Open Source in NLP research

Ann Copestake

Natural Language and Information Processing Group
Computer Laboratory
University of Cambridge

November 2016

Why Open Source in NLP research?

Research results should be open wherever possible:

- Public good from public money
- Reproducibility (serious NLP problem:
  see Fokkens et al (2013), ACL)
- More uptake and more citations . . .
- Researchers move institutions: retain access to your work!

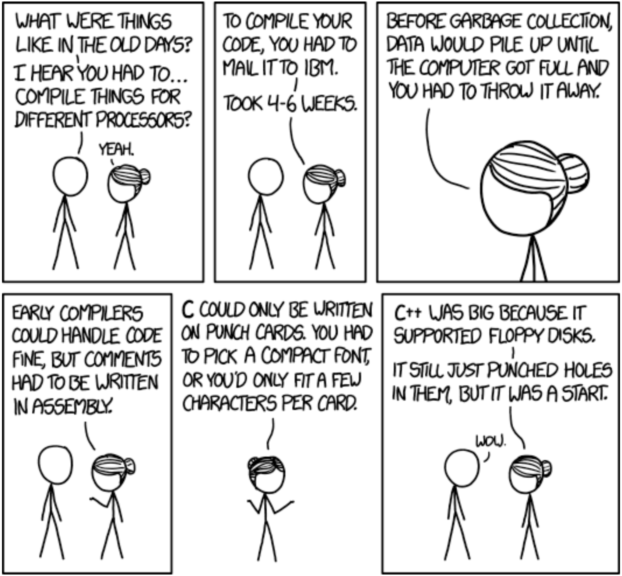
Terminology:

- Lingware: any computational linguistic resource other than software.
- Open Source (capitals): a convention from late 90s to distinguish truly open licences from others.

Research Software and Lingware

- Software may be released as Open Source at different levels of maturity and utility.
- This workshop: software and lingware which is:
  - genuinely reusable;
  - (ideally) used to build applications with end users;
  - (possibly) collaboratively developed.
- Easiest form of collaborative development is via interoperability: different pieces work together.
- Modern source control.

xkcd 1755


WHAT WERE THINGS LIKE IN THE OLD DAYS?
I HEAR YOU HAD TO... COMPILE THINGS FOR DIFFERENT PROCESSORS?
YEAH.
TO COMPILE YOUR CODE, YOU HAD TO MAIL IT TO IBM.
TOOK 4-6 WEEKS.
BEFORE GARBAGE COLLECTION, DATA WOULD PILE UP UNTIL THE COMPUTER GOT FULL AND YOU HAD TO THROW IT AWAY.
EARLY COMPILERS COULD HANDLE CODE FINE, BUT COMMENTS HAD TO BE WRITTEN IN ASSEMBLY.
C COULD ONLY BE WRITTEN ON PUNCH CARDS. YOU HAD TO PICK A COMPACT FONT, OR YOU'D ONLY FIT A FEW CHARACTERS PER CARD.
C++ WAS BIG BECAUSE IT SUPPORTED FLOPPY DISKS.
IT STILL JUST PUNCHED HOLES IN THEM, BUT IT WAS A START.
WOW.

LinGO project and DELPH-IN collaboration

- CSLI, Stanford: NLP project to develop software, grammars and lexicons. Funded by VerbMobil (Germany) and NSF. Some software previously developed at Cambridge under ACQUILEX.
- Substantial software and lingware development with over 30 people listed in LREC 2000 paper (mostly very part time).
- Open source by late 1990s (1998/9?) after some negotiation.
- DELPH-IN: international collaboration incorporating LinGO. Open source of some substantial contribution is a requirement for involvement.

Background: open source in NLP

- Up to early 1990s, much NLP research was done in companies: papers about highly complex software and lingware, very limited distribution.
- Results not reproducible (few numerical results anyway).
- Later: statistical NLP relied on somewhat restricted annotated corpora (LDC, ELRA), but software released or described well enough for (partial) replication.
- WordNet first substantial lingware with an open source licence (1991)?
  WordNets for other languages mostly not open, but now `http://globalwordnet.org/wordnets-in-the-world/`

Arguments against

- My software/lingware is too messy / I can't support it.
- Duty of reproducibility.
  Funder requirements, e.g., EPSRC:

  > *research data is a public good produced in the public interest and should be made freely and openly available with as few restrictions as possible in a timely and responsible manner*

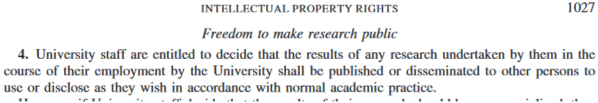
  `https://www.software.ac.uk/resources/guides/epsrc-research-data-policy-and-software`

  Policy encourages Open Source (but has provision for commercial exploitation)

  `https://www.software.ac.uk/resources/guides/adopting-open-source-licence`

Arguments against

- My boss/university/funder want to exploit the IP.
- In Cambridge, this is up to the researcher/research team:



Freedom to make research public

4. University staff are entitled to decide that the results of any research undertaken by them in the course of their employment by the University shall be published or disseminated to other persons to use or disclose as they wish in accordance with normal academic practice.

- The status of software etc should ideally be decided at the start of the project or collaboration, including when supervising students.
- Industry collaborators often have restrictions, but sometime can be convinced (especially if building on open software).

Arguments against

- Researcher from company X wants to use my software but can’t because it’s open source.
- Sell them a non-exclusive licence . . .
- Not sure whether this is still relevant: issue arose because of confusion between normal open source and copyleft.

Arguments against

- If I make my software open source, someone else can come along and make money from it.
- In theory, possible. Is it likely? Does it matter?
- Use a licence which requires attribution. Remember that you retain copyright.

Arguments against

- If I release my dataset, other people will get better results and I won't be able to publish my own work.
- Given current reviewing, it may make sense to hold back a dataset until paper with results has been accepted.
- Various attempts to ensure data/software is available for papers have had patchy success.