



**HAL**  
open science

## **Outlining a scholarly workbench – publication and data as a continuum**

Laurent Romary

### ► **To cite this version:**

Laurent Romary. Outlining a scholarly workbench – publication and data as a continuum. eScidoc days, The Royal Library, Nov 2010, Copenhague, Denmark. <inria-00537845>

**HAL Id: inria-00537845**

**<https://inria.hal.science/inria-00537845v1>**

Submitted on 19 Nov 2010

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

# Outlining a scholarly workbench – publication and data as a continuum

Laurent Romary

INRIA & Humboldt Univ. Berlin

# Overview

- A scientific information policy viewed from the point of view of research repositories
- Publication repositories
  - Where do we stand, where do we want to go?
  - Theory and practice
- Can this be a basis for a more global view of a research repository?
  - The case of textual information
- How can we shape the future of research repositories?

# A personal view

- Research bias
  - Computational linguistics
    - A multidisciplinary field
  - Publications: importance of conferences, long-standing culture of publication repositories
    - Cf. stats in HAL
  - Data: linguistic corpora, annotations, lexical databases, grammars, etc.
    - Standards...
- Scientific Information bias
  - Scientific information development in research organizations and research communities

# In the beginning was science...

- A scholar-centered perspective
  - Exploring new fields
    - Knowing what is new in his field: publications
    - Scrutinizing what the others are doing: experiments, data, sources
  - Making “discoveries”
    - Assessment by peers (certification)
    - Communicating to others
  - Organizing research
    - Setting up teams, projects, equipments
    - Applications, reports, assessments

# Scientific information management

- Providing the researcher with the means to work
  - Providing access to publications
    - Subscription policy
  - Giving him the means to record and disseminate his activity
    - Research repository
- Difficulties
  - Cope for the high costs of traditional scholarly publishing
  - Accommodate with the development of new technologies
  - Getting a comprehensive view on the researcher's production

# Scholarly publishing

- Certification
  - Management of the peer-reviewing process
- Dissemination
  - Reaching out libraries, scholars
- Long-term availability
  - Permanent reference and access
- Basic terminology
  - Stage 1: author's draft for review
  - Stage 2: author's draft after review
  - Stage 3: publisher's version after copy-editing

# Publication repositories

- Intended to deal with the dissemination and long-term availability functions
- Open access: a means for an end
  - Increasing the accessibility of scholarly results
  - Complementary to the certification process
- Components of a publication repository
  - Technical infrastructure – digital object management
  - Editorial support – content management, quality assessment (e.g. affiliations)
  - Political environment – who wants a repository and to which purpose

# To be or not to be central...

- Technical infrastructure (IT)
  - Need not be duplicated
  - Constant development of new services
- Editorial support (Library)
  - Needs to be close to research environments
  - Needs further functionalities (hidden to researchers)
- Political environment (Research management)
  - Needs to be concerted across institutions
  - Compromise between institutional visibility and coordination of available means
  - Research repository policy cannot be disentangled from SI policy (e.g. Springer-MPS)

But let's forget about concepts...

# Why do I use a publication archive?

- Record of my production
  - [My publications on HAL](#)
- Quick delivery to others
  - Write, deposit, give away
- Because I believe in open access?
  - Maybe a bad argument
    - Would I write without the perspective of an “official” publishing?
    - Would I want to avoid peer-review?
      - No. Relying on the recognition from my colleagues
      - Yes. If I would know my results would be used and attributed/recognized
  - Objective view
    - Happy to find papers from colleagues on google
    - Aware that putting my own work is an overhead
- Things are made easier thanks to a good infrastructure

# HAL – a quick overview

- Put together in the mid 90's as a mirror to ArXiv
  - Political independence, difficulty to get additional functionalities – arxiv as a close environment
  - Initiated by physicists, within CNRS
- Wider impact around since mid 2000's
  - Multidisciplinary: maths, human sciences, computer science
  - Multi-institutional: INRIA, INSERM, Universities
  - HAL has become a national publication repository

# Why do I use HAL(-INRIA)?

- Because it's visible
  - [Ranking Web of World Repositories](#)
  - My colleagues will easily find my publications: [Google search \[Laurent Romary standards\]](#)
- Because I feel at home
  - [HAL-INRIA](#)
  - Within one single instance of HAL: [Generic HAL](#)
- Because it has a couple of cool features
  - Online legibility: [Romary & Armbruster, 2010](#)
  - Facilitated deposit (affiliation): [HAL-Deposit](#)
  - Publication lists: [Haltools](#)
- Because INRIA has cool librarians...
  - Completion, correction, interaction, support

# What do I expect now?

- (even) Easier submission
  - What should I type in information which is already in the document I am depositing?
- Better statistics
  - [HAL - Stats](#)
  - Evolution of access over time
  - Source of download requests
- Better workspace functionality
  - Creating, managing and disseminating collections
  - Adding research material (e.g. TEI encoded dictionary samples)
- Better connection with other publication services
  - Google scholar, WoZ, Microsoft academic search
  - Duplicates, missing entries, bad affiliation, no link to HAL...

# Putting intelligence into the repository

I have a dream...

# Level 1 – getting started quickly

- Managing authors
  - One's own identity — default author, default affiliation(s)
  - Co-authors — favorite co-authors, favorite co-institutions
- Managing institutions
  - Reliable authority list of institutions and laboratories
  - Favorite co-institutions
- Managing publication loci
  - Journal list, conferences
- Managing publications
  - Duplicates, corrections, completions

# Level 2 – the repository as a tool

- Researcher workspace
  - Small scale (cf. dream)
- Institutional workspace
  - The repository as a reporting tool
  - (cf. HAL: exports for the annual report)
- Statistics
  - The repository as an indicator of scientific influence
  - From citation (in publications) to usage (downloads)
- Deep interoperability
  - High quality data for high quality services
  - Exports – imports, etc.
  - Harvesting, indexing: Beyond OAI-PMH
  - Anticipating the transition from metadata to full-text management

# Level 3 – bringing intelligence in the repository

- If only the repository had some knowledge about the data itself
  - Bringing-in data automatically
    - From publishers to repositories
  - Extracting information from documents
    - Typing-in information once and for all
  - Providing specific services for semi-structured datatypes
    - E.g. Synthetic views on a publication
- Two examples: the PEER project, the Darjah TEI demonstrator

# Intermezzo – the Text Encoding Initiative (TEI)

# The Text Encoding Initiative

- Initiated in 1987 by major international text centers
  - Adoption of SGML, than XML
  - Important contributions to the development of XML
- Organized as a membership consortium since 2000
  - 5 hosts (Virginia, Brown, Oxford, Nancy, Leithbridge)
  - Board (management) and council (technical content)
- Five editions of TEI guidelines (current P5)
  - Large community of users, continuous maintenance of content, evolution towards additional domains (e.g. manuscript transcription)

# Main technical features of the TEI

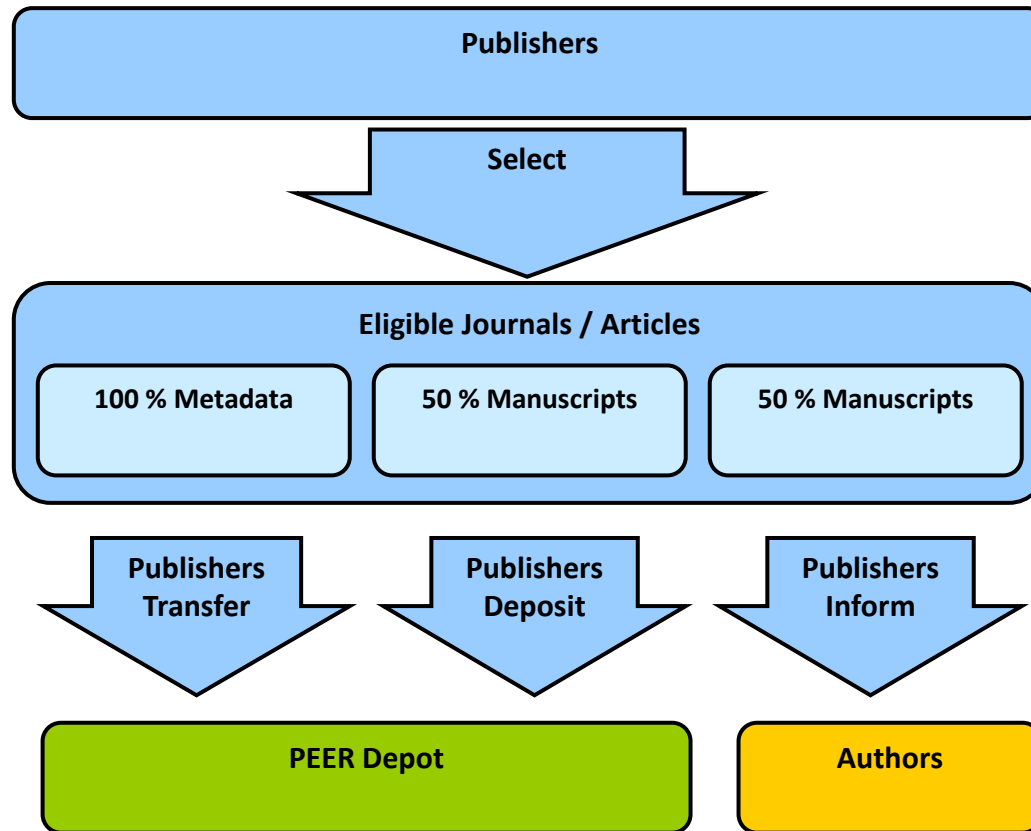
- More than 500 elements
- Modularity
  - Core modules
    - header text descriptions; bibliography
  - Thematic modules
    - drama; dictionaries; manuscript description
  - Additional components
    - time, names and dates; annotations;
- Customizability
  - ODD (one document does it all): specification language of the TEI
- Mime type: application/xml+tei

A project with a vision: PEER

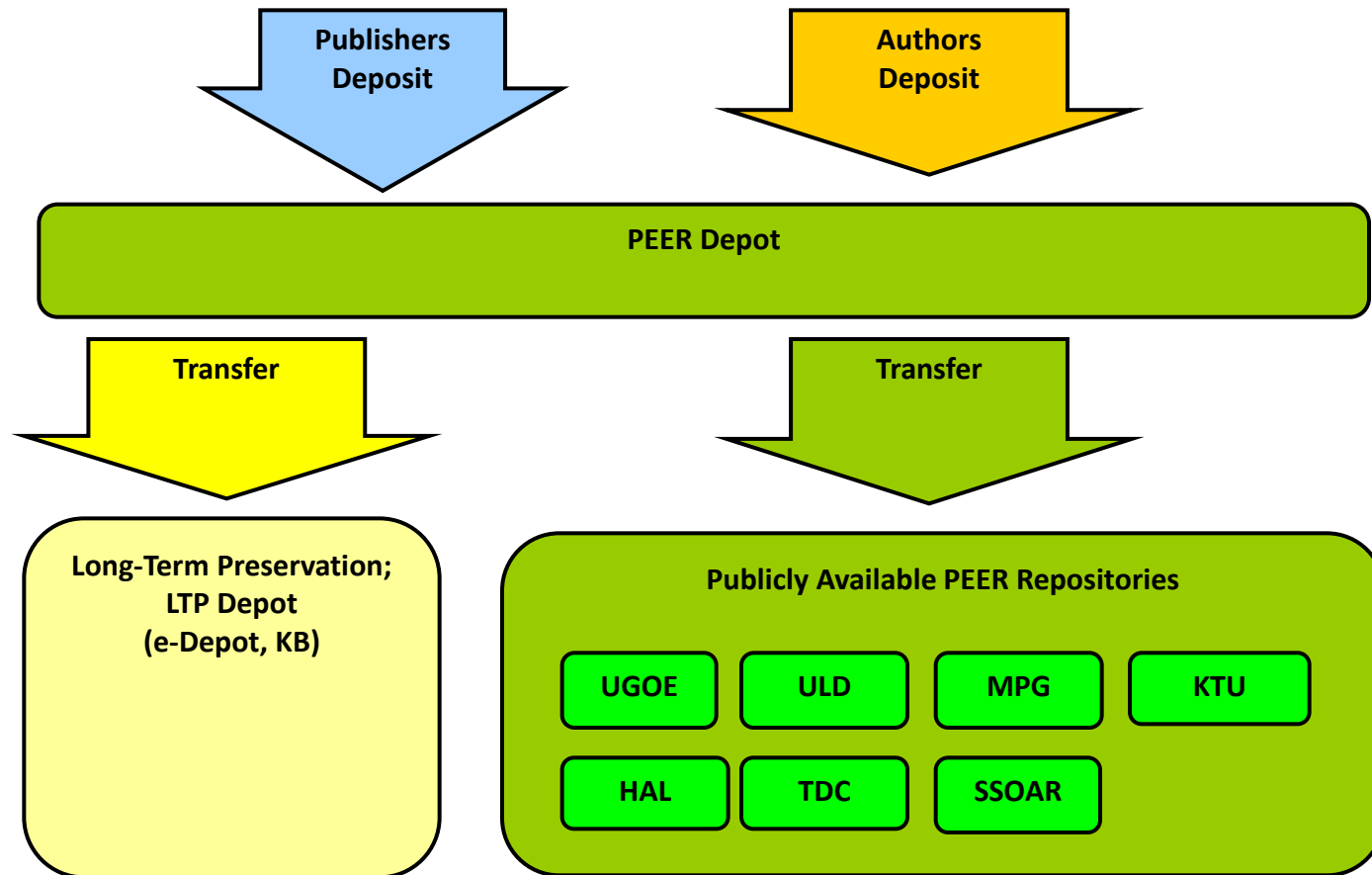
# The PEER project

- Initiated by the EU commission (DG INFSO)
- Objective: study the impact of systematically archiving *stage-two* outputs in “institutional repositories”
  - on journals and business models
  - on wider ecology of scientific research
- Consortium
  - STM, European Science Foundation (ESF), Goettingen State and University Library (UGOE), Max Planck Gesellschaft (MPG), INRIA

## Content submission - publishers



# Content submission – to repositories & LTP archive



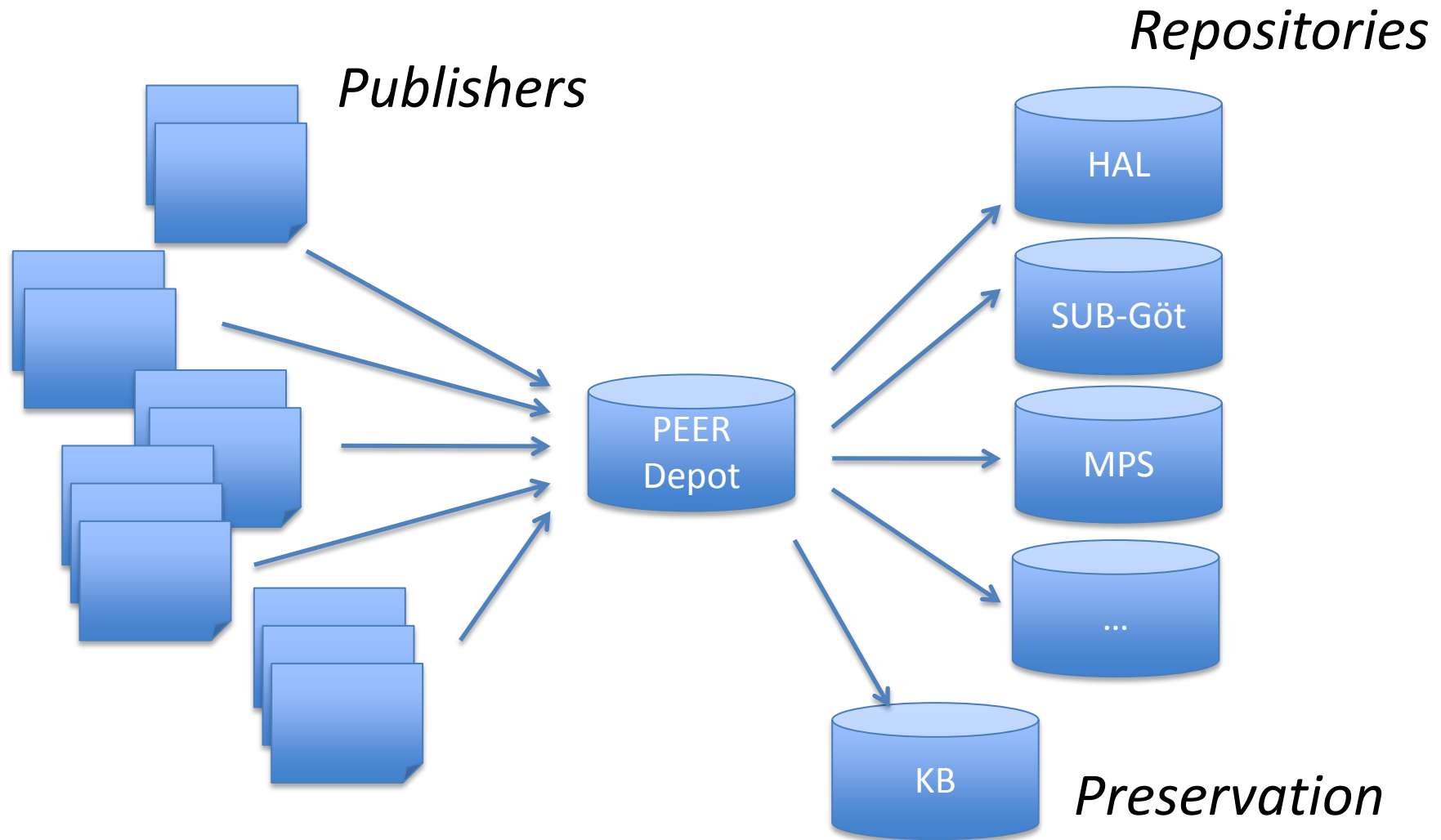
# Publishers involved the project

- BMJ Publishing Group (proprietary format)
- Cambridge University Press (NLM2.2)
- EDP Science (NLM3.0)
- Elsevier (proprietary format)
- IOP Publishing (NLM3.0)
- Nature Publishing Group (proprietary format)
- Oxford University Press (ScholarOne)
- Portland Press (NLM2.0)
- Sage Publications (proprietary format)
- Springer (proprietary format)
- Taylor & Francis Group (ScholarOne)
- Wiley-Blackwell (ScholarOne)

# The information chaos

- Article title
  - article-title/title | ArticleTitle | article-title | ce:title | art\_title | article\_title | nihms-submit/title | ArticleTitle/Title | ChapterTitle
- Journal title
  - j-title | JournalTitle | full\_journal\_title | jrn\_title | journal-title
- ISSN (print)
  - JournalPrintISSN | issn[@issn\_type='print'] | issn[@pub-type='ppub'] | PrintISSN | issn-paper
- First page of a paper
  - spn | FirstPage | ArticleFirstPage | fpage | first-page

# The PEER deposit workflow

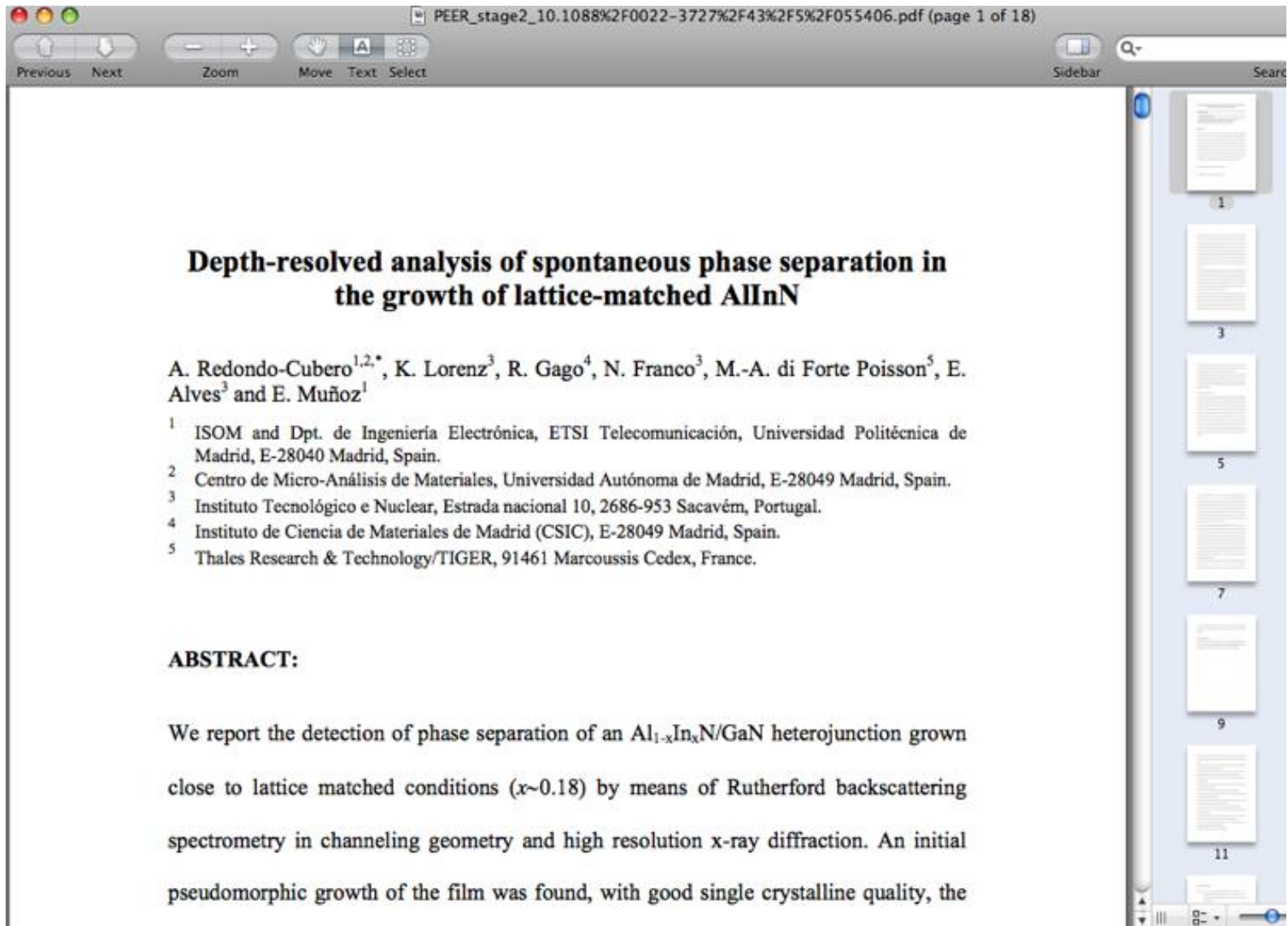


# TEI as a pivot format for interchange

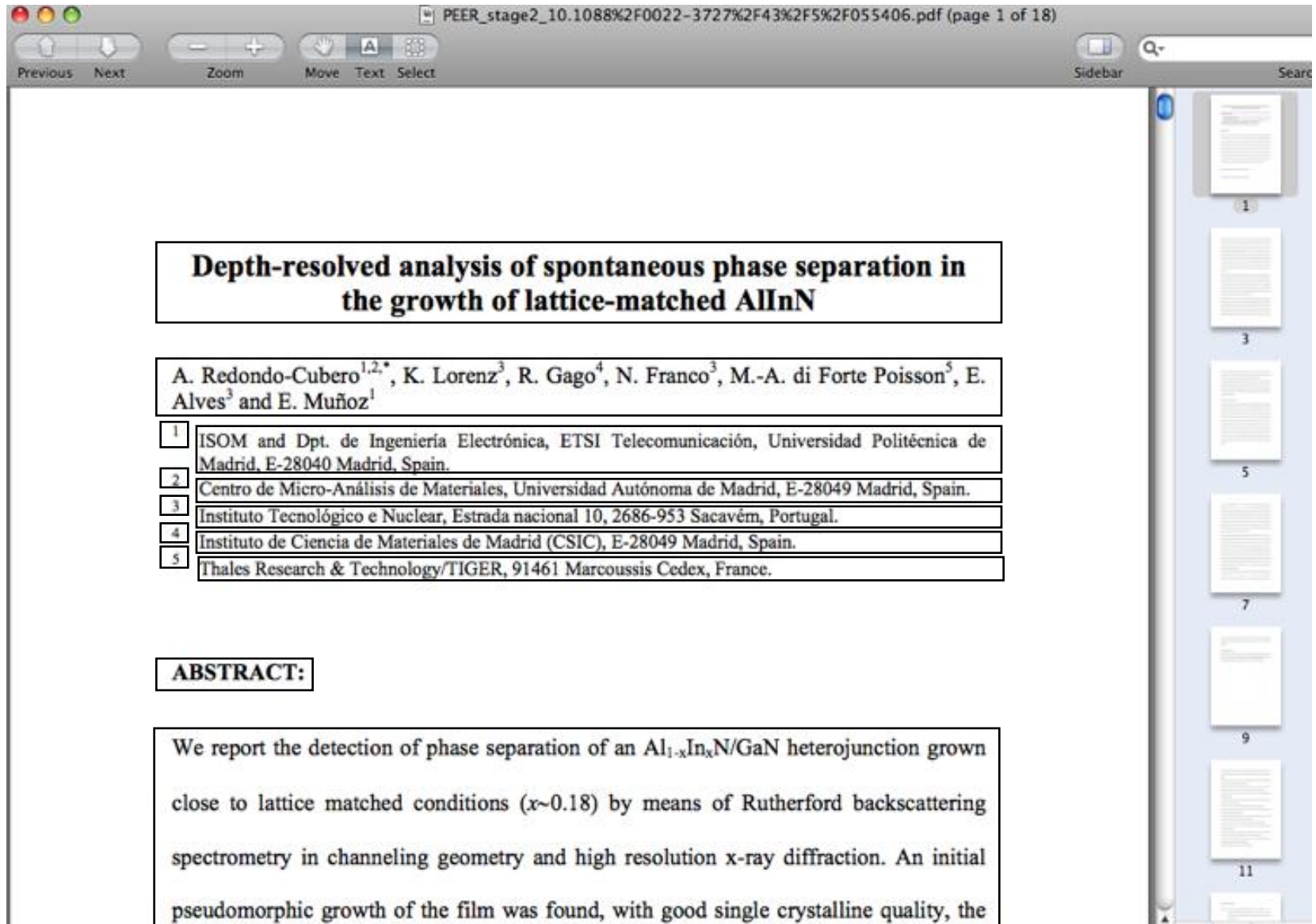
- General strategy: no information should be lost
  - Nearly everything in <sourceDesc>
  - + Keywords, Summary, Copyright
- Strict author description
  - Deep encoding of names
  - Deep encoding of affiliations (Web of Science - 3-level)
  - Deep encoding of addresses – getting the country right
- Precise publishing information
  - Pagination, DOIs, volume, issue, journals name(s)
  - Yes, <biblStruct> is cool!

... And when no metadata is  
available

# Metadata extraction from front page



# Layout & Block Analysis: XY-Cut algorithm



PEER\_stage2\_10.1088%2F0022-3727%2F43%2F5%2F055406.pdf (page 1 of 18)

Previous Next Zoom Move Text Select Sidebar Search

## Depth-resolved analysis of spontaneous phase separation in the growth of lattice-matched AlInN

A. Redondo-Cubero<sup>1,2,\*</sup>, K. Lorenz<sup>3</sup>, R. Gago<sup>4</sup>, N. Franco<sup>3</sup>, M.-A. di Forte Poisson<sup>5</sup>, E. Alves<sup>3</sup> and E. Muñoz<sup>1</sup>

- 1 ISOM and Dpt. de Ingeniería Electrónica, ETSI Telecomunicación, Universidad Politécnica de Madrid, E-28040 Madrid, Spain.
- 2 Centro de Micro-Análisis de Materiales, Universidad Autónoma de Madrid, E-28049 Madrid, Spain.
- 3 Instituto Tecnológico e Nuclear, Estrada nacional 10, 2686-953 Sacavém, Portugal.
- 4 Instituto de Ciencia de Materiales de Madrid (CSIC), E-28049 Madrid, Spain.
- 5 Thales Research & Technology/TIGER, 91461 Marcoussis Cedex, France.

**ABSTRACT:**

We report the detection of phase separation of an Al<sub>1-x</sub>In<sub>x</sub>N/GaN heterojunction grown close to lattice matched conditions ( $x \sim 0.18$ ) by means of Rutherford backscattering spectrometry in channeling geometry and high resolution x-ray diffraction. An initial pseudomorphic growth of the film was found, with good single crystalline quality, the

# Metadata extraction from front-page

```
<?xml-stylesheet type="text/xsl" href="xmlverbatimwrapper.xsl"?>
<biblStruct xml:lang="en" xml:id="b0">
  <analytic>
    <title level="a" type="main">Depth-resolved analysis of spontaneous phase separation in the growth of lattice-
matched AlInN</title>
    <author>
      <persName>
        <forename>A</forename>
        <surname>Redondo-Cubero</surname>
      </persName>
      <affiliation>
        <orgName type="department">ISOM and Dpt. de Ingeniería Electrónica, ETSI Telecomunicación</orgName>
        <orgName type="institution">Universidad Politécnica de Madrid</orgName>
        <address>
          <postCode>E-28040</postCode>
          <settlement>Madrid</settlement>
          <country key="ES">Spain</country>
        </address>
      </affiliation>
      <affiliation>
        <orgName type="department">Centro de Micro-Análisis de Materiales</orgName>
        <orgName type="institution">Universidad Autónoma de Madrid</orgName>
        <address>
          <postCode>E-28049</postCode>
          <settlement>Madrid</settlement>
          <country key="ES">Spain</country>
        </address>
      </affiliation>
    </author>
    <author>
      <persName>
        <forename>K</forename>
        <surname>Lorenz</surname>
      </persName>
      <affiliation>
        <orgName type="department">Instituto Tecnológico e Nuclear</orgName>
        <address>
          <addrLine>Estrada nacional 10</addrLine>
          <postCode>2686-953</postCode>
          <settlement>Sacavém</settlement>
          <country key="PT">Portugal</country>
        </address>
      </affiliation>
    </author>
  </analytic>
</biblStruct>
```

# Metadata extraction from front-page

```
<addrLine>Estrada nacional 10</addrLine>
<postCode>2686-953</postCode>
<settlement>Sacavém</settlement>
<country key="PT">Portugal</country>
</address>
</affiliation>
</author>
<author>
<persName>
<forename>E</forename>
<surname>Muñoz</surname>
</persName>
<affiliation>
<orgName type="department">ISOM and Dpt. de Ingeniería Electrónica, ETSI Telecomunicación</orgName>
<orgName type="institution">Universidad Politécnica de Madrid</orgName>
<address>
<postCode>E-28040</postCode>
<settlement>Madrid</settlement>
<country key="ES">Spain</country>
</address>
</affiliation>
</author>
</analytic>
<monogr>
<title level="j">Journal of Physics D: Applied Physics</title>
<title level="j" type="abbrev">J. Phys. D: Appl. Phys.</title>
<idno type="ISSN">0022-3727</idno>
<idno type="ISSNe">1361-6463</idno>
<imprint>
<biblScope type="issue">5</biblScope>
<date>2010</date>
</imprint>
</monogr>
<note>1. *. Corresponding author : andres.redondo@uam.es 2</note>
<keywords>RBS, channeling, AlInN, strain, XRD</keywords>
<idno type="doi">10.1088/0022-3727/43/5/055406</idno>
<div type="abstract">We report the detection of phase separation of an Al 1-
x In x N/GaN heterojunction grown close to lattice matched conditions (x?
0.18) by means of Rutherford backscattering spectrometry in channeling geometry and high resolution x-
ray diffraction. An initial pseudomorphic growth of the film was found, with good single crystalline quality, the nominal composi
</biblStruct>
```

# What do we have there?

- A coherent infrastructure to facilitate
  - The long-term management of scholarly content in research institutions
    - In-depth representation of bibliographical data
  - Smooth interaction between publishers and research institutions
    - Better understanding of what each of us can provide
    - E.g. Gold open access (cf. Springer-MPS)
  - Integration of legacy document within a repository
  - Pushing publications to other repositories
    - Institutional–thematic repositories

# Intelligent management of content

The “TEI repository”



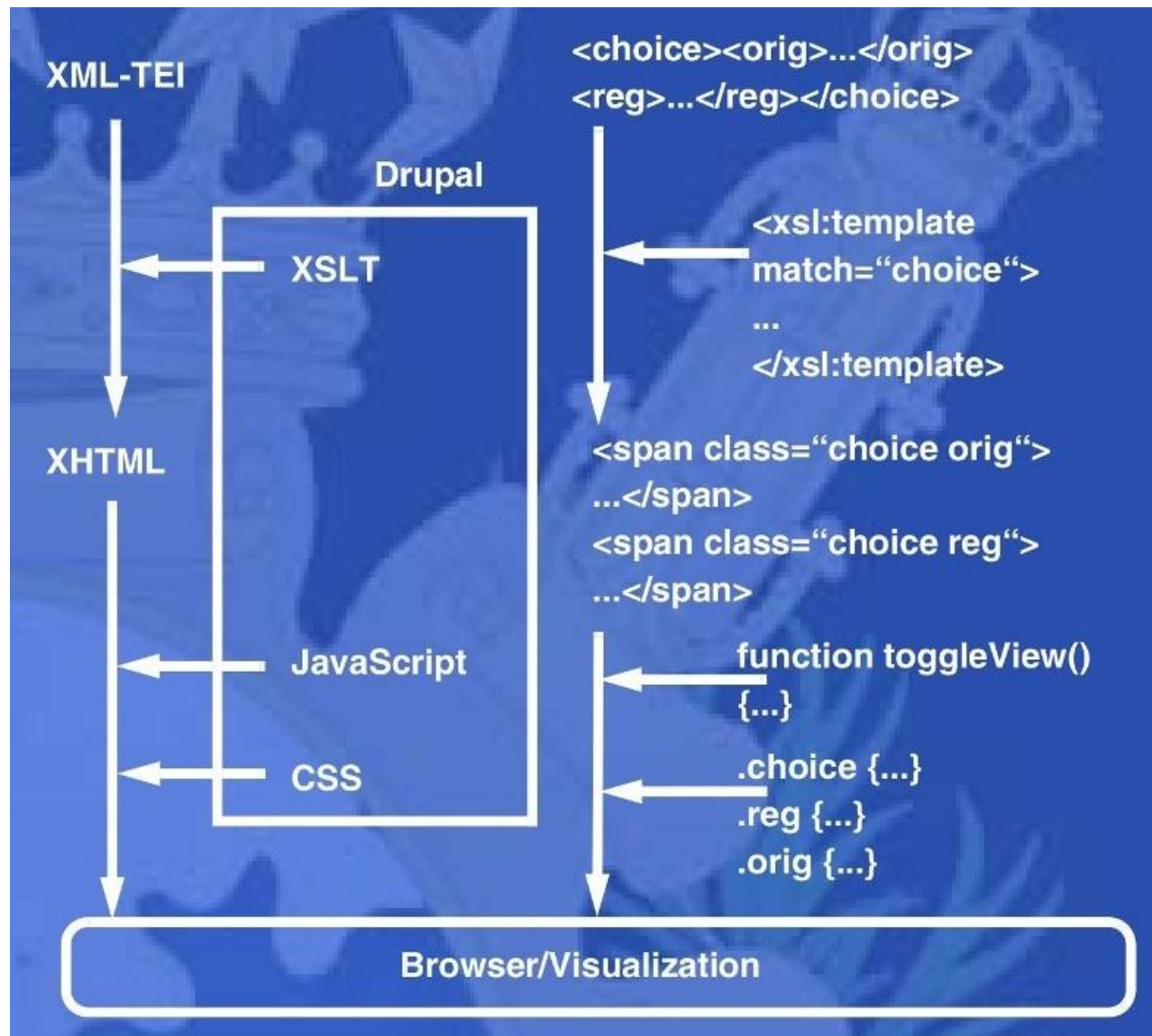
# Why a “TEI repository”?

- The continuum of full-text document
  - Publications – cf. Language Description Heritage
  - Primary sources
  - Further commentaries
- Various forms of intelligence
  - Manipulated like other items in the repository
    - Submit, publish, Meta-data search, presentation lists
    - Texts as accessible objects (decapsulation)
  - Basic understanding of data structure
    - Format checking, preview, content based search
  - Connection to external resources or tools
    - Decapsulation – limiting the intelligence

# Why a “TEI repository”? – cont.

- Because scholars need it!
- Isolated researchers
  - Sebastian Pape, Christof Schöch, Lutz Wegner: “Bringing Bérardier de Bataut's Essai sur le récit to the web: Editorial requirements and publishing framework”, TEI Member's Meeting and Conference 2010, University of Zadar, Kroatien.
  - [Bérardier de Bataut's Essai sur le récit](#)
  - [Online report](#)
- Research projects
  - Peter Stadler, “Building a historical social network from TEI documents”, TEI Member's Meeting and Conference 2010, University of Zadar, Kroatien.
  - <https://194.94.229.134/wega/xql/index.xql>

# Bérardier - transformation process



# An opportunity

- DARIAH – research infrastructure for the humanities
  - ESFRI roadmap
  - Preparation phase – coord. DANS (NL)
- Experimenting researchers’ environments within DARIAH
  - “Working for the poor”: offering a simple workspace for eScholars working on digital documents and collections
  - Deposit, describe, visualize, publish
- Demo: [TEI Repository](#)

Next step – virtual research

# Not a completely impossible idea

- Virtual astronomers
  - Most of them now are
  - Many do not even see a telescope
  - Huge databases of stellar objects, observations (multi-range) and publication data
- Virtual humanists
  - Progress in the humanities results from pooling together sources
  - Transcribing and studying sources are not necessarily part of the same research activity
  - Need for attribution-recognition mechanisms
    - Cf. report to DG INFSO: [Riding the wave](#)
  - Are we able to design the adequate environments for them?

# We can probably try conclude...

- The “Scholarly Workbench” never existed as an isolated entity – good thing
  - No separation between publication and data
  - Nothing like a generic research data environment
    - Specific datatypes: text, images, geo-temporal information
    - Specific scholarly communities
- Lessons to be learnt for a scientific information policy
  - No rush, be consequent
  - Keep all developments within a global strategy
  - Take benefits from available/demanding communities — be opportunistic
  - Services, services, services...
- Mühsam, mühsam ernährt sich das Eichhörnchen