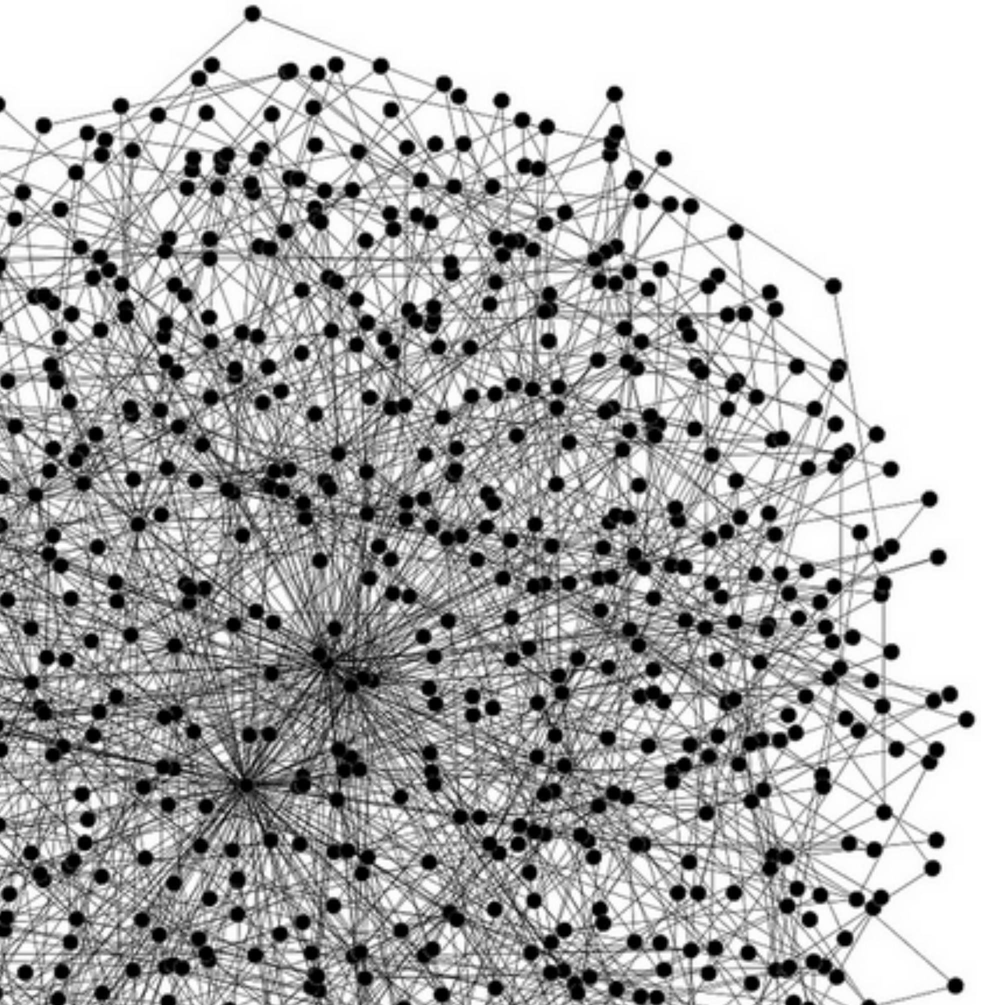
The References of References:

Enriching Library Catalogs via Domain-Specific Reference Mining



Giovanni Colavizza
Matteo Romanello (@mr56k)
Frédéric Kaplan (@frederickaplan)

Goal

Empowering scholars in the Humanities with better IR systems

Motivation - the Scholar

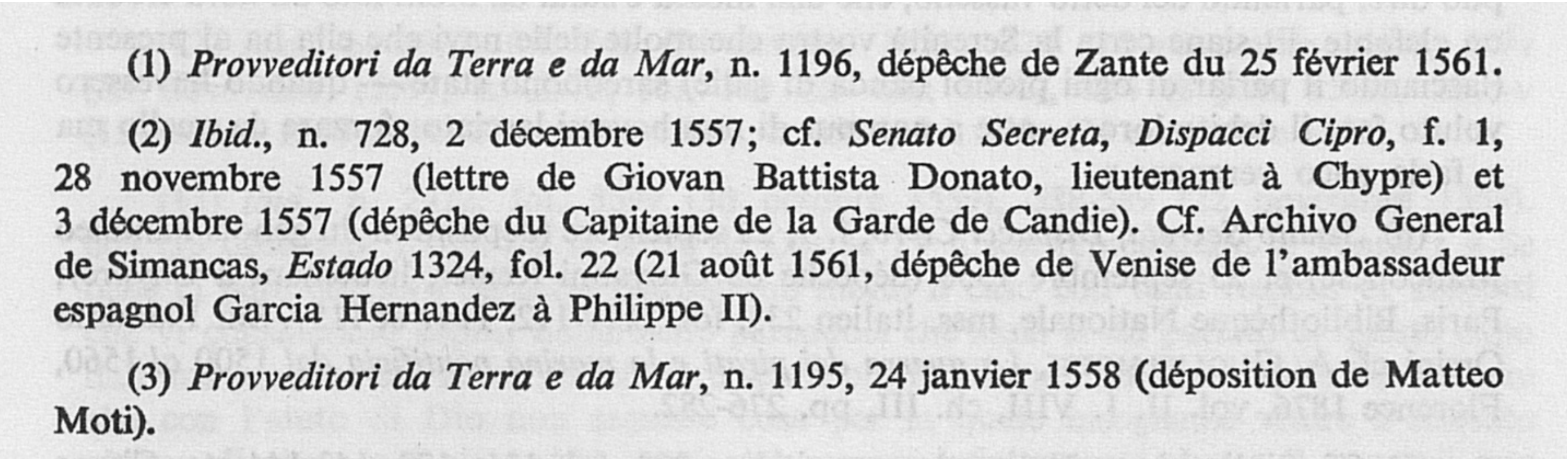
Sciences:
Google Scholar
English
mainly papers
Lower-cost information gathering

Humanities:
no Google Scholar-like system
multiple languages
mainly **monographs**
Higher-cost information gathering

Issues: **lack of data** [Sula and Miller, 2014] leads to **absence of services**: estimated coverage of Web of Science for Humanities circa 13% [Mingers and Leydesdorff, 2015].

Motivation - the Footnote

How humanists cite? **Footnotes** [see e.g. Hellqvist, 2009]



(1) *Provveditori da Terra e da Mar*, n. 1196, dépêche de Zante du 25 février 1561.

(2) *Ibid.*, n. 728, 2 décembre 1557; cf. *Senato Secreta, Dispacci Cipro*, f. 1, 28 novembre 1557 (lettre de Giovan Battista Donato, lieutenant à Chypre) et 3 décembre 1557 (dépêche du Capitaine de la Garde de Candie). Cf. Archivo General de Simancas, *Estado* 1324, fol. 22 (21 août 1561, dépêche de Venise de l'ambassadeur espagnol Garcia Hernandez à Philippe II).

(3) *Provveditori da Terra e da Mar*, n. 1195, 24 janvier 1558 (déposition de Matteo Moti).

Motivation - the Archive

Approximately half citations to **primary sources** [Wiberley Jr., 2009]

Motivation - the Scholar reloaded

Proposal: Enriching library catalogs

Use **reference monographs**, the “canon” of the domain, to **extract references** to the rest of the literature and **enrich library catalogs**.

Project: Linked Books

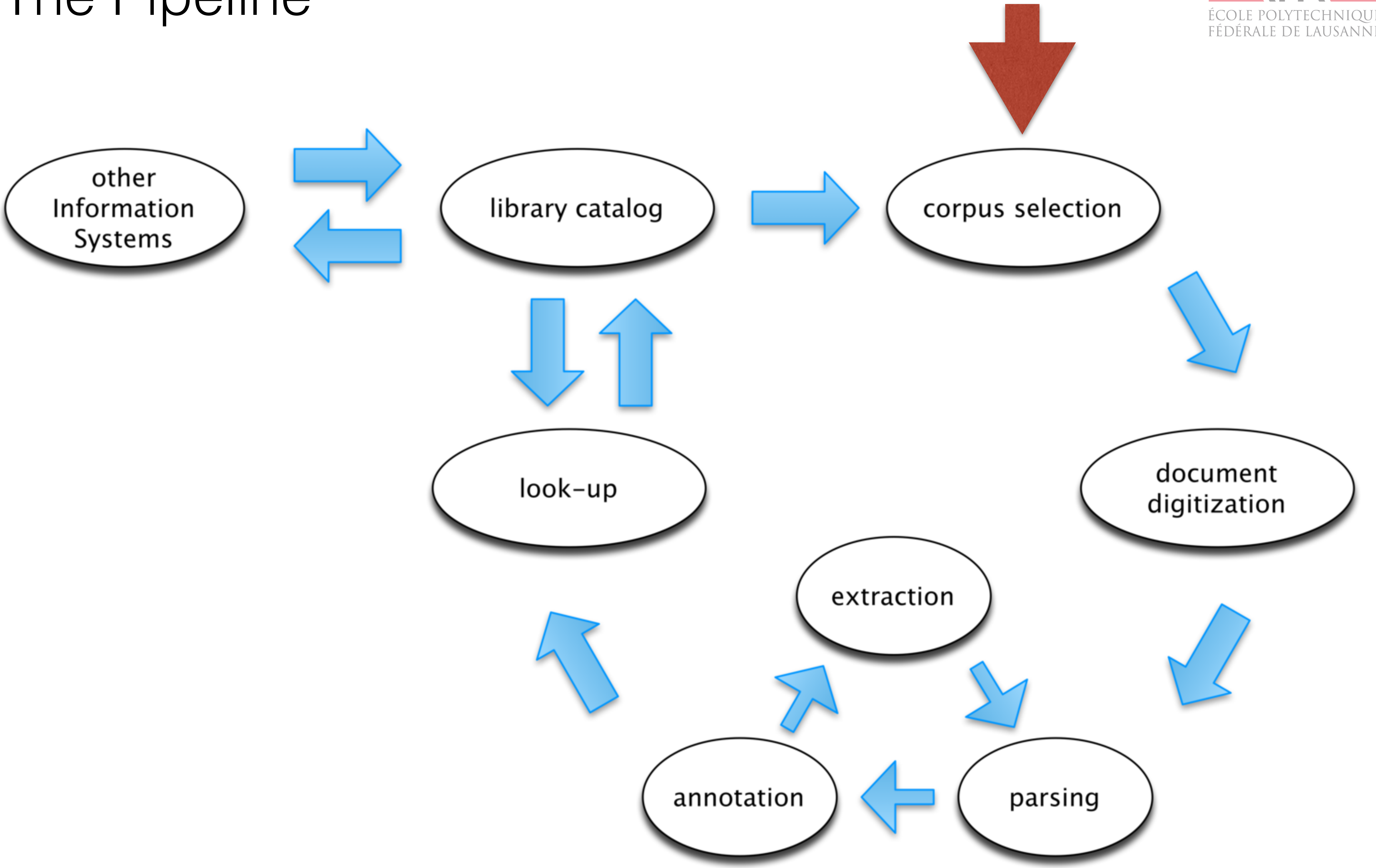
Focused on a case study/domain:

the **history of Venice**.

Partners so far:

- Ca’ Foscari University Library System
- Biblioteca Marciana
- Istituto Veneto di Scienze, Lettere ed Arti
- Archivio di Stato di Venezia
- EPFL

The Pipeline



Corpus selection

Use the means of the library:

1- Consultation shelves

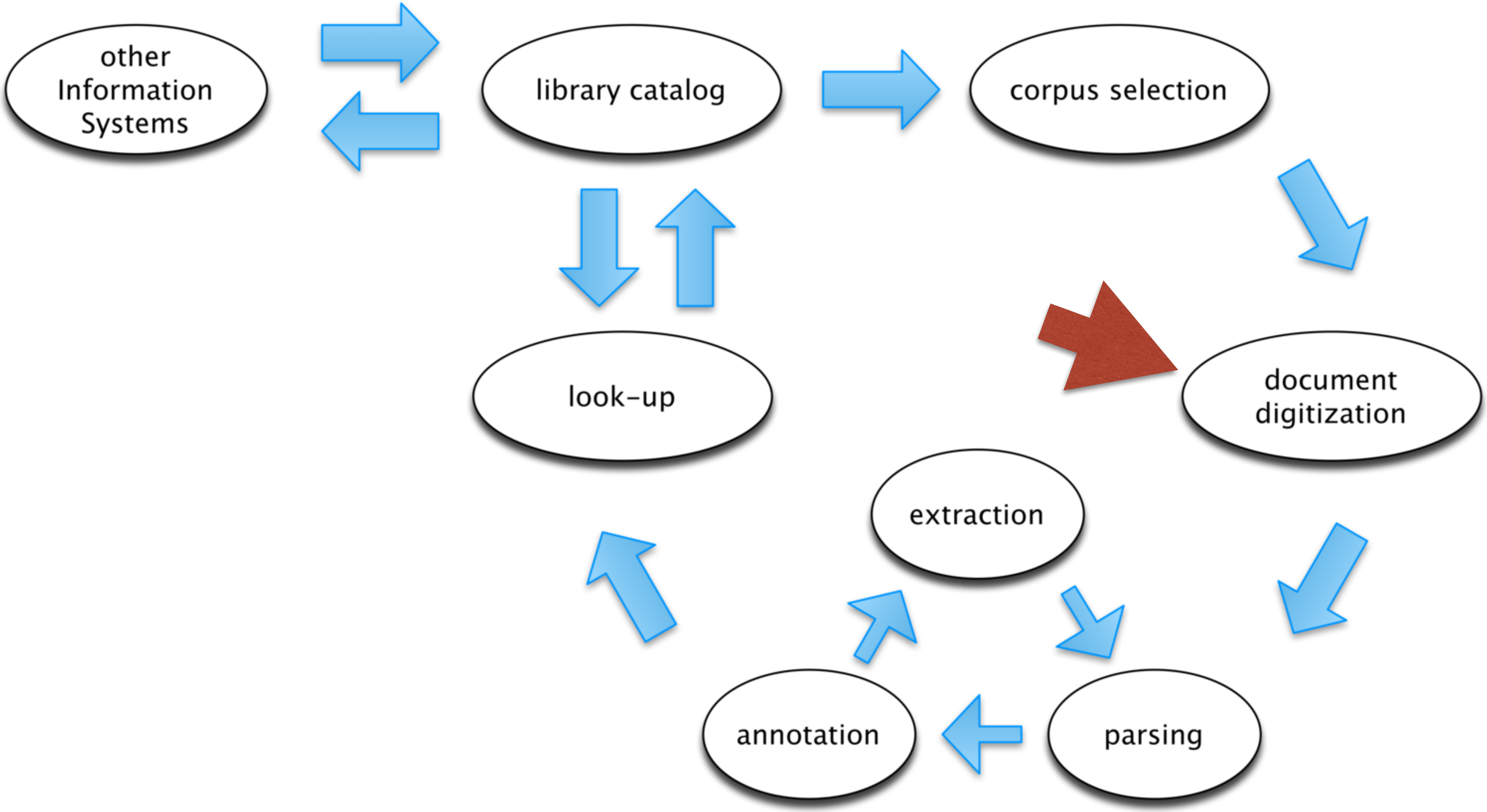
2- Dewey and subject classification

3- Scholarly bibliographies

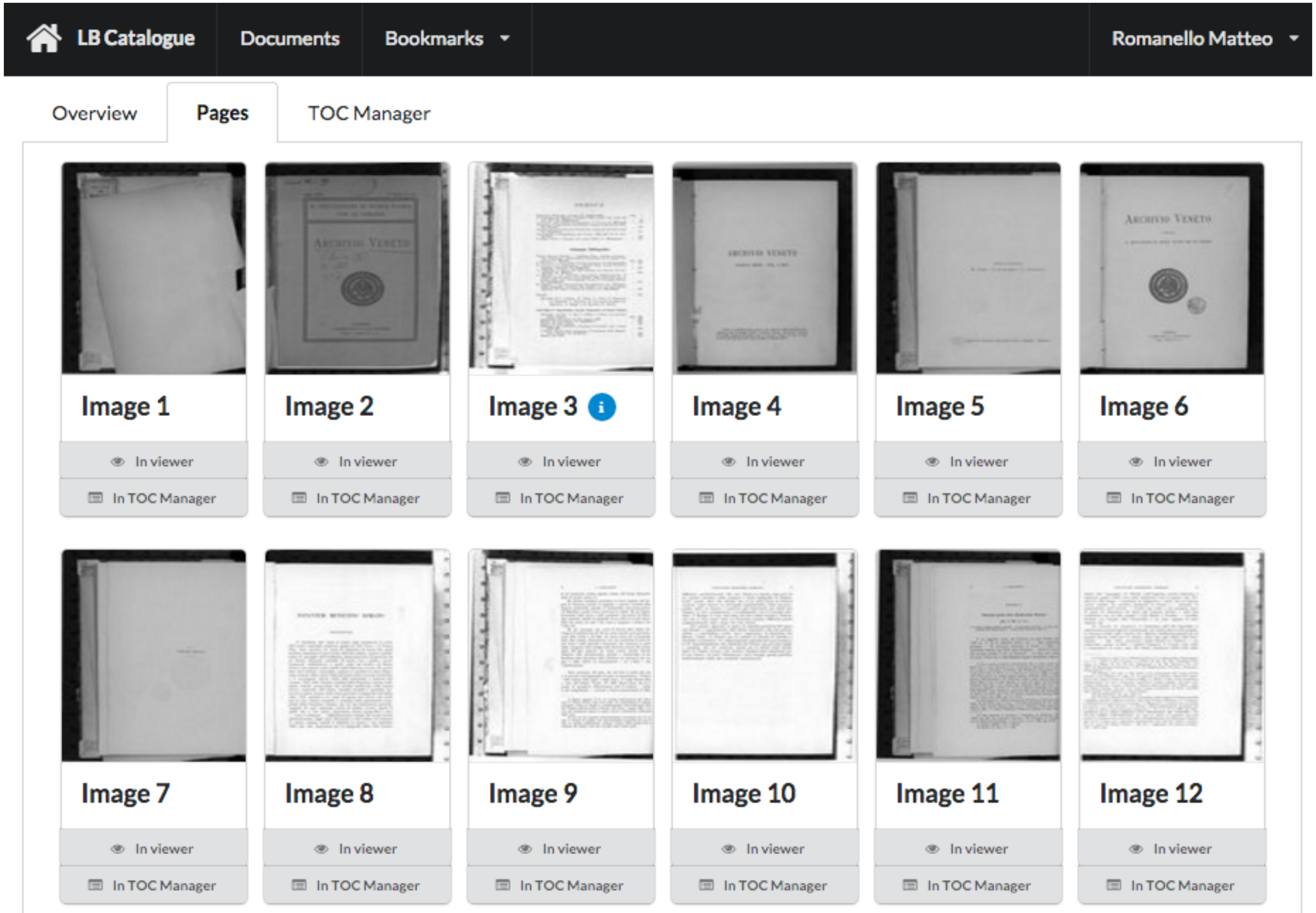
4- Keyword search

Result: 1904 monographs, 701 with a structured list of references.

The Pipeline - Digitization

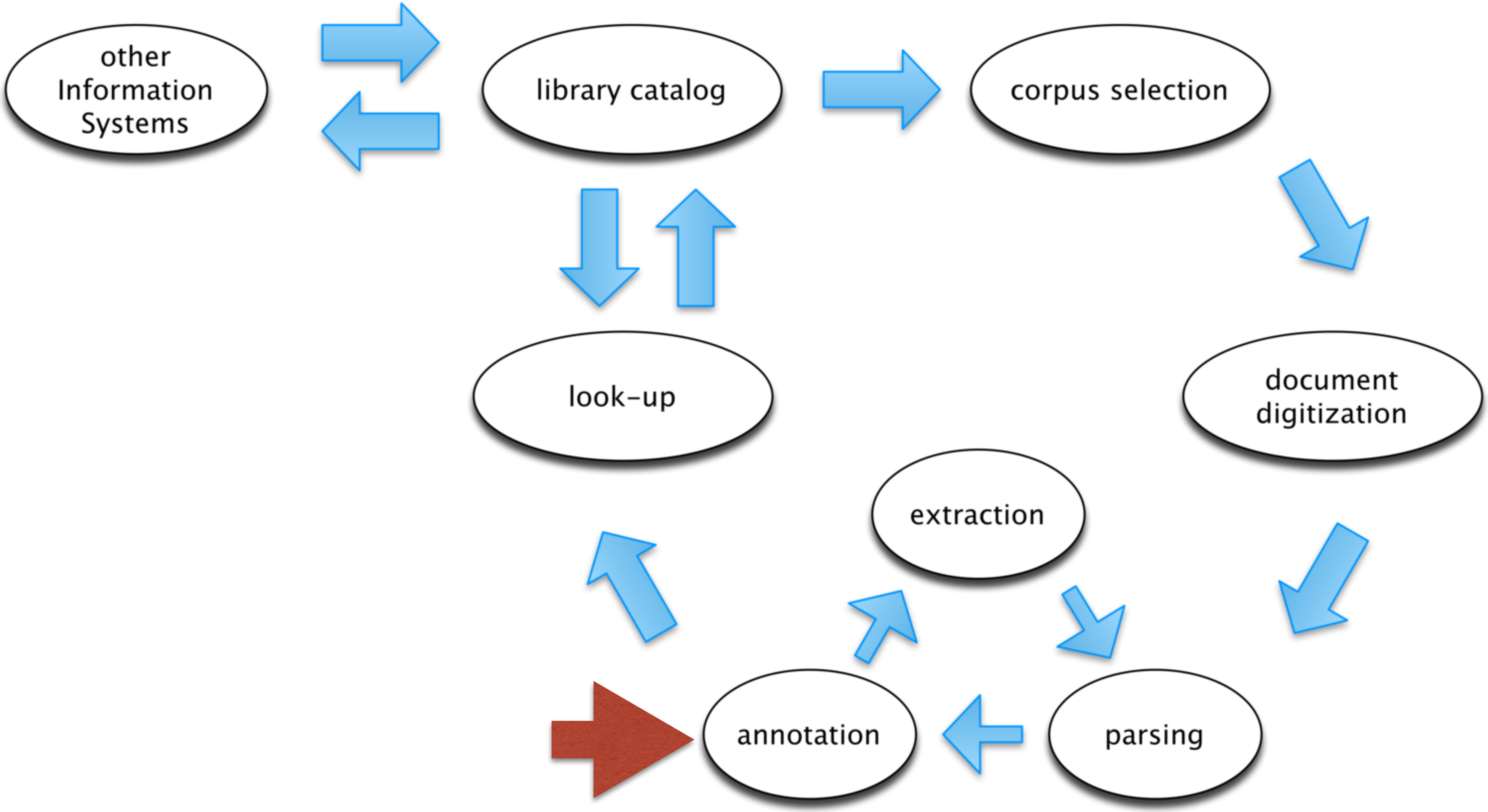


Digitization

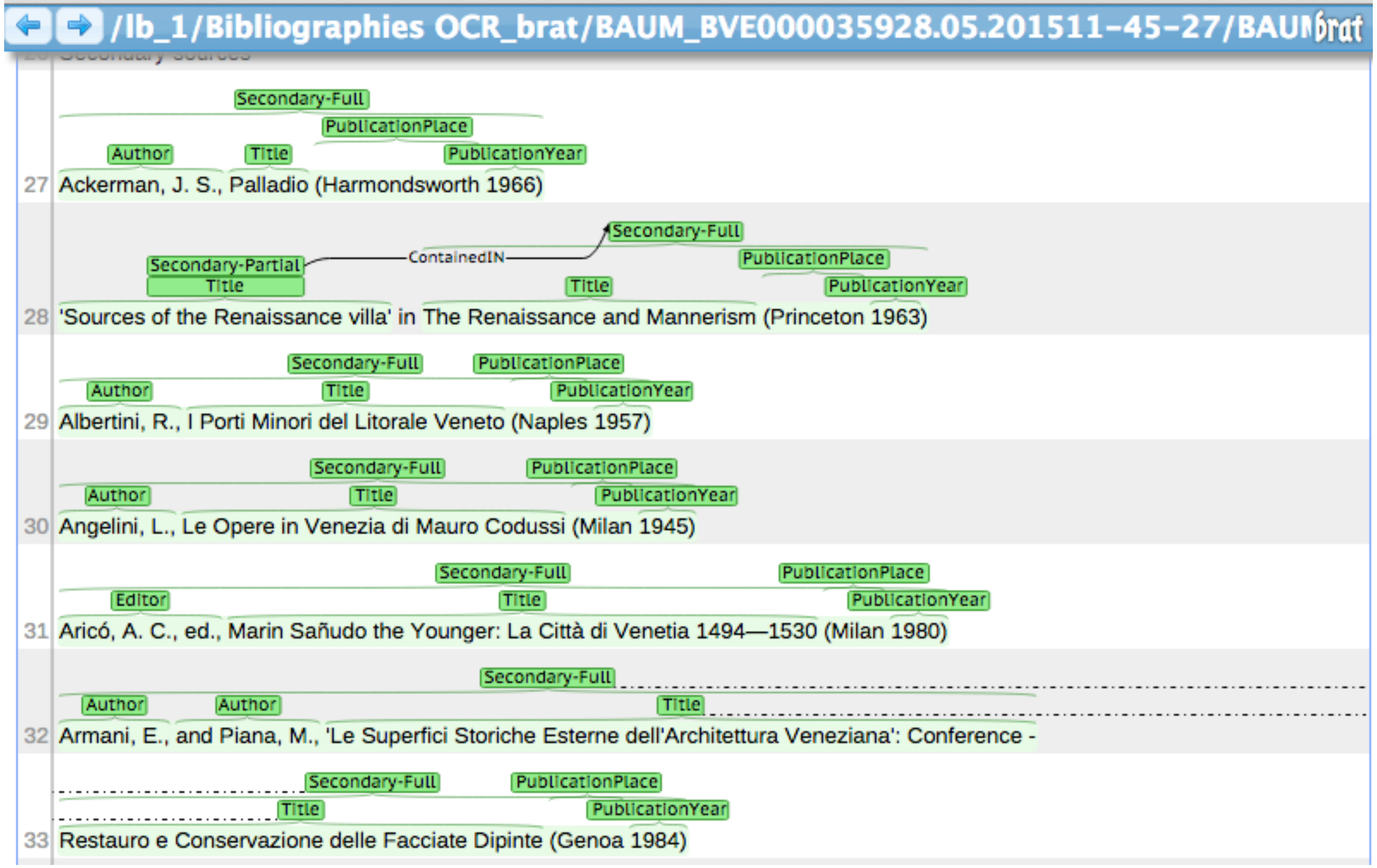


1,904 monographs + ~1,000 journal issues

The Pipeline - Annotation/Extraction/Parsing



Annotation



/lb_1/Bibliographies OCR_brat/BAUM_BVE000035928.05.201511-45-27/BAU

Ackerman, J. S., Palladio (Harmondsworth 1966)

'Sources of the Renaissance villa' in The Renaissance and Mannerism (Princeton 1963)

Albertini, R., I Porti Minori del Litorale Veneto (Naples 1957)

Angelini, L., Le Opere in Venezia di Mauro Codussi (Milan 1945)

Aricó, A. C., ed., Marin Sañudo the Younger: La Città di Venetia 1494—1530 (Milan 1980)

Armani, E., and Piana, M., 'Le Superfici Storiche Esterne dell'Architettura Veneziana': Conference -

Restauro e Conservazione delle Facciate Dipinte (Genoa 1984)

- annotated 27% of 701 monographs (with reference list)
    - 3.8% of all digitized pages (with references)
- annotators identified 33 citation styles, divided into 6 families
    - Yes, humanities scholars love customized reference styles!

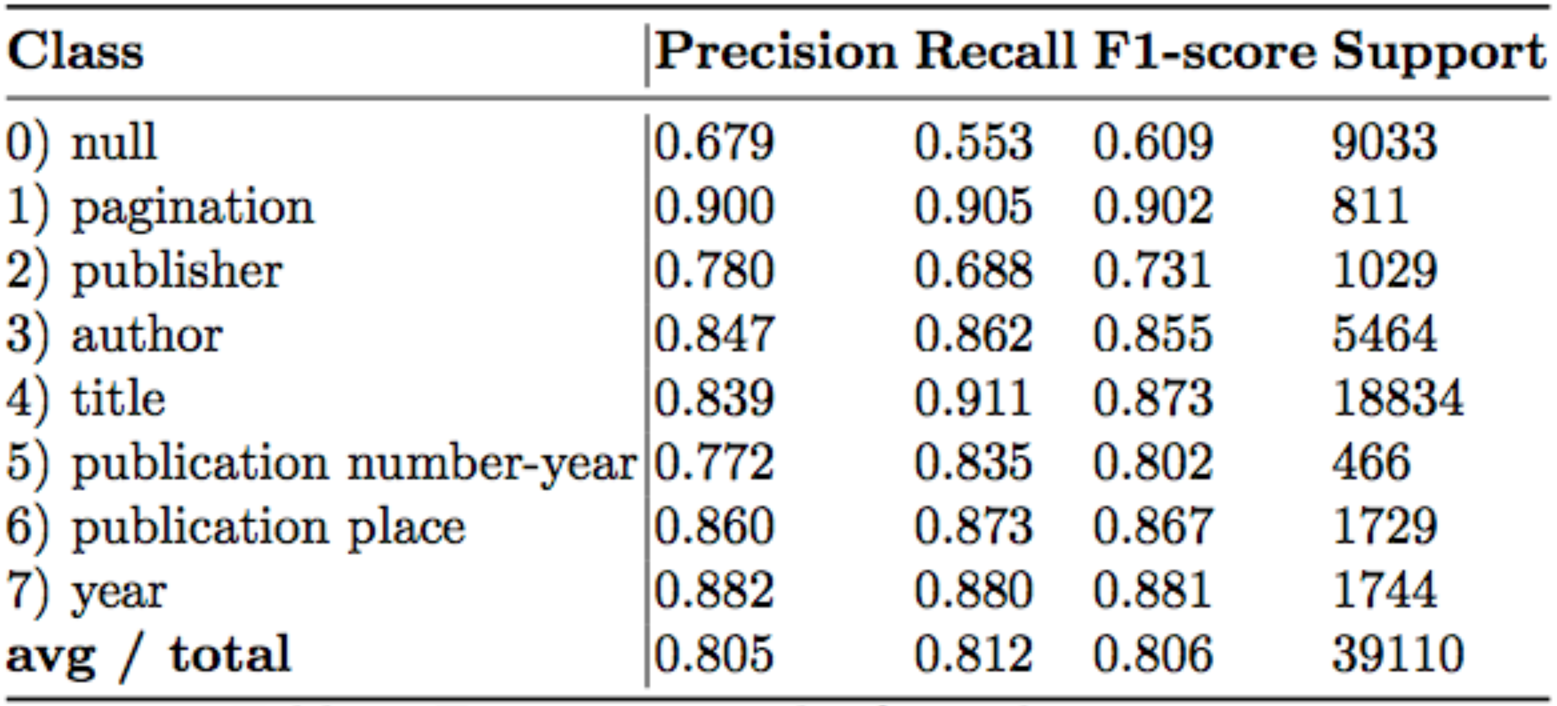
Reference Extraction/Parsing

Klinkhammer Lutz, *L'occupazione tedesca in Italia 1943-1945*, Torino, Bollati Boringhieri 1993.

[Klinkhammer author] [Lutz, author] [L'occupazione title] [tedesca title] [in title] [Italia title] [1943-1945, title] [Torino, publicationplace] [Bollati publisher] [Boringhieri publisher] [1993 publicationyear].

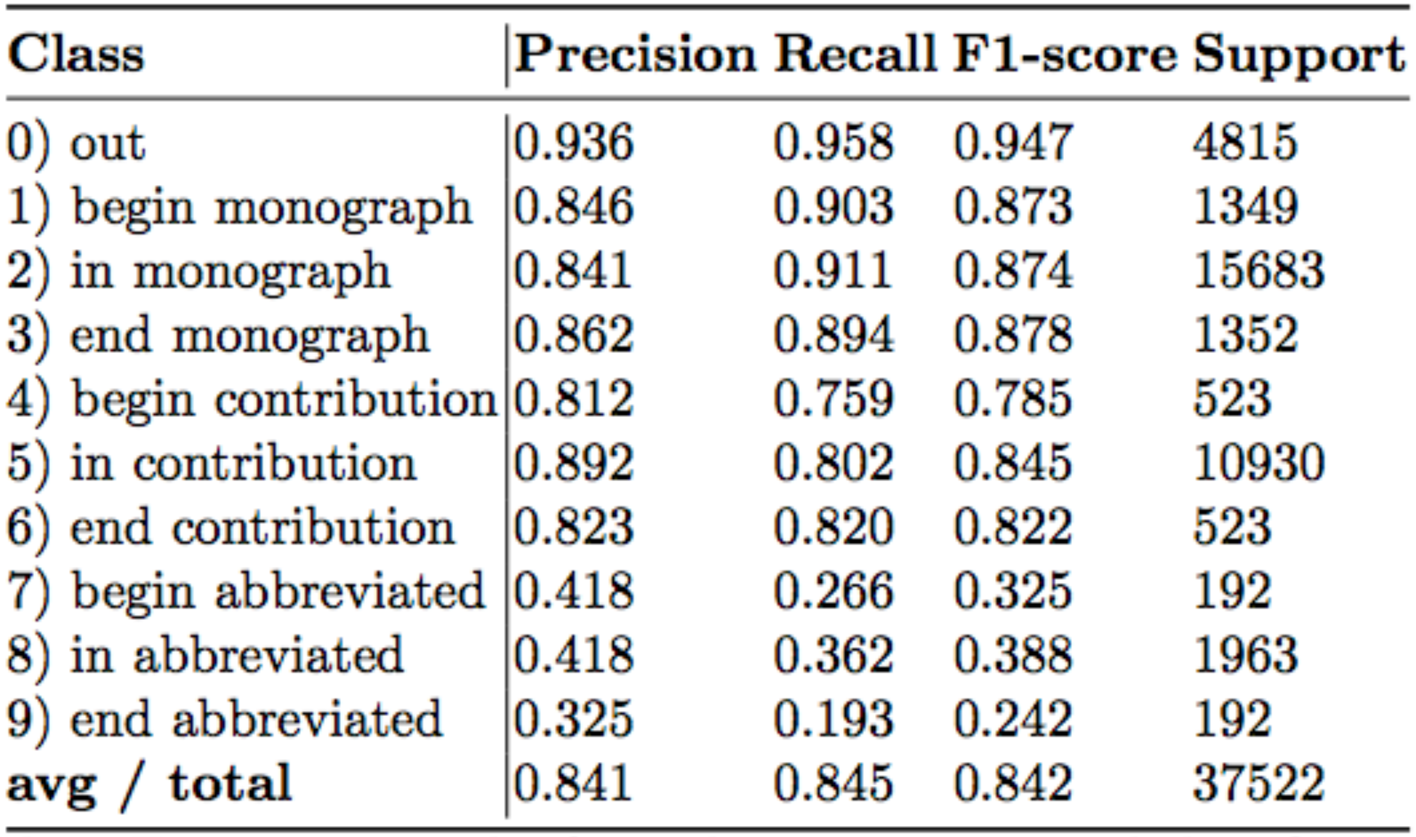
[Klinkhammer b-i-secondary-full] [Lutz, i-secondary-full] [L'occupazione i-secondary-full] [tedesca i-secondary-full] [in i-secondary-full] [Italia i-secondary-full] [1943-1945, i-secondary-full] [Torino, i-secondary-full] [Bollati i-secondary-full] [Boringhieri i-secondary-full] [1993 i-secondary-full].

Extraction/Parsing - Evaluation



Class	Precision	Recall	F1-score	Support
0) null	0.679	0.553	0.609	9033
1) pagination	0.900	0.905	0.902	811
2) publisher	0.780	0.688	0.731	1029
3) author	0.847	0.862	0.855	5464
4) title	0.839	0.911	0.873	18834
5) publication number-year	0.772	0.835	0.802	466
6) publication place	0.860	0.873	0.867	1729
7) year	0.882	0.880	0.881	1744
avg / total	0.805	0.812	0.806	39110

Table 1: Extraction results for task 1: parsing.



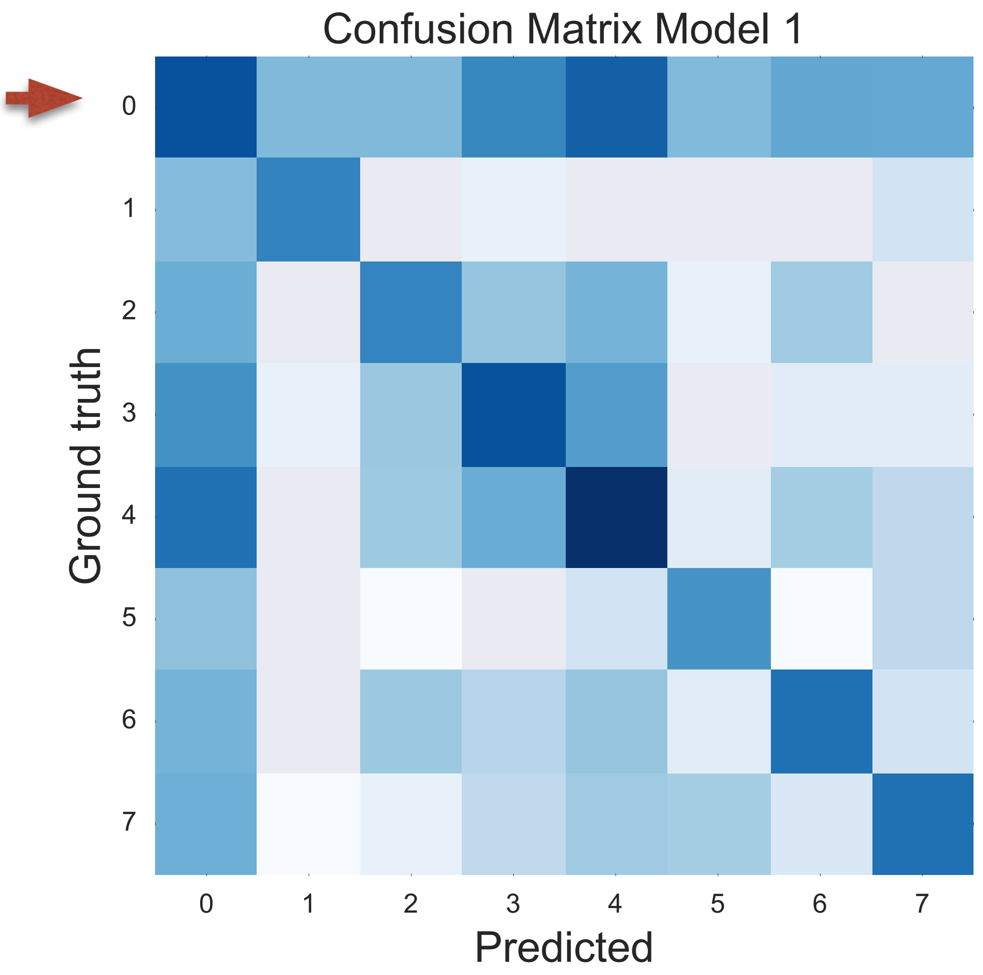
Class	Precision	Recall	F1-score	Support
0) out	0.936	0.958	0.947	4815
1) begin monograph	0.846	0.903	0.873	1349
2) in monograph	0.841	0.911	0.874	15683
3) end monograph	0.862	0.894	0.878	1352
4) begin contribution	0.812	0.759	0.785	523
5) in contribution	0.892	0.802	0.845	10930
6) end contribution	0.823	0.820	0.822	523
7) begin abbreviated	0.418	0.266	0.325	192
8) in abbreviated	0.418	0.362	0.388	1963
9) end abbreviated	0.325	0.193	0.242	192
avg / total	0.841	0.845	0.842	37522

Table 2: Extraction results for task 2: extraction and classification.

Extraction/Parsing - Confusion Matrix

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

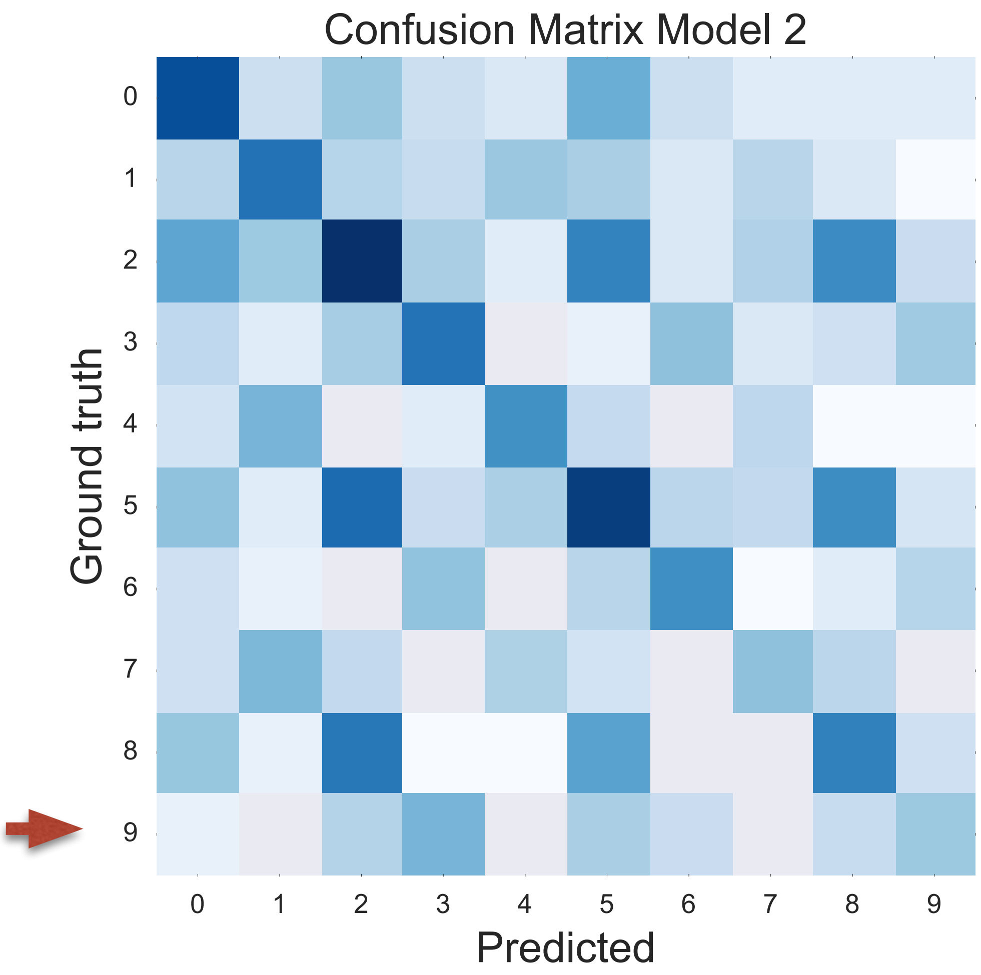
null



author

title

abbrev. (E)



monograph (E)

Task 1

F1 score

(avg) **0.806**

class="null" **0.609**

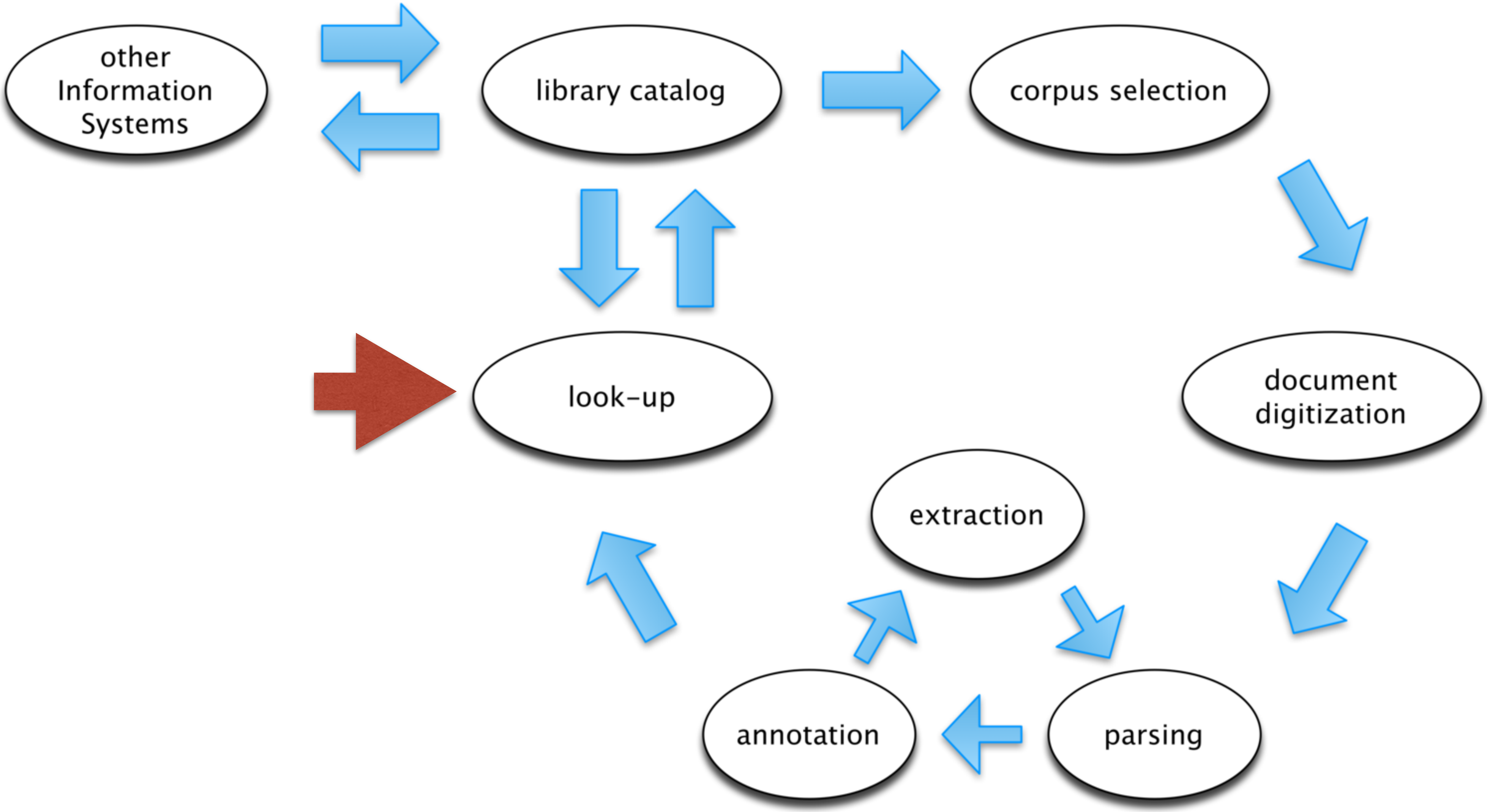
Task 2

F1 score

(avg) **0.842**

class="end abbreviated" **0.242**

The Pipeline - Lookup



Lookup

1. Against OPAC SBN (via API)

Goal: disambiguation of references

Steps:

1. search candidates by title
2. match reference metadata
3. assign each candidate a confidence score
4. return set of candidates

Evaluation:

- 2k references (out of 181k)
- 41.7% no candidates
- 58.3% with candidates:
  - 72.3% -> first candidate correct

Issues:

- OCR errors -> impact on search by title (low recall)
- API as a “black box” + bottleneck of search by title

Lookup

2. Against metadata of digitized books

Goal: verify cohesiveness of digitized corpus

Method:

- based on SBN lookup
- but lookup against digitization metadata
- tuned to maximize precision
- returns 1 or no matches

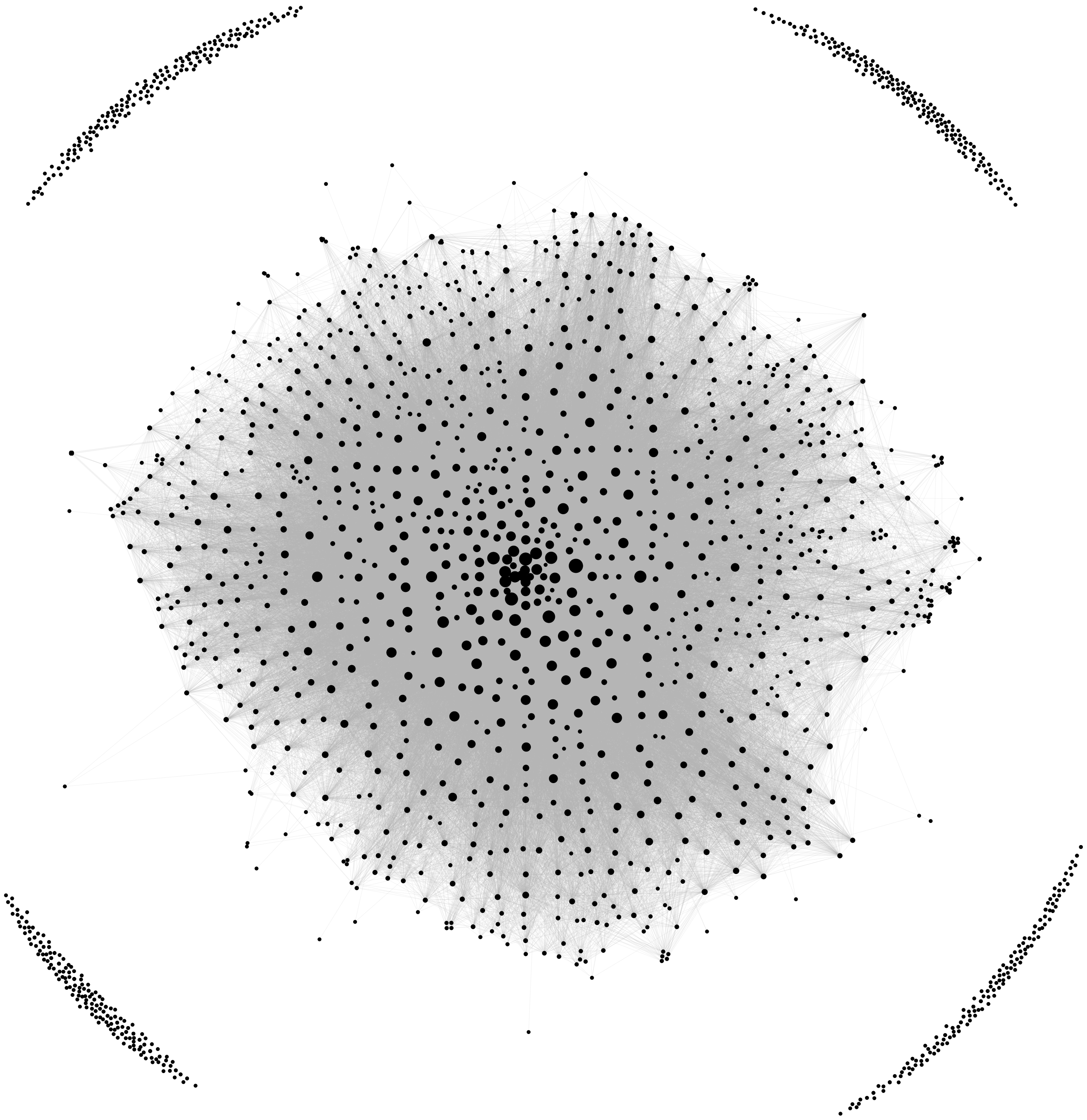
Evaluation*:

- 500 references (out of 181k)
- precision ~ 1.00
- recall > 0.95

Result:

- **only 7% of references** extracted from 701 monographs **point inwards** (i.e. towards the 1904 monographs)

Core of the discipline



co-citation network from extracted references*

giant component = 59% of selected corpus

books in the giant component -> **core of reference works** on history of Venice

giant component -> 32.5% with only works in consultation

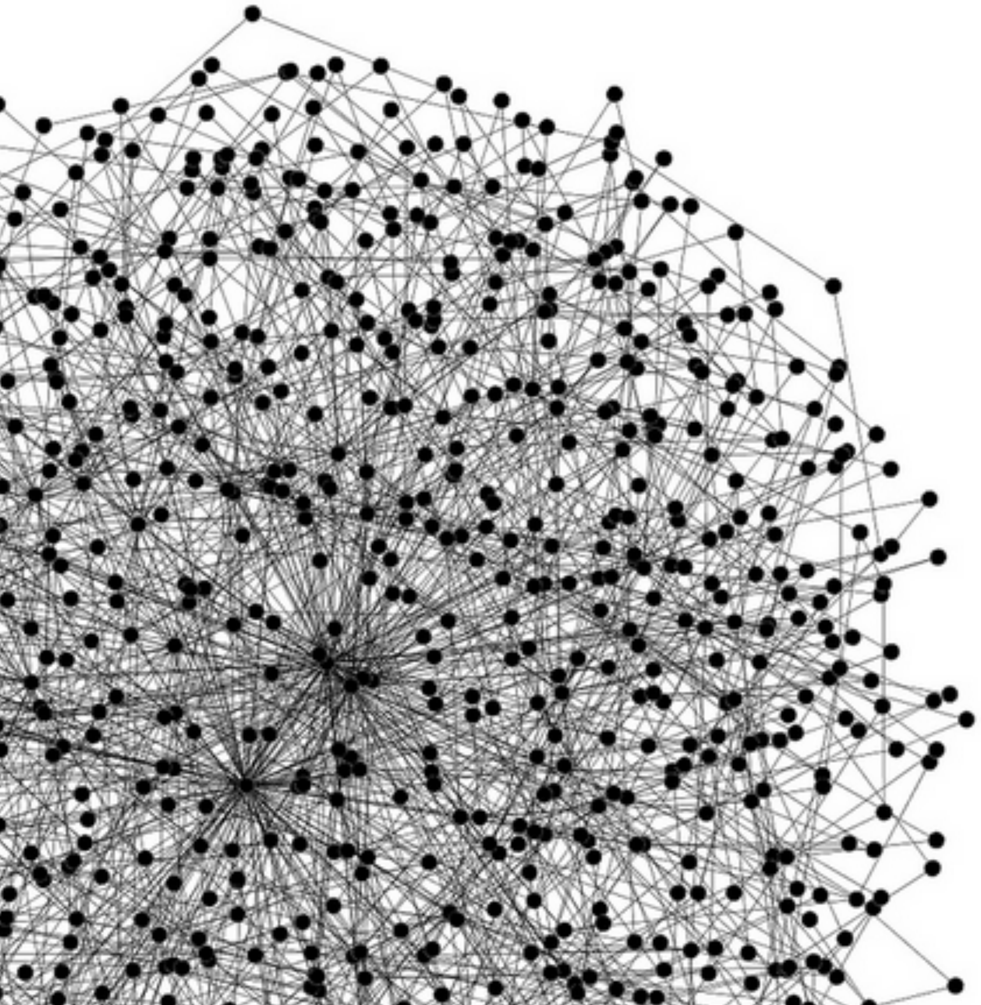
Conclusions and Outlook

data- and citation-driven approach to assess and exploit, from an IR point of view, domain-specific library holdings on the history of Venice

next big challenge: extraction, consolidation and disambiguation of references contained within footnotes (journals)

Thank you!

go.epfl.ch/linkedbooks



Giovanni Colavizza
Matteo Romanello (@mr56k)
Frédéric Kaplan (@frederickaplan)