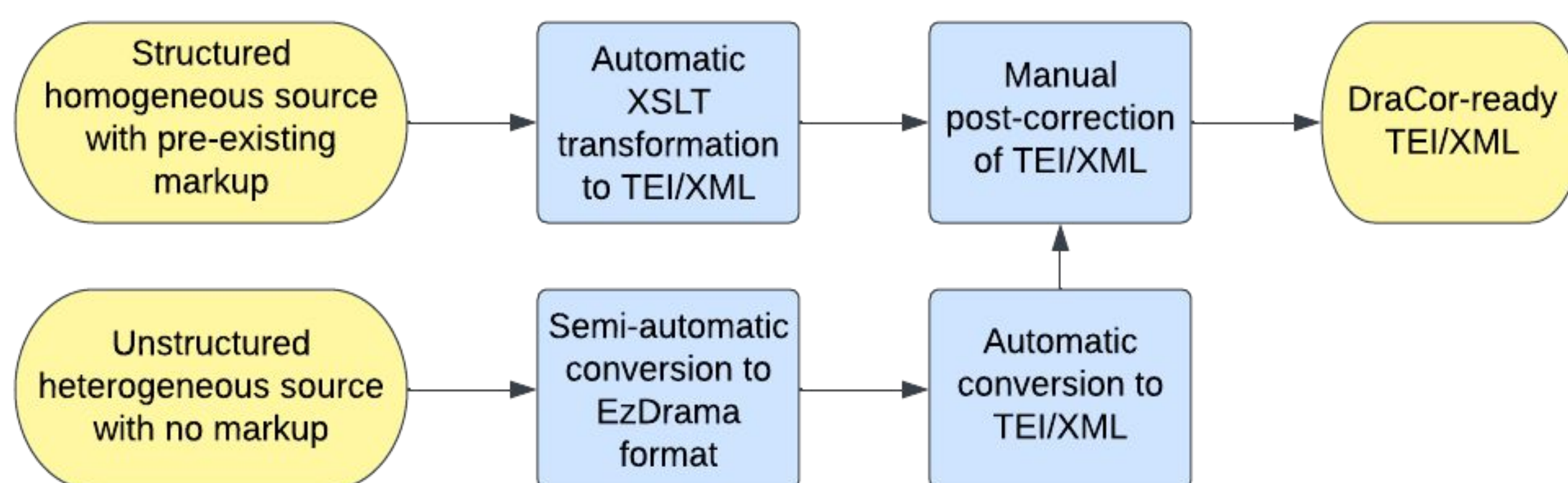# Distributed Corpus Building in Literary Studies: The DraCor Example

- ❖ **DraCor**: an open platform and a network of resources for hosting, accessing, and analysing theatre plays.
- ❖ 15 "programmable corpora" in 10 languages, about **3000 works**, all encoded **in TEI-XML**
- ❖ Ships with tools for quantitative analysis, API endpoints, Linked Open Data and much more!
- ❖ Further info: Fischer et al. 2017

Pipelines for **corpus-building** depend on the original markup of the sources:

- Structured homogeneous source with pre-existing markup → Automatic XSLT transformation to TEI/XML → Manual post-correction of TEI/XML → DraCor-ready TEI/XML
- Unstructured heterogeneous source with no markup → Semi-automatic conversion to EzDrama format → Automatic conversion to TEI/XML → Manual post-correction of TEI/XML

Map labels: TatDraCor, BashDraCor, SweDraCor, RusDraCor, ShakeDraCor, EPDraCor, GerShDraCor, UDraCor, GerDraCor, CzeDraCor, AlsDraCor, HunDraCor, FreDraCor, ItaDraCor, CalDraCor, RomDraCor, GreekDraCor, SpaDraCor

EPDraCor
- from the Early Print Project
- XSLT transformations
- metadata enhancement
- speaker disambiguation via the **who-is-#who** tool

CzeDraCor
- pilot project: 10 plays by Josef and Karel Čapek
- CLS INFRA research fellowship
- processing with **ezdrama** parser

UDraCor
- from various online sources
- volunteer-based encoding
- processing with **ezdrama** parser

More corpora (🇵🇱 🇳🇱 🇮🇱) are underway… and we're **open** for further **collaborations**!

Luca Giovannini [1,5] ○ Daniil Skorinkin [1] ○ Peer Trilcke [1]
Ingo Börner [1] ○ Frank Fischer [2] ○ Julia Dudar [3]
Carsten Milling [1] ○ Petr Pořízka [4]

[1] University of Potsdam ○ [2] Free University of Berlin ○ [3] University of Trier ○ [4] Palacký University Olomouc ○ [5] University of Padova

dracor.org