Towards Metadata-enriched Literary Corpora in Line with FAIR Principles

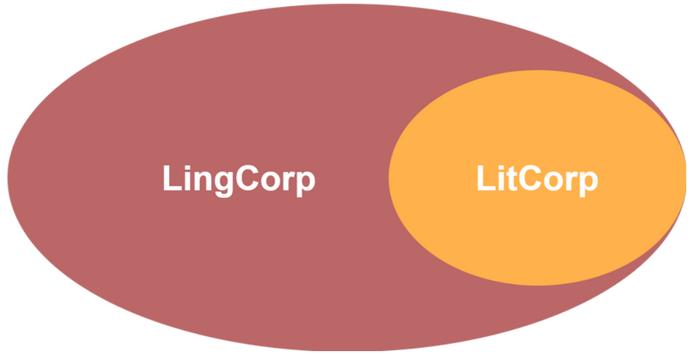
19/20MetaPNC

source: Charles d'Orbigny Dictionnaire Universel d’Histoire Naturelle

https://polona.pl/item/historja-naturalna-ptakow,OTI4OTUyNDg/

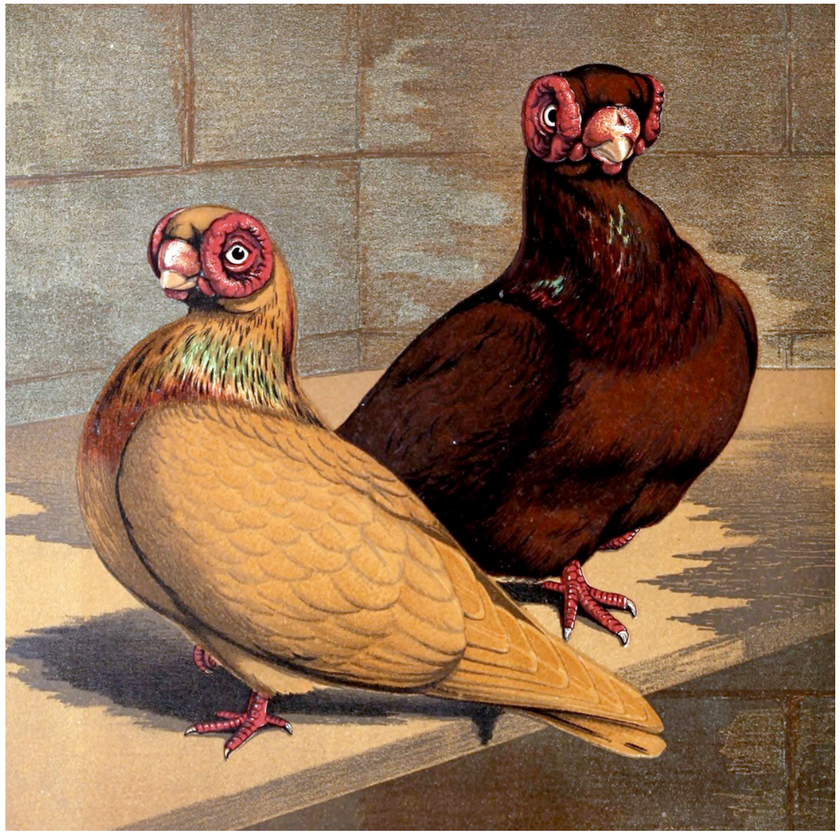
Towards Metadata-enriched Literary Corpora in Line with FAIR Principles

Literary corpora and/vs. linguistic corpora



Discipline rigours:

- material selection (literary genres)
- metadata categories specific to literary research (e.g. place of publication, author's place of origin, number of issues, literary genre)



source: Fulton, The Illustrated Book of Pigeons, 1876

Cf. How do you Compose your Literary Corpus or Literary Collection? Survey launched by Text+ Consortium, SPP-CLS and CLS INFRA

source: Borowski, Gemeinnüzzige Naturgeschichte des Thierreichs, 1780

Challanges of building a literary corpus from scratch

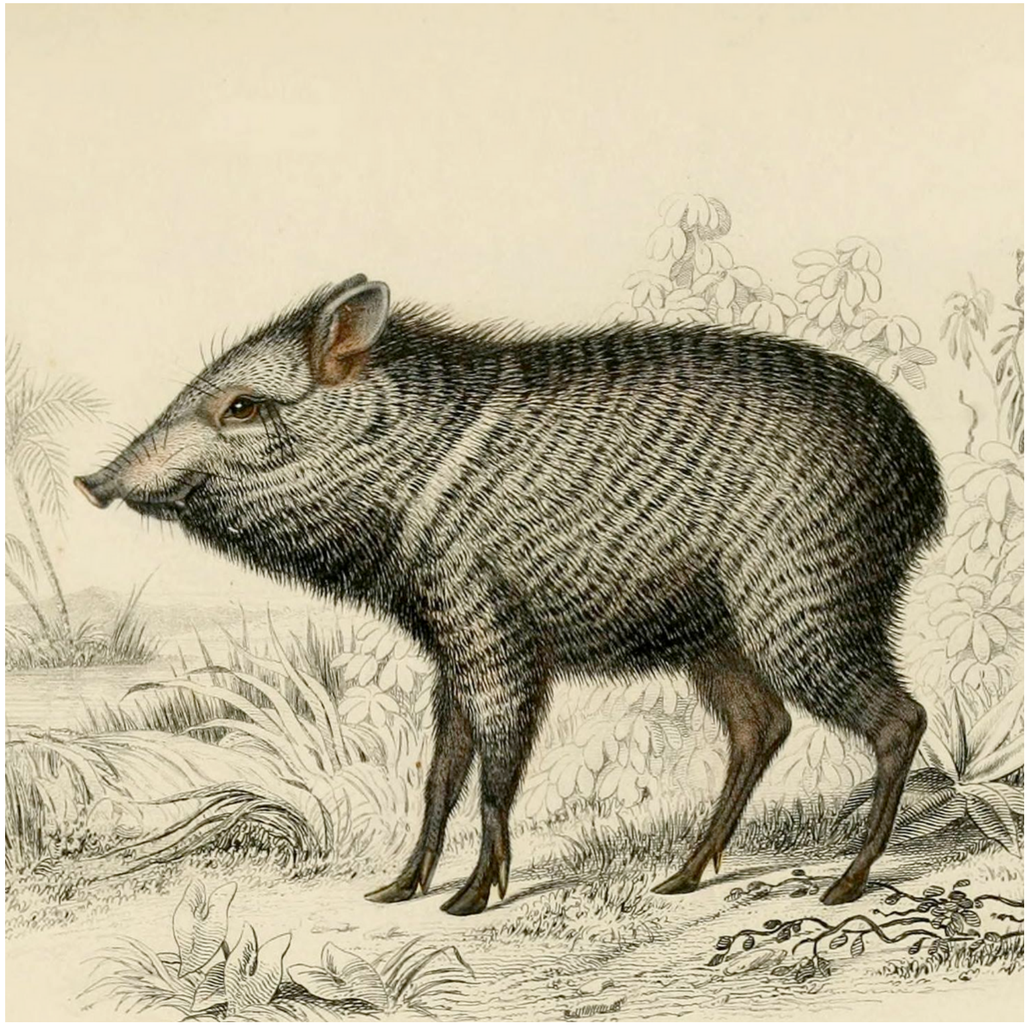
- assessing the representativeness of the data
- varying formats and text quality
- need to use catalogues vs. inability to compare data with other sources
- missing metadata, requiring manual searching
- limited author-related metadata (e.g. no socio-economic metadata)

19/20MetaPNC 1.0

Metadata-enriched Polish Novel Corpus from the 19th and 20th centuries

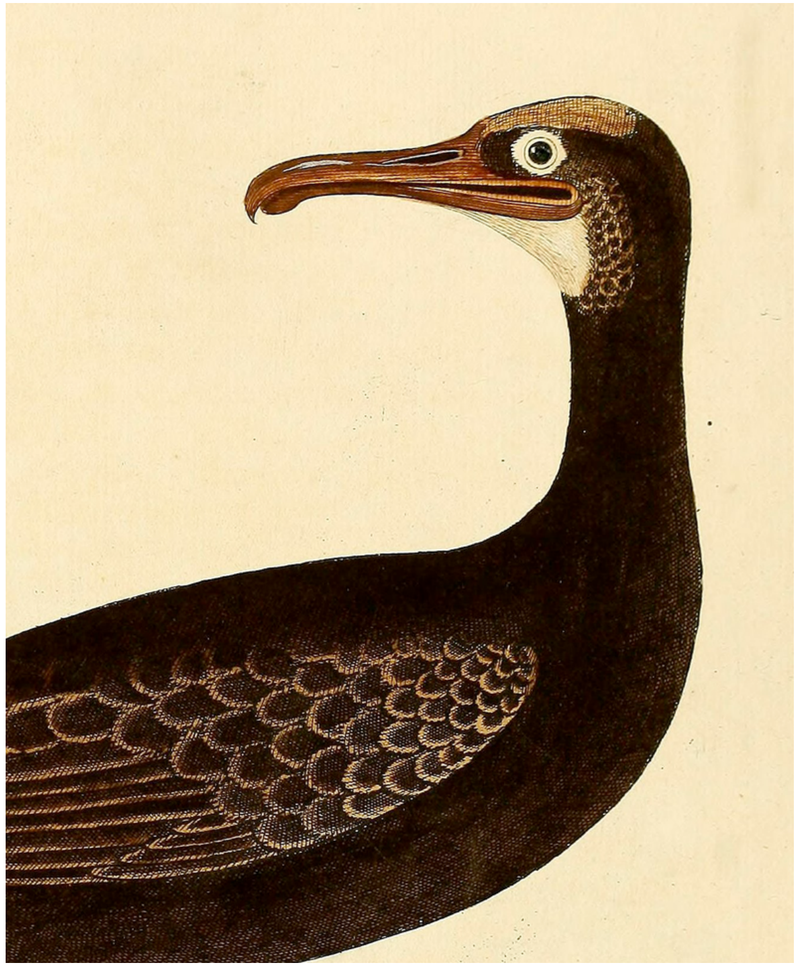
Research goal: to trace the impact of historical and spatial factors on the dynamics of literary processes

Case study: the transformation of urban/rural dichotomy in Polish fiction from 1864 to 1939



source: Charles d'Orbigny Dictionnaire Universel d’Histoire Naturelle

Corpus design



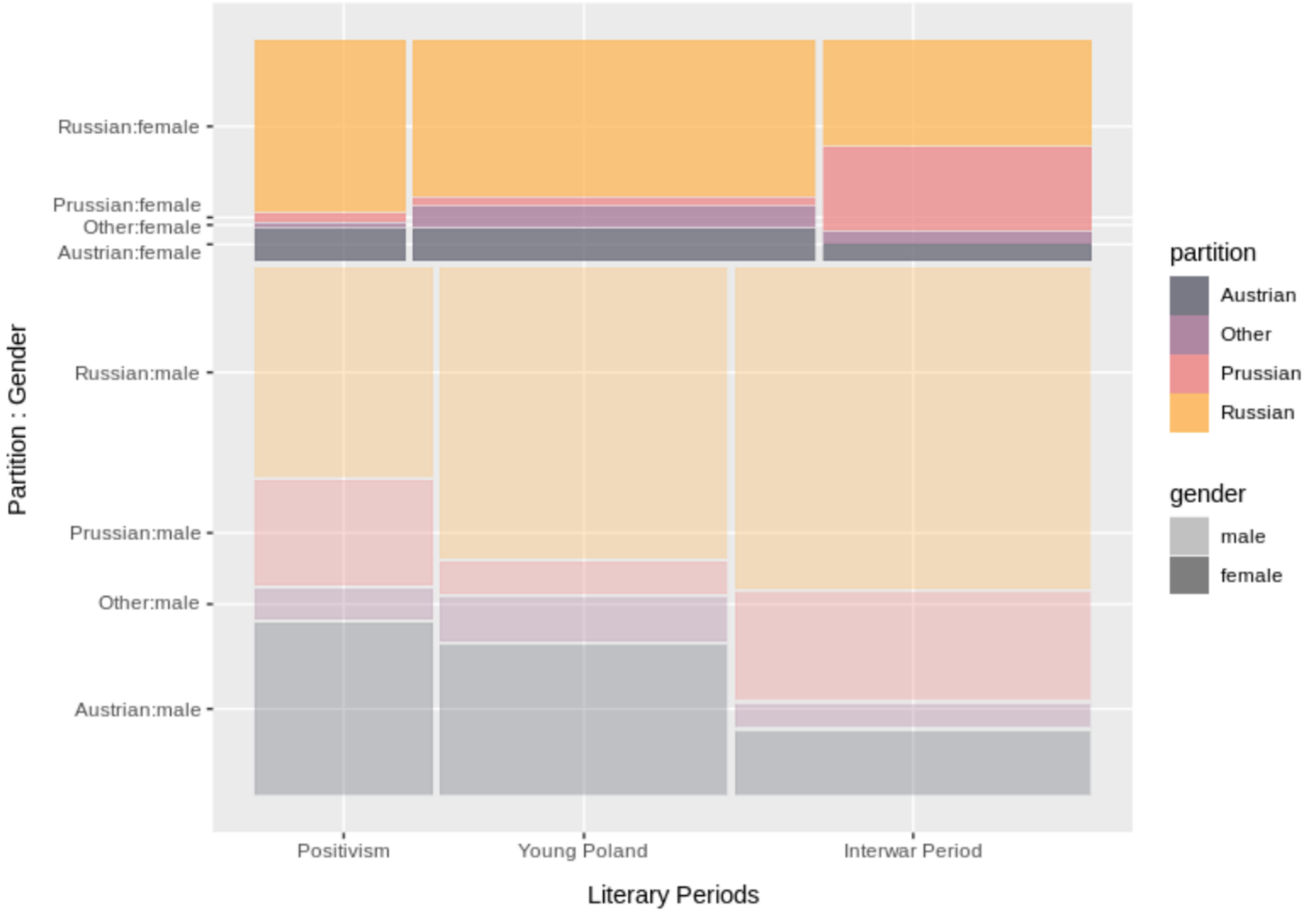
source: Albin, A natural history of birds, 1734

Selection criteria: novels originally written in Polish and first published as books between 1864 and 1939 with the time of the plot later than 1815

Balancing criteria:

- Date: three literary eras distinguished in Polish literary studies determined by the date of first publication (>= 20% each)
  - Positivism (1864–1890)
  - Young Poland (1890–1918)
  - the Interwar Period (1918–1939)
- Gender: female author 10%–50%
- Place of publication (three partitions): >=15% each
- Level of reception:
  - no more than 2 reprints >= 30%
  - more than 2 reprints >= 30%

Final result: 1,000 novels

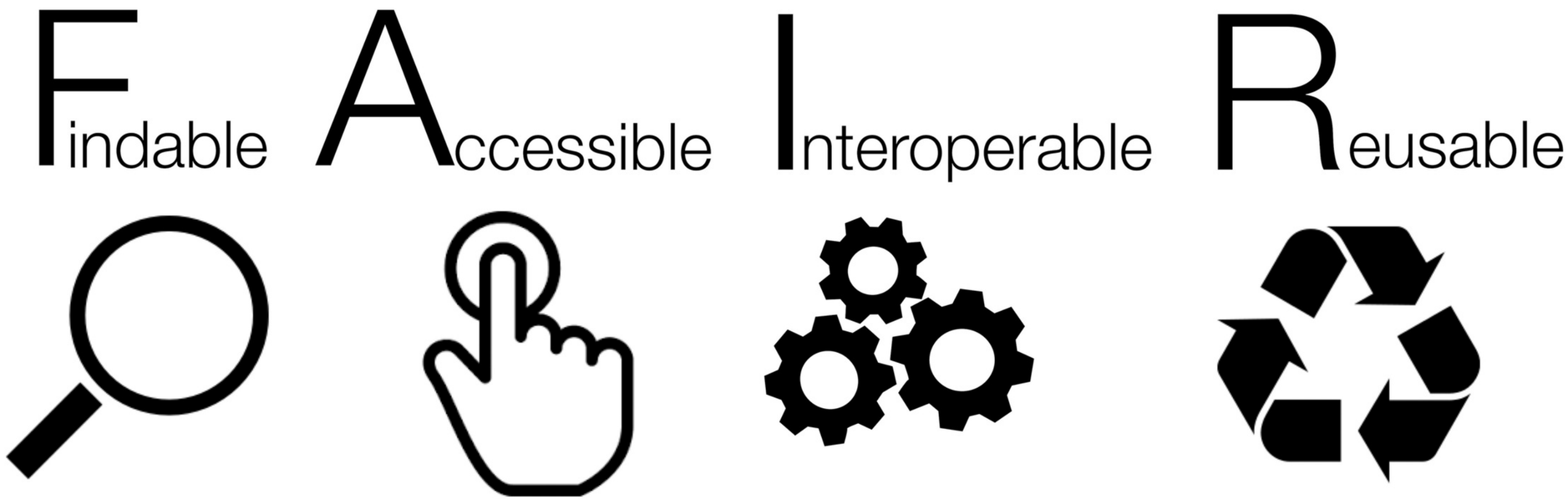


https://polona.pl/item/historja-naturalna-ptakow,OTI4OTUyNDg/

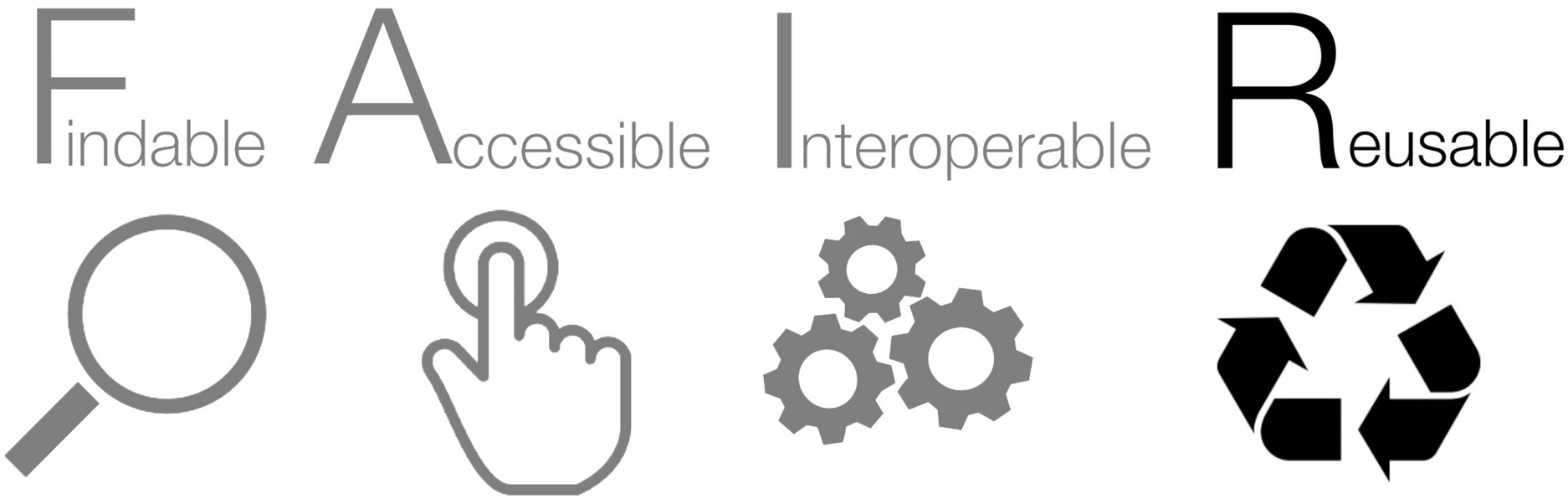
Towards
Metadata enriched
Literary Corpora
in Line with

FAIR Principles

Corpus FAIRification



Corpus FAIRification



Linked Data standards

5-star Linked Data

★	Make data available on the Web in whatever format.
★★	Make data available as structured data (e.g., Excel instead of an image scan of a table).
★★★	Use non-proprietary formats (e.g., CSV instead of Excel format).
★★★★	Use URIs to denote things, so that people can point at your data.
★★★★★	Link your data to other data to provide context.

7-star Linked Data Service

★★★★★★	Provide your data with a schema and documentation so that people can *understand and re-use* your data easily.
★★★★★★★	Validate your data and denote its provenance so that people can *trust the quality* of your data.

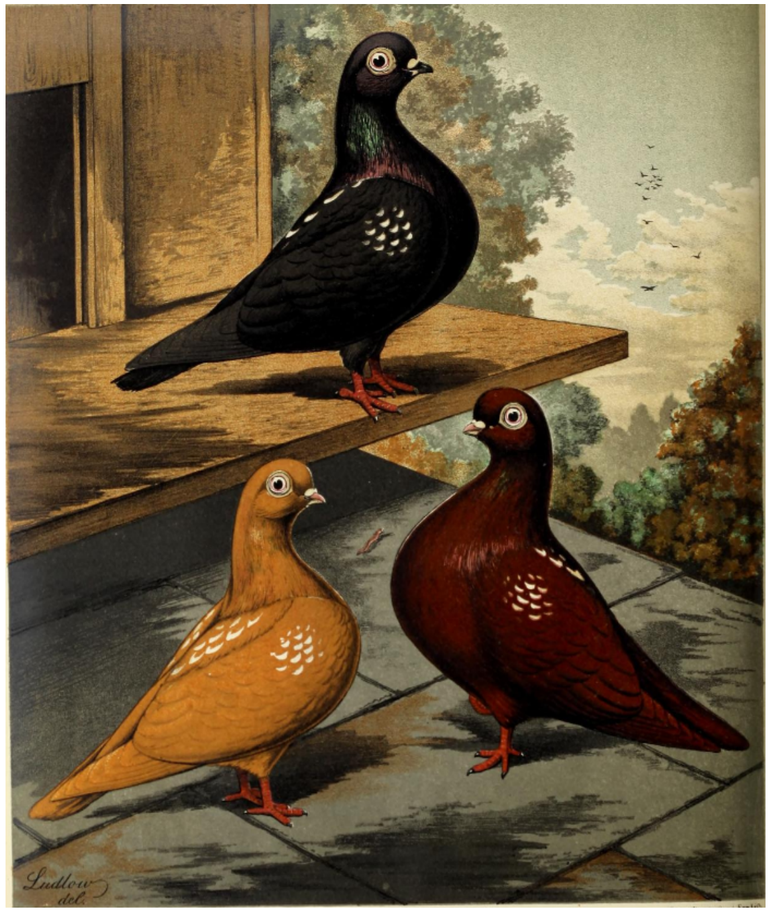
Source: Linked Data Finland. Living Laboratory Data Service for the Semantic Web (www.ldf.fi/)

https://polona.pl/item/historja-naturalna-ptakow,OTI4OTUyNDg/

Towards Metadata-enriched Literary Corpora in Line with FA R Princip es

Challanges of metadata description

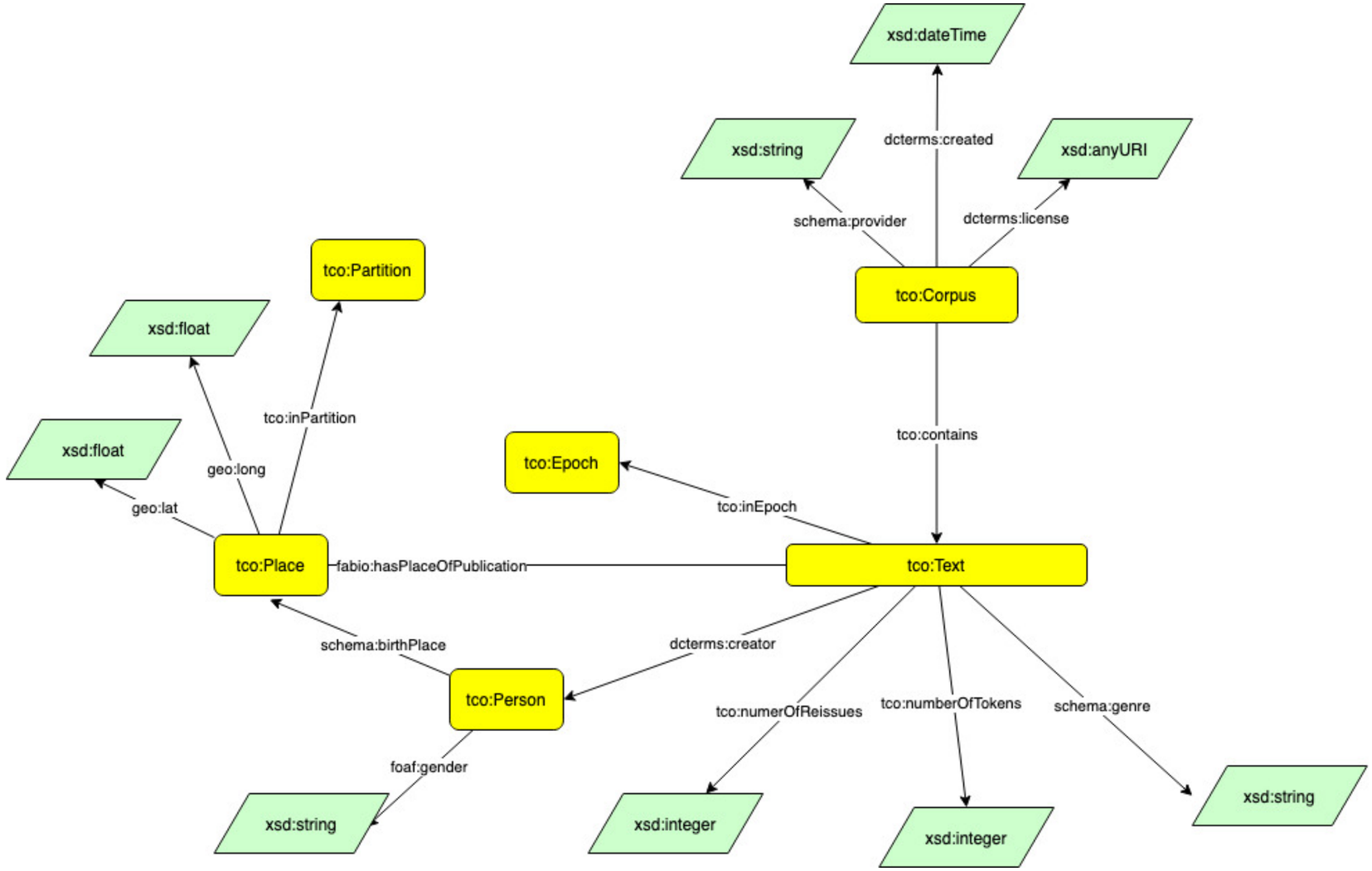
- traditional catalogue metadata vs. metadata for research purposes
- insufficient metadata formats
- how to make research-specific metadata easily comprehensible?



source: Fulton, The Illustrated Book of Pigeons, 1876

TCO: Text Corpora Ontology

(work in progress)



Related works:

- Europeana Data Model Primer (2013), Europeana. https://pro.europeana.eu/page/edm-documentation
- Jett, J., Cole, T. W., Maden, C., & Downie, J. S. (2016). The HathiTrust Research Center Workset Ontology: A Descriptive Framework for Non-Consumptive Research Collections. Journal of Open Humanities Data, 2(0)

Metadata enhancement

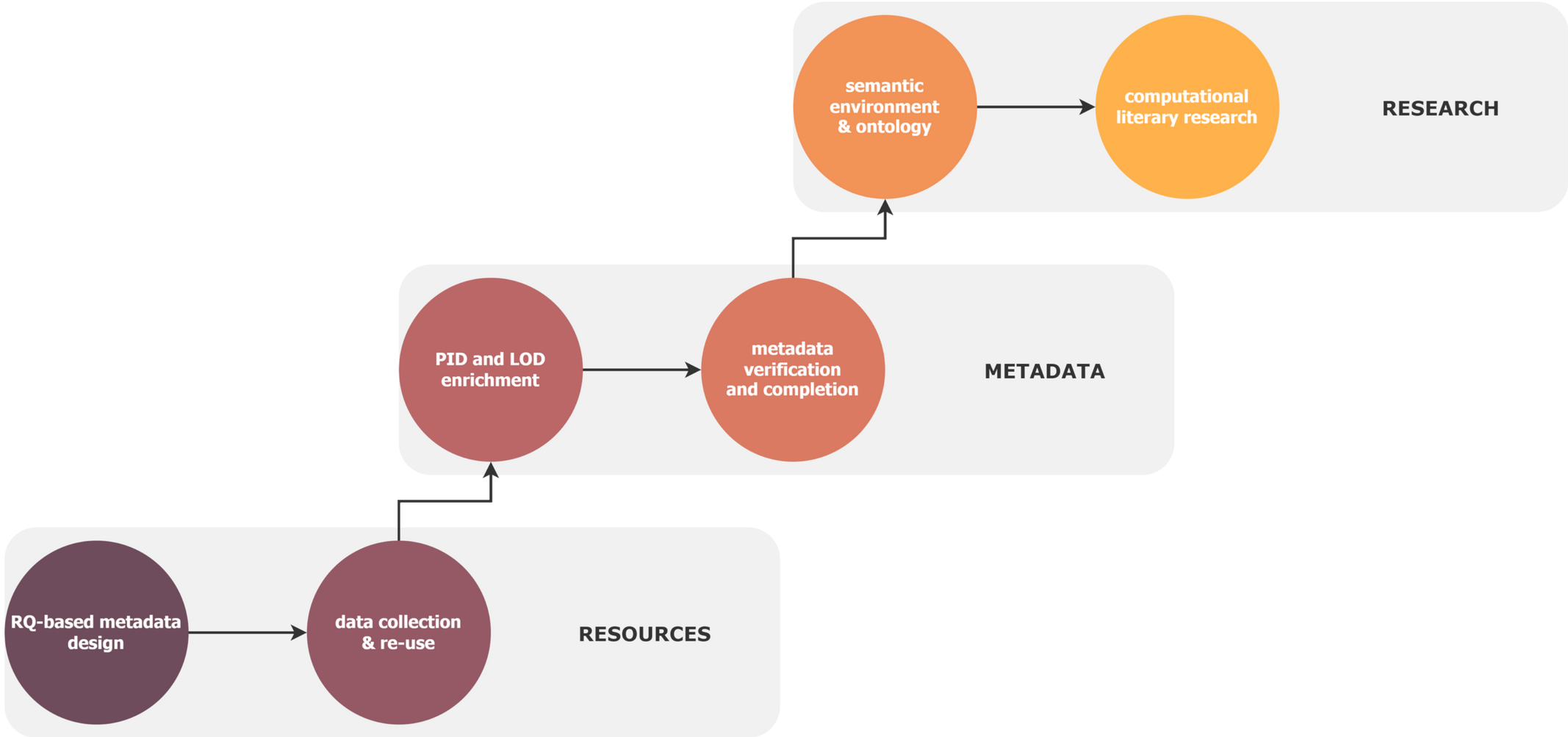
Jezierski, Edmund
– Wyspa Lenina.txt

Metadata enrichment

RDFization

```
tco:metapnc_b_246 a tco:Text,
        dcterms:BibliographicResource ;
    tco:inEpoch tco:metapnc_e_1572 ;
    tco:numberOfReissues 1 ;
    tco:numberOfTokens 76607 ;
    dcterms:creator tco:metapnc_p_1100 ;
    dcterms:date "1925"^^xsd:year ;
    dcterms:subject "Plot after the Congress of Vienna" ;
    dcterms:title "Wyspa Lenina" ;
    fabio:hasPlaceOfPublication tco:metapnc_g_1418 ;
    schema:contentUrl <https://polona2.pl/archive?uid=84911046&cid=87260474> ;
    schema:genre "Novel" ;
    owl:sameAs <http://polona.pl/item/84911046> .
```

The (meta)corpus creation workflow



The workflow in practice

source: Albin, A natural history of birds, 1734

19/20MetaPNC 1.0.1

- publicly available knowledge graph
- corpus documentation
- paper "Towards a contextualised spatial-diachronic history of literature: mapping emotional representations of the city and the country in Polish fiction from 1864 to 1939"

Towards computational literary research

Thank you!

Cezary Rosiński.....cezary.rosinski@ibl.waw.pl

Agnieszka Karlińska.....agnieszka.karlinska@nask.pl

Patryk Hubar.....patryk.hubar@ibl.waw.pl

Marek Kubis.....marek.kubis@amu.edu.pl

Jan Wieczorek.....jan.wieczorek@pwr.edu.pl

Computations

iBL
INSTYTUT BADAŃ LITERACKICH POLSKIEJ AKADEMII NAUK
Institute of Literary Research Polish Academy of Sciences

SCIENCE
NASK

UAM
Adam Mickiewicz University Poznań

Wrocław University of Science and Technology