MarcXimiL

A deduplication tool for library records... and more

- Alain Borel alain.borel@epfl.ch
- Jan Krause jan.krause@unige.ch

- http://marcximil.sourceforge.net

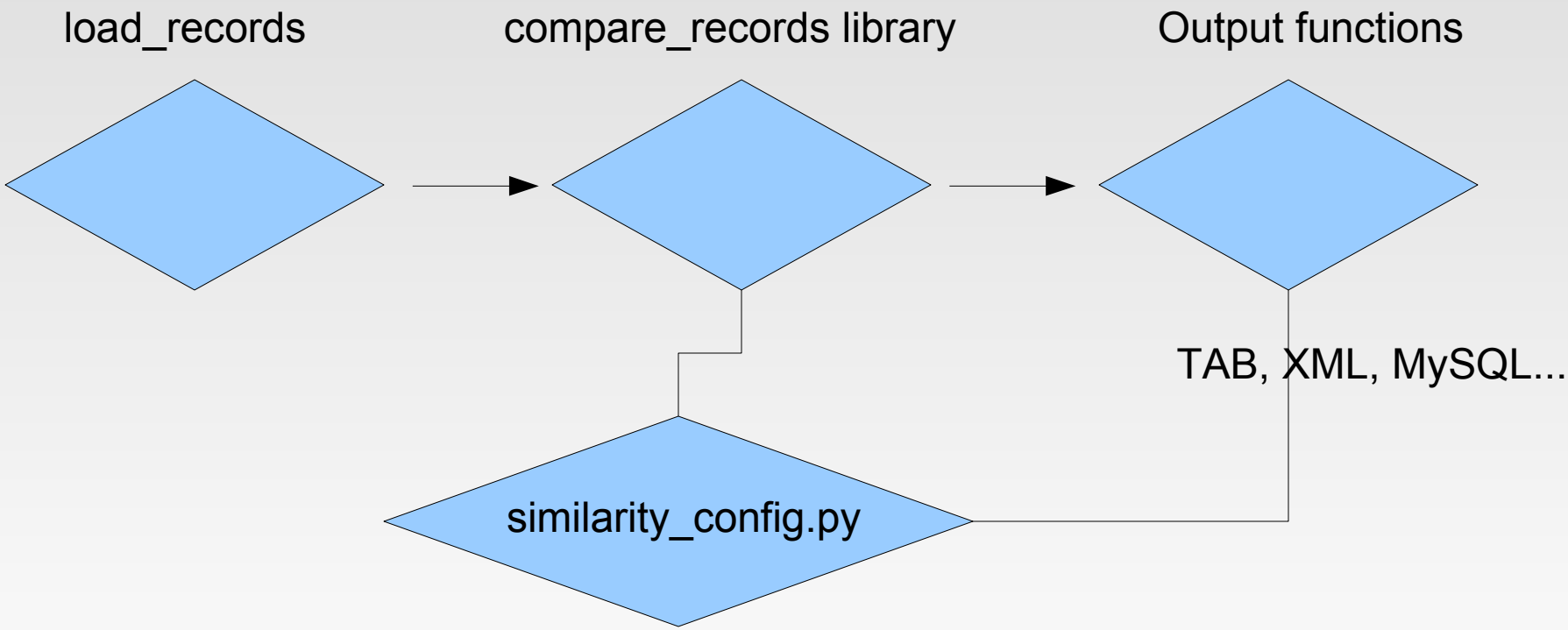
MARC is not dead yet!

MarcXimiL/MARC XML Similarity

- Open-source Python program
- Processes MARC XML record collections
- Calculates a record-record similarity score based on user-defined fields and methods
- Possible applications: **duplicate records detection**, FRBRisation, ”more like this...”, etc.

How does it work?

Simple console application: just type similarity.py



User-defined strategy: what fields? How to compare them? Which records do we look at?

similarity_config.py

```
records_comp = records_comp_single
report = report_tab

globalvars.output_threshold = 0.5
record_rules = geometric_mean_breakout

record_structure = { \
      '01recid'   :{'marc'      : '001',
                    'weight'    : 0,
                    'parse-func': parse_controlfield,
                    'comp-func' : fields_concat__raw },
      '02year'    :{'marc'      : '260  c',
                    'weight'    : 1,
                    'parse-func': parse_nonrep,
                    'comp-func' : years_comp__raw },
      '03authors' :{'marc'      : ['100  a', '700 a'],
                    'weight'    : 2,
                    'parse-func': parse_multi,
                    'comp-func' : authors_comp__raw },
 '04title'   : {'marc'      : ['245  a', '245  b'],
                'weight'    : 3,
                'parse-func': parse_concat,
                'comp-func' : okapibm25__wc }
'05title'   :{'marc'      : '245  a',
              'weight'    : 3,
              'parse-func': parse_nonrep,
              'comp-func' :
levenshtein__raw } }
```

So far...

- EPFL institutional repository: 2'000 suspiciously similar records detected among 60'000
- In process now: comparison of the CERN Document Server & INSPIRE archive thesis records

What's next?

- Further performance improvements (~1 h for 100'000 records)
- Better strategies for non-article records
- User-friendlier display of the results
- Input modules for cooler data formats?

Feel free to contact me: alain.borel@epfl.ch