Data Science Workshop
Lecture 13: State of the Art

Marcin Luckner, PhD
mluckner@mini.pw.edu.pl

Version 1.1
November 19, 2020

Warsaw University of Technology

MSc program in Data Science has been developed
as a part of task 10 of the project
„NERW PW. Science - Education - Development - Cooperation”
co-funded by European Union from European Social Fund.

State of the Art

A Time to Harvest by Chaosium Inc.

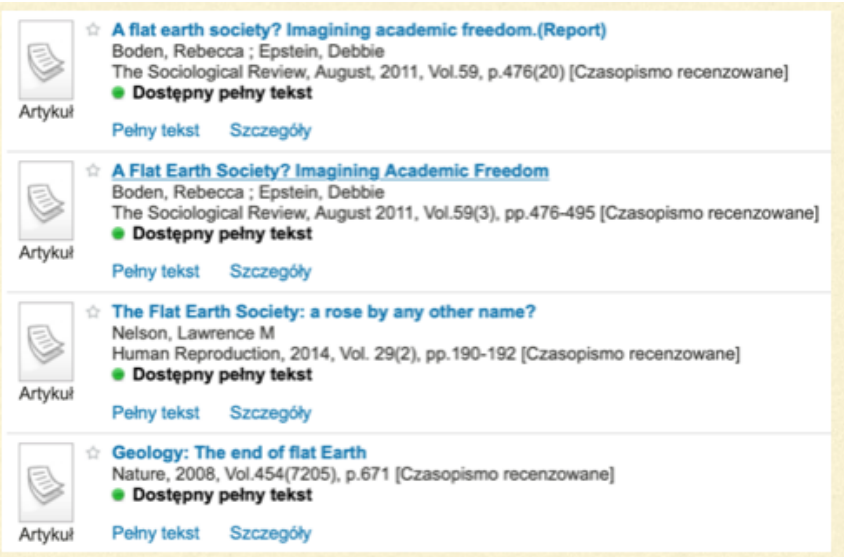
State of the Art review

- Working on some issue, we should make a state of the art review.
- The review allows us to answer the following questions:
  - Is it worth to take up this issue?
  - What has been done in this area?
  - What left to achieve?

State of the Art objectives

- With a state of the art we can:
  - Determine if our solution is innovative in comparison to the existing solutions.
  - Determine a benchmark.
  - Determine how to verify and assess our solutions.

Sources



- It is suggested to analyse scientific sources.
- Scientific works go through a revision process.
- Published works are a subject of a public academic discussion.
- That increases trust in published information.
- Works in Computer Science area are mostly published in an electronic version.

First steps of review

1. Determine the keywords that identify works in the examined area.
2. Search databases using the keywords.
3. Find **several dozen** of promising publications.
4. Carry out a primary selection of useful works.

Keywords

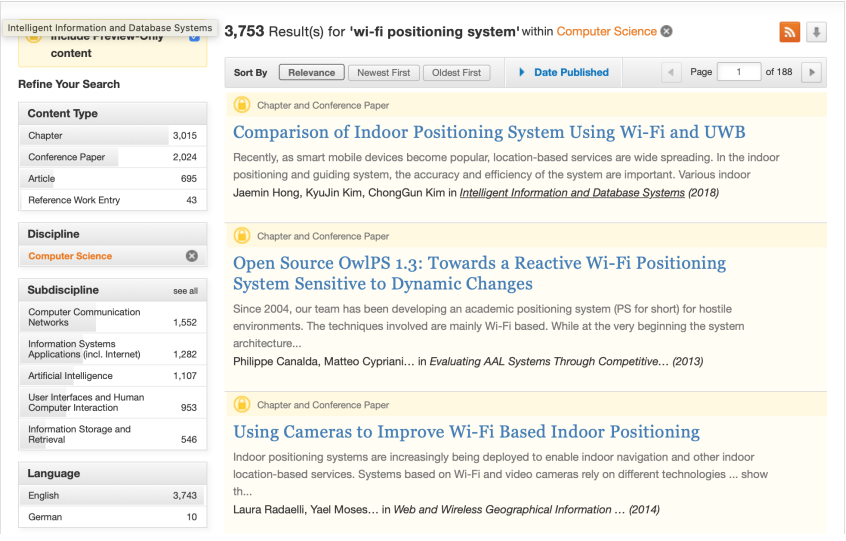
- Determine keywords that can be used to describe the examined area.
  - Find synonyms.
- When you find an article that covers the examined area, find the keywords inside the article and use them.

Publishers' databases

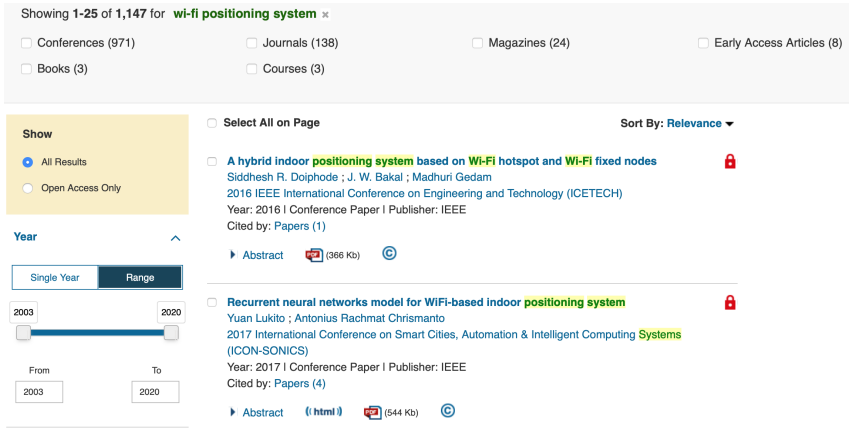
- Full-text databases.
- Usually limited to one publisher.
- The most extensive databases in the computer science area are Springer and IEEE.
  - Mostly, conference papers.

- ACM Digital Library
- IEEE/IEE Electronic Library
- Science Direct
- Springer Link
- Taylor and Francis Online
- Wiley Online Library

Sources



https://link.springer.com/



https://ieeexplore.ieee.org/

General databases

- Abstract databases.
- Collect entries from various sources.
- A good source of bibliographical data.
  - Respected databases in the computer science area are Web of Science and Scopus.

- The WUT Base of Knowledge
- DBPL Computer Science
- Google Scholar
- Scopus
- Web of Science

Results diversity

- An example of preliminary research for a given author.
  - Author: Marcin Luckner
  - Date: 31.01.2019
- The number of records
  - Google Scholar - 47
  - The WUT Base of Knowledge - 40
  - Scopus - 36
  - DBPL Computer Science - 35
  - Web of Science - 24

Database search guidelines

- Use more than one database.
- Use open sources and sources offered by the Main Library of WUT.
- Using full-text databases saves data access time.
- However, the abstract databases give us a more comprehensive view and sometimes link to the full tests.

Preliminary selection

- In case of a substantial number of articles covering the examined area, we can use the following criteria to narrow the set
  - A publication year
    - The newest publication should present the current State of the Art.
  - A citation number
    - Frequently cited articles may be important for the research area.
  - A publication place
    - Publications in respected journals (and in some conference materials) go through a multi-step review process.

Journal assessment

- The impact factor (IF) is an index that reflects the yearly average number of citations that articles published in the last two years in a given journal received.
- It is frequently used as a proxy for the relative importance of a journal within its field.
- Journals with higher impact factors are often deemed to be more important than those with lower ones.
- Journals with IF are grouped in Journal Citation Reports by Clarivate.

IF controversy

- IF is used to assess the authors.
- Some editors manipulate IP factor value
  - Self-citations and cross-citations.
  - The Polish journal *Przeglad Elektorotechniczny* was removed from JCR in 2013 with the self-citation index close to 77%.
- Predatory publishers use fake impact factors
  - Universal Impact Factor (UIF), Global Impact Factor (GIF), and Citefactor.

Conference assessment

- The CORE Conference Ranking provides assessments of significant conferences in the computing disciplines.
- Conferences are assigned to one of the following categories:
  - A* flagship conference, a leading venue in a discipline area
  - A excellent conference, and highly respected in a discipline area
  - B good conference, and well regarded in a discipline area
  - C other ranked conference venues that meet minimum standards

`https://www.core.edu.au/conference-portal`

Reference management

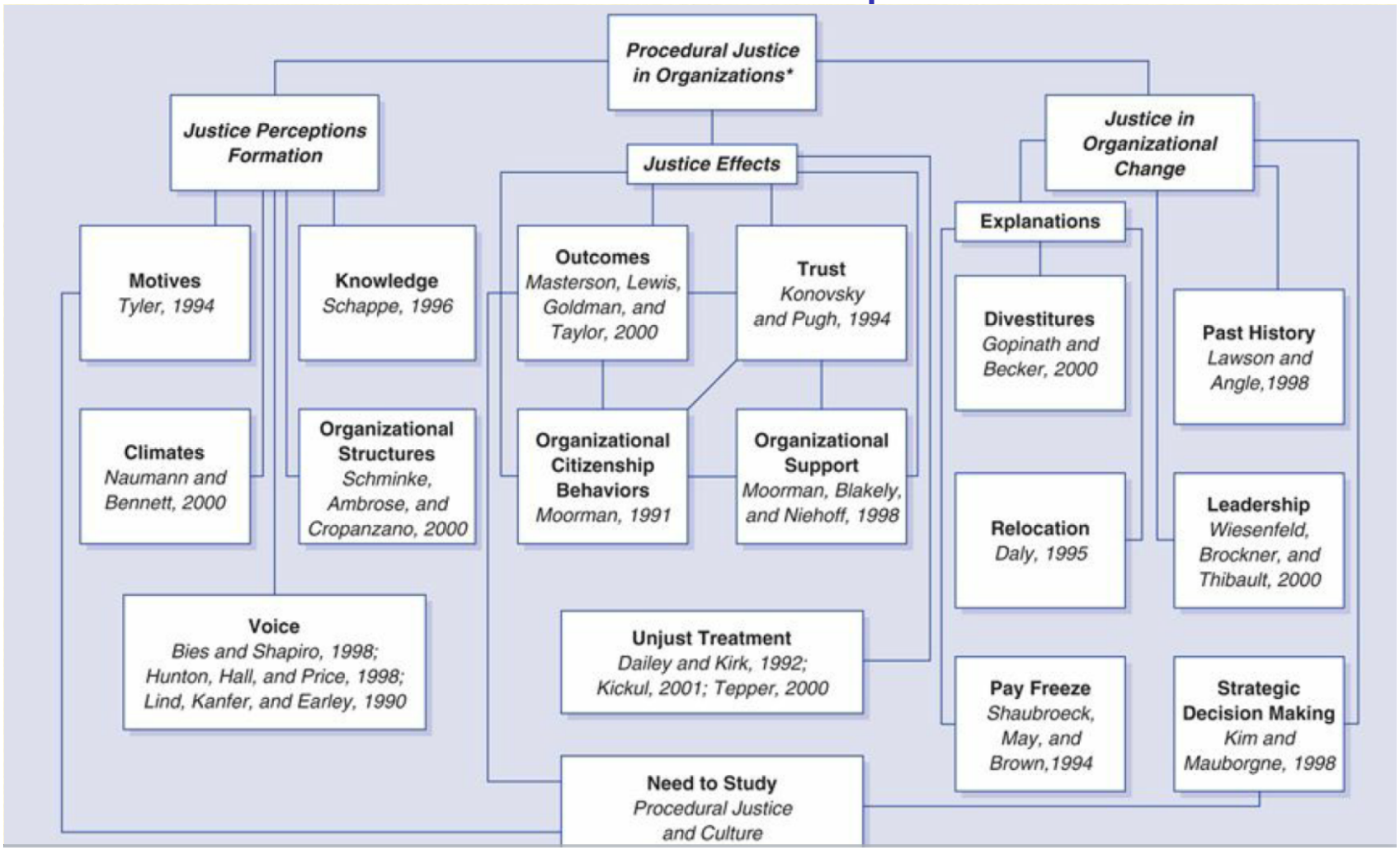
Using Reference Management Tools: EndNote and Zotero by Diarmuid Stokes

Reference management

- Working with several dozens of references, we have to keep them in order.
- An absolute minimum is to order bibliographical data and full texts in logical structure.
- Conceptually, we can be supported by hierarchical folder structure and literature map.
- A helpful tool is a reference manager.

Organised folder structure

Literature map



Research Design: Qualitative, Quantitative and Mixed Methods by John W. Creswell and J. David Creswell.

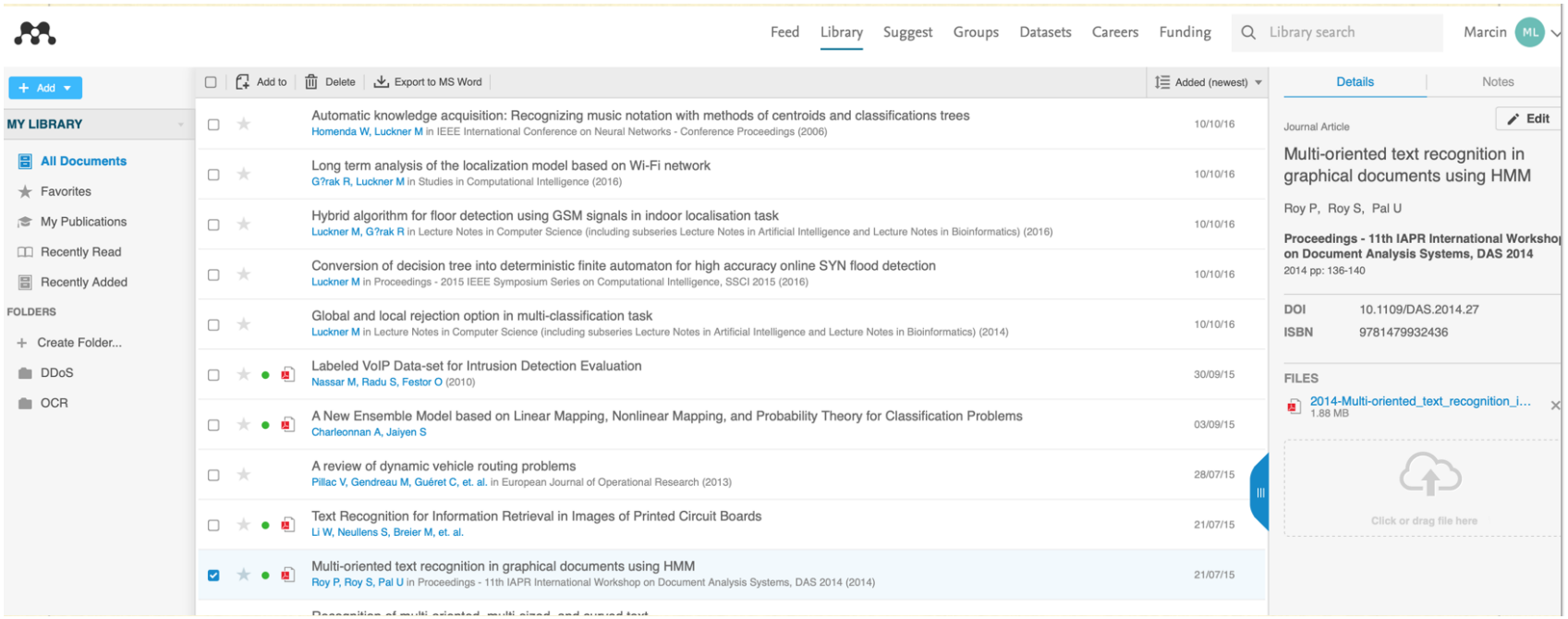
Organised bibliographical data

- Bibliographical data can be kept using bibtex format.
- That allows the author to create references quickly.
- A reference manager allows us to use this format comfortably.

```
@InProceedings{Bruckner2009,
  Title                    = {Nash Equilibria of Static Prediction
      Games},
  Author                   = {Michael Br{\"u}ckner and Tobias Scheffer},
  Booktitle                = {NIPS},
  Year                     = {2009},
  Pages                    = {171-179},
  Bibsource                = {DBLP, http://dblp.uni-trier.de},
  Crossref                 = {DBLP:conf/nips/2009},
  EE                       = {http://books.nips.cc/papers/files/nips22/
      NIPS2009_0534.pdf}
}

@inproceedings{Hao2016,
 author = {Hao, Shuang and Kantchelian, Alex and Miller, Brad and
     Paxson, Vern and Feamster, Nick},
 title = {PREDATOR: Proactive Recognition and Elimination of Domain
     Abuse at Time-Of-Registration},
 booktitle = {Proceedings of the 2016 ACM SIGSAC Conference on Computer
     and Communications Security},
 series = {CCS '16},
 year = {2016},
 isbn = {978-1-4503-4139-4},
 location = {Vienna, Austria},
 pages = {1568--1579},
 numpages = {12},
 url = {http://doi.acm.org/10.1145/2976749.2978317},
 doi = {10.1145/2976749.2978317},
 acmid = {2978317},
 publisher = {ACM},
 address = {New York, NY, USA},
 keywords = {domain registration, early detection, reputation system},
}
```

Reference manager



https://www.mendeley.com/

Reference managers comparison

Reference management software	Cost	Does it integrate with Microsoft Word?	Can it capture a webpage to create a record?
RefWorks	US$100/year or free at any subscribing academic institution	Yes (requires download of Write-N-Cite utility)	Yes
Zotero	Free/open source	Yes (also works with OpenOffice)	Yes
EndNote	US$249.95	Yes (also works with OpenOffice)	No
Mendeley	Free	Yes (also works with OpenOffice)	Yes
Citationsy	Free	Yes (also works with OpenOffice)	Yes

https://www.scribendi.com/academy/articles/reference_management_software_solutions.en.html

Task

- In teams
  - Determine the keywords to prepare the State of the Art for your project.

References


J. W. Creswell and J. D. Creswell.
Research Design: Qualitative, Quantitative and Mixed Methods.
SAGE, 2014.


M. E. Falagas and V. G. Alexiou.
The top-ten in journal impact factor manipulation.
Archivum Immunologiae et Therapiae Experimentalis volume, 2008.


M. Jalalian.
The story of fake impact factor companies and how we detected them.
Electron Physician, 2015.

Warsaw University of Technology

MSc program in Data Science has been developed
as a part of task 10 of the project
„NERW PW. Science - Education - Development - Cooperation”
co-funded by European Union from European Social Fund.