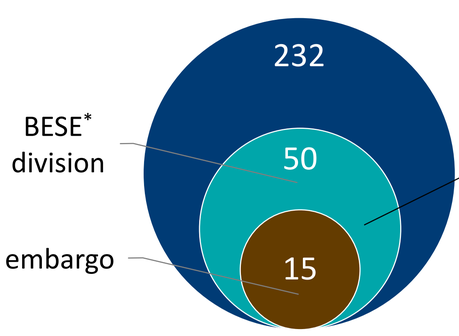
Data mining of Citations in Theses:
a workflow for automated analysis of Open Access and library holdings coverage

Jose I. Martín Cuesta (Systems Specialist), Lee Yen Han (Subject Specialist)

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology
University Library

KAUST theses 2016 - 2018



* Biological and Environmental Science and Engineering

sample of 35

determined librarian

Metadata from Crossref

7,429 citations…

…to **7,107 resources**

Objectives

- Assess library collection coverage of resources used in doctoral dissertations
- How many of those resources are Open Access?
- Complete and compare with other resource usage metrics

Book, book chapter, conference: 462

38% DOI

62% no DOI

✓ in holdings?
✗

✓ 15 **Summon** DOI?

Crossref add ISBN

✓ 10 ebooks static list?

✓ 102 **Summon**
✗ 31 ISBN?

✓ 23 **Summon**
✗ 125 title + 2 authors?

total:
✓ 150
✗ 156
? 156

Journal articles: 6,416

91% DOI

9% no DOI

✓ 5,213 **Summon**
✗ 809 title + 2 authors?

Crossref add ISSN

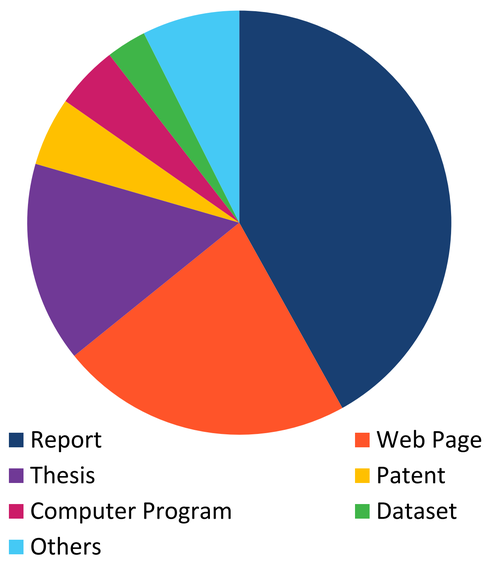
✓ 199 **Summon**
✗ 53 ISSN + year?

CORE

unpaywall

total:
✓ 5,412
✗ 862
? 142

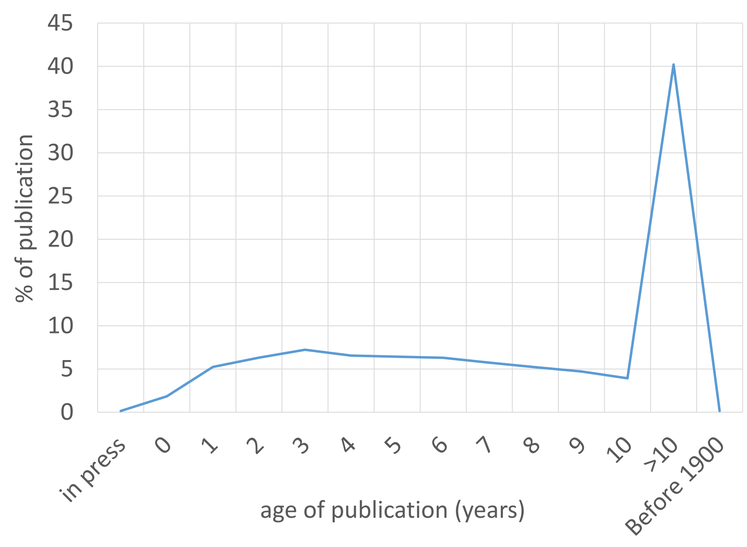
Others: 229



Findings

- Library holdings cover:
  - 84% of journal articles cited
  - 49% of the matched books, book chapters and conference proceedings (51% remained without conclusive coverage outcome)
- OA coverage of resources with DOI:
  - 49,8% of journal articles
  - 17,6% of books, book chapters and conference proceedings
- 20% of journal titles account for 73.6% of article citations (80/20% rule)
- Difficult to obtain accurate coverage information from the API of the discovery platform
- CORE and Unpaywall invaluable for OA versions and complement each other nicely (32% overlap for journal articles and 6% for books)
- Nearly 70% of cited resources published > 5 years ago

Age of Publication Cited



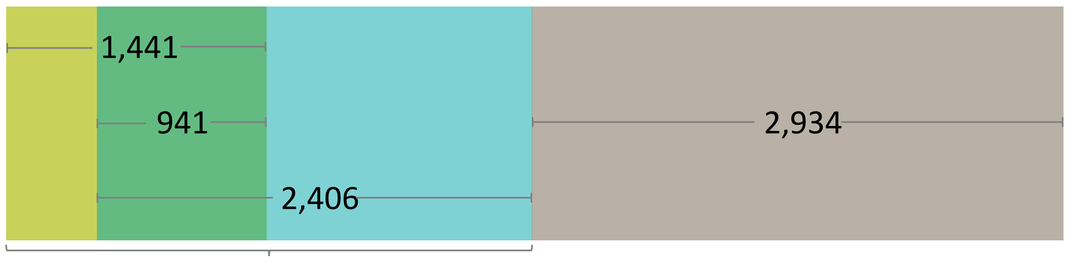
Open Access coverage

CORE Unpaywall not OA

Book, book chapter, conference with DOI: 176

31 (17.6%)

Journal articles with DOI: 5,840



2,906 (49.8%)

Future work

- Automation of Metadata extraction from Crossref
- Refine workflow to increase accuracy and minimize false matches
- Expand OA coverage analysis to resources without DOI using other metadata
- Extend study to all doctoral dissertations and other scholarly output