Data Mining Technologies at the service of Open Knowledge

Laurent Romary

Data Mining Technologies at the service of Open Knowledge

Laurent Romary

DARIAH & Inria (team ALMAnaCH)

Additional credits: Patrice Lopez, Luca Foppiano, Mohamed Khemakhem

Short personal presentation

- Research
  - Directeur de Recherche Inria, team ALMAnaCH
- Scientific information
  - Advisor at Inria (cf. open access policy)
- Infrastructures
  - Director General of the EU infrastructure DARIAH
- Standards
  - Chairman of ISO committee TC 37 (*language and terminology*; the ISO 639 people…)
  - Member of the TEI (Text Encoding Initiative) board

Today's menu

- Text and data mining of scientific documents
  - Scientific documents, TDM, legal issues
- Specific applications
  - Scientific quantities
  - Structuring information from texts
- Additional issues
  - Standards, openness…

ENTRÉE – SOME (NOT SO) LITE BITES

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 (leaving aside grey literature)
  - Scholarly publications
  - Standardisation documents
  - Patents
- Some common characteristics
  - Authorship: the basis of scientific attribution
  - Structure: usually a formal document organisation
  - Vocabulary: technical terms are essential to convey (or hide) meaning
  - Network of references: relating a document to the state of the art
  - Certification: workflow, responsibilities, metadata
  - Openness: how much do we have access to the documents?

Authorship

Publications - *The essence of publishing*

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

Standards - *Priority to the institution*

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

Patents - *A variety of roles*

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

Structure

Publications - *Semi-formal*

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

Standards - *Very formal*

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

Patents – *Very formal*

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

Language

Publications - *Semi-formal*

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

Standards - *Very formal*

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

Patents - *Obfuscating*

- Achieving widest coverage and preventing retrieval

Network

Publications - *Semi-formal*

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

Standards - *Very formal*

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

Patents - *Very formal*

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

Certification process

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

Openness

Publications – *on-going issue…*

- Traditional (vestigial?) concept of scholarly publishers
- Openness should be easy in the current digital context
- Debate on how to reach full openness to scholarly content…

Standards – *various models*

- ISO: pay per document
- W3C, TEI: full openness

Patents – *Open!*

- Openness is part of the patent design (18 months after first filing)
- Still, no open dissemination of content in standardised form
- WIPO, Google patent, etc.

Scientific documents - summary

- More commonalities than differences
  - Similar organisational principles
  - Joint referencing mechanisms
  - Related scientific and technical content
- Possible to foresee similar processing techniques
  - Now focussing on scientific papers
  - See (Lopez & Romary; 2009 and 2010) for similar works on patents
- There remains the issue of content accessibility…

TEXT AND DATA MINING

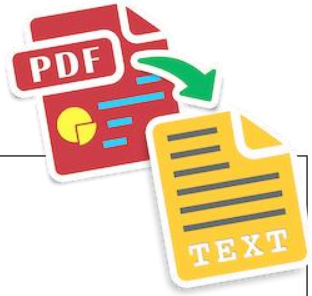
Data mining and machine learning

- Text and data mining
  - Identifying and qualifying information from a corpus of documents or a database
- Machine learning techniques
  - Implementation of mathematical (probabilistic) models
  - Trained on a corpus of existing documents or databases
    - *Unsupervised*: the machine learning algorithm recognizes patterns without prior knowledge of the content
    - *Supervised*: the training corpus is tagged for the categories to be recognised
- Models have been around since the 70's;
  - e.g. neural networks, Markov models
- Huge recent progress with the availability of increased computer capacity (storage, speed)

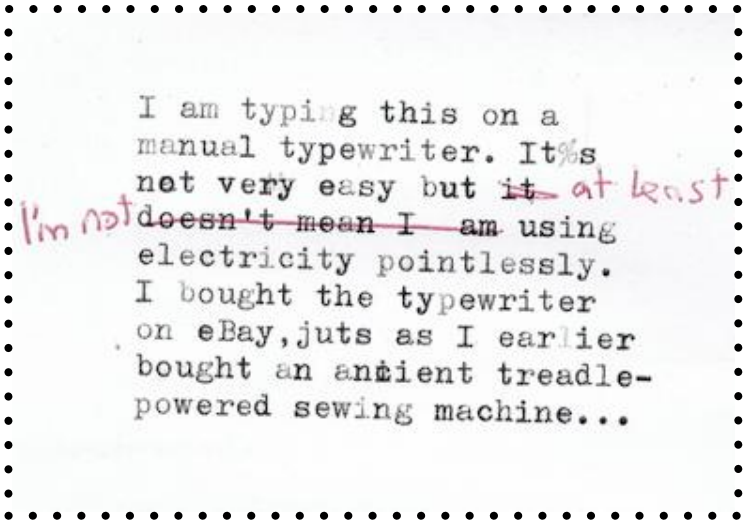
Text and Data mining of scientific documents

- Exploring all aspects of a document
  - Meta-data (title, authors, abstract, keywords)
  - Bibliographic references (list, anchors in the text)
  - Full text structure (divisions, paragraphs, figures, tables)
  - Scientific objects (stellar objects, molecules, etc.)
- A wide range of possible applications
  - Improving information retrieval tools
  - Linking documents (bibliography, similar content)
  - Building up databases of scientific knowledge
  - Automatic construction of scientific hypotheses

Input is often noisy!



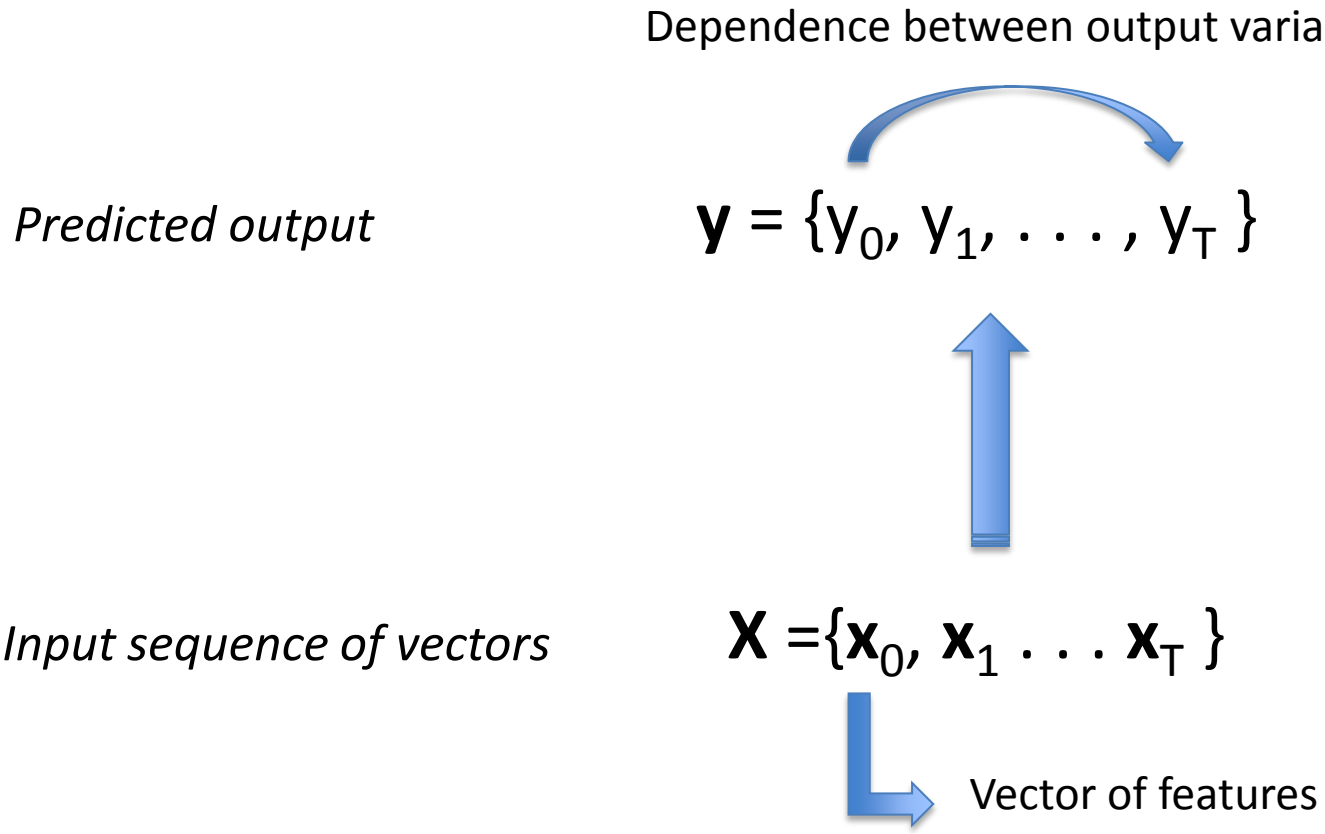
\LLT.HE; Eisfos
.'SEETHE
OF OUFIGOD"ISa!uuo.
THE OUTLOOK.
AN EARNEST WORD.
N 1880 we had but 35 Missionaries, all told, in South America. Now we have 76, or more than double the number that we had 19 years ago. It would naturally be supposed that the income of the S.A.M.S. would have increased in something like the same proportion. Has this been the case? No!
On the contrary, the income from all sources remains almost i stationary. For instance, the average income for the past 18 years has been ,£12,305, while the actual income for the now distant year iSSo was ,£11,836, or only .£669 less than tint average. The income for 181/1 was but .£12,361, or only ^525 more thr.n in 1880 ' In 1X97 it was £'7i573, but that was most exceptional, as it included one noble donation of ,£2,600.
To state the present position in another way. the work of the Society for Cod and South America has increased 117 per cent., while the money entrusted to the Society, with which to sustain that work, has increased very little. It is manifest that such a result has not been attained
APRIL, 1891,1



Conditional random fields (CRF)

- Segmenting and assigning labels to a set of observations sequences
  - E.g. part of speech tagging
  - [Le DET][petit ADJ][chat NOUN][est COP][mort ADJ][. PCT]
- Part of the general class of *graphical models*
  - Expression of local factors between groups of variables
  - General distribution results in the combination of these local factors
    - E.g. Bayesian networks, neural networks, factor graphs, Markov random fields, Ising models
- CRFs are centred on solving the classification problem
  - Modelling the conditional distribution
  - CRFs typically outperform Hidden Markov Models

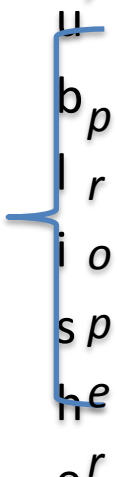
Predicting variables from an input sequence of feature vectors



CRF: Modelling p(**y**|**x**) for large set of features with complex dependencies

Variables and features in practice

Osselton, N. E. 1979. 'Some Problems of Obsolescence in Bilingual Dictionaries'. In R. R. K. Hartmann (ed.), *Dictionaries and their Users: Papers from the 1978 B.A.A.L. Seminar on Lexicography*. Exeter: University of Exeter, 120–126.



Training workflow

- Manually annotated examples
  - Possible bootstrap from previous iterations
- Automatically generated features
- Further training campaigns when new types of data come in
  - E.g. bibliography: Chicago etc. styles, various publisher's styles, …
- Quick convergence when variables and features are well suited to the problem
  - A time-consuming activity: *Feature engineering*

ARE WE JUST ALLOWED TO DO THIS?

Is the right to read a right to mine?

- Problem:
  - Wide quantity of produced scientific information
    - Text and data mining as a way to identify useful content
  - Fragmentation of the corpus
    - Necessity to copy and gather content
    - Heterogeneous re-use rights attached to each asset

The European context

- European Copyright Directive (2001/29/EC)
  - Unclear re-use right for TDM
    - Even for legally acquired material (e.g. subscription)
    - Indirect protection of publishers through the EU database regulation
    - Looser situation in the US
  - The current situation “leaves authors in a vacuum not knowing whether what they are doing is legal or illegal”: Lidia Borrell-Damián, European University Association (EUA)
- Towards a TDM exception for research at EU level?
  - Non for-profit organisations?
  - Publishers fight to prevent a re-examination of the directive
    - Even introduce clauses preventing re-use rights in their licences ruling out mining
    - Foster restrictions even in the context of an exception (using publisher’s software, against the payment of a fee…)
    - Specific clause in article 3 of the current project for a future revision
  - Germany and UK already have an exception
  - And France?...

Article 3 – Text and Data Mining

1. Member States shall provide for an exception to the rights provided for in Article 2 of Directive 2001/29/EC, Articles 5(a) and 7(1) of Directive 96/9/EC and Article 11(1) of this Directive for **reproductions and extractions made by research organisations in order to carry out text and data mining** of works or other subject- matter to which they have **lawful access** for the purposes of scientific research.
2. Any contractual provision contrary to the exception provided for in paragraph 1 shall be unenforceable.
3. Rightholders shall be allowed to apply **measures to ensure the security and integrity of the networks and databases** where the works or other subject-matter are hosted. Such measures shall not go beyond what is necessary to achieve that objective.
4. Member States shall encourage rightholders and research organisations to define commonly-agreed best practices concerning the application of the measures referred to in paragraph 3.

The French context

- Loi pour une République Numérique (article 38)
  - *« Les copies ou reproductions numériques réalisées à partir d'une source licite, en vue de l'exploration de textes et de données incluses ou associées aux écrits scientifiques pour les besoins de la recherche publique, à l'exclusion de toute finalité commerciale. »*
  - Legally acquired material – copies for the purpose of TDM – public research (no commercial purpose)
- *Waiting for Godot*
  - The article is not valid until application guidelines (*décret d'application*) are published
  - The French government declared that they would wait for the EU to finalize its text…

In practice…

- How much can we trace from the source?
  - Cf. Machine Learning models
    - Data hidden behind numbers
  - Individual information extracted from the source
    - E.g. Named Entities, quantities
    - Anchors to the source text or individual information piece
      - Can this be seen as citations?
  - Full extraction of content
    - E.g. GROBID extracts
    - Impossible to re-publish content :-}

PLAT PRINCIPAL – A TWO COURSE MENU

APPLICATION 1 – MINING QUANTITIES

A variety of measurement forms

A 20kg ingot is made in a high frequency induction melting furnace and forged to 30mm in thickness and 90mm in width at 850 to 1,150°C. Specimens No.2 to 4, 6 and 15 are materials embodying the invention. Others are for comparison. No.1 is a material equivalent to ASTM standard A469-88 class 8 for generator rotor shaft material. No. 5 is a material containing relatively high Al content.

These specimens underwent heat treatment by simulating the conditions for the large size rotor shaft centre of a large capacity generator. First, it was heated to 840°C to form austenite structure and cooled at the speed of 100°C/hour to harden. Then, the specimen was heated and held at 575 to 590°C for 32 hours and cooled at a speed of 15°C/hour. Tempering was done at such a temperature to secure tensile strength in the range of 100 to 105kg/mm2 for each specimen.

GROBID - Quantities

- CRF-based analysers of scientific quantities in texts
- Recognizes 120 base units and expressions of atomic, interval and lists of values
- Standardisation in SI units (JSR-363, ISO 80000 series, ...)
- Identifies and attaches the “quantified” substance or objects to the measurements
- Processes 1000 words per seconds (one thread)

GROBID – Quantities (cont.)

- Independent from the scientific field
- Fills in a “hole” in the data mining landscape
- Open source under an Apache 2 licence
- Already used in large scale applications
  - NASA Jet Propulsion Laboratory
  - Istex: corpus of scientific articles with a national subscription (France)

Ex.: simple expressions of quantities

A mixture of **10kg** of silicon nitride powder was charged into the mixing chamber 20 of the mixing vessel 18.

Atomic value

quantity type: **mass**

raw value: **10**

raw unit name: **kg**

normalized value: **10**

normalized unit name: **kg**

quantified (experimental):

raw: **silicon nitride powder**

normalized: **silicon nitride powder**

Ex.: Interpreting intervals

Introduction

There are an increasing number of severely injured patients who present to hospital each year. Trauma is the leading cause of death in all ages from 1 to 44 years. Haemorrhagic shock accounts for 80% of deaths in the operating theatre and up to 50% of deaths in the first 24 h after injury. Only 16% of major emergency departments in the UK use a massive haemorrhage guideline [1].

The management of massive haemorrhage is usually only one component of the management of a critically unwell patient. These guidelines are intended to

1153

Interval

quantity type: **time**

raw: from **1** to **44**

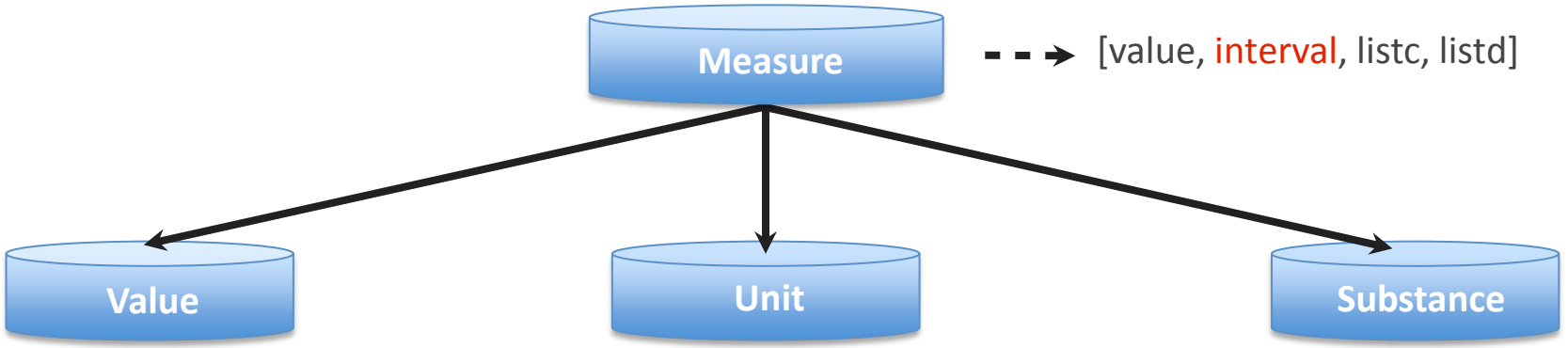
raw unit name: **years**

quantified (experimental):
raw: **all ages from**

normalized: **age**

CRF models in grobid-quantities

... and 200–700 µmol/l of sodium valproate with ...



200

700

prefix: µ

unit: mol

unit: liter

sodium valproate

type: amount of substance

normalization: [0.2, 0.7]

normalized unit (SI): mol/m3

normalized: sodium valproate (W 650844)

Input text and features

A 20 kg ingot is…

A a A A A A A A A A ALLCAPS NODIGIT 1 NOPUNCT X X 0 0
20 20 2 20 20 20 0 20 20 20 NOCAPS ALLDIGIT 0 NOPUNCT dd d 0 0
kg kg k kg kg kg g kg kg kg NOCAPS NODIGIT 0 NOPUNCT xx x 1 0
ingot ingot i in ing ingo t ot got ngot NOCAPS NODIGIT 0 NOPUNCT xxxx x 0 0
is is i is is is s is is is NOCAPS NODIGIT 0 NOPUNCT xx x 0 0

token ; lowercase ; prefix (1,2,3,4) char ; suffix (1,2,3,4) char ; capitalisation ; digit ; single char (0/1) ; punctuation ; word shape (Xxxx) ; word shape trimmed; is known unit token; is number token

And there goes the training…

Ex.: Interpreting percentages

Fifty-three journals were collected: 13 were eliminated from analysis, because they were incomplete, unclear or unreadable. 40 journals were analysed: 19 were journals of subjects of race Z (4 women and 15 men, 30 ± 10 years, 176 ± 7 cm, 70 ± 9 kg, 15 ± 5 % of fat mass, VO 2max : 50 ± 8 ml · kg −1 · min −1 and 21 of race A (6 women and 15 men, 40 ± 7 years, 176 ± 7 cm, 72 ± 10 kg, 18 ± 8 % fat mass, VO 2max : 58 ± 8 ml · kg −1 · min −1). Energy, macronutrients (CHO, fat and proteins) and liquid intakes were analysed.

Interval

quantity type: **fraction**

raw: from **10** to **20**

raw unit name: **%**

quantified (experimental):

raw: **fat mass**

normalized: **fat mass**

Ex.: Unit normalisation

II. EXPERIMENT

The experiments are done in a near IR optical tweezers setup described in [8, 10]. The trapping laser wavelength is 975 nm and it is focused by a 100×/1.25 NA microscope objective with a transmitted power of 62 mW. The microparticles are magnetic beads (Promag Bangs) with a mean diameter of 3.16 μm immersed in water. The aqueous sample is placed between two microscope coverslides with a separation $\sim$ 100 μm. The dynamics are captured with high speed video recorder at 2,000 and 300,000 frames per second (fps). The slower recording speed is used to capture the overall dynamics for a few seconds while the higher speed can capture the explosions in a single frame to measure the maximum sizes of the cavitation bubbles.

The trapping beam waist is raised above the bottom coverslip at a height of between 15 and 25 μm. Initially, the particles are near the bottom and within a few micrometers (radial) from the center of the beam waist. The particles are pulled to the waist and later are pushed away by the explosion, the total displacement and direction are random as these depend on the size of the bubble and the location where the bubble is created. Typical displacements are on the order of 10 μm in the axial z direction and a few microns in the transverse plane xy

Interval

quantity type: **length**

raw: from **15** to **25**

raw unit name: **µm**

normalized: from **0.000015** to **0.000025**

normalized unit: **m**

quantified (experimental): raw: **a height of**

normalized: **height**

Demo time!

http://quantity.science-miner.com

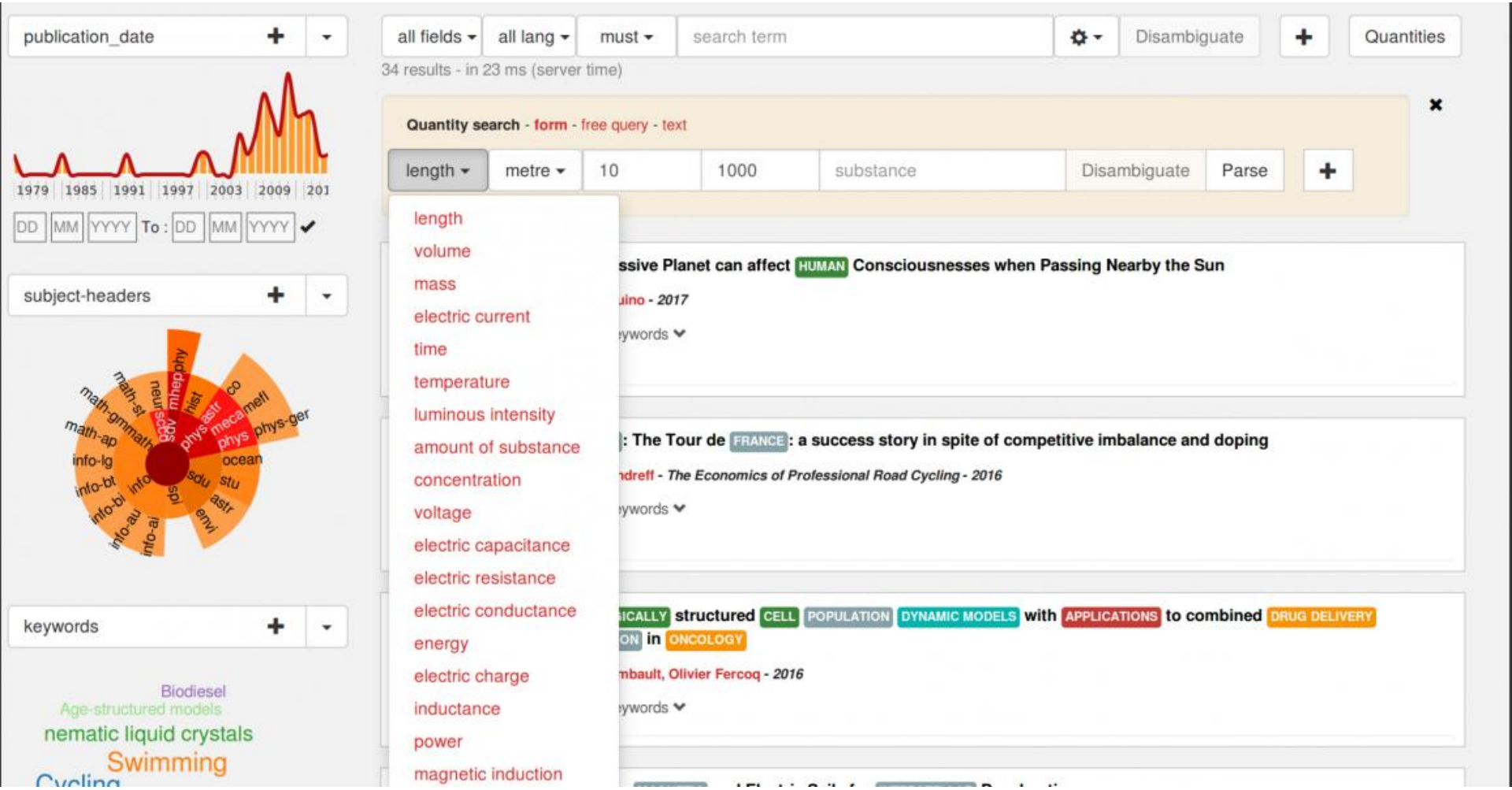
IMPLEMENTING GROBID-QUANTITIES IN A REAL-LIFE PROJECT: ISTEX

The national Istex project

- ANR funded project (*Investissement d'avenir*)
- Currently 15 million objects (target: 20M)
  - Specific requirements concerning archival and re-use (text and data mining)…
- Production line
  - Conversion of publishers' formats
    - Metadata, unstructured full-text, structured full-text
    - TEI (Text Encoding Initiative) as a pivot format
  - Automatic meta-data extraction from PDF (Grobid)
- anHALytics as a content browsing platform

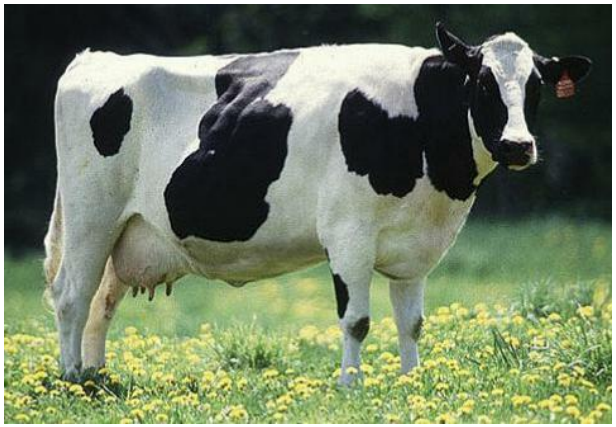
The Istex document repository

Searching quantities in a corpus of scientific text (anHALytics)



APPLICATION 2 – STRUCTURING CONTENT FROM PDF DOCUMENTS

Beyond PDF documents…



Cow (structured data)

publishing

text-mining

Hamburger
(unstructured data)

"Converting PDF to XML is a bit like converting hamburgers into cows. You may be best off printing it and then scanning the result through a decent OCR package."
Michael Kay (http://lists.xml.org/archives/xml-dev/200607/msg00509.html)

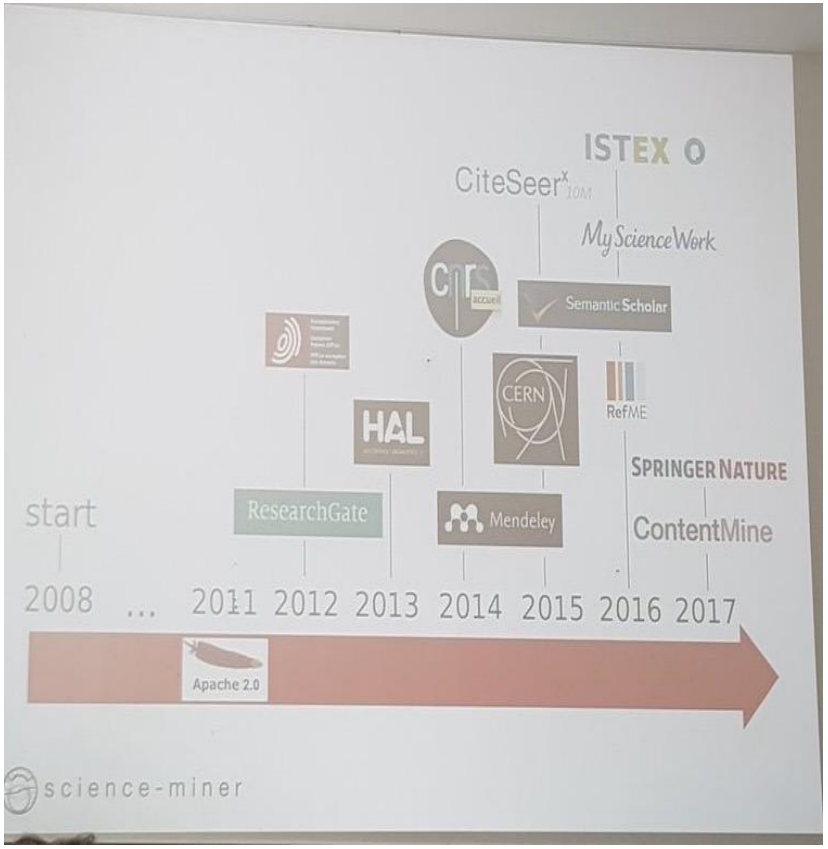
Inspired from: Duncan Hull

Getting acquainted to GROBID

GROBID (GeneRation Of Bibliographic Data)

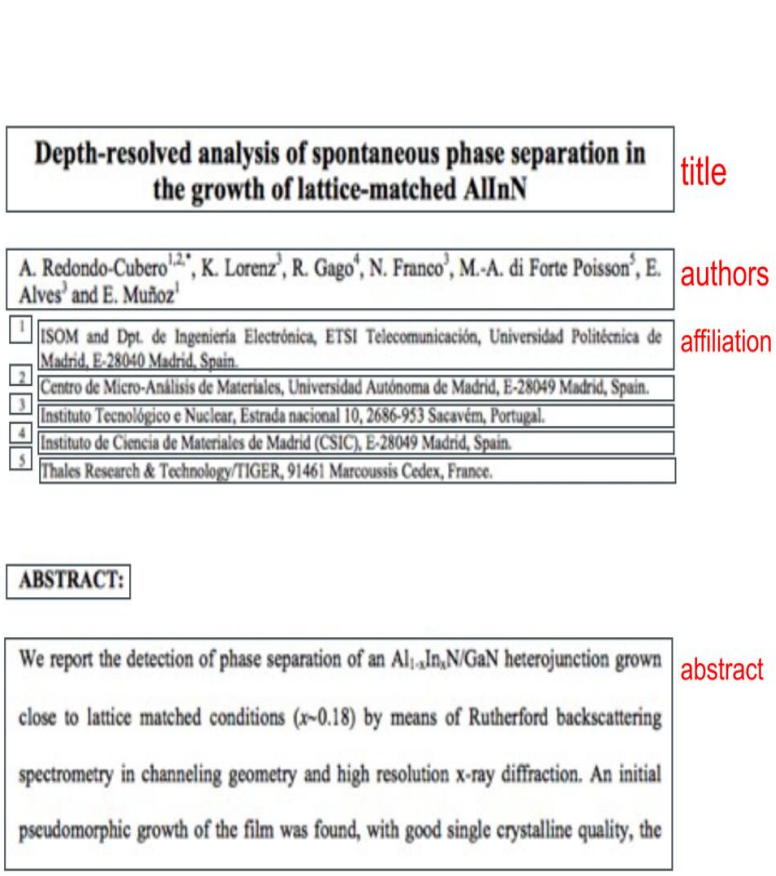
(*Lopez et al. 2015*)

- Cascading content extraction from PDF
- CRF: Conditional Random Fields
- TEI: corpus annotation and final output



Example: GROBID for meta-data extraction

GROBID (GeneRation Of Bibliographic Data) (*Lopez et al. 2015*)



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

title

A. Redondo-Cubero[1,2,*], K. Lorenz[3], R. Gago[4], N. Franco[3], M.-A. di Forte Poisson[5], E. Alves[3] and E. Muñoz[1]

authors

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.

affiliation

ABSTRACT:

We report the detection of phase separation of an $Al_{1-x}In_xN$/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

abstract

Grobid

About TEI PDF Patent Admin Doc

Service to call: Process Citation

Consolidate

Laurent Romary, Mike Mertens, Anne Baillot. Data fluidity in DARIAH – pushing the agenda forward. BIBLIOTHEK Forschung und Praxis, De Gruyter, 2016, 39 (3), pp.350-357. <hal-01285917v2>

Submit

Bibliographic reference

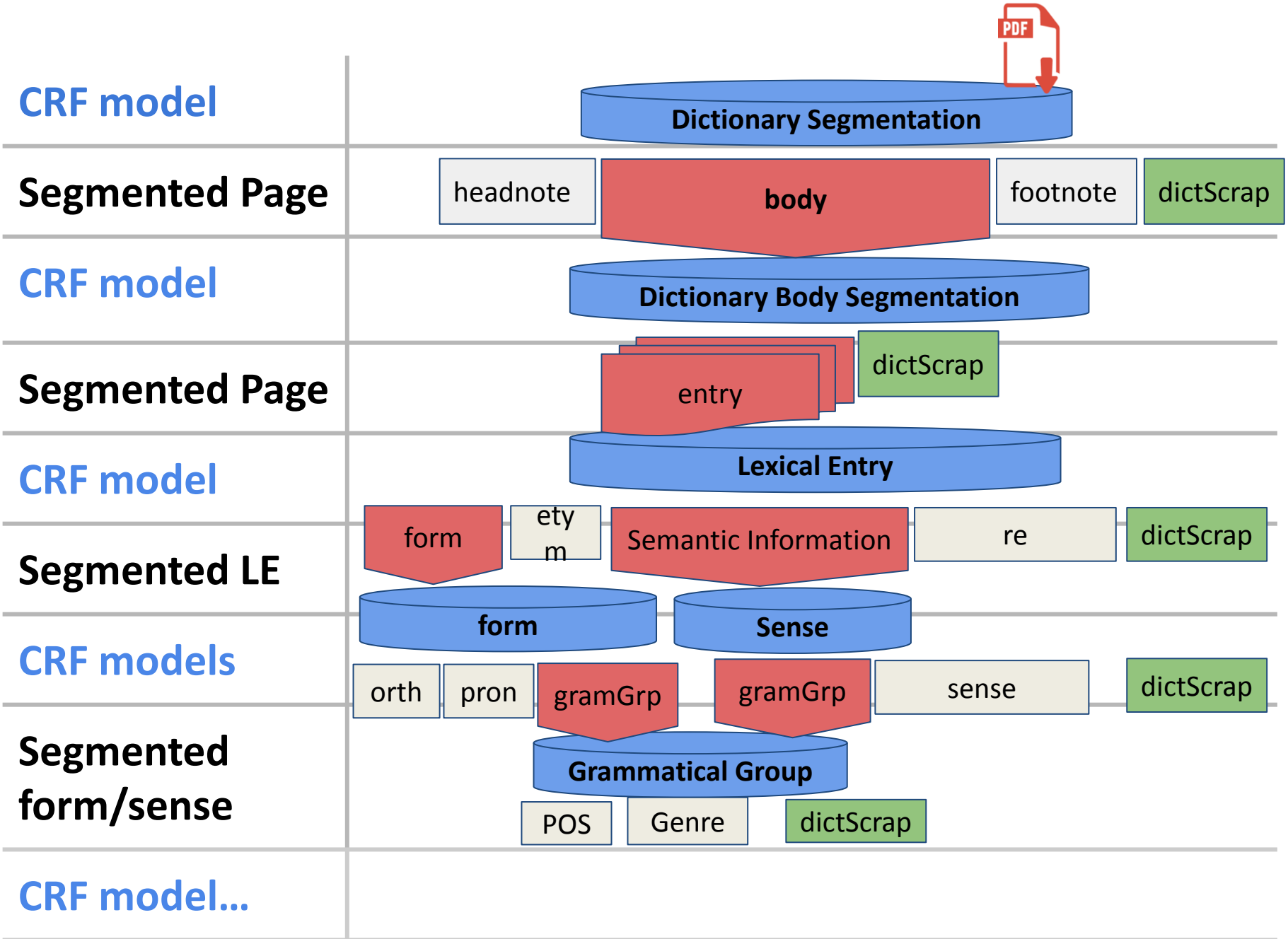
```
<biblStruct >
    <analytic>
        <title level="a" type="main">Data fluidity in DARIAH â pushing the agenda forward</title>
        <author>
            <persName
                xmlns="http://www.tei-c.org/ns/1.0" coords="">
                <forename type="first">Laurent</forename>
                <surname>Romary</surname>
            </persName>
        </author>
        <author>
            <persName
                xmlns="http://www.tei-c.org/ns/1.0" coords="">
                <forename type="first">Mike</forename>
                <surname>Mertens</surname>
            </persName>
        </author>
        <author>
            <persName
                xmlns="http://www.tei-c.org/ns/1.0" coords="">
                <forename type="first">Anne</forename>
                <surname>Baillot</surname>
            </persName>
        </author>
    </analytic>
    <monogr>
        <title level="j">BIBLIOTHEK Forschung und Praxis</title>
        <imprint>
            <biblScope unit="volume">39</biblScope>
            <biblScope unit="issue">3</biblScope>
            <biblScope unit="page" from="350" to="357" />
            <date type="published" when="2016" />
        </imprint>
    </monogr>
</biblStruct>
```

From GROBID to GROBID-Dictionary

- Numerous projects dealing with legacy (unstructured) dictionaries
  - Cf. Borchmann et alii, Widmann & Buchanan @eLex 2018
  - Monolingual, Bilingual
  - Old, modern
- Possible transition?
  - from costly manual and rule based
  - to Machine Learning techniques
- Need for exchangeable lexical resources
  - Using the Text Encoding Initiative guidelines (again)

Approach: a complex cascading of CRF models

GROBID-Dictionaries: LI Processing



Orkney Scots dictionary

chap *v.* **1** knock, '*He chappid fower or five times at the door and got no reply.*' **2** mash potatoes, '*I like me tatties chappid.*' **3** chop wood, '*Ah'll go and chap twa three sticks for the fire.*' **chappeen tree** potato masher.

```
<entry>
  <form type="lemma">
    <orth>chap</orth>
  </form>
  <gramGrp><pos>v.</pos></gramGrp>
  <sense>
    <sense>1 knock, 'He chappid fower or five times at the door and got no reply.'</sense>
    <sense>2 mash potatoes, 'I like me tatties chappid.'</sense>
    <sense>3 chop wood, 'Ah'll go and chap twa three sticks for the fire.' chappeen tree
potato masher</sense>
  </sense>
</entry>
```

Mueller (1878) (“Etymologisches Wörterbuch der englischen Sprache”, 2nd ed., Cöthen 1878/1879)

Cabbage 1. *kohl; altengl.* cabage, *bei* Hal. 226 cabes, cabishes: *mlat.* gabusia, *fr.* cabus, *it.* cappuccio; *vgl. ndl.* cabuis, cabuyscoole, *nhd.* kappes, *worüber* Weigand 1, 562: „*Im* vocab. incip. teut. ante lat. kabbas, *mhd.* der kapaʒ, kapeʒ, *spätahd.* kabuʒ, capuʒ. *Aus fr. der* cabus, *it.* capúccio, *welches wie russ. die* kapusta *kohl, aus mlat.* caputium *kapuze hervorging und der geschlossene kohl schien einer mönchskappe ähnlich;*“ *vgl.* Diez 1, 110 *und unter den nhd.* kabisz, kabis Grimm 5, 9.

```
<entry>
   <form>
      <orth>Cabbage</orth><label>1.</label>
   </form>
   <etym>
      <seg><def>kohl</def>;<lang>altengl.</lang>
         <mentioned>cabage</mentioned>,<seg>bei</seg>
         <bibl>Hal. 226</bibl><mentioned>cabes</mentioned>,
         <mentioned>cabishes</mentioned>: <lang>mlat.</lang>
         <mentioned>gabusia</mentioned>, <lang>fr.</lang>
         <mentioned>cabus</mentioned>, <lang>it.</lang>
         <mentioned>cappuccio</mentioned>;
         <seg>vgl.</seg><lang>ndl.</lang>
         <mentioned>cabuis</mentioned>,
         <mentioned>cabuyscoole</mentioned>, <lang>nhd.</lang>
         <mentioned>kappes</mentioned>, <seg>worüber</seg>
         <bibl>Weigand 1, 562</bibl>:
      </seg>
      <quote>„Im <bibl>vocab. incip. teut.</bibl> ante
         <lang>lat.</lang> <mentioned>kabbas</mentioned>,
         <lang>mhd.</lang> der <mentioned>kapaʒ</mentioned>,
         <mentioned>kapeʒ</mentioned>,<lang>spätahd.</lang>
         <mentioned>kabuʒ</mentioned>,
         <mentioned>capuʒ</mentioned>.Aus <lang>fr.</lang>
         der <mentioned>cabus</mentioned>, <lang>it.</lang>
         <mentioned>capúccio</mentioned>, welches wie
         <lang>russ.</lang> die <mentioned>kapusta</mentioned>
         <def>kohl</def>, aus <lang>mlat.</lang>
         <mentioned>caputium</mentioned> <def>kapuze</def>
         hervorging und der geschlossene kohl schien einer
         mönchskappe ähnlich;“
      </quote>
      <seg><seg>vgl.</seg> <bibl>Diez 1, 10</bibl>
         <seg>und unter den</seg> <lang>nhd.</lang>
         <mentioned>kabisz</mentioned>,
         <mentioned>kabis</mentioned> <bibl>Grimm 5, 9</bibl>.
      </seg>
   </etym>
</entry>
```

Grobid - perspectives

- A generic machinery for text document analysis
  - Online services, Open source software
- Still some work ahead
  - Tables, figures, etc.
  - More models for more forms of documents (books, PhD theses) or dictionaries (related entries)
- Application: Re-publishing content
  - E.g. public domain material
    - *Petit Larousse illustré* – joint project with Univ. Montpellier
  - The XML-TEI as a pivot format
    - Generation of HTML, ePub, simplified versions (e.g. learners' dictionary)
    - … or Braille output for visually impaired persons
      - Specific exception in France since 2006 or from the Marrakech treaty (2013)

LE DESSERT – CAFÉ GOURMAND

STANDARDS – AN ESSENTIAL ASPECT OF OPEN KNOWLEDGE

Why do we actually need standards?

- A paradox in science
  - Freezing knowledge, overhead in a creativity phase, forcing scholars to produce comparable results…
- Some obvious arguments
  - Long term legibility of data
  - Facilitate re-use and, yes, comparison of results
  - Simplifies the development of generic tools (queries, visualisation, etc.)

The text and language standardisation landscape

- Generic horizontal standards
  - ISO 639 series and BCP 47 for languages
  - Scripts, times and dates, etc.
  - XML for document representation
- Specific vertical standards
  - Text Encoding Initiative (TEI) for textual content
    - Covers a variety of textual forms: prose, drama, manuscript transcription, dictionaries
    - 569 XML elements: shared culture for digital projects in the humanities
    - Open licence: CC-BY + BSD 2 clauses
    - Continuous revision: 2 releases per year

How to be standards-oriented?

- Advocate
  - As an implementer, but also as a research manager … or a funder (DFG, ANR)
- Train
  - The role of DARIAH (cf. #DARIAHTeach)
  - Introducing basic standards literacy in university curricula
- Participate
  - Standards are not frozen and must reflect on-going practices
  - Contributing to standards as a way to transfer knowledge
    - E.g. elected members of the TEI technical council

THE UNDERLYING NEED OF AN OPEN ACCESS POLICY/STRATEGY

Open science - Acting at all levels

- Principles and policies
  - Data seal of approval, FAIR principles, benefits of open science
- Formats and standards
  - Documented, re-usable and comparable data
- Attribution
  - Acknowledging authorship in the wide sense: authors, institutions, projects
- Licences
  - Eliciting re-use possibilities
- Repositories
  - Making data visible and re-usable

Open access issues and questions

- How do we gain maximal access and re-use rights on scientific publications?
- Do we want to leave the publication landscape unchanged?
  - When the digital turn has already changed so much…
- Do we let the private sector take care of it for us?
- How do we preserve our digital sovereignty as scholars?

The French context

- Inria – a strong open access policy since 2014
  - Deposit mandate in the HAL national repository (annual report)
  - Forbidding hybrid APCs (Article Processing Charges)
  - Central budget for APCs corresponding to full open access journal
- Loi pour une République Numérique
  - Article 30: for publicly funded research (50%), right to make post-print available online, 6/12 month embargo, publisher's version if available online
- Appel de Jussieu - http://jussieucall.org
  - Reducing subscription budgets to invest in new publication models
- HCERES - Haut Conseil de l'évaluation de la recherche et de l'enseignement supérieur
  - Assessment institution for higher education and research
  - Announced (Jan. 2018) that HAL will be the only source for publications

Digestif

- Where do we go from here?
  - Be open
    - Publications, data, software, services
    - This is the role of the EU infrastructure DARIAH to facilitate this…
  - Understand the magic behind models
    - Increasing computer literacy is a must in the digital turn
      - But this does not necessarily mean “programming”
  - Re-use and cite
    - A necessary condition for the success of the openness agenda

Merci de votre attention!

laurent.romary@inria.fr

@laurentromary

References

DARIAH

Jennifer Edmond, Frank Fischer, Michael Mertens, Laurent Romary. The DARIAH ERIC: Redefining Research Infrastructure for the Arts and Humanities in the Digital Age. *ERCIM News*, ERCIM, 2017, Digital Humanities. 〈hal-01588665〉

Openness

Laurent Romary. Wissenschaftliche Arbeit und Open Access. *GEGENWORTE*, Berlin-Brandenburgischen Akademie der Wissenschaften, 2013, 30. 〈hal-00854014〉

Laurent Romary, Michael Mertens, Anne Baillot. Data fluidity in DARIAH – pushing the agenda forward. *BIBLIOTHEK Forschung und Praxis*, De Gruyter, 2016, 39 (3), pp.350-357. 〈10.1515/bfp-2016-0039〉. 〈hal-01285917v2〉

Grobid

Laurent Romary, Patrice Lopez. GROBID - Information Extraction from Scientific Publications. *ERCIM News*, ERCIM, 2015, Scientific Data Sharing and Re-use, 100, 〈https://ercim-news.ercim.eu/en100/r-i/grobid-information-extraction-from-scientific-publications〉. 〈hal-01673305〉

Grobid dictionary

https://digilex.hypotheses.org/250

Mohamed Khemakhem, Luca Foppiano, Laurent Romary. Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. *electronic lexicography, eLex 2017*, Sep 2017, Leiden, Netherlands. 〈hal-01508868v2〉

TEI

Laurent Romary. Questions & Answers for TEI Newcomers. *Jahrbuch für Computerphilologie 10*, Mentis Verlag, 2009, Jahrbuch für Computerphilologie. 〈hal-00348372v2〉

Digital turn

Sandra Collins, et al.. Going Digital: Creating Change in the Humanities: ALLEA E-HUMANITIES WORKING GROUP REPORT. [Research Report] ALLEA. 2015. 〈hal-01154796〉

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