Exuberanter: A Versatile Python Tool for Literature Data Extraction

Email: wenne@chalmers.se
Phone: +46 725 039 395

Teo Lovmar[1], Johan Bengtsson-Palme[1, 2, 3] and Marcus Wenne[1, 2]

CHALMERS

Motivation

- **Efficiency Need:** Literature reviews are time-consuming and demand substantial manual effort.
- **Reproducibility Concern:** Manual literature review outcomes can vary and may lack reproducibility.
- **Knowledge Expansion:** The rapid growth of scientific literature amplifies these challenges.
- **Our Solution:** We provide a tool for efficient, reproducible literature reviews, with semi-automatic data extraction capabilities and an intuitive user interface.

Features

- **Efficient Literature Review:** Optimize your literature surveying with our semi-automated approach.
- **Downloading articles:** Simplify the task of obtaining, and assessing research articles from PubMed.
- **User customization:** Specify what type of information you would like to extract from the different sections of the articles.
- **Data Extraction:** Extract data using regex and plot data extraction tools in a semi-automatic fashion.
- **User Interface:** Navigate and interact with the processed data in a straightforward, user-friendly manner.

Workflow

Pre defined search query

First the user specifies one or multiple query terms which will be used to search PubMed.

eg. '"antibiotic resistance genes" AND q-PCR'

Download articles and their metadata

Using the NCBI API Efetch, Exuberanter downloads all available open access articles, and their accompanying metadata.

Articles not available through the API can be manually downloaded and accessed by Exuberanter.

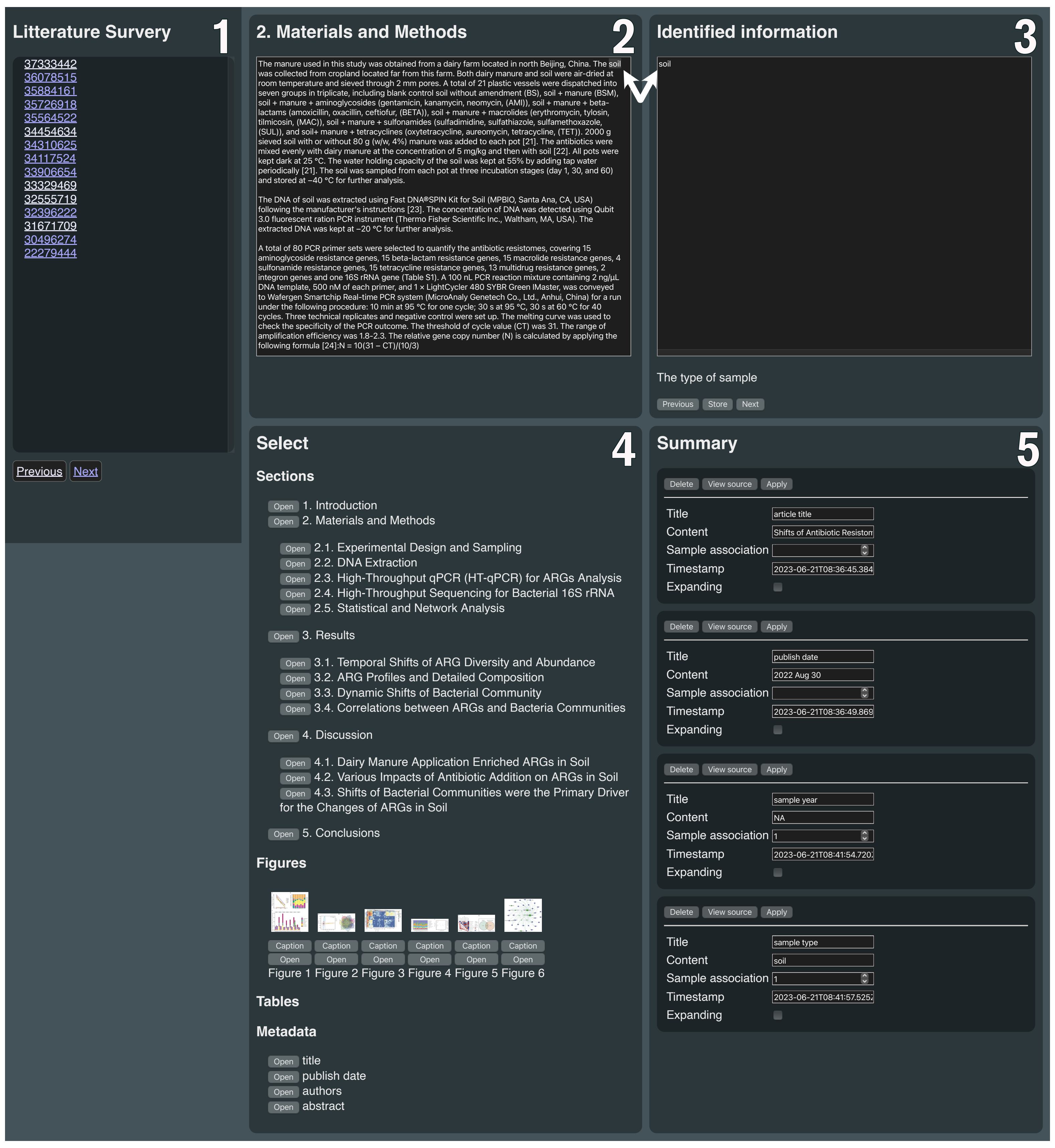
Extract pre specified data from articles

The user specifies what type of information Exuberanter should extract using regex or an NLM API (to come). All data is stored in a JSON file.

Verify or modify the extracted data via a UI

Via the UI the user can manually verify the pre-extracted data, or manually change it. The tool has WebPlotDigitizer (*https://automeris.io/WebPlotDigitizer*) integrated which facilitates plot data extraction.

Interface



Litterature Survery 1

37333442
36078515
35884161
35726918
35564522
34454634
34310625
34117524
33906654
33329469
32555719
32396222
31671709
30496274
22279444

Previous | Next

2. Materials and Methods 2

The manure used in this study was obtained from a dairy farm located in north Beijing, China. The soil was collected from cropland located far from this farm. Both dairy manure and soil were air-dried at room temperature and sieved through 2 mm pores. A total of 21 plastic vessels were dispatched into seven groups in triplicate, including blank control soil without amendment (BS), soil + manure (BSM), soil + manure + aminoglycosides (gentamicin, kanamycin, neomycin, (AMI)), soil + manure + beta-lactams (amoxicillin, oxacillin, ceftiofur, (BETA)), soil + manure + macrolides (erythromycin, tylosin, tilmicosin, (MAC)), soil + manure + sulfonamides (sulfadimidine, sulfathiazole, sulfamethoxazole, (SUL)), and soil+ manure + tetracyclines (oxytetracycline, aureomycin, tetracycline, (TET)). 2000 g sieved soil with or without 80 g (w/w, 4%) manure was added to each pot [21]. The antibiotics were mixed evenly with dairy manure at the concentration of 5 mg/kg and then with soil [22]. All pots were kept dark at 25 °C. The water holding capacity of the soil was kept at 55% by adding tap water periodically [21]. The soil was sampled from each pot at three incubation stages (day 1, 30, and 60) and stored at −40 °C for further analysis.

The DNA of soil was extracted using Fast DNA®SPIN Kit for Soil (MPBIO, Santa Ana, CA, USA) following the manufacturer's instructions [23]. The concentration of DNA was detected using Qubit 3.0 fluorescent ration PCR instrument (Thermo Fisher Scientific Inc., Waltham, MA, USA). The extracted DNA was kept at −20 °C for further analysis.

A total of 80 PCR primer sets were selected to quantify the antibiotic resistomes, covering 15 aminoglycoside resistance genes, 15 beta-lactam resistance genes, 15 macrolide resistance genes, 4 sulfonamide resistance genes, 15 tetracycline resistance genes, 13 multidrug resistance genes, 2 integron genes and one 16S rRNA gene (Table S1). A 100 nL PCR reaction mixture containing 2 ng/μL DNA template, 500 nM of each primer, and 1 × LightCycler 480 SYBR Green IMaster, was conveyed to Wafergen Smartchip Real-time PCR system (MicroAnaly Genetech Co., Ltd., Anhui, China) for a run under the following procedure: 10 min at 95 °C for one cycle; 30 s at 95 °C, 30 s at 60 °C for 40 cycles. Three technical replicates and negative control were set up. The melting curve was used to check the specificity of the PCR outcome. The threshold of cycle value (CT) was 31. The range of amplification efficiency was 1.8-2.3. The relative gene copy number (N) is calculated by applying the following formula [24]:N = 10(31 − CT)/(10/3)

Identified information 3

soil

The type of sample

Previous | Store | Next

Select 4

Sections

Open 1. Introduction
Open 2. Materials and Methods
Open 2.1. Experimental Design and Sampling
Open 2.2. DNA Extraction
Open 2.3. High-Throughput qPCR (HT-qPCR) for ARGs Analysis
Open 2.4. High-Throughput Sequencing for Bacterial 16S rRNA
Open 2.5. Statistical and Network Analysis
Open 3. Results
Open 3.1. Temporal Shifts of ARG Diversity and Abundance
Open 3.2. ARG Profiles and Detailed Composition
Open 3.3. Dynamic Shifts of Bacterial Community
Open 3.4. Correlations between ARGs and Bacteria Communities
Open 4. Discussion
Open 4.1. Dairy Manure Application Enriched ARGs in Soil
Open 4.2. Various Impacts of Antibiotic Addition on ARGs in Soil
Open 4.3. Shifts of Bacterial Communities were the Primary Driver for the Changes of ARGs in Soil
Open 5. Conclusions

Figures

Caption Open Figure 1 | Caption Open Figure 2 | Caption Open Figure 3 | Caption Open Figure 4 | Caption Open Figure 5 | Caption Open Figure 6

Tables

Metadata

Open title
Open publish date
Open authors
Open abstract

Summary 5

Delete | View source | Apply

Title	article title
Content	Shifts of Antibiotic Resiston
Sample association	
Timestamp	2023-06-21T08:36:45.384
Expanding	□

Delete | View source | Apply

Title	publish date
Content	2022 Aug 30
Sample association	
Timestamp	2023-06-21T08:36:49.869
Expanding	□

Delete | View source | Apply

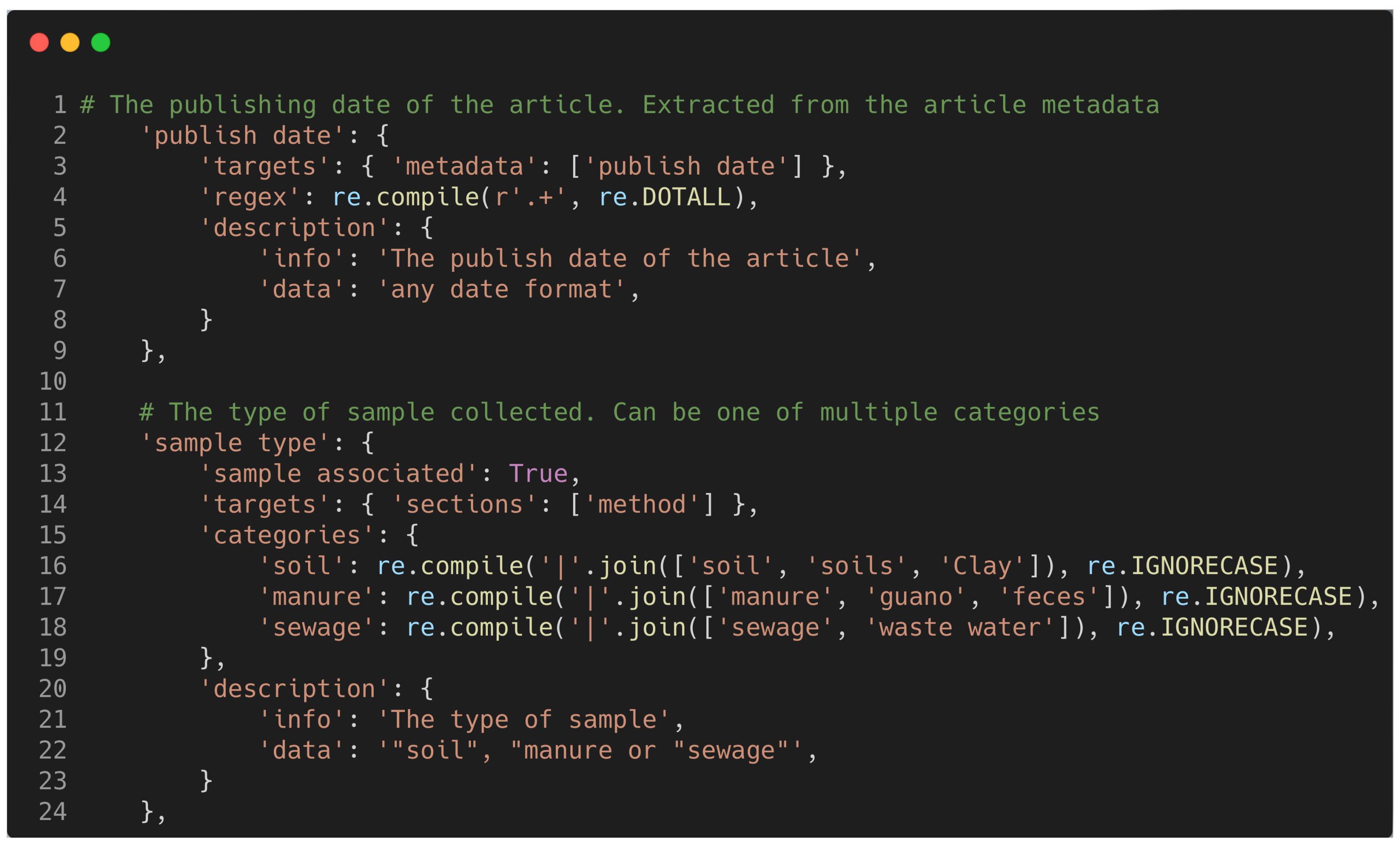
Title	sample year
Content	NA
Sample association	1
Timestamp	2023-06-21T08:41:54.720
Expanding	□

Delete | View source | Apply

Title	sample type
Content	soil
Sample association	1
Timestamp	2023-06-21T08:41:57.525
Expanding	□

1. Navigation between individual articles (PID).
2. Article text where data points identified by the tool are highlighted (left arrow).
3. Either automatically detected, or manually specify values for each datapoint (right arrow).
4. The various article sections, figures or tables can be selected here.
5. Individual datapoints collected during the literature survey.

Code examples



```python
# The publishing date of the article. Extracted from the article metadata
    'publish date': {
        'targets': { 'metadata': ['publish date'] },
        'regex': re.compile(r'.+', re.DOTALL),
        'description': {
            'info': 'The publish date of the article',
            'data': 'any date format',
        }
    },

    # The type of sample collected. Can be one of multiple categories
    'sample type': {
        'sample associated': True,
        'targets': { 'sections': ['method'] },
        'categories': {
            'soil': re.compile('|'.join(['soil', 'soils', 'Clay']), re.IGNORECASE),
            'manure': re.compile('|'.join(['manure', 'guano', 'feces']), re.IGNORECASE),
            'sewage': re.compile('|'.join(['sewage', 'waste water']), re.IGNORECASE),
        },
        'description': {
            'info': 'The type of sample',
            'data': '"soil", "manure or "sewage"',
        }
    },
```

GitHub

https://github.com/mnyt-aqw/Exuberanter/

Help!

We are currently having problems **extracting data from tables** due to the large variability in how journals format them.

Tips on how to solve this are greatly appreciated!

1. Division of Systems and Synthetic Biology, Department of Life Sciences, SciLifeLab, Chalmers University of Technology, Gothenburg, Swede
2. Centre for Antibiotic Resistance Research in Gothenburg (CARe), Gothenburg, Sweden
3. Department of Infectious Diseases, Institute of Biomedicine, The Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden