Navigating the (Systems) Literature

David Cock

8 May 2014
Why Read Papers?

At least 3 reasons:

• Novelty — To convince yourself that people care about the problem, and that it hasn't already been solved.

• Expertise — Aim to become the local (or even global) expert in your topic. By 3rd year, expect other group members (including researchers) to ask whether X has been done before.

• Communication — In your papers, and your thesis, you need to direct the reader to your sources. Aim to extend and improve the literature.
Why Read Papers?

At least 3 reasons:

- Novelty — To convince yourself that people care about the problem, and that it hasn’t already been solved.
Why Read Papers?

At least 3 reasons:

• Novelty — To convince yourself that people care about the problem, and that it hasn’t already been solved.

• Expertise — Aim to become the local (or even global) expert in your topic. By 3rd year, expect other group members (including researchers) to ask you whether X has been done before.
Why Read Papers?

At least 3 reasons:

- **Novelty** — To convince yourself that people care about the problem, and that it hasn’t already been solved.

- **Expertise** — Aim to become the local (or even global) expert in your topic. By 3rd year, expect other group members (including researchers) to ask you whether X has been done before.

- **Communication** — In your papers, and your thesis, you need to direct the reader to your sources. Aim to extend and improve the literature.
Outline

- The Way Things Are
- Publishers are Definitely Stupid
It’s 2025. You’ve found a great new paper on verifying Angry Birds XII, A.
It cites $E$, saying “It is well-recognised[1], that all commercially-successful kernels are formally verified.”.
You put $E$ into Google Scholar, and click on “cited by”, and find $B$ and $C$. $B$ is a paper on verified GC, while $C$ is a paper on timewasting. You decide that $C$ is directly relevant, while $B$ less so.
You see that $E$ cites another paper, as the first to show that large-scale verification works. This paper is clearly pivotal, but is now mostly of historical interest.
Citation Chains

Chasing citations of this paper is pretty futile. $D$ is a paper in a naval engineering journal on the use of seL4 in the control systems of on-board fusion reactors.
Citation Chains

The literature is huge DAG, and you need to walk it efficiently. Identifying ‘sink’ articles, like \( E \), and ‘landmarks’, like Klein et. al. is crucial, as they either cite or are cited by everything worth reading.
Citation Chains

You can’t read everything, so you need to learn to prune effectively.
Finding Things

You need to know where the work in your field is published, and how to find it. Generally, this comes down to knowing three things:
Finding Things

You need to know where the work in your field is published, and how to find it. Generally, this comes down to knowing three things:

- Venues
You need to know where the work in your field is published, and how to find it. Generally, this comes down to knowing three things:

- Venues
- Publishers
You need to know where the work in your field is published, and how to find it. Generally, this comes down to knowing three things:

- Venues
- Publishers
- Aggregators
Venues (Systems)

Publisher are Definitely Stupid

- Prestige venues: SOSP, OSDI, . . .
  - The big-story results will be here.
- The second tier: HotOS, Eurosys, . . .
  - Higher volume, not necessarily lower quality. Often more focused.

Workshops
- Incremental or early work, highly focused.

Journals
- Longer papers, not a thing in systems.
Prestige venues: SOSP, OSDI, …
The big-story results will be here.
Venues (Systems)

Prestige venues: SOSP, OSDI, …
The big-story results will be here.

The second tier: HotOS, Eurosys, …
Higher volume, not necessarily lower quality. Often more focused.
Prestige venues: SOSP, OSDI, …
The big-story results will be here.

The second tier: HotOS, Eurosys, …
Higher volume, not necessarily lower quality. Often more focused.

Workshops
Incremental or early work, highly focused.
Venues (Systems)

Prestige venues: SOSP, OSDI, …
The big-story results will be here.

The second tier: HotOS, Eurosys, …
Higher volume, not necessarily lower quality. Often more focused.

Workshops
Incremental or early work, highly focused.

Journals
Longer papers, not a thing in systems.
Publishers

- In systems, mostly ACM, IEEE & Usenix.
- Springer and Elsevier for some journals (sometimes proceedings).
- Some open-access or independents.
- Sometimes (e.g. ACM DL) index each other’s pubs — don’t trust citation details. Go to the source.
The publisher’s database is the canonical source for citation information:

- ACM Digital Library (DL)\(^1\)
- IEEE Xplore\(^2\)
- Usenix\(^3\)
- SpringerLink\(^4\)
- Elsevier ScienceDirect\(^5\)

\(^1\)http://dl.acm.org
\(^2\)http://ieeexplore.ieee.org/
\(^3\)https://www.usenix.org/publications/proceedings
\(^4\)http://link.springer.com
\(^5\)http://www.sciencedirect.com
Aggregators

Index articles across publishers.

6http://citeseerx.ist.psu.edu/
7http://scholar.google.com/
Aggregators

Index articles across publishers.

- Citeseer\(^6\) — Slow, crappy bibtex, don’t bother.

---

\(^6\)http://citeseerx.ist.psu.edu/
\(^7\)http://scholar.google.com/
Aggregators

Index articles across publishers.

- Citeseer\(^6\) — Slow, crappy bibtex, don’t bother.
- Google Scholar\(^7\) — Better, basically a narrow google search, noisy. Use for cited-by.

\(^6\)http://citeseerx.ist.psu.edu/
\(^7\)http://scholar.google.com/
Aggregators

Index articles across publishers.

- Citeseer\textsuperscript{6} — Slow, crappy bibtex, don’t bother.
- Google Scholar\textsuperscript{7} — Better, basically a narrow google search, noisy. Use for cited-by.
- ACM DL — Indexes most of CS, pretty well cultivated. Don’t trust for citation details other than ACM.

\textsuperscript{6}http://citeseerx.ist.psu.edu/
\textsuperscript{7}http://scholar.google.com/
Outline

• The Way Things Are

• Publishers are Definitely Stupid
Many publications are still paywalled, although some publishers have got the message (e.g. Usenix).

- Your university library will usually have a subscription.
- At UNSW, use the library proxy by appending to the URL e.g.
  http://dl.acm.org.wwwproxy0.library.unsw.edu.au/
- This can be automated, see next slide.
- Some very old papers aren’t digitized, but that’s rare in CS.
- Most authors put a prepress version on their own website. At NICTA this is automatic.
There are \( n + 1 \) of these, I used Zotero\(^8\), and found it useful.

- Integrated with Firefox.
- Detects ACM DL, IEEExplore, \ldots
- Saves reference details with a click.
- Free cloud sync for references and PDFs.
- Detects library proxy and automatically redirects.

\(^8\text{https://www.zotero.org/}\)
It’s possible to write a thesis without LaTeX (e.g. in Word), but I don’t recommend it. BibTeX is your friend, learn it, love it, use the group bib files\(^9\).

Notes:

- **Don’t trust bibtex you found on teh Intertubes.** Proofread it, and cross-check against the publisher.
- Check if something’s already in the group bib files. If not, add it, after first proofreading it (see above).
- BibTex has quirks. If you can’t get an entry to compile, ask someone experienced.
- Include extra data e.g. DOI if possible.
- Learn the local quirks (e.g. address field for conference location).

\(^9\)hg clone /data/hg_root/bibtex
Questions?