Supplementary Material 2: Data Extraction Supplement

Science Citation Knowledge Extractor

Heather Lent, Gustave Hahn-Powell, Asher Haug-Baltzell, Sean Davey, Mihai Surdeanu, Eric Lyons\*

**\* Correspondence:** Eric Lyons: ericlyons@email.arizona.edu

# Data Extraction Workflow



**Supplemental Figure**: Data Extraction Workflow. After a user submits a query (1 or more PubMed IDs to analyze), SCKE uses the Entrez API to obtain bibliographic data (e.g. citations, dates, etc.) and textual data (full text manuscripts and/or abstract). The BioNLP Processor annotates the text with features such as named entities (NER), part-of-speech tags, and lemmas. These annotations are used to generate the final datasets, which are presented through interactive visualizations.

# Extracted Data Examples

The following tables illustrate the data that is collected and stored in the database. Each example corresponds to one row in a different table.

|  |
| --- |
| Input Table |
| post\_id | 50 |
| datestamp | 2018-11-12 |
| pmid | 9108111 |
| title | Motor-driven gene expression |
| author | Jarvis ED, Nottebohm F |
| journal | Proceedings of the National Academy of Sciences of the United States of America |
| pubdate | 1997 Apr 15 |
| url | <http://www.ncbi.nlm.nih.gov/pubmed/9108111> |
| num\_citations | 111 |
| abstract | NA |
| whole\_article | NA |
| pmcid | NA |

|  |
| --- |
| Citations Table |
| post\_id | 26538 |
| datestamp | 2018-10-23 |
| pmicd | 6152990 |
| title | Vocal practice regulates singing activity–dependent genes underlying age-independent vocal learning in songbirds |
| author | Hayase S, Wang H, Ohgushi E, Kobayashi M, Mori C, Horita H, Mineta K, Liu WC, Wada K |
| journal | PLoS Biology |
| pubdate | 2018 Sep 12 |
| citesPmid | 9108111 |
| url | https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6152990 |
| abstract | yes |
| whole\_article | yes |
| sents | 350 |
| tokens | 11580 |
| annotated | yes |

|  |
| --- |
| Queries Table |
| post\_id | 145 |
| datestamp | 2018-04-18 |
| query | 9108111 |
| range\_years | 1997+2018 |
| total\_pubs | 108 |
| unique\_pubs | 108 |
| unique\_journals | 55 |
| num\_abstracts | 107 |
| num\_whole\_articles | 107 |
| num\_sents | 10391 |
| num\_tokens | 372182 |