27-29 March 2019

National Library of Australia, Canberra

Australian Academy of the Humanities' 2nd Humanities, Arts and Culture Data Summit

and

3rd international DARIAH Beyond Europe workshop

DARIAH-EU
Digital Research Infrastructure for the Arts and Humanities

Australian Research Data Commons

#DARIAHBeyondEurope #HACDS2019

Infrastructural Challenges for Large Scale Digital Text Corpora:
A View From the European Margins

Dr Antonija Primorac, Associate Professor

Faculty of Humanities and Social Sciences

University of Rijeka, Croatia

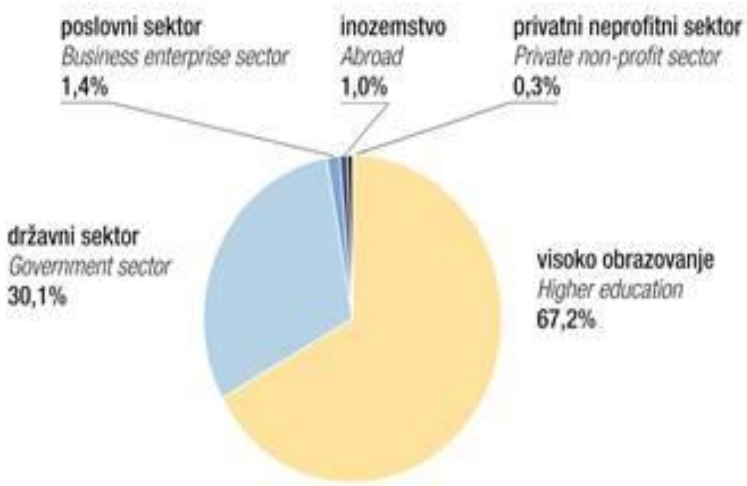
COST Action *Distant Reading for European Literary History*

Distant Reading

UNIRI

Why ‘margins’?

G-1. UDIO DRŽAVNIH PRORAČUNSKIH SREDSTAVA ZA IR PREMA SEKTORIMA (STVARNI IZDACI) U 2017.
SHARE OF GOVERNMENT BUDGET APPROPRIATIONS OR OUTLAYS FOR R&D, BY SECTORS (ACTUAL PAYMENTS), 2017



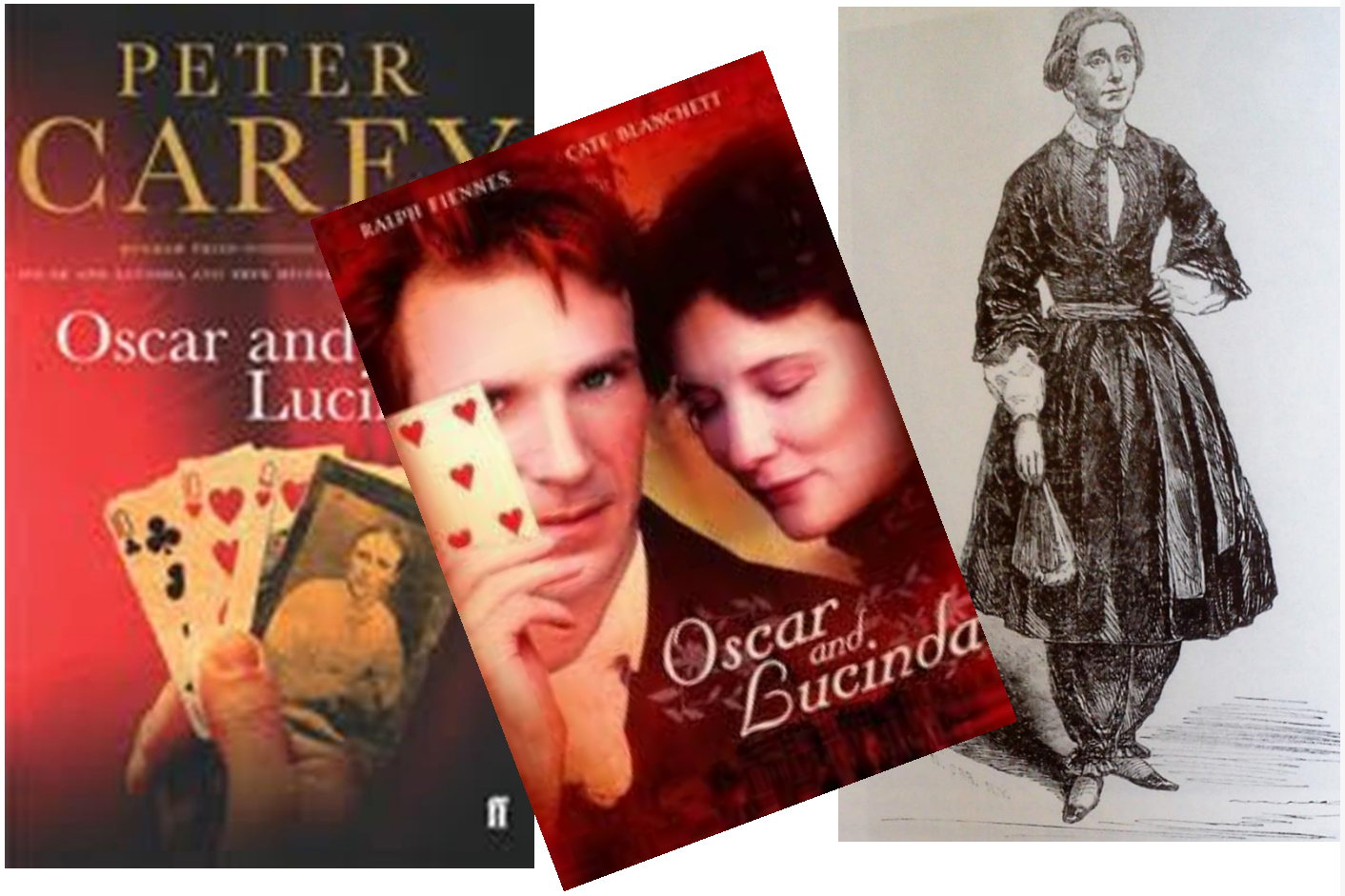
Only 0,7% of the national budget for research* infrastructure in Croatia → Increased importance of OA for researchers!

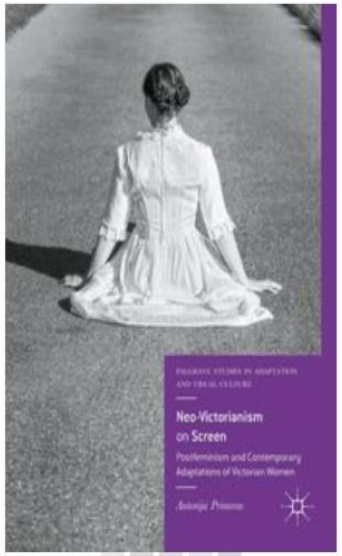
**Source:*
Croatian Bureau of Statistics’ Report for 2017-18, https://www.dzs.hr/Hrv_Eng/publication/2018/08-02-02_01_2018.htm

Personal PoV:

Using TROVE's Digitised newspapers for research

published as Chapter 3 in *Neo-Victorianism on Screen: Postfeminism and Contemporary Adaptations of Victorian Women* (Palgrave Macmillan 2018)





CHAPTER 3

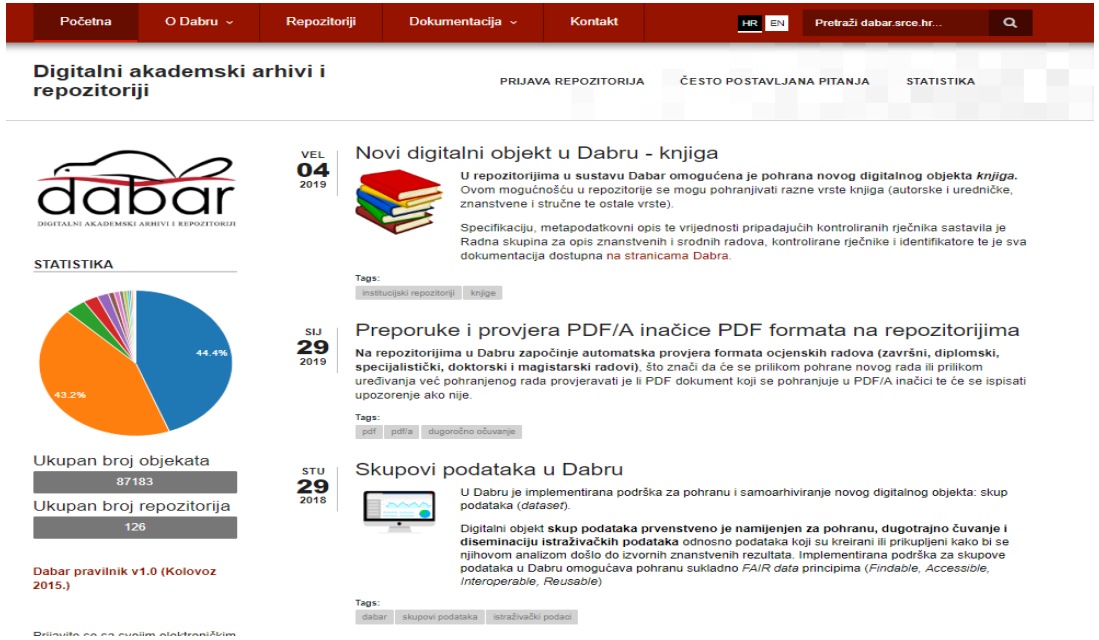
Neo-Victorianism on Screen

Re-presenting the Past: Gender, Colonial Space and Cultural Nostalgia in Neo-Victorianism on Screen

Authors: Antonija Primorac

This chapter examines the relationship

Open access in Humanities Research: a Croatian Perspective



Most academic output (from BA and MA theses to articles and books) available through different repositories (e.g. Hrčak, or Dabar:

https://dabar.srce.hr/repozitoriji

Issues:

- Access (via institutional e-mail for some repositories)
- Fragmentary nature
  - no unified/centralised approach to metadata or format, no one depository or centralised network, each institution has its own depository, which affects funding *and* access
- Organisation
  - Two ministries in charge (Ministry of Culture and Ministry of Science)
  - Disjunction between relevant civil servants & academics
- Strategy
- No continuity of policies (political)
- Management

Infrastructural Challenges for Large Scale Digital Text Corpora in Croatia

Burning issues:

- **Funding:** no more funding for infrastructural needs in research projects ('fundamental research only' policy), either **for digitisation** or **for maintenance of digitised data**; exception: digitising 'cultural heritage' (purview of the Ministry of Culture)
- **Copyright limitations** (e.g. issues with a project on digitised publications in the National and University Library, Zagreb)

Previously:

- No systematic funding for the digitalisation of literature (instead: project related).
- Result: lack of books, plethora of periodicals:
  - some manuscripts; literature for use in schools (classics and the Croatian canon), and large swathes of newspaper archives have been digitised
  - **Just because there isn't *digitised* data doesn't mean data doesn't exist!**

COST Action
Distant Reading for European Literary History

Distant Reading

Grounded in the Distant Reading paradigm (i.e. using computational methods of analysis for large collections of literary texts).

Set to create a shared theoretical and practical framework to enable innovative, sophisticated, data-driven, computational methods of literary text analysis across at least 10 European languages (currently 27 European and non-European countries involved).

Aims:

- To foster an insight into cross-national, large-scale patterns and evolutions across European literary traditions,
- To facilitate the creation of a broader, more inclusive and better-grounded account of European literary history and cultural identity.

COST Action
Distant Reading for European Literary History

Goals:

- build a multilingual European Literary Text Collection (ELTeC), ultimately containing around 2,500 full-text novels in at least 10 different languages, permitting to test methods and compare results across national traditions;
- establish and share best practices and develop innovative computational methods of text analysis adapted to European multilingual literary traditions;
- consider the consequences of such resources and methods for rethinking fundamental concepts in literary theory and history.

COST Action
Distant Reading for European Literary History

- FAIR Guiding Principles: (meta)data should be "findable, accessible, interoperable, re-useable“
- Every software and data creating framework we use is open source and that every member of each working group (Action as a whole) can use them freely.
  - Everything we do should be open source and freely available, hence everybody inside and outside the Action can access our work. Our corpus is licenced with CC-BY, a licence which everybody allows to reuse our corpus in different contexts.
- Github: provides versioning control, documentation and communication features

COST Action
Distant Reading for European Literary History

- Identified releases of European Literary Text Collection / ELTeC (including all sub-corpora) to be published on Zenodo.
- For defining and validating encoding levels of our corpus: use the modelling framework of the TEI (Text encoding initiative); this is called ODD (One Document does it all)
  - The TEI framework is established within the DH community;
  - It allows participants to handle different levels of encoding;
  - It allows adding metadata describing the resource, the encoding schema, as well as the contributing persons.

Infrastructural Challenges facing COST Action Distant Reading for European Literary History:

- One of the plans was to use freely available library catalogue metadata (WorldCat data) to estimate the ‘canonicity’ of any given novel.
- Experience so far:
  - it has proven very difficult to get hold of a WorldCat Search API key for the project, without paying a substantial fee (for which there is no budget in the Action).
- Desirable change:
  - Open Access to library catalogue data on a large scale
  - A research infrastructure requirement for our project.
  - Also useful far beyond the COST Action.

Infrastructural Challenges: COST Action *Distant Reading for European Literary History*

Desirables:

- Access to an **Open Access OCR pipeline that is able to deal with multiple languages**, including *historical states of languages* and *historical typesetting conventions*, *fonts* etc.
- Making a powerful OCR web service available for research in order to enable digitization in multiple locations across Europe.
- Suggestions:
  - DARIAH hosting an "OCR4all" instance on a sufficiently-powerful server, for example.
  - *Plus* support for creating gold standard transcriptions and OCR models for various languages, as well as a model library gathering existing models, and offering training to use the OCR pipeline.

Infrastructural challenges in COST Action Distant Reading for European Literary History:

Linguistic annotation pipelines for multiple languages, with multiple layers of annotation (tokenization, lemmatization, POS-tagging, Named Entity Recognition as a minimum), using a shared minimal tagset (such as UniversalPOS for POS).

Currently:

- CLARIN Weblicht service as a guiding light?

! More work needs to be done to increase coverage in terms of languages and annotation layers beyond English or German

Infrastructural suggestions from COST Action Distant Reading for European Literary History:

Project management infrastructure (i.e. toolset or package 'one-stop-shop') for managing projects offered by DARIAH-EU?

- E. g. Wordpress for a website, Mediawiki for project management and documentation, Mailman mailing lists for communication.

DARIAH partnering up with an initiative like Framasoft in France?

- to help them further develop, maintain, translate and run their services (which offer many alternatives to Google Services, like: Etherpad, Polls, Forms, Mailing lists, etc.)

Blue-sky thinking from COST Action Distant Reading for European Literary History:

- An alternative to Github.com for collaborative code and data development

  I.e., a large, open Github-like service free for research and innovation across Europe, with private and public repositories, and with a connection to a long-term archiving solution such as Zenodo (i.e. a version of Github.com+Zenodo= one large research-oriented "gitlab.eu" instance to replace Github.com?)

- More open source tools / services
- More open access to (meta)data
- Open Access to research publications
- Fostering Open Science in general.

Distant Reading

Sources and individuals cited and consulted:

- COST Action *Distant Reading for European Literary History*
  - Professor Christoph Schöch, Trier University, Germany
  - Dr Carolin Odebrecht, Humboldt/Würzburg University, Germany
  - Dr Mike Kestemot, Antwerp University, Belgium
  - Memorandum of Understanding: https://e-services.cost.eu/files/domain_files/CA/Action_CA16204/mou/CA16204-e.pdf
  - GitHub Organisation Distant Reading: https://github.com/distantreading
  - COST Action 16204 Distant Reading: http://www.cost.eu/COST_Actions/ca/CA16204
  - Cost Action Distant Reading: https://www.distant-reading.net/ (work in progress)
- Croatian Bureau of Statistics, https://www.dzs.hr/Hrv_Eng/publication/2018/08-02-02_01_2018.htm
- DABAR repozitoriji, https://dabar.srce.hr/repozitoriji
- DARIAH-HR
  - Dr Koraljka Kuzman Šlogar, Institute for Ethnology and Folklore Studies
- European Open Science Cloud https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud