Onboard onto DraCor!

Prototyping Workflows to Homogenize Drama Corpora

Texts with no markup
(e.g. TXTs from OCRed PDFs)

Matrosen.
Wie? Kommt ihr denn nicht selbst an Bord?

Texts with basic markup (e.g. HTMLs)

<i>*Matrosen.*<i>
Wie? Kommt ihr denn nicht selbst an Bord?

Texts with advanced markup (e.g. XMLs)

```
<sp who="#matrosen">
  <speaker>Matrosen.
  </speaker>
    <l>Wie? Kommt ihr denn
    nicht selbst an Bord?</l>
</sp>
```

ezdrama
a simple, markdown-like language for conversion to XML

Python scripts
ad-hoc transformation pipelines

XSLT
(or Python scripts)

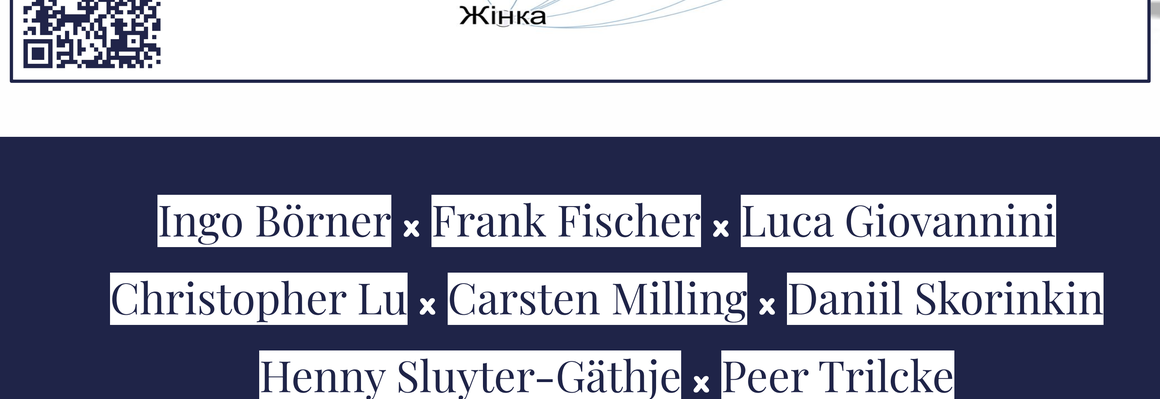
Showcase #1 UDraCor 🇺🇦

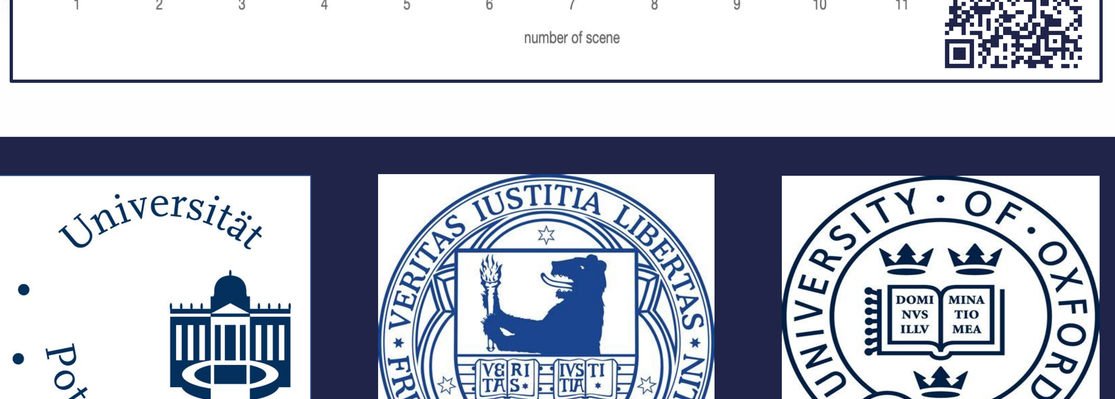
- from a curated list of canonical Ukrainian drama (~150 items)
- plain texts fetched from non-scholarly online libraries
- *ezdrama* markup applied by a group of volunteers
- sneak preview: staging.dracor.org/u

Showcase #2 EPDraCor 🇬🇧

- from the *Early Print Library* (~850 items, earlyprint.org)
- automated pipeline for importing texts and removing linguistic markup
- manual speaker disambiguation through a dedicated web service
- sneak preview: staging.dracor.org/ep

DraCor-ready TEI-XML





Ingo Börner × Frank Fischer × Luca Giovannini
Christopher Lu × Carsten Milling × Daniil Skorinkin
Henny Sluyter-Gäthje × Peer Trilcke

Project homepage: dracor.org 🦜