Unscrambling the omelette:

turning free-text citations into Wikipedia format citations

Kerry Raymond

kerry.raymond@wikimedia.org.au

File:Mad scientist caricature 2.png. (2014, December 31). *Wikimedia Commons, the free media repository*. Retrieved 14:35, June 11, 2019 from https://commons.wikimedia.org/w/index.php?title=File:Mad_scientist_caricature_2.png&oldid=144870516.

“I would like, if I may, to take you on a strange journey” (The Criminologist, Rocky Horror Picture Show, 1975)

- There were very few computer science journals in the library
  - CS not indexed by the major indexing publications
- Staff & students personally subscribed to journals by sea mail
- The citations in the papers were the main source of knowledge of other publications
- Letters written to authors to request copies of their papers (using electric typewriters & air mail postage)
- Rolled-my-own bibliographic database with simple key word search capability

“Don’t dream it! Be it!”

(Dr Frank N Furter, Rocky Horror Picture Show, 1975)

- Computers on every desk, on every phone, and everywhere else
- Internet: email, WWW, social media
- Subscription databases (mostly behind paywalls)
- Google Scholar
- Trove
- ResearchGate, Academia …
- ORCID, VIAF, …
- Meta-data
- New concepts: “smallest publishable unit of research”, “citebait”

“Say! Any of you guys know how to Madison?”

(Brad Majors, RHPS, 1975)

- What didn’t change:
  - “Dear Sir”
  - Rendering of citations
    - Despite all the cataloguing, identities, meta-data, we strip it all out and deliver the reader:
      - a set of words (often abbreviated)
      - a sprinkling of punctuation (particularly commas and full stops)
      - just like we did with the typewriter
      - a light dusting of fonts (to prove we aren’t using typewriters)

“He had a certain naive charm, but no muscle”

(Furter, F.N., RHPS, 1975)

- Despite WP:NOTPAPER Wikipedia also *renders* citations like it was 1975 all over again
- Under the hood, Wikipedia may have structured citations (e.g. the {{cite}} templates) or just be random text and punctuation
- How to reverse the rendering process and restore structure etc?

“ I hope you're adaptable, Dr. Scott. I know Brad is.”

(ibid)

- People are quite good at restoring structure to citations:
- Mary Smith, David Jones, Frogs adapting to climate change, 2011
- People adapting to citations, 2013, Mary Smith, David Jones
- Mary Smith, David Jones, Department stores adapting to online retailing, 2015
- Mary Smith, David Jones, 2017

“Janet, they're obviously foreigners, and this must be one of their national dances.” (Majors op. cit.)

- Being able to read helps
  - *read*: convert squiggles into concepts and connect them meaningfully
- But computers can’t read, just as we cannot read foreign languages

- maHvaD ghu'vam poH warp qa‘ : nagh jISuvvIpbe' mIllogh cha‘ : 1975

- Klington to English: Lets do the time warp again, Rocky Horror Picture Show, 1975

“So come up to the lab and see what's on the slab”

(Furter *loc. cit.*)

- Exercise time: 5 minutes or so
  - Take a lucky dip for a Wikipedia article with free-text citations
  - Look at that Wikipedia article and its unstructured citations
  - Wikipedians: Try convert some into {{cite}} format, if you can
  - Non-Wikipedians: Just do it as mental exercise to identify the different parts of the citation
  - Feel free to use the resources of the Internet if it will help
  - Feel free to take a 2nd lucky dip if you are fabulous at this!

“If only we were amongst friends… or sane persons!” (Weiss J 1975)

- Genuine question:
  - What strange or difficult or confusing things did you encounter?
  - Queensland State Archives, 1894, Cleveland Divisional Board Minutes, 2 July 1894, QSA Item ID869236, Minutes
  - Bundling one citation; with another citation; and another; but unfortunately some citations contain semicolons
  - Ibid, op cit, loc cit
  - Commentary not a citation (originally both footnotes)

“There's a light, over at the Frankenstein place”

(ibid)

- Interpreting citations is easier
  - If they come from a common source and use a house style
  - If they are in discipline/domain of which you have prior knowledge
  - You can read the language
  - You wrote them

“This sonic transducer, it is I suppose some kind of audio-vibratory-physio-molecular transport device?” (Scott, E.V., op cit)

- Machine interpretation of free-text citations is not easy!
- Rules to interpret a known set of citations in a known house style

```
([A-Za-z .'-]+), \"([A-Za-z0-9 .?:&;',-]+)\", in
([A-Za-z .'-]+) and ([A-Za-z .'-]+) \(eds\),
\"([A-Za-z0-9 .?:&;,-]+)\", (\d+) \((\d\d\d\d)\)
```

- 1 or 2 authors with a chapter title (possibly with commas) in a monograph with 2 editors and a volume number and a year”
- 30+ other rules to recognise *almost all* citations

“I'm lucky, he's lucky, we're all lucky!”

(Magenta 1975)

- Machine processing needs heuristics to compensate for the inability to read:
  - But heuristics are not perfect rules, they gamble on the likely probabilities
  - E.g. recognising human names (probably authors & editors)
  - But documents can be authored by organisations too, so need to know words that suggest organisations e.g. in the New South Wales State Heritage Register:
    - Architect, Archive, Branch, Council, Company, Consultant, Department
    - Want to avoid having first/last name processing of non-human authors, e.g. “Wales, State Records of New South”
    - What’s the heuristic for “David Jones” vs “Jones, David”?

“And crawling, on the planet's face, some insects, called the human race. Lost in time, and lost in space… and meaning”

Title: Rocky Horror Picture Show, Date: 1975

- Big question: Why do we render citations in cryptic formats from a pre-computer era instead of giving the reader the structure they need to interpret and re-use them?
- In the meantime, how do build tools and develop heuristics to make a good guess at the structure?