Open Access to Research Data

Accessing, using and publishing collections of textual data in digital literary studies

Christof Schöch
Junior Research Group "Computational Literary Genre Stylistics" (CLiGS)
Department for Literary Computing, University of Würzburg, Germany

DARIAH-IT Summer School: *Models for Open Access Publishing*
Villa Vigoni, Italy, June 9-11, 2015

Introduction

Overview

- Introduction
  - What is open access to research data?
  - Related issues
- A closer look
  - What can you do with open research data?
  - What is required from research data?
  - Do resources fulfil these requirements?
  - Why is getting access good, but giving access better?
- Current issues
  - Some challenges
  - Main hindrance: Legal issues
- Conclusions
- Recommended readings

What is open access to research data in the humanities?

- Open Access (reminder): The possibility to access, read, download, modify, mix, analyse, share, store, republish materials without legal, economic, or technical hindrances; see opendefinition.org
- Research data in the humanities: "Structured collections of digital, selectively constructed, machine-actionable abstractions representing some aspects of a given object of humanistic inquiry."
- Examples for research data
  - Primary sources: literary texts, correspondence, newspapers, language corpora, historical documents, paintings, other human artefacts
  - Factual data relevant to humanistic inquiry: data about people, places, cultures, and primary sources
  - Research publications: journal articles, books, blog posts, twitter (state of the art, history of science)

Some related issues

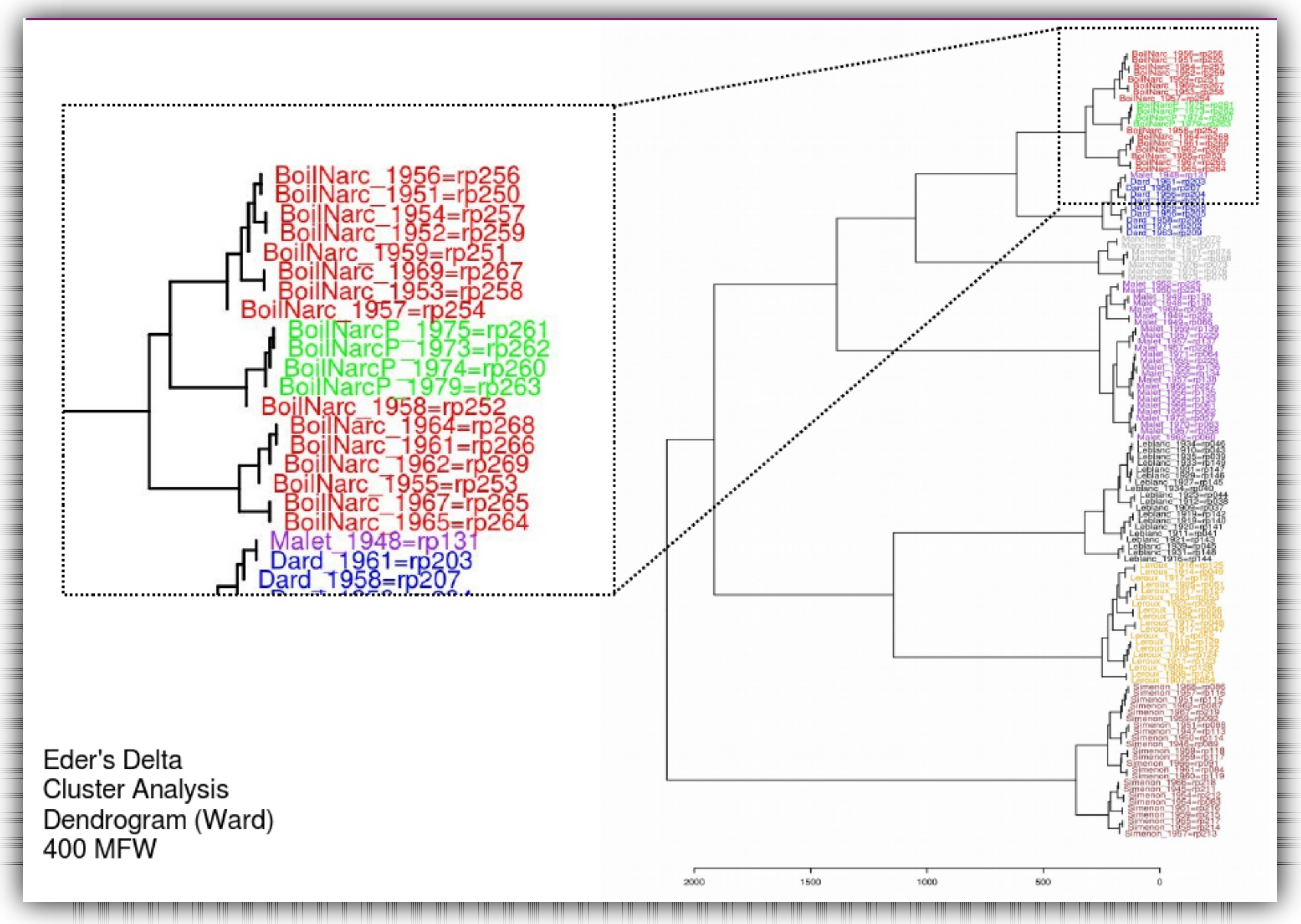
- Open Access Publishing (of course)
- Linked Open Data
- Open Source Software
- Collaborative Writing (platforms, data formats)
- Long-Term Archiving
- Dataset Peer Review
- Citeability and academic credit

A closer look

What can you do with open research data?

- Authorship Attribution / Stylistics: who wrote this anonymous text? Is this pastiche successful?
- Genre Stylistics: What are the stylistic features distinctive of two related literary genres?
- Topic Modeling: which themes are there, and how are they distributed, in a text collection
- Literary Network Analysis: who speaks how often with whom, in a play?
- Citation network analysis: who cites whom, and is cited by whom?
- ...many more things!

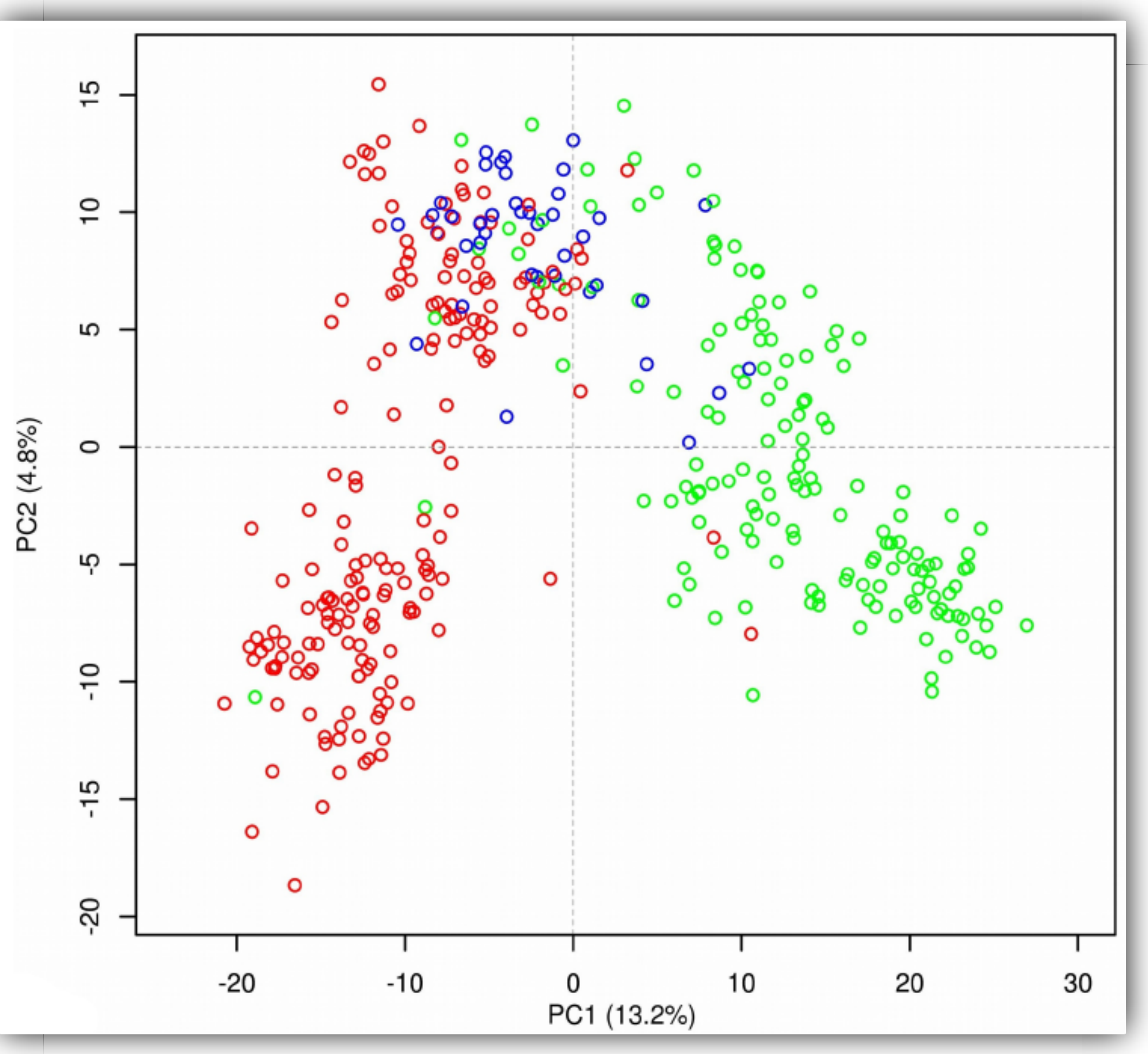
Authorship Attribution / Stylistics



Eder's Delta
Cluster Analysis
Dendrogram (Ward)
400 MFW

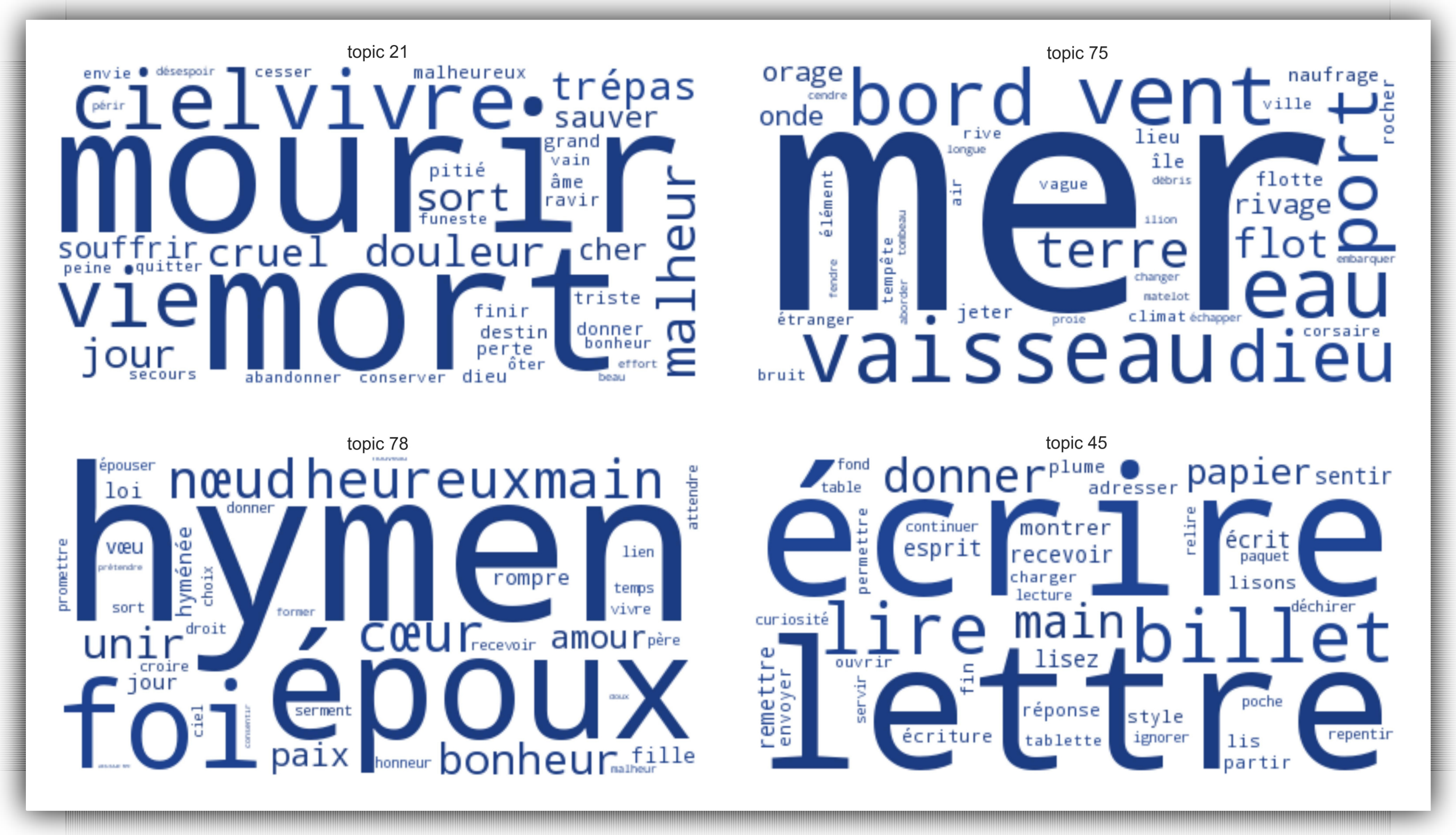
- French crime fiction Boileau-Narcejac: Arsène Lupin pastiches

Genre Stylistics



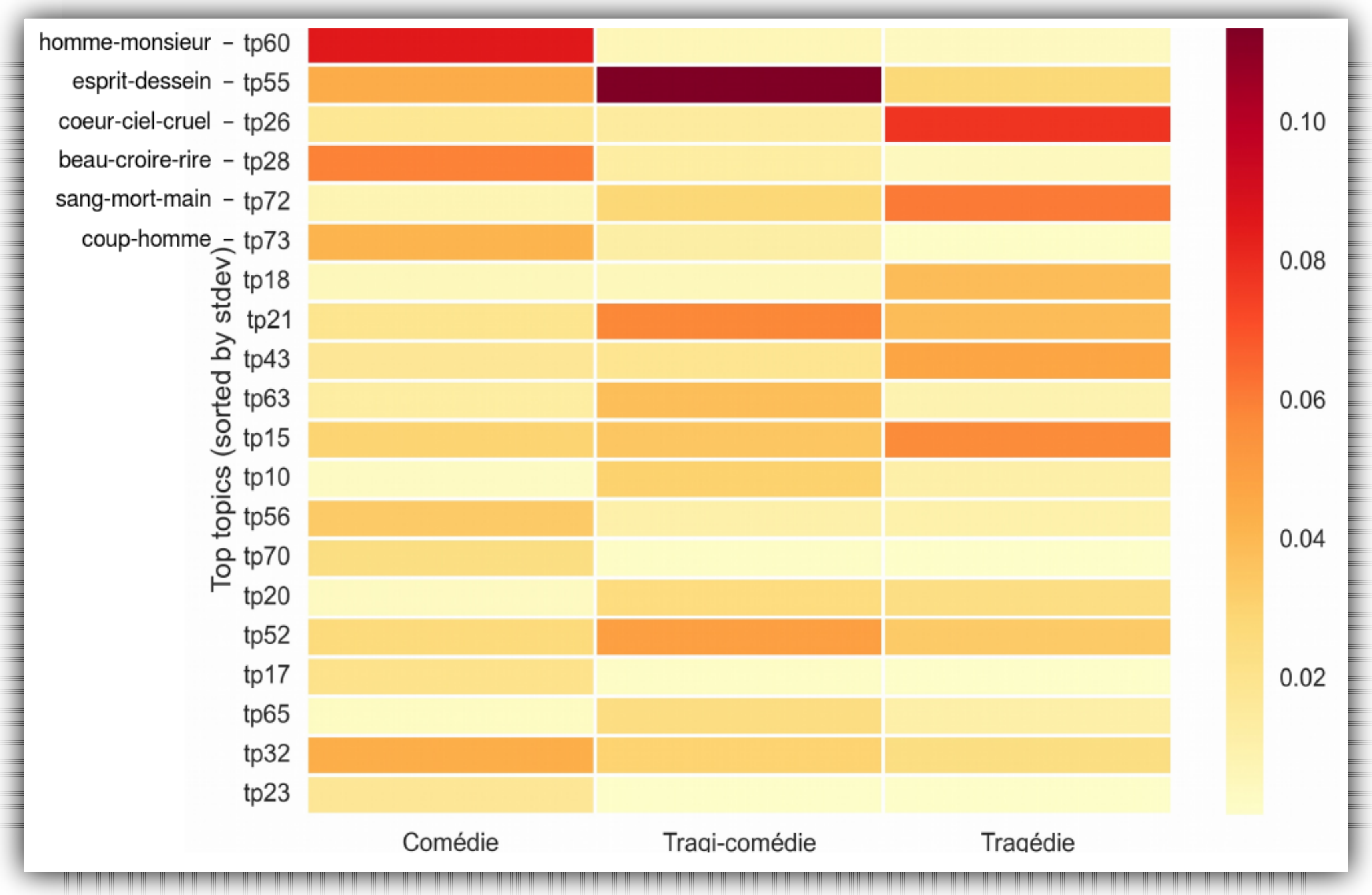
- French drama: comedies (red), tragedies (green) and tragi-comedies (blue)

Topic Modeling



- French drama: four out of 80 topics in 375 plays

Topics in Subgenres



- French drama: distinctive topics for three dramatic subgenres

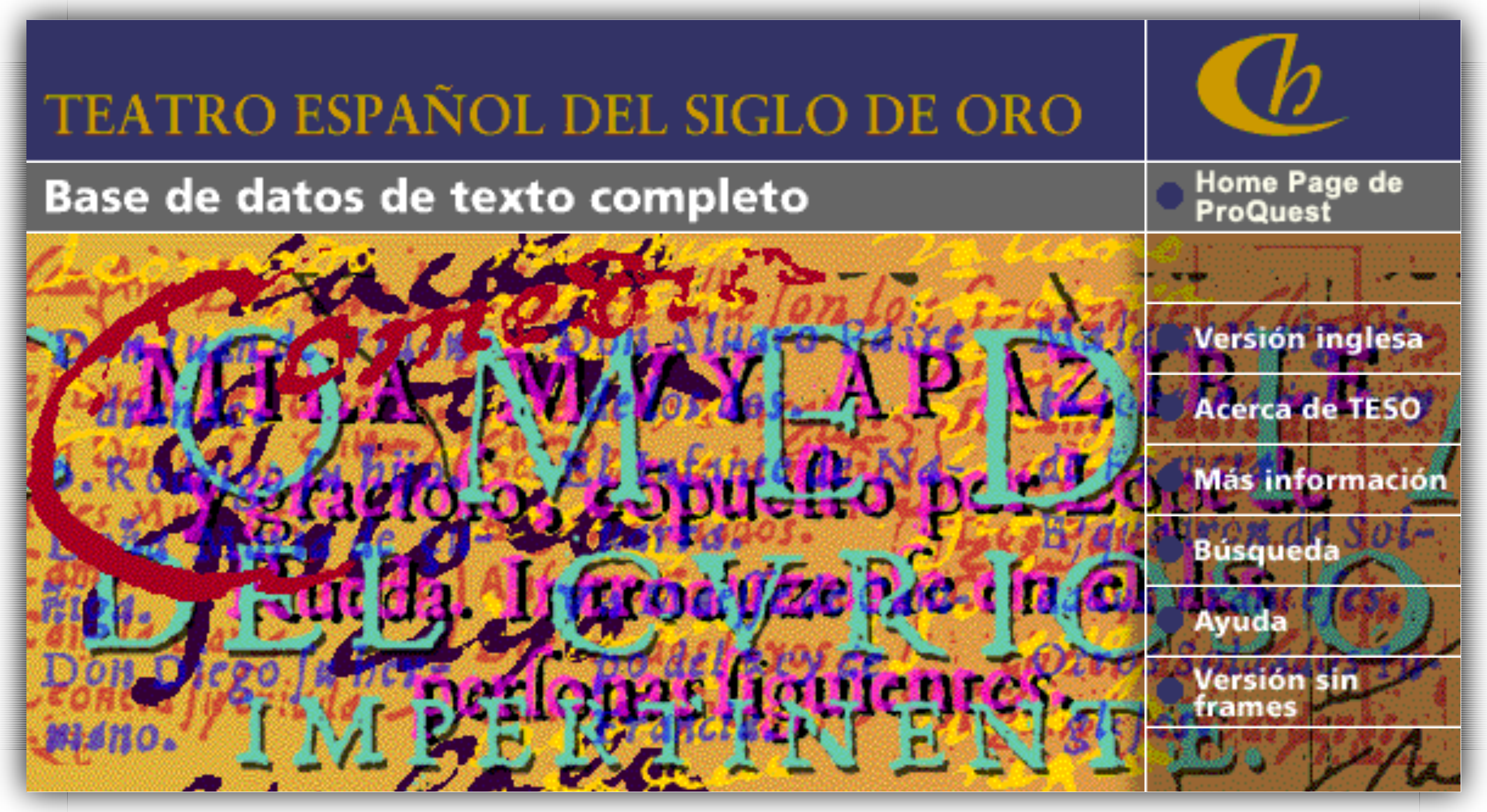
What are the requirements?

- Quantity of data: the more, the better; digital methods scale and sometimes need scale
- Quality of data: OCR / transcription with few errors; transparency about this
- File formats: semi-structured / structured, consistent, interoperable, non-proprietary, e.g. XML-based
- Quality of metadata: Detailed, complete, consistent
- Availability: selective bulk download
- Rights: access, read, download, modify, mix, analyse, share, store, republish

Are these requirements being fulfilled?

Some examples

Teatro español del siglo de oro (Chadwyck)

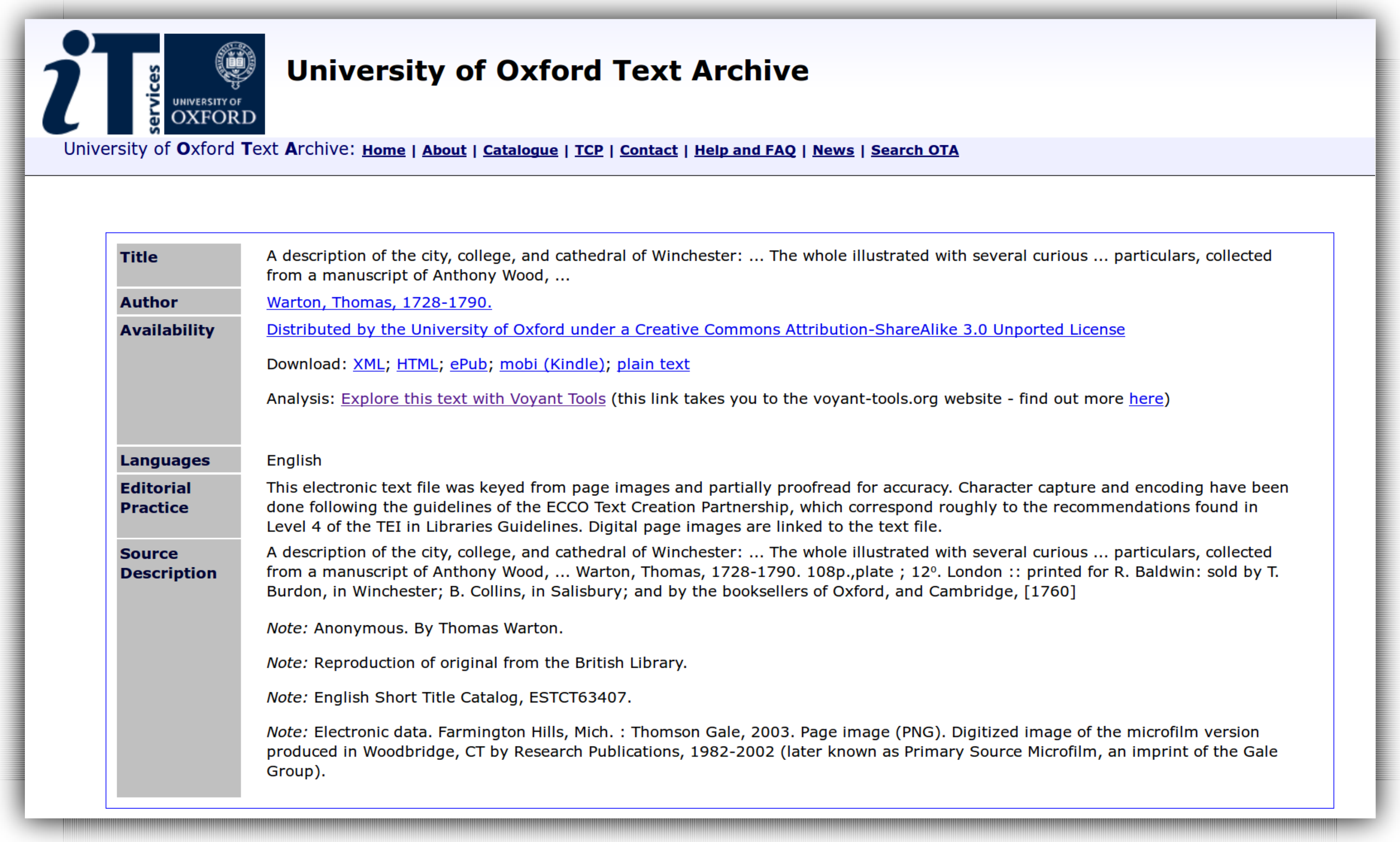


TEATRO ESPAÑOL DEL SIGLO DE ORO

Base de datos de texto completo

- Home Page de ProQuest
- Versión inglesa
- Acerca de TESO
- Más información
- Búsqueda
- Ayuda
- Versión sin frames

Oxford Text Archive

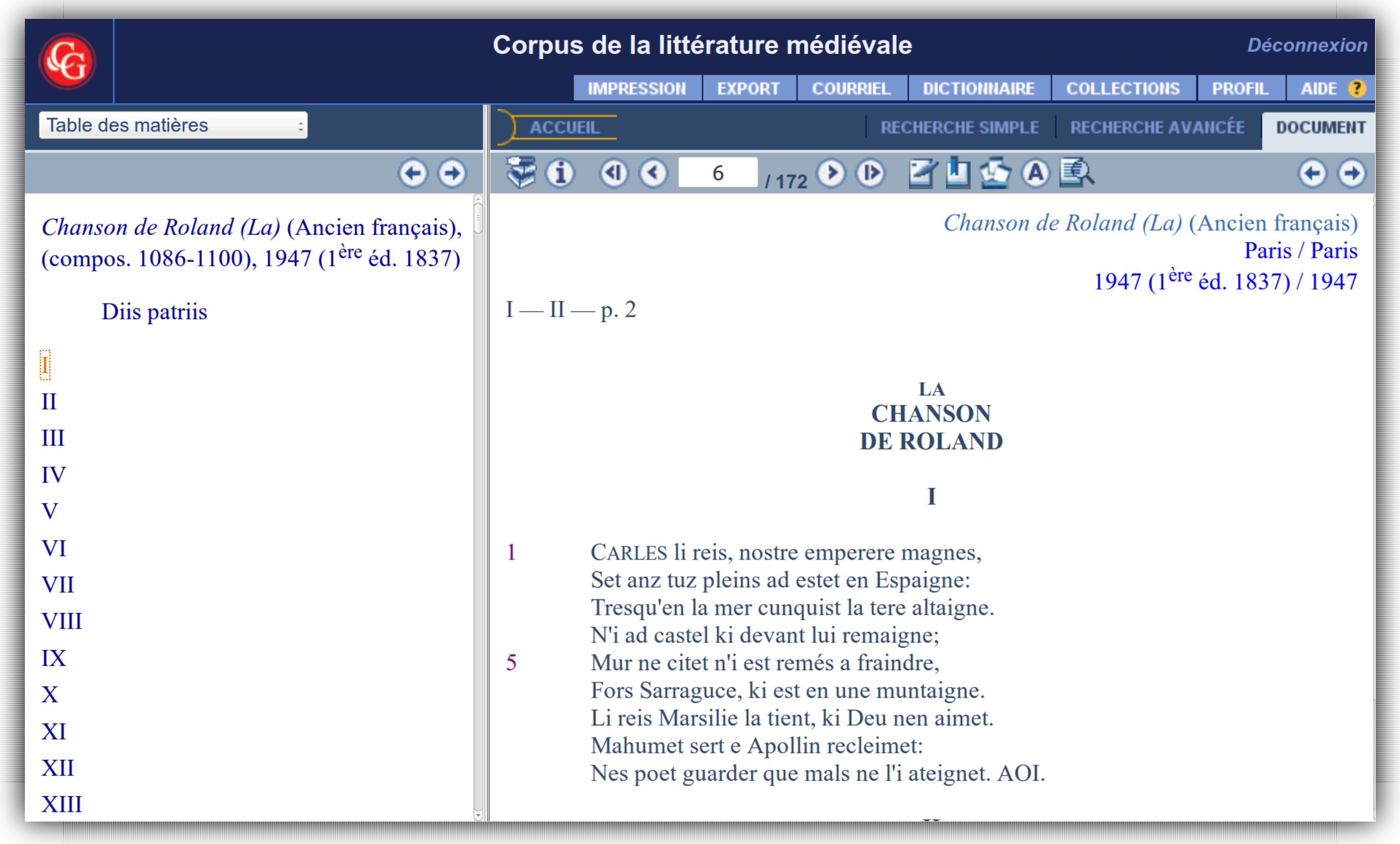


University of Oxford Text Archive

University of Oxford Text Archive: Home | About | Catalogue | TCP | Contact | Help and FAQ | News | Search OTA

Title	A description of the city, college, and cathedral of Winchester: ... The whole illustrated with several curious ... particulars, collected from a manuscript of Anthony Wood, ...
Author	Warton, Thomas, 1728-1790.
Availability	Distributed by the University of Oxford under a Creative Commons Attribution-ShareAlike 3.0 Unported License Download: XML; HTML; ePub; mobi (Kindle); plain text Analysis: Explore this text with Voyant Tools (this link takes you to the voyant-tools.org website - find out more here)
Languages	English
Editorial Practice	This electronic text file was keyed from page images and partially proofread for accuracy. Character capture and encoding have been done following the guidelines of the ECCO Text Creation Partnership, which correspond roughly to the recommendations found in Level 4 of the TEI in Libraries Guidelines. Digital page images are linked to the text file.
Source Description	A description of the city, college, and cathedral of Winchester: ... The whole illustrated with several curious ... particulars, collected from a manuscript of Anthony Wood, ... Warton, Thomas, 1728-1790. 108p.,plate ; 12º. London :: printed for R. Baldwin: sold by T. Burdon, in Winchester; B. Collins, in Salisbury; and by the booksellers of Oxford, and Cambridge, [1760] *Note:* Anonymous. By Thomas Warton. *Note:* Reproduction of original from the British Library. *Note:* English Short Title Catalog, ESTCT63407. *Note:* Electronic data. Farmington Hills, Mich. : Thomson Gale, 2003. Page image (PNG). Digitized image of the microfilm version produced in Woodbridge, CT by Research Publications, 1982-2002 (later known as Primary Source Microfilm, an imprint of the Gale Group).

Corpus de littérature médiévale (Garnier)



Chanson de Roland (La) (Ancien français), (compos. 1086-1100), 1947 (1ère éd. 1837)

Diis patriis

I
II
III
IV
V
VI
VII
VIII
IX
X
XI
XII
XIII

Chanson de Roland (La) (Ancien français)
Paris / Paris
1947 (1ère éd. 1837) / 1947

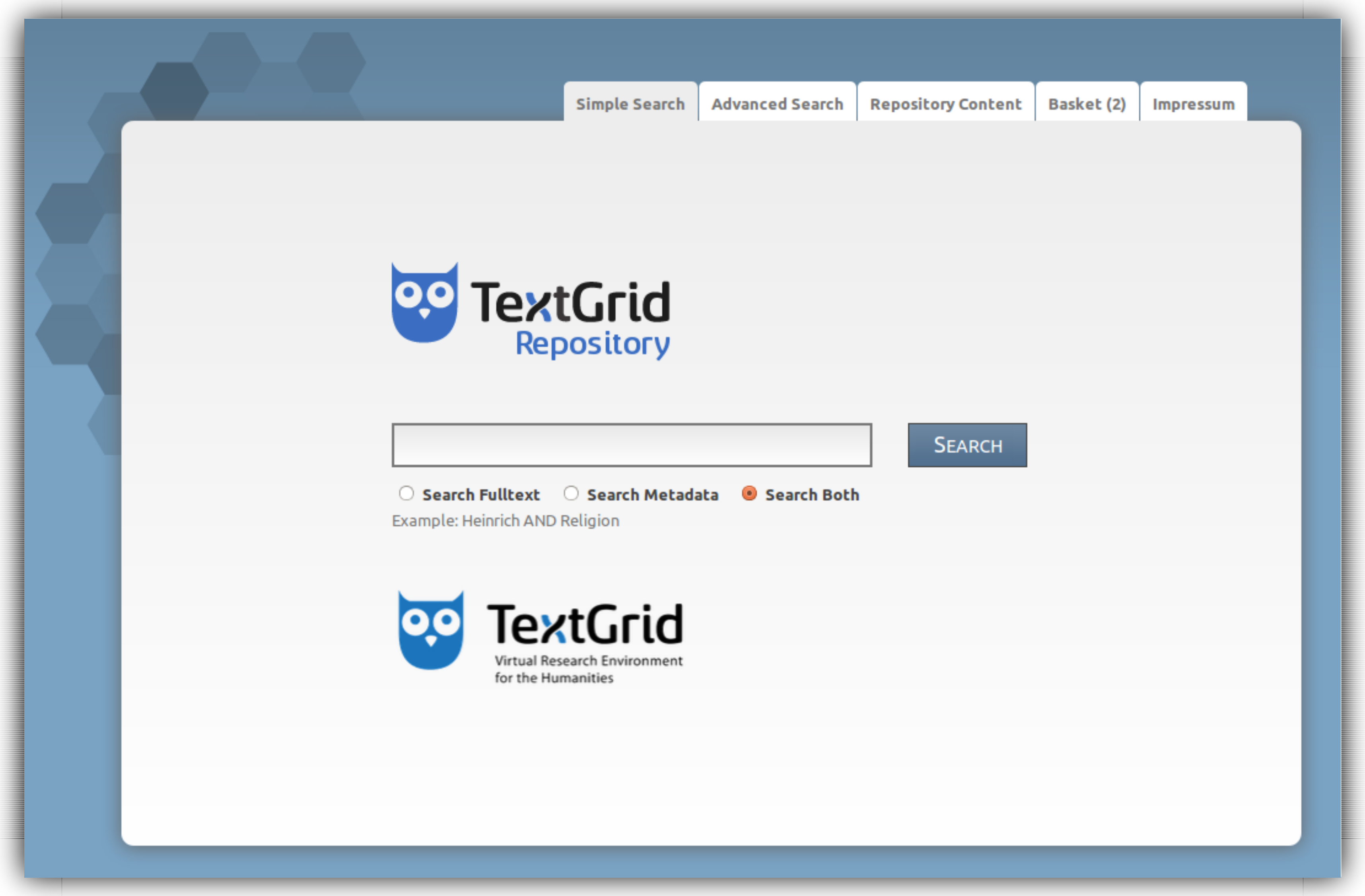
I — II — p. 2

LA
CHANSON
DE ROLAND

I

1 CARLES li reis, nostre emperere magnes,
Set anz tuz pleins ad estet en Espaigne:
Tresqu'en la mer cunquist la tere altaigne.
N'i ad castel ki devant lui remaigne;
5 Mur ne citet n'i est remés a fraindre,
Fors Sarraguce, ki est en une muntaigne.
Li reis Marsilie la tient, ki Deu nen aimet.
Mahumet sert e Apollin recleimet:
Nes poet guarder que mals ne l'i ateignet. AOI.

Digitale Bibliothek (TextGrid)



○ Search Fulltext ○ Search Metadata ● Search Both

Example: Heinrich AND Religion

Why is getting access good, but giving access better?

- Getting Access:
    - Efficiency: do analyses faster / differently
    - Innovation: ask new questions, use new research methods
    - Avoid redundancy of efforts: digitize once, and once only
- Giving Access:
    - Transparency: show others the foundations of your analysis
    - Reproducibility: let others do your analysis again
    - Reusability: Promote research into the data you are interested in

Current issues

Some challenges

- Availability of useful data (OCR & copyright)
- Standards for data and metadata formats
  - PDF and Word are not good analysis formats (reading vs. archiving vs. mining)
  - Plain-text formats like XML, JSON, CSV, Bibtex are better
- Competencies in two domains at a time
  - Humanistic domain (research questions, contextual knowledge)
  - Computational / methodological domain (methods, tools, algorithms, programming language)

Main hindrance: legal issues

- Copyright law
- Database Protection
- Subscription contracts
- Licenses
- More than anything else: uncertainty about these issues

"All in all, our evaluation at this stage of our analysis is that rightholders and publishers are in a more comfortable position than research institutions, let alone commercial companies willing to engage in data analysis projects. Whether this is a desirable result or not is a policy option, for which the objectives of the European Union to promote research and innovation should be taken into account." (Triaille, Study on the Legal Framework of Text and Data Mining, 2014)

Conclusion

What can infrastructures (like DARIAH) do?

- Infrastructures concern technical solutions, but also competencies and community building
- Therefore, infrastructures can explain, recommend or offer...
  - Standards for data and metadata
  - Legal issues
  - Trustworthy repositories (e.g., Zenodo, DARIAH Repository)
  - Persistent Identifier (PID) services (DOI, handle)
  - Discovery services

What can each of us do?

- Follow the rule: "love your data and help others love it, too." (Goodman et al. 2014)
- If you find useful research data that is not open access, tell the world that it isn't, and ask the provider why it isn't
- If you find freely available, useful research data, use it in your research to show people what wonderful things can be done with it
- If you produce research data, publish it with a DOI, in a standard format, with an open license whenever possible (Creative Commons "BY" or "0"), and add as much information to it as you can (metadata)

Thank you!

Recommended Readings

- Christine L. Borgman. “The Conundrum of Sharing Research Data.” *Journal of the American Society for Information Science and Technology*, 2012. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1869155.
- Alyssa Goodman et al. “Ten Simple Rules for the Care and Feeding of Scientific Data.” *PLoS Computational Biology* 10.4, 2014. doi:10.1371/journal.pcbi.1003542, http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1003542.
- Matthew L. Jockers. *Macroanalysis. Digital Methods and Literary History*. Champaign, IL: University of Illinois Press, 2013.
- *Riding the wave - How Europe can gain from the rising tide of scientific data*. Final Report. Brussels: EU, 2010. http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf