**Final Project Proposal**

Yarroll, La Monte Henry Piggy / Osuga, Shintaro / Murphy, Christopher / Balasi, Jennifer

We propose conducting our final project on a comprehensive collection of scientific articles focused on mycology. Our dataset consists of over 6,000 species descriptions, compiled through extensive research by the Imperial Institute of Agricultural Research in collaboration with the Mycological Society of America and other notable institutions. These articles were annotated semiautomatically by La Monte Yarroll. Each article provides detailed nomenclature of mycelial species, accompanied by descriptions in English and often in Laten. These articles follow a standardized notation system, enabling us to efficiently tokenize the descriptive content and associate it with formal species names.

The objective of our project is to develop an AI-based tool that can analyze input descriptions and accurately retrieve relevant articles containing matching or closely aligned descriptions. This approach will allow for precise querying across our dataset. Below is a representative sample of the text format we will be working with. Notably, each research article we intend to use is structured according to this consistent notational format, ensuring uniformity and facilitating our data processing efforts.

The data can be found at <https://github.com/piggyatbaqaqi/skol.>

We plan to use the framework from <https://github.com/autonlab/dr-drafts-sota-literature-search.> The author of the package has offered to help us if we need additional guidance.

The annotation syntax comes from YEDDA, <https://github.com/jiesutd/YEDDA/.>

**Data Set Example:**

[@Ectrogella Besseyi sp. nov.#Nomenclature\*]

[@Sporangia irregulariter tubularia, saepe

lobata, interdum ellipsoidalia, longitudine 27-42 µm, diametro 13-17 a, singula

vel pauca in cellulis Olpidiopsidis Schenkianae, tubulas zoosporis emittienti- °

bus singulas vel multas formantia longitudine variantes; zoosporis numerösis,

zoosporis biflagellatis, 5 µm longis, 2.5 µm crassis; sporis perdurantibus non visis.#Description\*]

[@Parasitica in thallis et sporangiis Olpidiopsidis Schenkianae Zópf, prope P

Ann Arbor, Michigan, Novem. 1945 et Feb. 1948,#Habitat-distribution\*]

[@Sporangium irregularly tubular, often lobed, sometimes lip:

soidal, 27-42 µm long by 13-17 µm in diameter, one to several in a

host cell, forming one to several discharge tubes of varying length;#Description\*]

[@34 Mycotoaia, Vol. 41, 1949#Misc-exposition\*]

[@zoospores numerous, delimited within the sporangium, upon dis-

charge from the sporangium encysting and forming motionless

clusters at the orifice of the discharge tube, emerging from the

cyst as a laterally biflagellate zoóspore, 5 X 2.5. Resting spores

not observed.#Description\*]