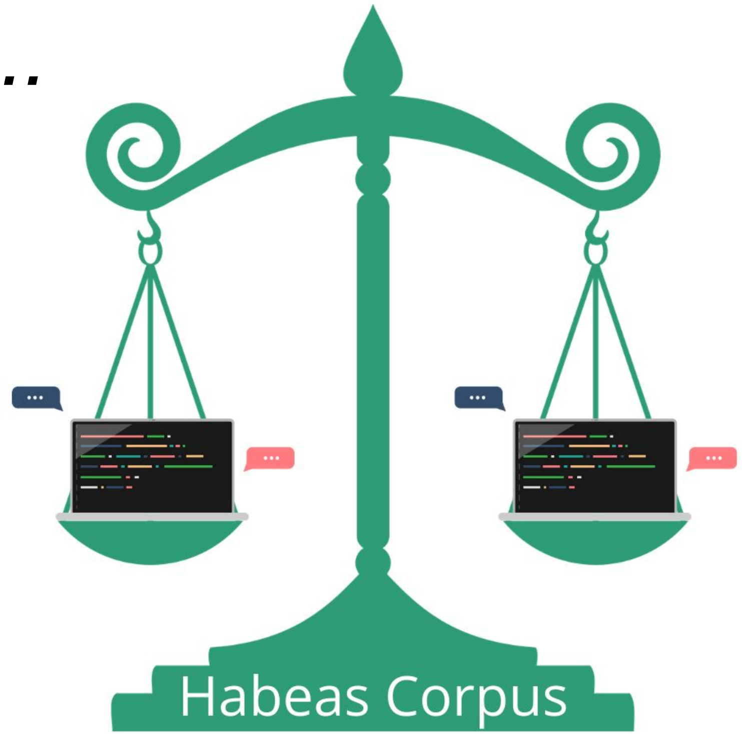
Introducing...

Alexander Konovalov
Hao Ye
Louise Chisholm
Mark Turner
Neil Chue Hong
Sammie Buzzard
Stephan Druskat



A Collaborations Workshop 2021 HackDay Project

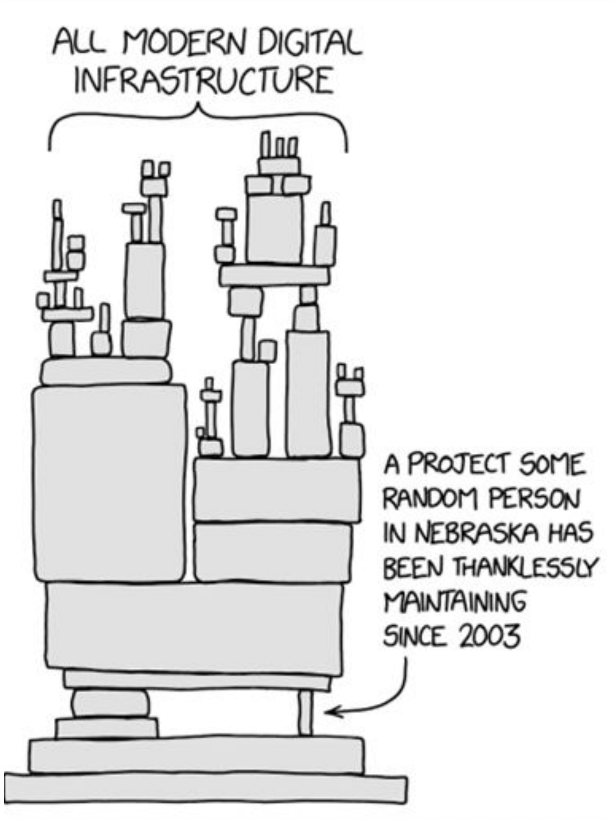
https://github.com/softwaresaved/habeas-corpus

What is the problem?

There are some catalogues and lists of research software, but there are no ***comprehensive*** directories.

This obscures the time and effort spent on this critical component of the research ecosystem.

How can we increase the value, use, and impact of software if we don't know what is there?


ALL MODERN DIGITAL INFRASTRUCTURE
A PROJECT SOME RANDOM PERSON IN NEBRASKA HAS BEEN THANKLESSLY MAINTAINING SINCE 2003

"Dependency" by XKCD
https://xkcd.com/2347/ (CC BY-NC 2.5)

Why do we need this?

	RSE	Researchers	Students	Community Managers	Funders
Enables research **on** research software, creating new opportunities	●	●	●		
Informs decisions **about** research software regarding what to fund, develop and maintain in the long term	●	●		●	●
Increases visibility of time, labor, money invested in essential software programmes	●	●	●	●	●
Empowers users and contributors to find research software, outside of their networks	●	●	●	●	
Informs discussions on FAIR software	●	●	●	●	●

Team work

Shared why we chose this project, and work to our strengths

Decision making by consensus

Quickly established shared co-working space and infrastructure

Split coding tasks using pair programming, along with “crowdsourcing” sessions with all members

Inclusion of non-coding team member → giving the perspective of client for the data, discussing which metadata to collect, developing the logo & slides

Collaboration tools

zoom

slack

GitHub

Approach

Focussed on CORD19 dataset (https://doi.org/10.5061/dryad.vmcvdncs0)

- Initial qualitative analysis:
    - Sampling the dataset and examining existing machine learning approaches to software recognition.
        - Have formulated feedback for ChanZuckerburg's team
    - Dataset cleaning 🧹
    - Identifying Github URLs from software mentions in text.
    - Extracting software licenses from Github URLs
    - Automating the above using software scripts.

Implementation and infrastructure

Challenges

- Extract Data set
- Develop metadata
- Need to manually add the github repositories
- Machine learning has identified incorrect software names, requiring manual QA/QC

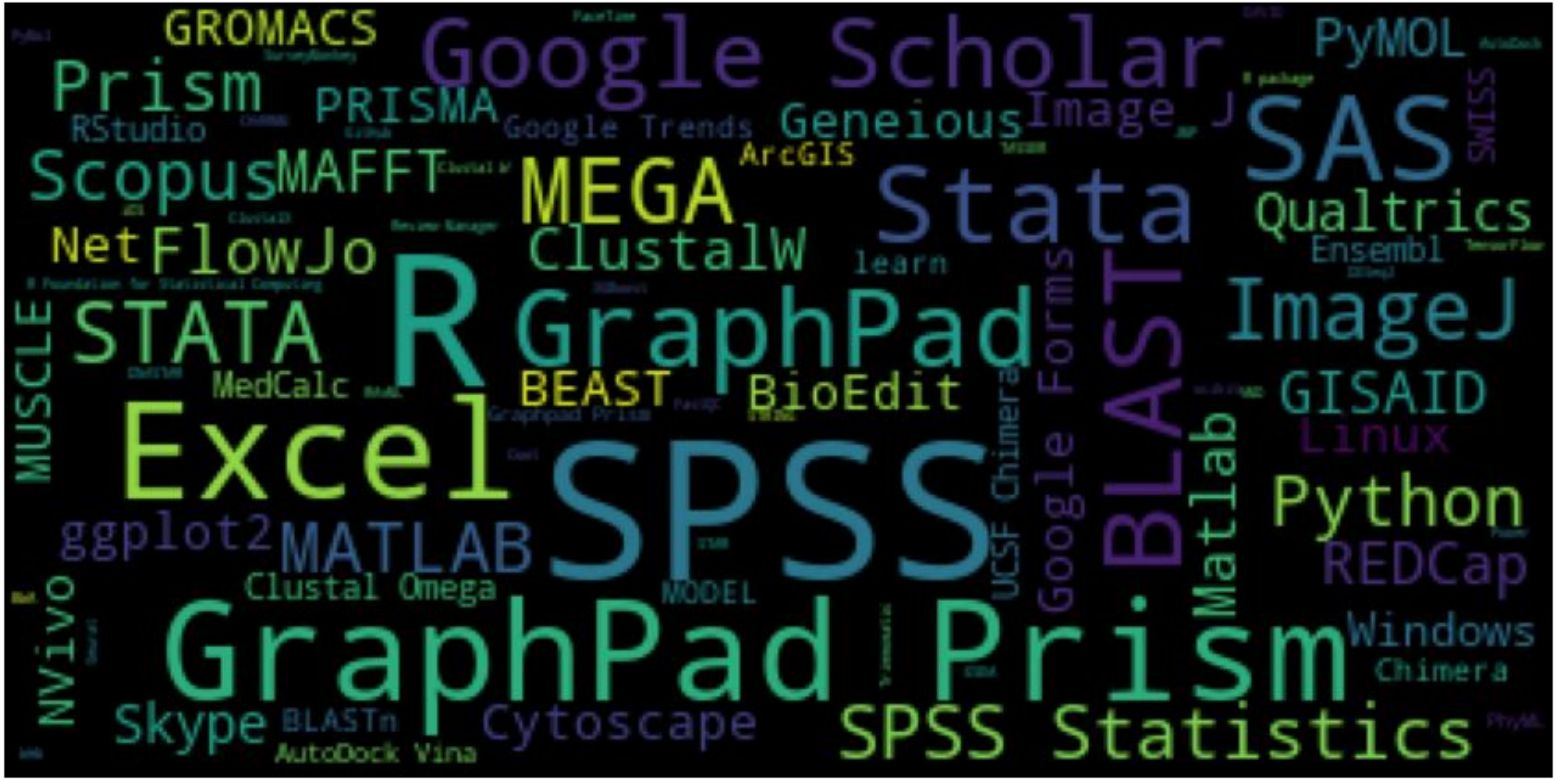
Collaboration

- Agree on metadata fields and format
- Modular data processing scripts
- GitHub collaboration workflow
    - issues, branches, PRs

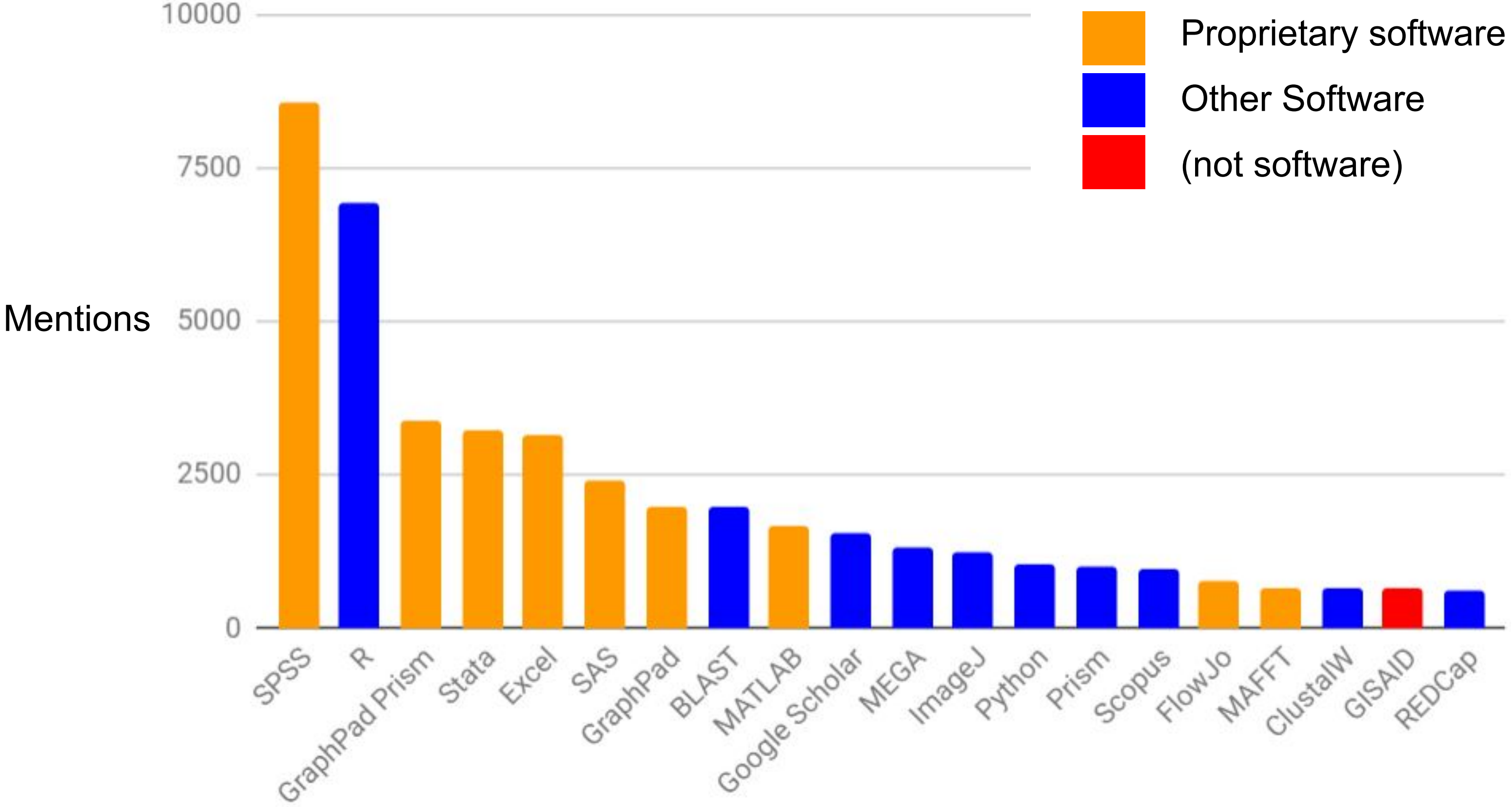
Code released under MIT license, documentation under CC BY 4.0 and data under CC0.

Binder Demo

What software are researchers using to study COVID-19?



Top 20 software used to study COVID-19



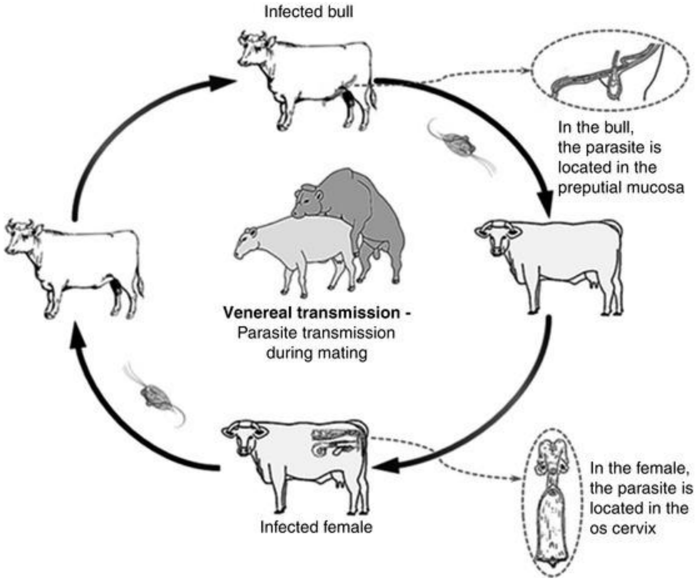
using google sheets for these plots

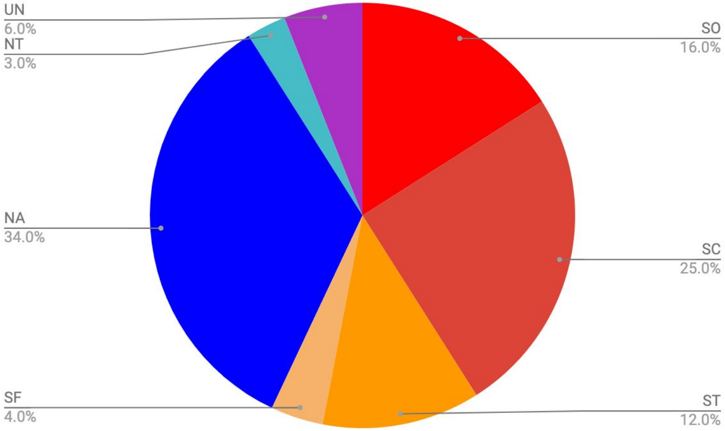
Quality control in original dataset

37% of the 100 “software” mentions in the random sample weren’t software, including a bovine STD!

19% were typos or duplicates

Sometimes different software has the same name


Infected bull
In the bull, the parasite is located in the preputial mucosa
Venereal transmission - Parasite transmission during mating
Infected female
In the female, the parasite is located in the os cervix



Code	Description
SO	Software where a link to a code repository can be found
SC	Software but no link to a code repository can be found
ST	Typo but the mention is to software
SF	Specific function / subroutine in a larger software package or library with a different name
NA	Not software but correctly spelt
NT	Not software but incorrectly spelt
UN	Other classification - unknown / needs further investigation

What’s next

- Create crowdsourcing instructions for contributors, software reviewers, users
- Release dataset and HackDay work as its own software & data package
- Inform development of training programmes to meet the needs of disciplines
- Support long-term projects led by the Research Software Alliance and SSI
- Contribute to Stephan’s PhD research
- Give feedback to Chan Zuckerburg Initiative about their dataset
- Come up with a better bacronym for Habeas than *“Helpfully Annotated Basic Encyclopedia of Academic Software”* - send us your suggestions!

Extra Slides

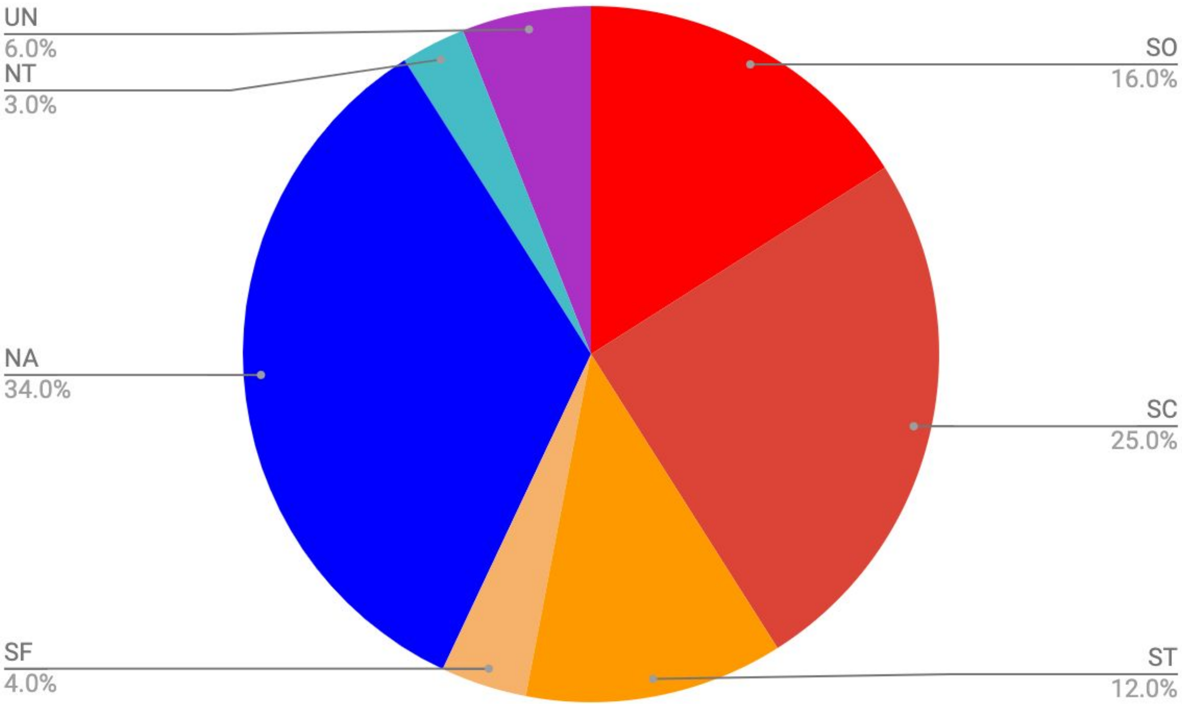
Open questions

What about publishing a programme as part of the package?

Reviewers and editors don't check for:

- correctness of software is cited or that it is available at the specified source

Distribution of software types



Code	Description
SO	Software where a link to a code repository can be found
SC	Software but no link to a code repository can be found
ST	Typo but the mention is to software
SF	Specific function / subroutine in a larger software package or library with a different name
NA	Not software but correctly spelt
NT	Not software but incorrectly spelt
UN	Other classification - unknown / needs further investigation

Limitations of cleaning algorithms:

Name clashes Spectra!

- Typo, multiple meanings so clumping together do not work

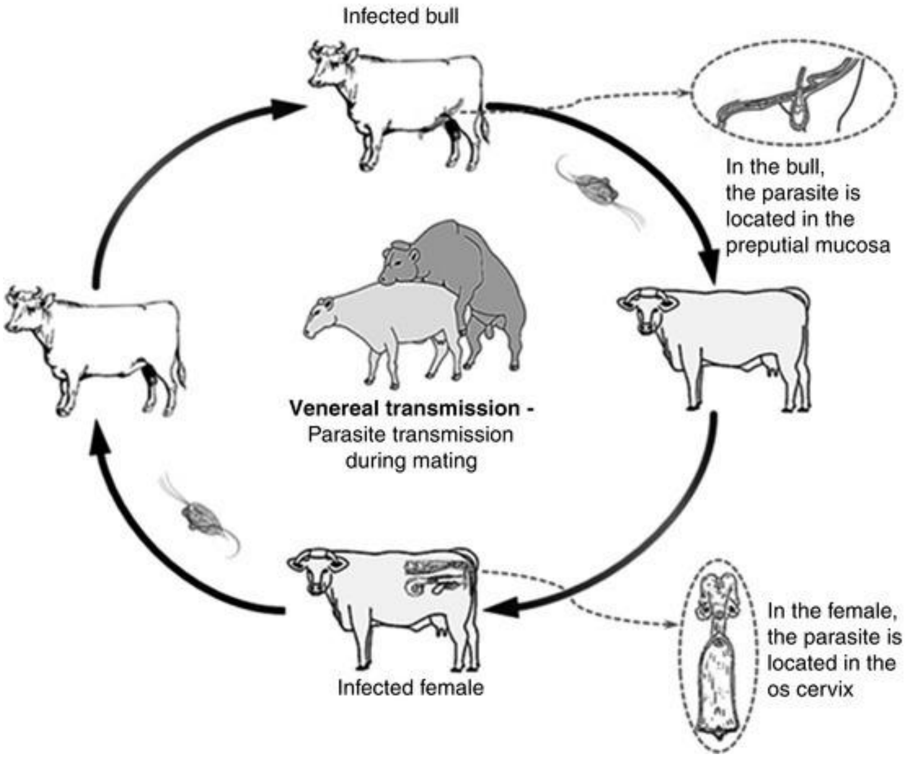
Same software

- https://github.com/Ensembl
- https://github.com/Ensembl/ensembl-client

Different software

Not a software

Quality Control issues identified


Infected bull
In the bull, the parasite is located in the preputial mucosa
Venereal transmission - Parasite transmission during mating
In the female, the parasite is located in the os cervix
Infected female

Life cycle of bovine Tritrichomonas foetus
DOI : 10.1007/978-3-319-70132-5_14

It’s not always software identified:

We caught Tritrichomonas, a bovine STD, moonlighting as a software!

We also found reports, methods, EU Funding Programmes (FP7), typos.

Also queries about how black holes contribute to COVID19…