High Performance
Language Technologies

Reproducibility in Natural Language Processing

Aurélie Névéol

HPLT Winter School, February 6, 2024

LISN
LABORATOIRE INTERDISCIPLINAIRE
DES SCIENCES DU NUMÉRIQUE
université
PARIS-SACLAY
cnrs

Where am I speaking from?

- ITN MiRoR "Methods in Research on Research" (2016-2020)
- Contributions to French research group « Éthique et TAL »
- Interest in reproducibility for clinical NLP
    - Involvement in "shared tasks" as a participant and organizer
    - Literature surveys and studies to further understanding of reproducibility

Research questions

- What are the challenges of reproducibility?
- How can reproducibility be increased in NLP systems?
- How can NLP help with reproducibility?

Challenges of reproducibility

Reproducibility is the essence of science

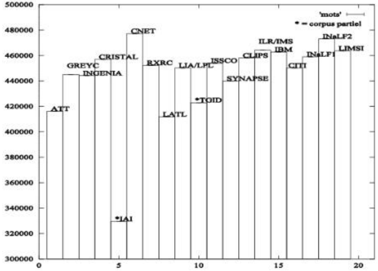
Reproducibility: Independently running a research experiment and yielding the same results on each iteration

Reproducibility has many perks

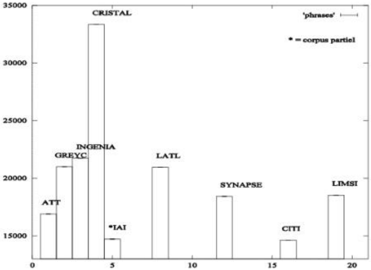
- Helps avoid disaster and move science forward
- Makes it easier to publish papers
- Helps you get your point across
- Enables continuity of your work
- Helps build your reputation, e.g. attracts more citations

Piwowar HA, Day RS, Fridsma DB. Sharing detailed research data is associated with increased citation rate. PLoS One. 2007 Mar 21;2(3):e308.
Markowetz F. Five selfish reasons to work reproducibly. Genome Biol. 2015 Dec 8;16:274. .

And yet...



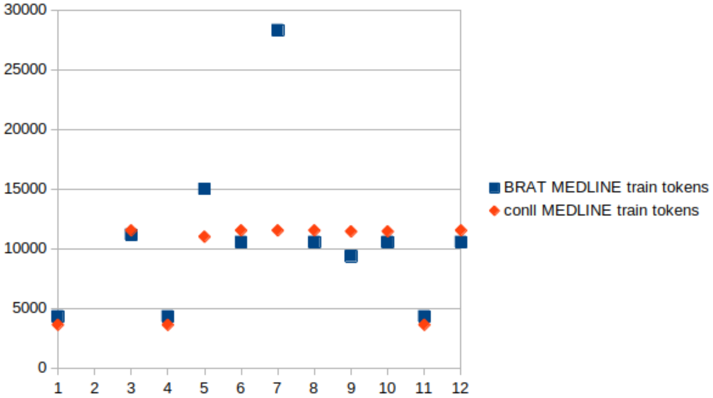
number of words



number of sentences

Adda G, Mariani J, Lecomte J, Paroubek P, Rajman M. 1998. The GRACE Part-Of-Speech Tagging Evaluation Task. Proc. LREC 1998

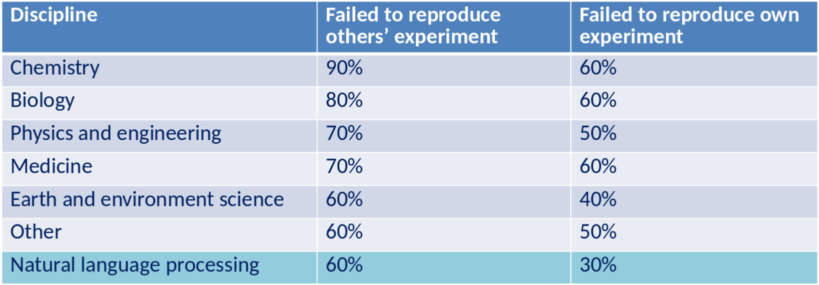
(Try it, you will be surprised!)



M2 Student lab assignment, "count tokens in corpus". 2024

A "reproducibility crisis"

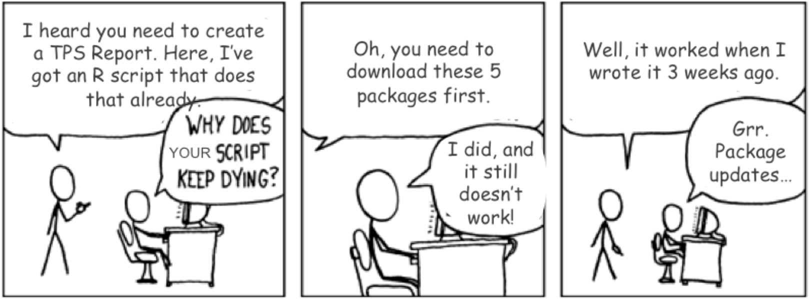
Surveys of 1,500 scientists (2016) and 225 NLP researchers (2019)



Discipline	Failed to reproduce others' experiment	Failed to reproduce own experiment
Chemistry	90%	60%
Biology	80%	60%
Physics and engineering	70%	50%
Medicine	70%	60%
Earth and environment science	60%	40%
Other	60%	50%
Natural language processing	60%	30%

Baker M. 1,500 scientists lift the lid on reproducibility. Nature. 2016 May 25;533(7604):452-4 .
Mieskes M, Fort K, Névéol A, Grouin C, Cohen KB. NLP Community Perspectives on Replicability. Proc. RANLP. 2019:768–775.

Reproducibility in practice


I heard you need to create a TPS Report. Here, I've got an R script that does that already.
WHY DOES YOUR SCRIPT KEEP DYING?
Oh, you need to download these 5 packages first.
I did, and it still doesn't work!
Well, it worked when I wrote it 3 weeks ago.
Grr. Package updates...

- The Zigglebottom Tagger: a prototypical reproducibility story
- ... without a villain!

Pedersen T. 2008. Empiricism is not a matter of faith. Computational Linguistics:34(3):465-470
Comic: inspired by XKCD

Sources of variability

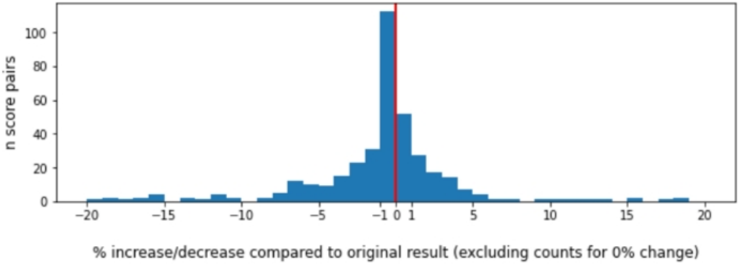
- Preprocessing, e.g. tokenization, stop-word lists
- “Data cleaning”, e.g. normalization of case, diacritics
- Software versions, system variations
- Parameters, including training/test split

Fokkens A, Van Erp M, Postma M, Pedersen T, Vossen P, Freire N. 2013. Offspring from Reproduction Problems: What Replication Failure Teaches Us. Proc ACL: 1691-1701

Documentation and archiving are key

- Availability of research material
    - Data protection (GDPR, copyright...)
    - Software protection (raise of closed source models)
    - Protocols lack details
- Reporting bias
    - page limits
    - Novelty valued over reproducibility

Overview of reproducibility in NLP



- Data and code are elusive (40% of papers with data in 2016)
- 14% of results reproduced

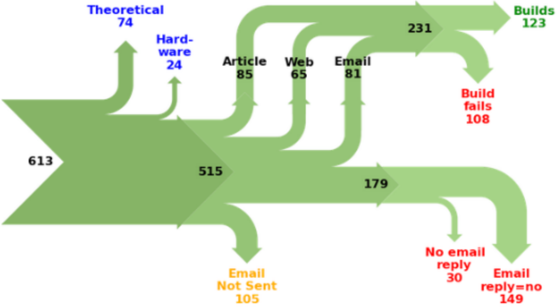
Belz A, Agarwal S, Shimorina A, Reiter E. A Systematic Review of Reproducibility Research in Natural Language Processing. EACL 2021:381–393
Mieskes M. A quantitative study of data in the NLP community. Proc ACL Workshop on Ethics in NLP. 2017

Material availability in biomedical NLP

- Study of BioNLP 2016 proceedings
  - 48% of papers contained links to data, 61% links to code, 21% to both data and code
  - Inter-annotator agreement was .57 for data, .63 for code

Cohen KB, Névéol A, Xia J, Hailu N, Hunter L ,Zweigenbaum P. Reproducibility in Biomedical Natural Language Processing. Proc AMIA Annu Symp. 2017.

Code usability



- Studied 613 articles from 8 ACM conferences
- Steps towards using the code to reproduce experiments:
  - locate, obtain, install « 25 %

Christian Collberg and Todd Proebsting. “Repeatability in Computer Systems Research,” CACM 59(3):62-69.2016.

Dimensions of reproducibility

- Reproducibility of a **value**
  - some experiments are non deterministic, e.g. using deep learning models

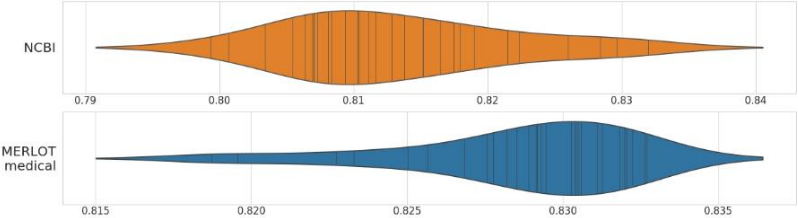


Image source: Tourille et al. LOUHI 2018

Cohen KB, Xia JB, Zweigenbaum P, Callahan T, Hargraves O, Goss F, Ide N, Névéol A, Grouin C, Hunter LE. Three Dimensions of Reproducibility in Natural Language Processing. LREC 2018. 2018:156-165.

Dimensions of reproducibility

- Reproducibility of a **value**
  - some experiments are non deterministic, e.g. using deep learning models
- Reproducibility of a **finding**
  - different values can yield the same finding, e.g. A > B
- Reproducibility of a **conclusion**
  - conclusions are an interpretation of findings

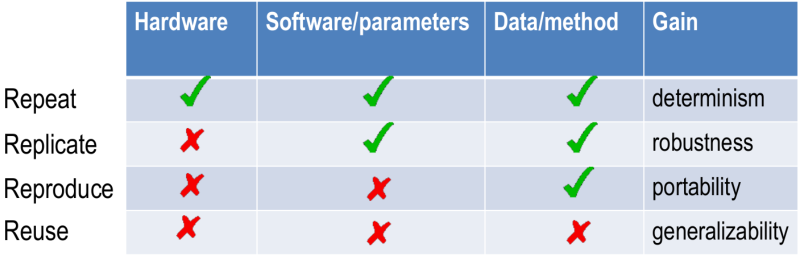
Cohen KB, Xia JB, Zweigenbaum P, Callahan T, Hargraves O, Goss F, Ide N, Névéol A, Grouin C, Hunter LE. Three Dimensions of Reproducibility in Natural Language Processing. LREC 2018. 2018:156-165.

Increasing reproducibility

Concrete steps

- Understanding the stakes
  - Experiments, literature reviews
- Facilitating levers
  - Sharing data, tools, workflows...
  - Shared tasks
- Documentation
  - detailing protocols, elliciting audits
  - reporting guidelines

Contributions of experiment features to reproducibility



	Hardware	Software/parameters	Data/method	Gain
Repeat	✓	✓	✓	determinism
Replicate	✗	✓	✓	robustness
Reproduce	✗	✗	✓	portability
Reuse	✗	✗	✗	generalizability

Cohen-Boulakia S, Belhajjame K, Collin O, et al.Scientific workflows for computational reproducibility in the life sciences: status, challenges and opportunities. Future Gen Comput Syst 2017; 75: 284–98.

Shared tasks foster reproducibility

- Primary goal is to provide a forum for direct comparison of approaches
- Research material shared with the community
  - Definition of a "task"
  - Annotated corpus with train/dev/test splits
  - Evaluation metrics and scripts or framework

Reproducibility study at CLEF eHealth 2016

Protocol

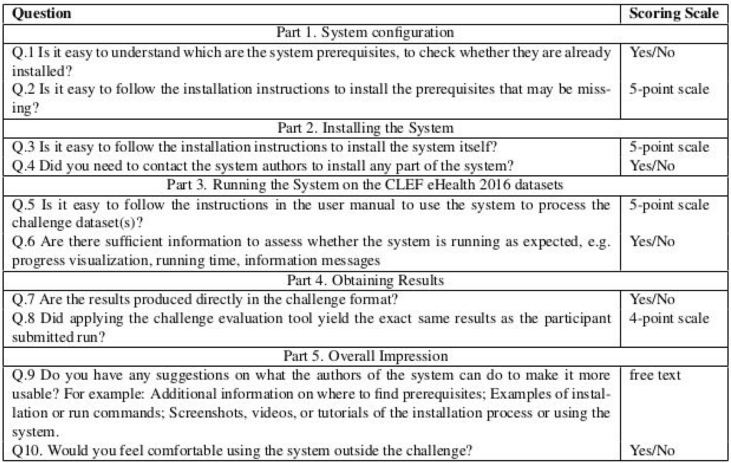
CLEF eHealth

- Participants to an ICD10 coding task could submit their system for replication of results
  - 3 system submissions (out of 7 participants)
  - 4 analysts committed to reproduce results in their usual working environment
- Evaluation criteria
  - Evaluation check-list covering install/run/results
  - Replication time

Névéol A, Cohen KB, Grouin C, Robert A. Replicability of Research in Biomedical Natural Language Processing: a pilot evaluation for a coding task. Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis, LOUHI. 2016:78-84.

Reproducibility study at CLEF eHealth 2016

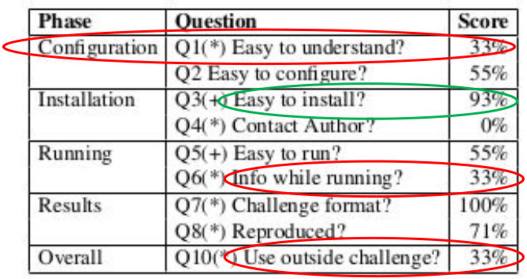
Scoring sheet



Question	Scoring Scale
Part 1. System configuration	
Q.1 Is it easy to understand which are the system prerequisites, to check whether they are already installed?	Yes/No
Q.2 Is it easy to follow the installation instructions to install the prerequisites that may be missing?	5-point scale
Part 2. Installing the System	
Q.3 Is it easy to follow the installation instructions to install the system itself?	5-point scale
Q.4 Did you need to contact the system authors to install any part of the system?	Yes/No
Part 3. Running the System on the CLEF eHealth 2016 datasets	
Q.5 Is it easy to follow the instructions in the user manual to use the system to process the challenge dataset(s)?	5-point scale
Q.6 Are there sufficient information to assess whether the system is running as expected, e.g. progress visualization, running time, information messages	Yes/No
Part 4. Obtaining Results	
Q.7 Are the results produced directly in the challenge format?	Yes/No
Q.8 Did applying the challenge evaluation tool yield the exact same results as the participant submitted run?	4-point scale
Part 5. Overall Impression	
Q.9 Do you have any suggestions on what the authors of the system can do to make it more usable? For example: Additional information on where to find prerequisites; Examples of installation or run commands; Screenshots, videos, or tutorials of the installation process or using the system.	free text
Q10. Would you feel comfortable using the system outside the challenge?	Yes/No

Reproducibility study at CLEF eHealth 2016

Scoring results



Phase	Question	Score
Configuration	Q1(*) Easy to understand?	33%
	Q2 Easy to configure?	55%
Installation	Q3(+) Easy to install?	93%
	Q4(*) Contact Author?	0%
Running	Q5(+) Easy to run?	55%
	Q6(*) Info while running?	33%
Results	Q7(*) Challenge format?	100%
	Q8(*) Reproduced?	71%
Overall	Q10(*) Use outside challenge?	33%

Reproducibility study at CLEF eHealth 2016

Results

CLEF eHealth

- Results were reproduced! But:
  - No single analyst was able to reproduce all results
  - Reproduction time greatly varied
- Reproducibility is challenging
  - Everyone should experience it
- Reproducibility requires resources
  - For authors to produce quality, documented systems
  - For users to understand and efficiently deploy

Névéol A, Cohen KB, Grouin C, Robert A. Replicability of Research in Biomedical Natural Language Processing: a pilot evaluation for a coding task. Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis, LOUHI. 2016:78-84.

Towards actionable reproducibility in clinical NLP

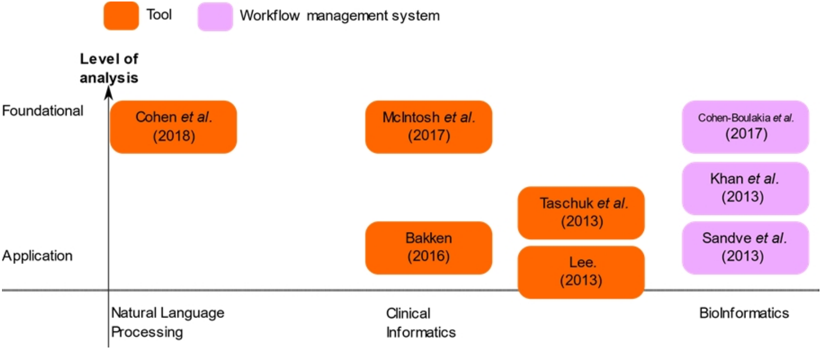
From research to hospital operations

- Need for standardization, traceability, automation
  - NLeveraging expertise and experience accross disciplines
- Reproducibility criteria/desiderata expressed by the bioinformatics, medical informatics, NLP communities Characterize clinical NLP w. r. t. reproducibility
- Analysis of 7 clinical NLP systems (for English)

Digan W, Névéol A, Neuraz A, Wack M, Baudoin B, Burgun A, Rance B. Can reproducibility be improved in clinical natural language processing? A study of 7 clinical NLP suites. J Am Med Inform Assoc. 2021. Mar 1;28(3):504-515.

Identifying reproducibility criteria from the litterature

MEDLINE search and snowballing



🟧 Tool 🟪 Workflow management system

Level of analysis

Foundational

Application

Cohen *et al.* (2018)

McIntosh *et al.* (2017)

Bakken (2016)

Taschuk *et al.* (2013)

Lee. (2013)

Cohen-Boulakia *et al.* (2017)

Khan *et al.* (2013)

Sandve *et al.* (2013)

Natural Language Processing

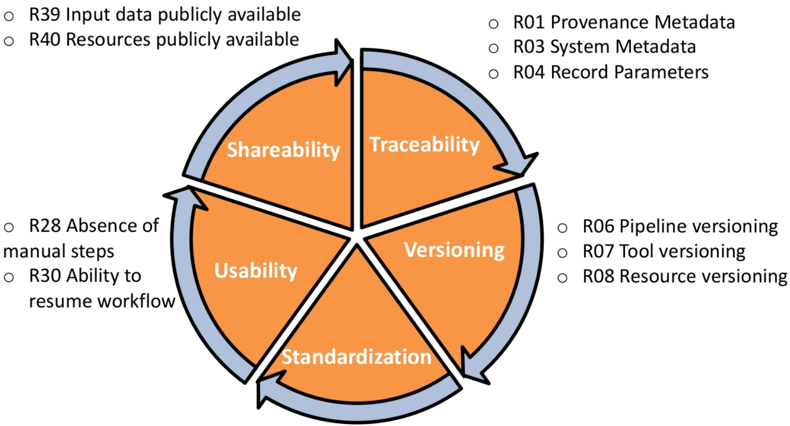
Clinical Informatics

BioInformatics

Digan W, Névéol A, Neuraz A, Wack M, Baudoin B, Burgun A, Rance B. Can reproducibility be improved in clinical natural language processing? A study of 7 clinical NLP suites. J Am Med Inform Assoc. 2021. Mar 1;28(3):504-515.

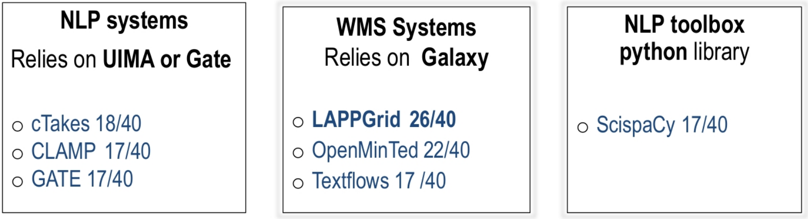
Reproducibility criteria from the litterature in three fields

bioinformatics, medical informatics and NLP


○ R39 Input data publicly available
○ R40 Resources publicly available
○ R01 Provenance Metadata
○ R03 System Metadata
○ R04 Record Parameters
Shareability
Traceability
○ R28 Absence of manual steps
○ R30 Ability to resume workflow
Usability
Versioning
○ R06 Pipeline versioning
○ R07 Tool versioning
○ R08 Resource versioning
Standardization

Digan W, Névéol A, Neuraz A, Wack M, Baudoin B, Burgun A, Rance B. Can reproducibility be improved in clinical natural language processing? A study of 7 clinical NLP suites. J Am Med Inform Assoc. 2021. Mar 1;28(3):504-515.

Evaluating clinical NLP systems



NLP systems

Relies on **UIMA or Gate**

- ○ cTakes 18/40
- ○ CLAMP 17/40
- ○ GATE 17/40

WMS Systems

Relies on **Galaxy**

- ○ **LAPPGrid 26/40**
- ○ OpenMinTed 22/40
- ○ Textflows 17 /40

NLP toolbox

python library

- ○ ScispaCy 17/40

Reproducibility can be improved

- Especially versioning, standardization and shareability
- Experience from bioinformatics suggests modularity and workflows can help

Medkit library for French clinical NLP

- `https://github.com/TeamHeka/medkit`

Digan W, Névéol A, Neuraz A, Wack M, Baudoin B, Burgun A, Rance B. Can reproducibility be improved in clinical natural language processing? A study of 7 clinical NLP suites. J Am Med Inform Assoc. 2021. Mar 1;28(3):504-515.

Use of reporting guidelines in health research

MiRoR
Methods in
Research on Research

- Reporting Guidelines are recent tools
  - Majority have not been assessed for efficiency of reporting improvement
  - CONSORT have been shown to improve completeness of reporting
  - A systematic review reports that overall adherence to guidelines is suboptimal
- Impact of Reporting Guidelines
  - Before/ After conducting a study
  - Training, Understanding, Implementing, Monitoring, Collaborating

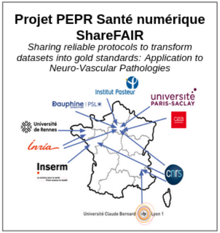
Blanco D, Altman D, Moher D, Boutron I, Kirkham JJ, Cobo E. Scoping review on interventions to improve adherence to reporting guidelines in health research. BMJ Open. 2019 May 9;9(5):e026589.

Natural Language Processing and Reporting Guidelines

- NLP could facilitate adherence to reporting guidelines
  - Automatically assess guideline compliance
  - Match guideline item with implementation in manuscript
- Guidelines for reporting (bio)NLP research?
  - ACL reproducibility checklist
  - Beware of checklists...

Using NLP to analyze worflow reproducibility

On-going work by C. Sebe and ShareFAIR project members

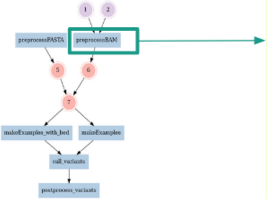


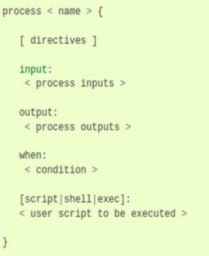
Projet PEPR Santé numérique ShareFAIR

Sharing reliable protocols to transform datasets into gold standards: Application to Neuro-Vascular Pathologies

Using NLP to analyze worflow reproducibility

A workflow is a sequence of processes that enables bioinformatics analysis, through tool and file manipulations.

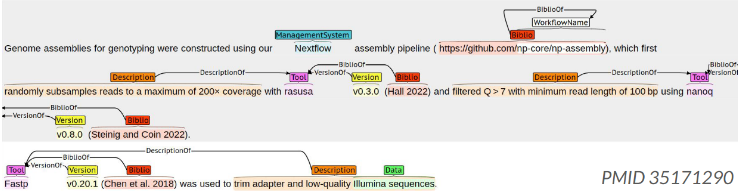




```
process < name > {

   [ directives ]

   input:
    < process inputs >

   output:
    < process outputs >

   when:
    < condition >

   [script|shell|exec]:
   < user script to be executed >

}
```

- Goal: integration of data extracted from code and article

Modeling workflow information through annotation



Genome assemblies for genotyping were constructed using our Nextflow assembly pipeline (https://github.com/np-core/np-assembly), which first randomly subsamples reads to a maximum of 200× coverage with rasusa v0.3.0 (Hall 2022) and filtered Q > 7 with minimum read length of 100 bp using nanoq v0.8.0 (Steinig and Coin 2022).

Fastp v0.20.1 (Chen et al. 2018) was used to trim adapter and low-quality Illumina sequences.

PMID 35171290

- annotation scheme includes 16 entities, 10 relations
- 24 articles annotated by 3 annotators with complementary expertise
- inter-annotator agreement .70
- Entity recognition using NlStruct and SciBERT yields .72 F1

Sebe C, Névéol A, Cohen-Boulakia S, Gaignard A. Extraction d'informations sur les workflows scientifiques à partir de la littérature. EGC RNTI-E-39. 2023:313-320

Acknowledgements

Funding and support: anr, MARIE CURIE ACTIONS, CLEF ASSOCIATION

Collaboration and inspiration:

- Andreas Rauber (Vienna University of Technology)
- Kevin B. Cohen (University of Colorado)
- Cyril Grouin (LISN-CNRS), Aude Robert (INSERM/CépiDC)
- Patrick Paroubek and Pierre Zweigenbaum (LISN-CNRS)
- William Digan and Bastien Rance (APHP/Inserm)
- Margot Mieskes (Darmstat), Karën Fort (LORIA- Sorbonne)
- Sarah Cohen-Boulakia, Olivier Ferret (CEA), Clémence Sebe (LISN, U Paris-Saclay)

Summary

Reproducibility is hard to achieve!

It's not just the result... also the journey

NLP can help improve reproducibility