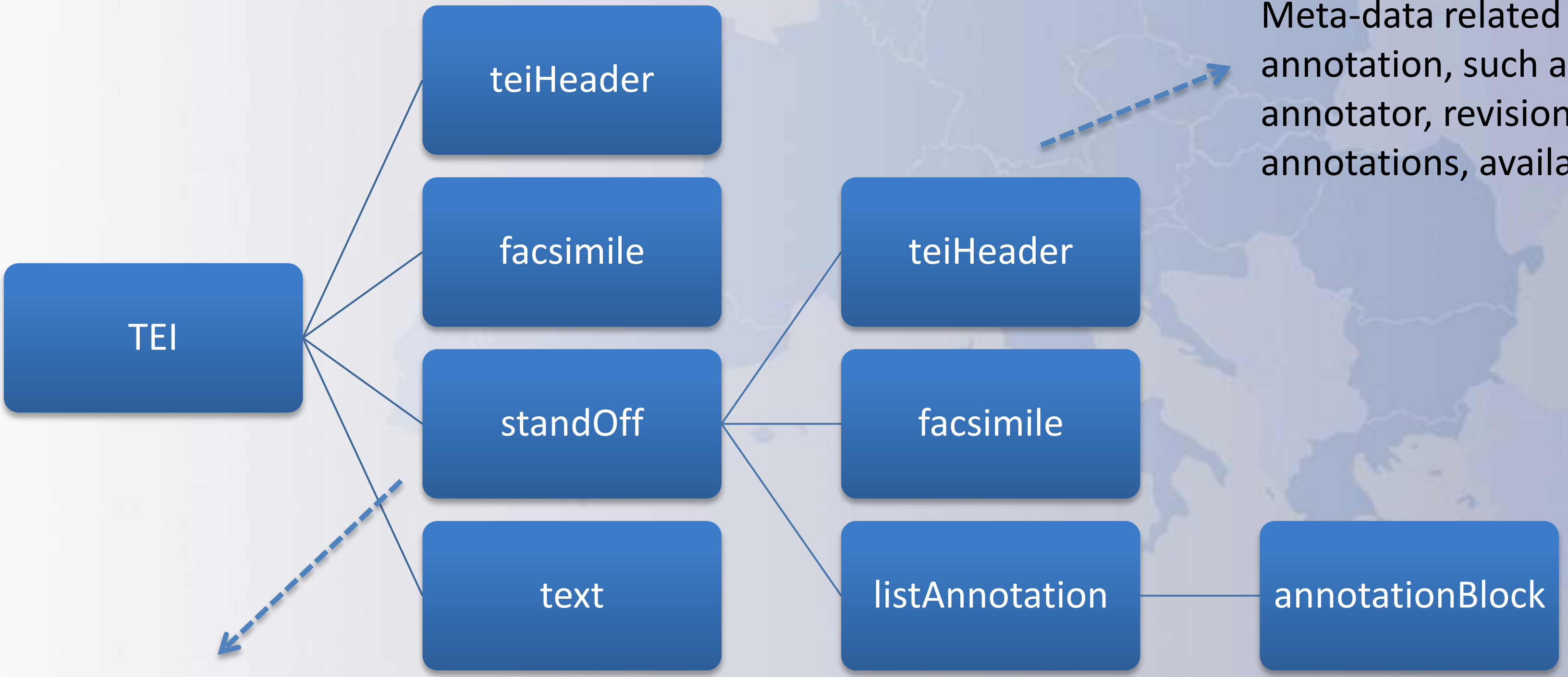
# Embedded standoff: Basic concept

- 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

# 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
- March 2016: TEI release 6.0.0
  - New element <annotationBlock> for interlinear annotation
- August 2016: publication of ISO 24624 Transcription of Spoken Language

# 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

<div>-like component for structuring complex series of annotations

Elementary annotation unit

DARIAH-EU

# Application: interlinear annotation

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

```
<annotationBlock 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>
</annotationBlock>
```

# Standoff interlinear annotation

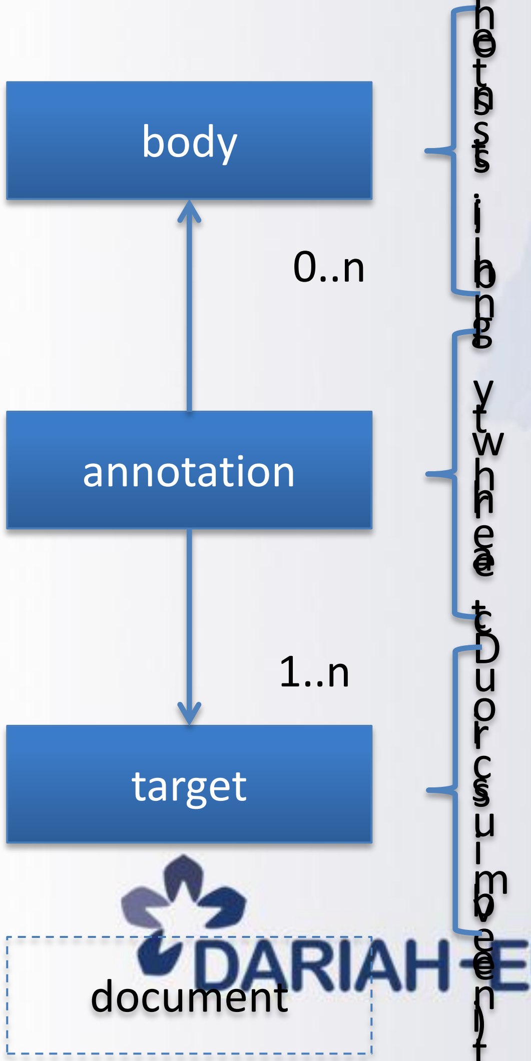
- Encoding interlinear annotation as stand-off markup
  - In <standOff>

```
<annotationBlock inst="#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>
</annotationBlock>
```

  - 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>
```

# Going further: mapping the Open Annotation model



[illegible]

<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>

# 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

# Issues (many)

- Which header do we need?
  - Standoff annotation usually requires very restricted 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>?)
  - Aligning with the various ISO standards: MAF, SynAF and SemAF series

# Next steps

- Finalising the content model of <annotationBlock>
  - Completely open model?
  - Constrained with specific model classes? (OA)
  - Alternation between the two (or more) options
- Gathering reference example from existing implementations
  - Istex, Termith, EPO, IDS
- Finalising the graft in the guidelines
  - Section in chapter 16 Linking, Segmentation, and Alignment?
- Don’t give up the fight…

DARIAH-EU

DARIAH-EU

Digital Research Infrastructure
for the Arts and Humanities

# JOINING EFFORTS WITH OTHERS: TEI AND LMF

www.dariah.eu

# A divided landscape

- The TEI print dictionary chapter
  - Available since more than 20 years
    - See http://www.tei-c.org/Vault/Vault-GL.html
  - Used in a wide variety of dictionary projects
    - 6 entries just in http://www.tei-c.org/Activities/Projects/
    - Disseminated at quick pace within the COS E-NEL network (credits: Toma Tasovac)
- ISO 24613:2008 Language resource management - Lexical markup framework (LMF)
  - Shorter life span
  - Mostly implemented in NLP related activities
- Is it worth reconciling the 2?
  - Yes: for the sake of combining a well-defined model with a rich XML infrastructure
  - A need for the TEI to have a terser model
    - Curation, interchange, tools, automatic generation of TEI constructs
- Is it just possible.
  - Yes: and now!

DARIAH-EU

# The need for a revision

- Main assets of ISO 24613 LMF
  - Comprehensive core model + series of annexes for additional modules
  - Perfectible XML serialisation…
- Going towards a multi-part standard
  - Simplifying the editorial process (drafting, decision making, revising; various tempi)
  - Reflecting the needs of specific communities (modules, serialisation)

DARIAH-EU

# Overview of the current plans

- Resolution 2016-04.2 (WG 4) Multi-part development of LMF
  - Part 1: Core model
  - Part 2: MRDs
  - Part 3: Diachrony-Etymology
  - Part 4: TEI serialisation
  - Part 5: LBX serialisation

DARIAH-EU

# (Part 4) A TEI serialisation for LMF

- Objective
  - Preventing re-inventing element that already exist
  - Eliciting constraints on the TEI model
- Method
  - Covering core model and a selected number of extensions
    - Remaining in the scope of the Print dictionary chapter
    - Extending scope if we feel there is a need from the potential TEI applications (e.g. syntax)
  - Sub-setting the TEI guidelines
    - Associating a definite TEI construct for each component of the LMF Meta-model
    - Adding constraints when necessary
      - (e.g. @xml:lang mandatory on <entry>?)
  - Complementing the TEI
    - Defining new constructs (or elements?) if necessary
      - We are not bound to the existing chapter, even if we have to abide to the Birnbaum principle

# Gathering mapping proposals

| Component | TEI construct |
|---|---|
| Lexical Entry | <entry>…</entry> |
| Form | <form>…</form> |
| Lemma | <form type="lemma">…</form> |
| Word Form | <form type="inflected">…</form> |
| Syntactic Behaviour | ?? |
| ?? | <etym> |

| Data category | TEI construct |
|---|---|
| /PartOfSpeech/ | <pos> |
| /Gender/ | <gen> |
| … | |

How far should we go here?

# Once upon a time, the clergyman…

```
<entry xml:lang=“en”>
  <form type="lemma">
    <orth>clergyman</orth>
    <gramGrp>
      <pos>commonNoun</pos>
    </gramGrp>
  </form>
  <form type="inflected">
    <orth>clergyman</orth>
    <gramGrp>
      <number>singular</number>
    </gramGrp>
   </form>
   <form type="inflected">
     <orth>clergymen</orth>
     <gramGrp>
       <number>plural</number>
     </gramGrp>
   </form>
</entry>
```

# (Part 3) The case of etymology

- A flat model in the current TEI chapter
  - No sense of etymon: <mentioned>
  - No sense of etymological process
    - Typed and recursive <etym>
  - No grouping of etymon related information
    - Usage, grammatical constraints, source, date, language, etc.
- A need for revision
- Pushing a fully fledged model

DARIAH-EU

# Before-after example

## Old school

```
<etym>
  <lang>Ahd.</lang>
<mentioned>âband</mentioned>,
<lang>mhd.</lang>
<mentioned>âbent</mentioned>;
<bibl>zur Etym. s. Kluge Mitzka 18. Aufl.
unter ,,Abend'', ferner Schwäb. Wb. 1,
11ff. Schweizdt. Wb. 1,34ff.</bibl>
</etym>
```

## Structured

```
<etym type="inheritance">
  <cit type="etymon" xml:lang="goh">
    <oRef>âband</oRef>
    <lang>Ahd.</lang>
  </cit>
  <etym type="inheritance">
    <cit type="etymon" xml:lang="gmh">
      <oRef>âbent</oRef>
      <lang>mhd.</lang>
    </cit>
  </etym>
  <bibl>zur Etym. s. Kluge Mitzka 18.
Aufl. unter ,,Abend'', ferner Schwäb. Wb.
1, 11ff. Schweizdt. Wb. 1,34ff.</bibl>
</etym>
```

DARIAH-EU

# An interesting moment

- Time to complement and consolidate the existing practices
  - TEI as reference framework
  - ISO as a precise standardisation background
- Various ongoing projets and groups
  - Clarin Standards committee, DARIAH WG Lexical Resources, TEI LingSIG
  - COST E-NEL WG2, EU Parthenos
- Joining efforts
  - Towards a single information space
    - Basecamp, GitHub, Blog
    - Exchanging information
  - Increasing participation as experts
    - ISO-TEI in particular

DARIAH-EU

DARIAH-EU

Digital Research Infrastructure
for the Arts and Humanities

# WHITHER TEI?

www.dariah.eu

# 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
  - Numerous training events (cf. DiXiT)
- Taken up by larger organisations
  - Academies, Dictionary projects, EPO… especially in Europe

DARIAH-EU

# Consolidating our conceptual model

- TEI as a rich space of elementary constructs
- 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
- TEI as a data modelling infrastructure

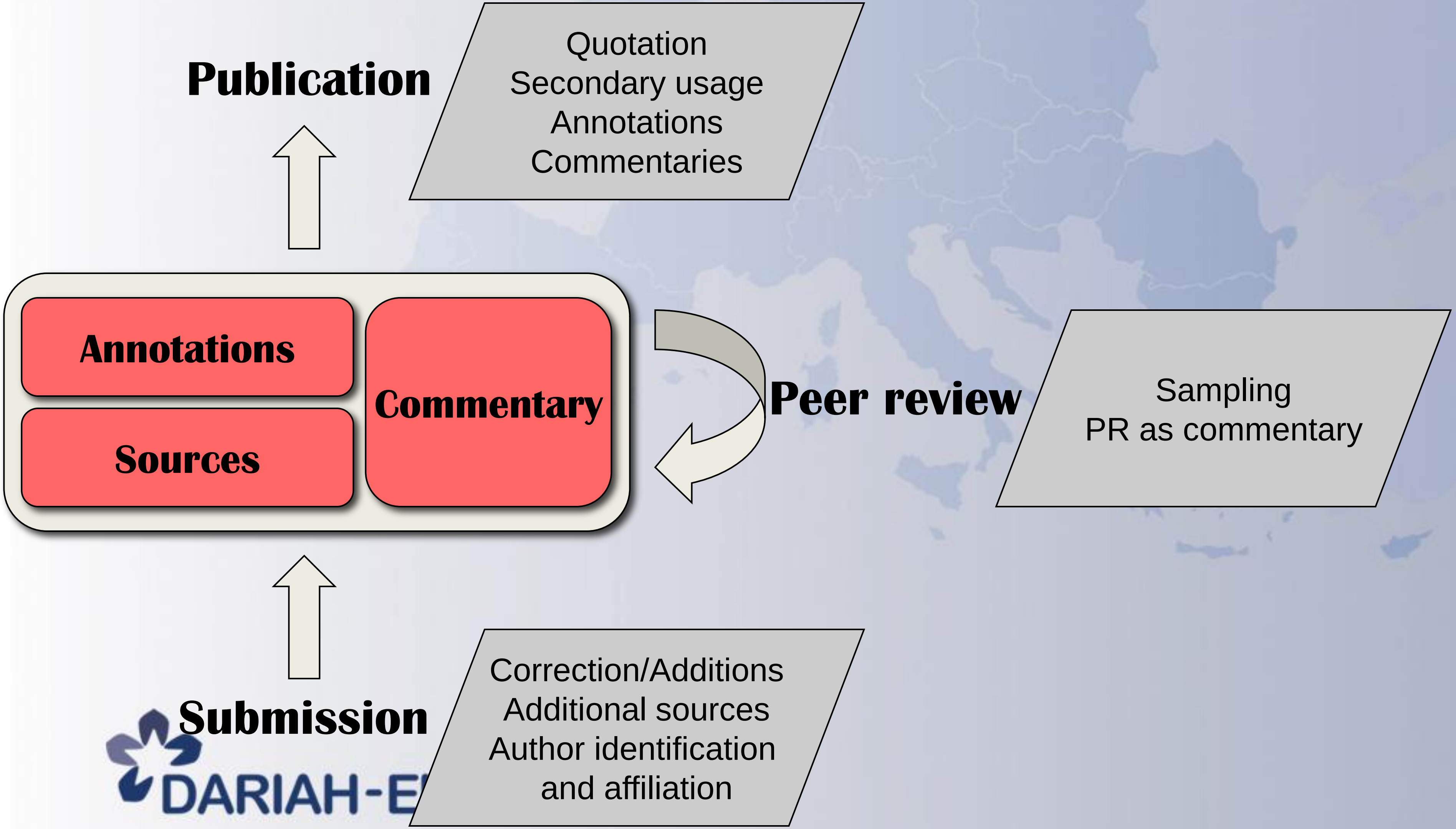
DARIAH-EU

# Focusing, enlarging?

- Enlarging our expert basis
  - 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
- A basis for data publications?
  - Code name: *Living sources*

DARIAH-EU

# *Living sources*



**Publication**

Quotation
Secondary usage
Annotations
Commentaries

**Annotations**

**Sources**

**Commentary**

**Peer review**

Sampling
PR as commentary

**Submission**

Correction/Additions
Additional sources
Author identification
and affiliation

DARIAH-EU

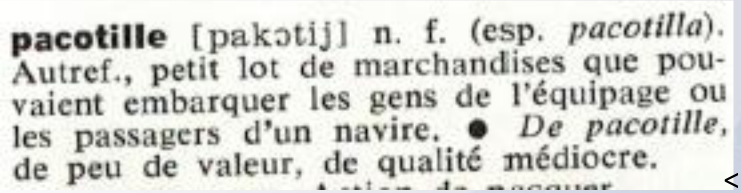
Digital Research Infrastructure
for the Arts and Humanities

# MERCI !

www.dariah.eu

# Automatic dictionary structure recognition

PhD theses by Mohamed Khemakhem (Inria, projet H2020 Parthenos)



**pacotille** [pakɔtij] n. f. (esp. *pacotilla*). Autref., petit lot de marchandises que pouvaient embarquer les gens de l'équipage ou les passagers d'un navire. ● *De pacotille*, de peu de valeur, de qualité médiocre.

*CRF (Conditional Random Fields) based data mining*

```
<entry>
        <form type="lemma">
          <orth>pacotille</orth>
          <pron><pc>[</pc>pakɔtij<pc>]</pc></pron>
          <gramGrp><pos>n.</pos><gen>f.</gen></gramGrp>
        </form>
        <etym><pc>(</pc><lang norm="es">esp.</lang>
          <oRef>pacotilla</oRef><pc>)</pc></etym>
        <sense>
          <usg type="time">Autref.</usg> <pc>,</pc>
          <def>petit lot de marchandises que pouvaient embarquer les gens de
l'équipage ou les passagers d'un navire</def>
          <pc>.</pc>
        </sense>
        <re>
          <form type="compound">
            <orth>De pacotille</orth>
          </form><pc>,</pc>
          <sense><def>de peu de valeur, de qualité médiocre.</def></sense>
        </re>
      </entry>
```

*Fine grained recognition of the various component of an entry in a legacy dictionary*

*Using the TEI as reference output format (coordination with ENEL recommendations)*

**Perspectives: Creating step by step a large-scale network (diachronic and synchronic) of our lexical patrimony**

- LAURENT ROMARY, MIKE MERTENS, ANNE BAILLOT, „DATA FLUIDITY IN DARIAH – PUSHING THE AGENDA FORWARD“, MARCH 2016 (HTTPS://HAL.INRIA.FR/HAL-01285917/)

DARIAH-EU

DARIAH-EU

Digital Research Infrastructure
for the Arts and Humanities

Data “journals”

# CODE NAME: *LIVING SOURCES*

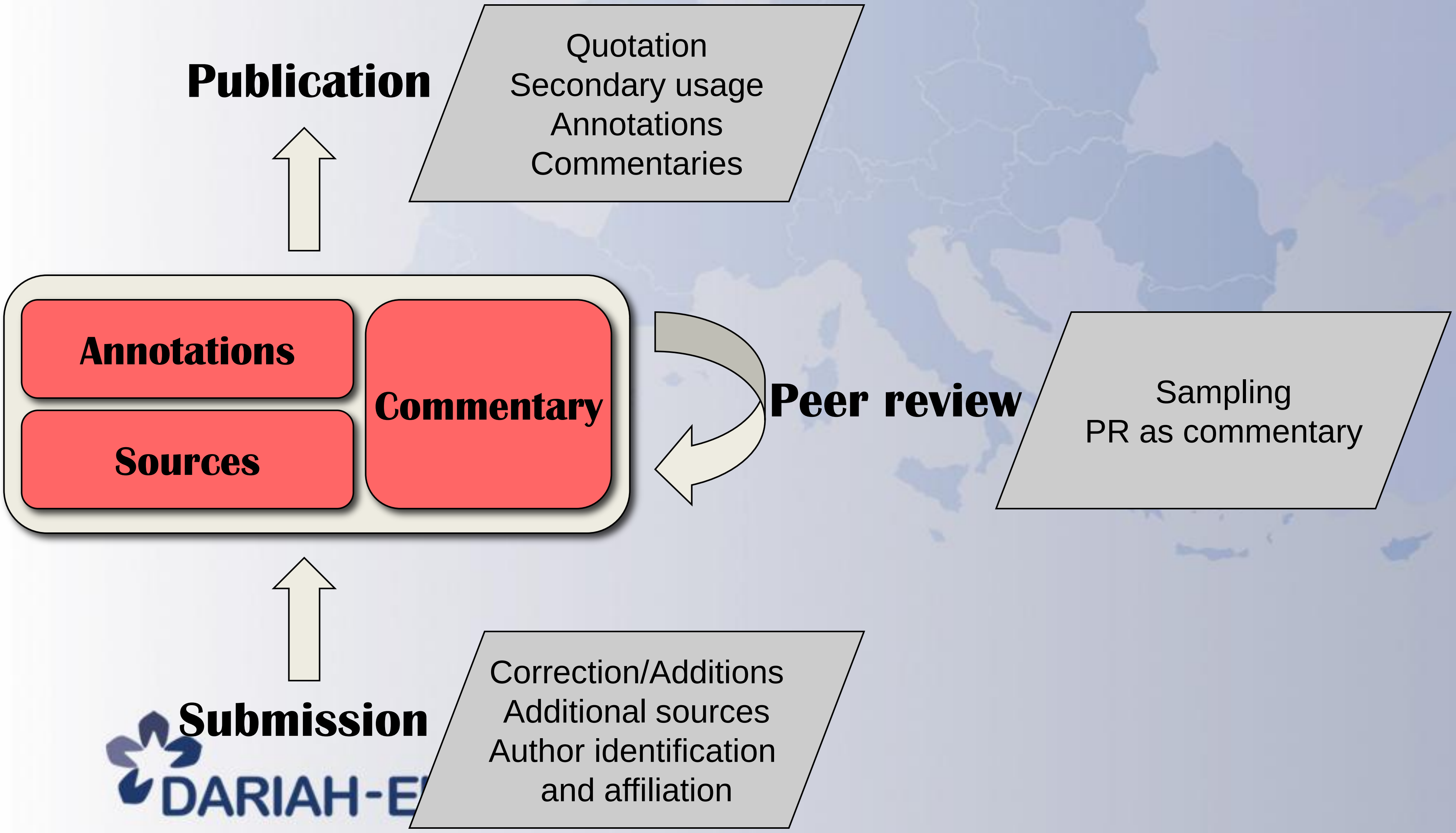
www.dariah.eu

# Exploring new scholarly communication models

- Open Access
  - Open access to scholarly papers is on its way (nearly!)
  - Encouraging scholars to disseminate their research data in the same conditions
    - Despite fears (plagiarism?) or unstable technical settings
- Scholarly recognition
  - Ensuring acknowledgement for the work carried out on digital sources
- A necessity to anticipate an evolving landscape
  - Imagining what the certification of research data could be
  - Building upon what other research fields have accheives (e.g. genomics)

DARIAH-EU

# *Living sources*



**Publication**

Quotation
Secondary usage
Annotations
Commentaries

**Annotations**

**Sources**

**Commentary**

**Peer review**

Sampling
PR as commentary

**Submission**

Correction/Additions
Additional sources
Author identification
and affiliation

# Thought experiment: living sources in lexicography

- A wide variety
  - Of domains
    - Psycholinguistics, Field linguistics, Comparative linguistics, Lexicography, Natural Language Processing, Dialectology
  - … and forms
    - Frequency lexica, association lexica, word lists, full form lexica, pronunciation, morphosyntactic lexica, semantic lexica (e.g. Wordnet, Framenet)
  - and sources
    - Hand designed, corpus based, automatic extraction, compilation of sources, etc.

DARIAH-EU

# Submissions: we need more than data

- (Lexical) Data cannot be evaluated in isolation
  - A certain scientific perspective on the material, whether observed or compiled
- A preface will be requested with all submissions
  - Scientific background
    - Research field, role of the data proper in the research activity
    - Scientific theories and hypotheses at stake
  - Editorial background
    - Rationale for compiling the data
    - Selection criteria (data sampling, descriptive features)
  - Underlying lexicographic model
    - Onomasiological/Semasiological, Data categories, specification of orthography
  - Links to other databases/sources
    - Full context

DARIAH-EU

# Reviewing lexical content

- Methodology
  - Characterizing the scientific contribution
    - Data model, accuracy of descriptors, data gatjhering methodology
- Technology
  - Conformance to existing standards
    - LMF-TEI, TMF-TBX, SKOS (?)
- Usefulness
  - Exhaustivity of the resources vs. sampling
  - Licences, rights to re-use

DARIAH-EU

DARIAH-EU

Digital Research Infrastructure
for the Arts and Humanities

A multi-stage process

# PEER-REVIEW

www.dariah.eu

# Stage 1 - Technical check

- (Closed) submission to editors
- Validation of technical issues (data structure, preface)
- Possible retraction for scientific check at this stage
- Editorial/technical support provided to authors
- Outcome:
  - Technical publication

DARIAH-EU

# Stage 2 - Content check

- Open peer-review submission (time-restricted)
  - Cf. Copernicus model
- Critical submission about the submission as a whole decide on acceptance
  - Vs. On individual entries
  - Issue: sampling strategy (randomized, left to reviewer, entry or feature based)
- Separate commentaries on individual entries
- Outcome
  - Publication="scientifically relevant database"
- Various publication status:
  - word list, wordform collection, word field, language particular dictionary, comparative dictionary, etc.
  - Correspond to various quality levels

# Stage 3 - Living commentary and growth of data

- Addition of more data, corrections, versions by the author
  - Acceleration validation (consistency check)
- (Identified) third party contributions
  - Commentaries
  - Additional entries or features?
    - Issue: validation by authors and editorial board?

DARIAH-EU

DARIAH-EU

Digital Research Infrastructure
for the Arts and Humanities

# ISSUES RELATED TO IMPLEMENTATION

www.dariah.eu

# Repository and services

- Building-up of a technical infrastructure which enhances the usability of datasets (one stop shop, comparability, searchability, persistence, etc.)
  - Envisioned user group: scientists who look for a hosting environment
  - Example: Tapas initiative for TEI documents
- Standards of interoperability of data portals/journals/archives with a common search engine/browser-like tool
  - Envisioned user group: scientists who want to keep a strong hold on their data
- Persistence of data is secured for data submitted to the system (grid-like backup)

DARIAH-EU

# Certification as overlay

- Cf. Overlay journal model
  - Publication ("making public") in a publication repository
  - Certification by an overlay editorial committee
  - Ex. Episciences.org – a public infrastructure
- Implementing data journals as overly journals
  - Stabilizing the data repository infrastructures
  - Allowing a variety of certifications (cf. technical vs. Scientific)

DARIAH-EU

# Interoperability and Standards

- Through Living Sources, data is offered a second-life
  - Data will be searched, modified, crossed with other sources
    - Interoperability is thus a central issue => standards
- Scientific freedom vs. Limitation provided by standards
  - Need of flexible representation formats or models
- Documenting one’s own practices and data semantics is essential
  - Cf. ODD mechanisms with TEI resources
  - Importance of helping scholars to retrieve legacy data

DARIAH-EU

# Citability of data

- Various levels of granularity
  - Complete submission seen as a scientific contribution
  - Preface, or any of its components
  - Full lexical resources, or any entry thereof, down to specific feature
- Various types of usage
  - Referencing
  - Further processing (cross dictionary search, sub-selection of data, etc.)
- Technical answers
  - Unique identifier for the full submission
  - Selection recipes (à la XQuery) also stored in the repository
    - E.g. all intransitive verbs

# Accessibility - copyright

- Main assumption
  - Data must be fully and freely available for acces but also re-use
    - Cf. current debates on TDM
- Distribution under a simple licence
  - Recommendation: Creative commons with attribution. CC-BY
  - (Limited) Right for Living Sources (to store) and distribute the data

# Versioning

- During the review process
  - Cf. editorial validation, peer review
- After publication
  - « living »
- A major sub-issue: granularity
  - Versioning full submission, down to elementary entries and/or fields
    - E.g. correcting a word qualifier

DARIAH-EU

# Long term business model

- Archiving and access
  - Needed anyhow for the research community
- Certification
  - Should this be taken up by professional stakeholders ?
- Outreaching
  - Easy to achieve through powerful academic dissemination forums (e.g. linguist List)
- Importance of a community-based (public) infrastructure

DARIAH-EU

# A feasible endeavour

- Science driven
  - Long standing expectation from various communities
- Technologically mature
  - Standards and tools
  - Emergence of data repositories
  - Overlay journal platforms
- Politically timely
  - Cf. Riding the wave report, RDA initiative… and DARIAH
  - Understanding the ecology of research data publishing
- Setting up things as a portfolio of initiatives
  - Scholarly communities/domains or communities of practice (working on similar objects) should now be creative in moving ahead…

DARIAH-EU