# DOCX Standard Operating Procedure (SOP) Parser for Life Science Experiments

This repository contains a set of files to parse SOP from lab experiments.

# Why?

As research lab usually has their own set of SOP to conduct experiments, a tool to extract metadata from an editable document (e.g., DOCX) would be handy. The metadata is helpful in documenting the research and hence improves the reproducibility of the conducted research. To do this, the SOP should follow some annotation rules (described later below).

# How is this repo structured?

* The base directory contains the metadata extraction script
* *input* directory contains the docx SOPs example for extraction
* *output* directory contains extracted steps order – key – value in both JSON and XLSX format

# What is extracted from the SOP, and how is it represented in the docx document?

The parser extracts:

1. The *order* of the steps, based on the order of the paragraph in the docx SOP document.
2. The *key* for the metadata, based on the value represented in the curly bracket after the pipe character {value|key}. Example: {sequence alignment|stage} with ‘stage’ as the key.
3. Comments are allowed within the key, represented within regular brackets after the pipe symbol. Example: {receptor residue|(minimization) target}, with ‘minimization’ as the comment.
4. The *value* of the metadata is based on the first value represented in the curly bracket before the pipe character {value|key}. Example: {sequence alignment|stage} with ‘sequence alignment’ as the value.
5. Specific *flow control* type, such as for each, while, if, else if, and for. Some examples:
   1. <for each| generated pose> which corresponds to <flow type| iterated value>
   2. <while|pH|lte|7> which corresponds to <flow type|key|logical operator|value>. This is continued with e.g., Iterate over <+|1> which corresponds to <increment/decrement operation|increment/decrement value>
   3. <if|pH|lte|7> which corresponds to <if|key|logical operator|value>
   4. <else if|pH|between|[8-12]> which corresponds to <else if|key|logical operator|value>
   5. <else> which corresponds to <flow type>
   6. <for|pH|[1-7]|+|1> which corresponds to <for|key|[range]|iteration\_operation|magnitude>

The overall example of the SOP document is available in the *input/sop2.docx* file. The color in the *sop2.docx* does not play any role in the order/key/value extraction.

# How the parser should be run?

1. Create SOP according to the above annotation rules.
2. Change the input directory/file name in the python script (2nd last line)
3. Change the output directory/filename (last line)
4. Run the script.

# What are the further plans?

1. Align the used keys with terms from an ontology, or if the term does not exist, create a new term by extending an ontology or creating a term within a new ontology.