Goldschmidt 2021 Abstract
https://doi.org/10.7185/gold2021.7293

Making our resources more minable: Web-scraping and standardization of data formatting

DR. ERIN L MARTIN[1], VITOR BARROTE[2] AND PETER A CAWOOD[1]

[1]Monash University

[2]Ruhr-Universität Bochum

Presenting Author: erin.martin@monash.edu

Scientific journal publishers are continuously increasing the volume of supplementary data in their catalogues hosted on composite data repository websites. These repositories are an invaluable resource, particularly with the restrictions on data collection from physical samples, imposed by the COVID-19 pandemic. Furthermore, repositories can be web-scraped in minutes using simple python scripts, saving substantial time in data gathering. Our python web-scraping code connects to a data repository and carries out a keyword search. Articles satisfying the search criteria are returned and their metadata are input into a relational database while all associated supplementary files are downloaded. The program collects metadata and supplementary files of ~100 articles in <2 minutes.

Cleaning of the collected data occurs prior to data-basing. Data cleaning is an important step in data collection but substantially slows down the process due to inconsistent formatting, leading to issues in automated data extraction coding.

The main obstacles include but are not limited to: 1) supplementary files in different file formats, including tables in portable document files (pdfs), which can introduce errors, especially if the pdfs are not composed of editable text (e.g. scanned images). 2) Metadata that should be incorporated in big-data collections are often not included within the supplementary material. Information such as sampling coordinates, analytical method, analysed phase, QA/QC, and the type of uncertainty reported are often limited to methods sections of articles. 3) Formatting of data tables is inconsistent, and unsuitable for automated data extraction. For example, when extracting U-Pb age data from a table, a piece of code may search for the term $<^{206}Pb/^{238}U$ Age> as a column header in a specified row. But, tables are often formatted so that “age” is included in a separate cell, the ratio is split over two rows, or a shorthand version of the ratio is written (i.e. 6/8), yielding errors.

Given the recent efforts dedicated towards data-mining and machine learning in the geosciences, and the sheer volume of new geochemical data produced annually, it is important that we adopt a community-driven standardized format for data reporting, in order to maximize the information extracted from a given dataset.