Improving reproducibility by automating key resource tables

Zubair Afzal, Marleen Sta, Marialaura Martinico, Daniel Gregory,
Lekhraj Sharma, Ramsundhar Baskaravelu, George Tsatsaronis
Elsevier, The Netherlands

Keywords: STAR Methods, Key resource table, Reproducibility, Accessible Research

INTRODUCTION

Reproducibility in research has gained a lot of interest since the publication of John Ioannidis' landmark article [1]. In a 2016 study [2], 90% of the researchers who participated in the survey showed concerned about 'reproducibility crisis' in research. This is largely due to incomplete and at times poor reporting of the methods and materials used in the research. About 60% of researchers in Biology and Chemistry domain failed to reproduce their own results and in Chemistry, about 87% researchers failed to reproduced results published by others [2]. Similar reproducibility trends are observed in other fields as well [3]. The cost of irreproducible preclinical research is estimated to be USD 28 billion per year in the United States alone [4]. To help address concerns from the research community around rigor and reproducibility, Cell Press launched STAR Methods [5] in 2016. The STAR (Structured, Transparent, Accessible Reporting) Methods aims at promoting sharing of resources, data, code, and software by making the methods more complete, transparent, easy to follow and understand. An important element of the STAR Methods is the Key Resource Table (KRT), which provides a structured way of reporting resources or reagents with catalogue and standard identifiers. At the time of writing, the STAR Methods participating journals requires authors to provide a KRT next to the original submission, which is largely a manual effort and takes time and effort on authors' end. In this study, we looked at helping the authors by generating the KRT in an automated fashion and asking the authors only to validate the table and enter corrections where necessary. We hypothesize that semi-automating the KRT using machine learning techniques will increase uptake for authors by saving the author valuable time. This would result in having more reproducible research available to the researchers.

METHODS

A Key Resource Table (KRT) summarizes the critical resources and materials used in the manuscript, such as antibodies, bacterial and virus strains, biological samples, chemical, peptides, and recombinant proteins, cell lines, oligonucleotides, recombinant DNAs, and, software and algorithms. These reagents and resources are typically described in the methods section of an article. Therefore, for an automatic KRT system to be efficient and effective, it is important that the method section in the article is first identified properly. Since a method section may have different headings in different journals, we developed a heuristic based *Method Section Extractor* which identifies the beginning and the ending of the method section in the article. Identifying mentions of reagents or resources in the method text is essentially a Named Entity Recognition (NER) task. In this study, we focused on five main entities of the KRT, namely: (1) antibodies, (2) cell lines, (3) chemicals, (4) proteins, and, (5) peptides. Figure 1 shows the automatic KRT generation pipeline developed in this study.

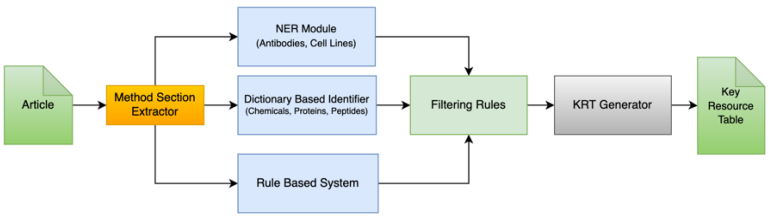

Figure 1: Automatic KRT generation pipeline

The NER module contains two machine learning models based on CRF (Conditional Random Fields) that were trained to recognize antibodies and cell lines. To train the models, we first asked two Subject Matter Experts (SMEs) to manually label the mentions of antibodies in 800 randomly selected articles. 80% of the labeled set was used to train the model and 20% for testing. Similarly, for cell lines, SMEs labeled 1,000 randomly selected articles, where 90% of them were used to train the model and 10% for testing. For identifying chemicals, proteins, and peptides, we used an indexing engine with several relevant dictionaries. Lastly, we also developed a rule-based system which captures the most straightforward cases of all the above entities. To deal with the false positives, a post processing filtering module was developed to improve system precision across all entities. Once all the entities are identified, we generated the KRT table which is then presented to the author for validation by the new submission wizard in the Elsevier Submission System (EES). The pilot ran for 6 months on 10 different journals.

RESULTS

We evaluated our NER models for antibodies and cell lines on an independent test set and calculated precision, recall, and f1-scores. The antibodies model had a precision of 0.93, recall of 0.75, and an f1-score of 0.83. Similarly, the cell lines model had a precision of 0.92, recall of 0.76, and f1-score of 0.84. In total, 9,291 submissions were processed during the 6 month pilot period. The system extracted a method section for 83% of the articles correctly. Of all the generated KRTs, 53% were accepted by the authors. Authors added additional entities to 9% of the generated KRTs and removed some entities from 17% KRTs.

CONCLUSION

STAR Methods is an important step in making the research more accessible and reproducible. The Key Resource Table (KRT), which is at the heart of the STAR Methods, requires authors to enter all reagents and resources used in the study manually. In journals where the KRT is an optional submission item, the current author uptake was only 17%. To increase author uptake, we implemented an automated KRT generation system where authors are only required to validate the entries which requires significantly less effort than creating a new table manually. All components of the system showed high precision and recall. In a 6-month live pilot with 10 subscribed journals, we observed an average author uptake of 53%, an increase of more than 300%. The pilot results indicate that the authors appreciate having a semi-automated approach to the KRT rather than having to manually create one from scratch. Apart from the obvious advantage of enabling clear reporting of methods and results and facilitating better study designs, such automation steps by scientific publishers can improve reproducibility and replicability of published research.

REFERENCES

[1] J. P. A. Ioannidis, “Why Most Published Research Findings Are False,” *PLoS Med.*, vol. 2, no. 8, p. e124, Aug. 2005, doi: 10.1371/journal.pmed.0020124.

[2] M. Baker, “1,500 scientists lift the lid on reproducibility,” *Nature*, vol. 533, no. 7604, 2016, doi: 10.1038/533452a.

[3] J. R. Stevens, “Replicability and reproducibility in comparative psychology,” *Front. Psychol.*, vol. 8, no. MAY, 2017, doi: 10.3389/fpsyg.2017.00862.

[4] L. P. Freedman, I. M. Cockburn, and T. S. Simcoe, “The economics of reproducibility in preclinical research,” *PLoS Biol.*, vol. 13, no. 6, 2015, doi: 10.1371/journal.pbio.1002165.

[5] C. Press, “STAR Methods Guide for Authors,” *Cell*, no. 1, 2016.