Mobilizing Biodiversity Data with Frugal Tools

Jorrit H. Poelen

Independent Biodiversity Informatics / Software Engineer
jhpoelen@xs4all.nl

Frugal Tools

Minimalist

Specialized

Independent

Compatible with other tools

Reusable

Budget friendly

In line with ... Unix Philosophy

"Write programs that do one thing and do it well.

Write programs to work together.

Write programs to handle text streams, because that is a universal interface."

- Doug McIlroy, inventor of Unix Pipes in 1973, quoted in Peter H. Salus. A Quarter-Century of Unix. Addison-Wesley. 1994. ISBN 0-201-54777-5.

Discussion Prompts (coming up ...)

Reflection on how to promote, reuse and improve frugal tools in biodiversity research.

On what **action do you spend most of your time** when mobilizing digital biodiversity data? And if you removed a bottleneck, how?

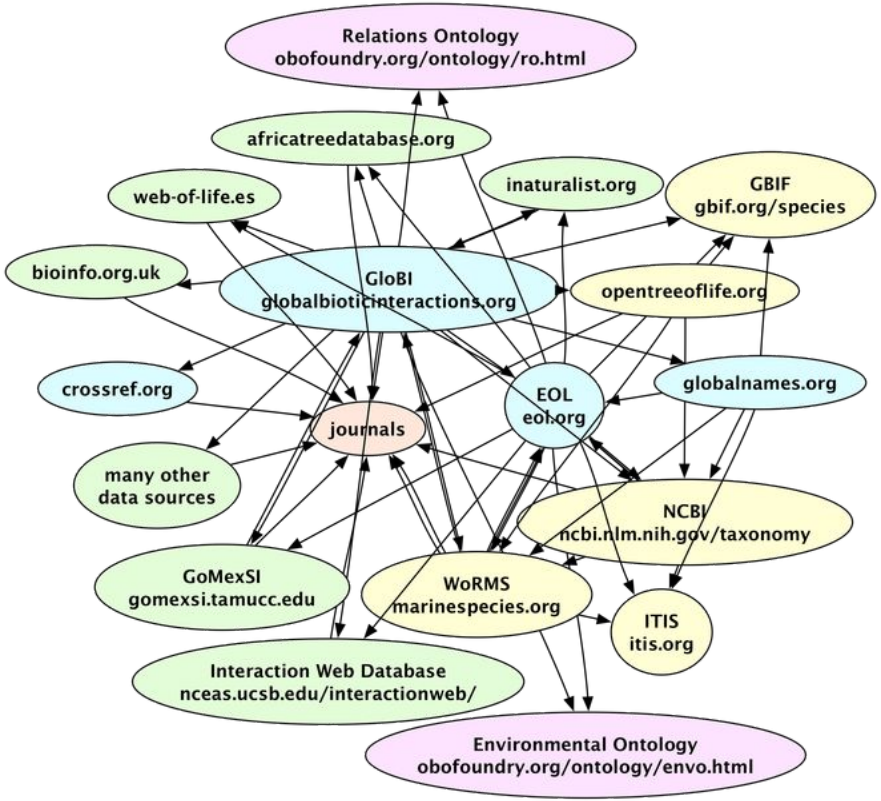
What are the **input** and **output** of this action?

What **tools** do you use to perform this action?

What are the computational building blocks in biodiversity research?

Use Case: Global Biotic Interactions

. . . a fully automated infrastructure that continuously translates existing species interaction datasets into open, **linked**, aggregated data archives.



https://globalbioticinteractions.org

Use Case: Global Biotic Interactions

Challenge:

Make the continuous taxonomic name translation / linking process reliable, reproducible, scalable (~millions) in a context of rate-limited web apis, outages, and unforeseen changes.

A solution:

Build a specialized offline-enabled term-mapping tool: Nomer.

https://github.com/globalbioticinteractions/nomer

Nomer, a term matcher

input: tab-separated values (tsv) file stream

action: match id, name columns to associated terms

output: tab-separated values (tsv) file stream

	Homo sapiens
	Enhydra lutris

NCBI:9606	Homo sapiens
EOL:328583	Enhydra lutris

Nomer support various matching methods including **online** only (e.g., resolver.globalnames.org) and **offline-enabled matchers** (e.g., globi-cache).

Nomer, a term matcher

Return first match for identifier NCBI:9606 using echo (1971), pipes (1973) and nomer (2018).

$ echo -e "NCBI:9606\t" | nomer replace

GBIF:2436436 Homo sapiens

Nomer, a term matcher

Append all match types and matches to id NCBI:9606 using echo (1971), pipes (1973) and nomer (2018).

$ echo -e "NCBI:9606\t" | nomer append

NCBI:9606	SAME_AS	GBIF:2436436	Homo sapiens ...
NCBI:9606	SAME_AS	IRMNG:10857762	Homo sapiens ...
NCBI:9606	SAME_AS	ITIS:180092	Homo sapiens ...

Nomer, a term matcher

List all the unique identifiers associated with Homo sapiens on a single line using echo (1972), nomer (2018), cut (1974), sort (1972), uniq (1973), and tr (1977).

$ echo -e "\tHomo sapiens" | nomer append | cut -f4 | sort | uniq | tr "\n" " "

EOL:327955 GBIF:2436436 INAT_TAXON:43584 IRMNG:10857762 ITIS:180092 NBN:NHMSYS0000376773 NCBI:741158 NCBI:9606 OTT:770315 OTT:933436 WD:Q15978631

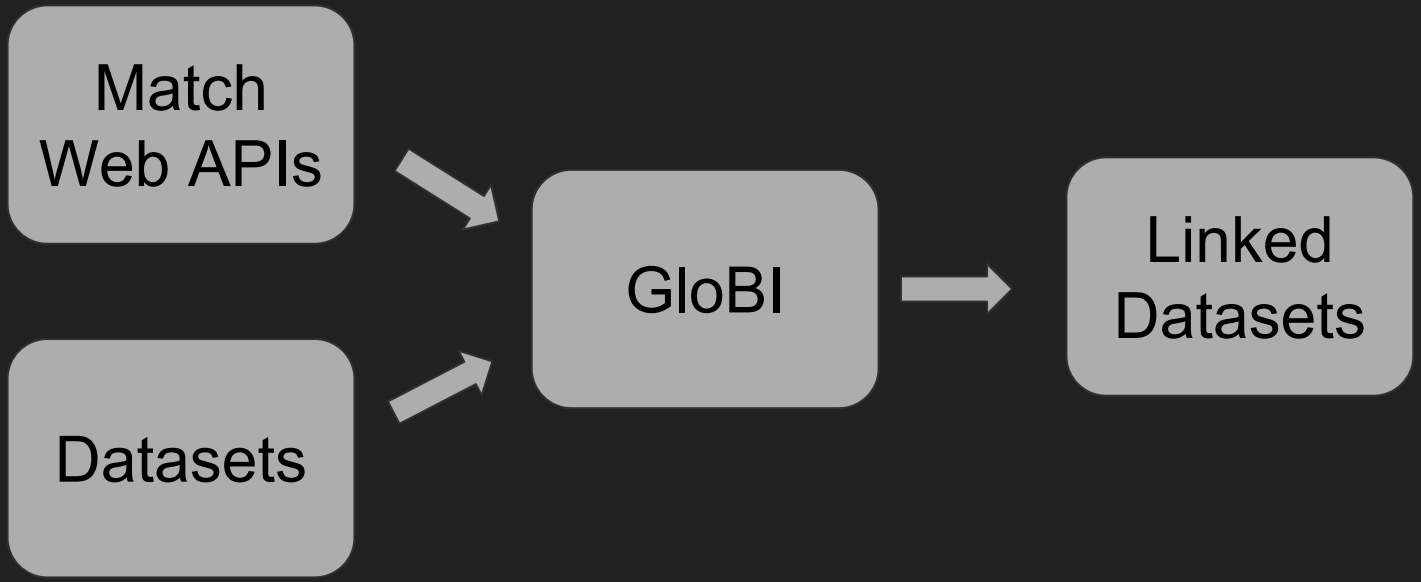
Nomer, a term matcher

Match all names from GloBI interaction datasets, count all unique unmatched names and put them in a file using elton (2017), nomer (2018), cut (1986), grep (1977), sort (1972), uniq (1973), tee (1974) and wc (1971).

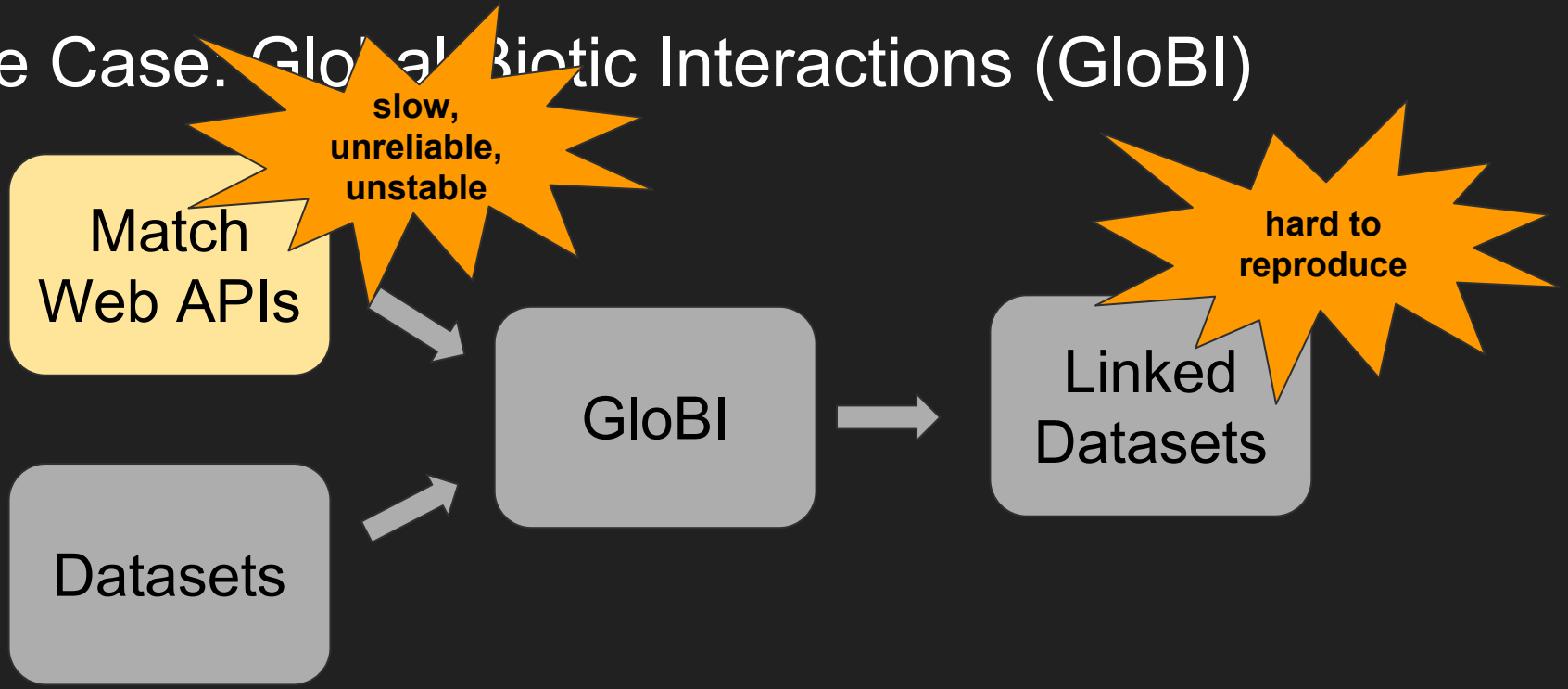
$ elton names | cut -f1,2 | nomer append | grep "NONE" | cut -f1,2 | sort | uniq | tee unmatched_names.tsv | wc -l

22848

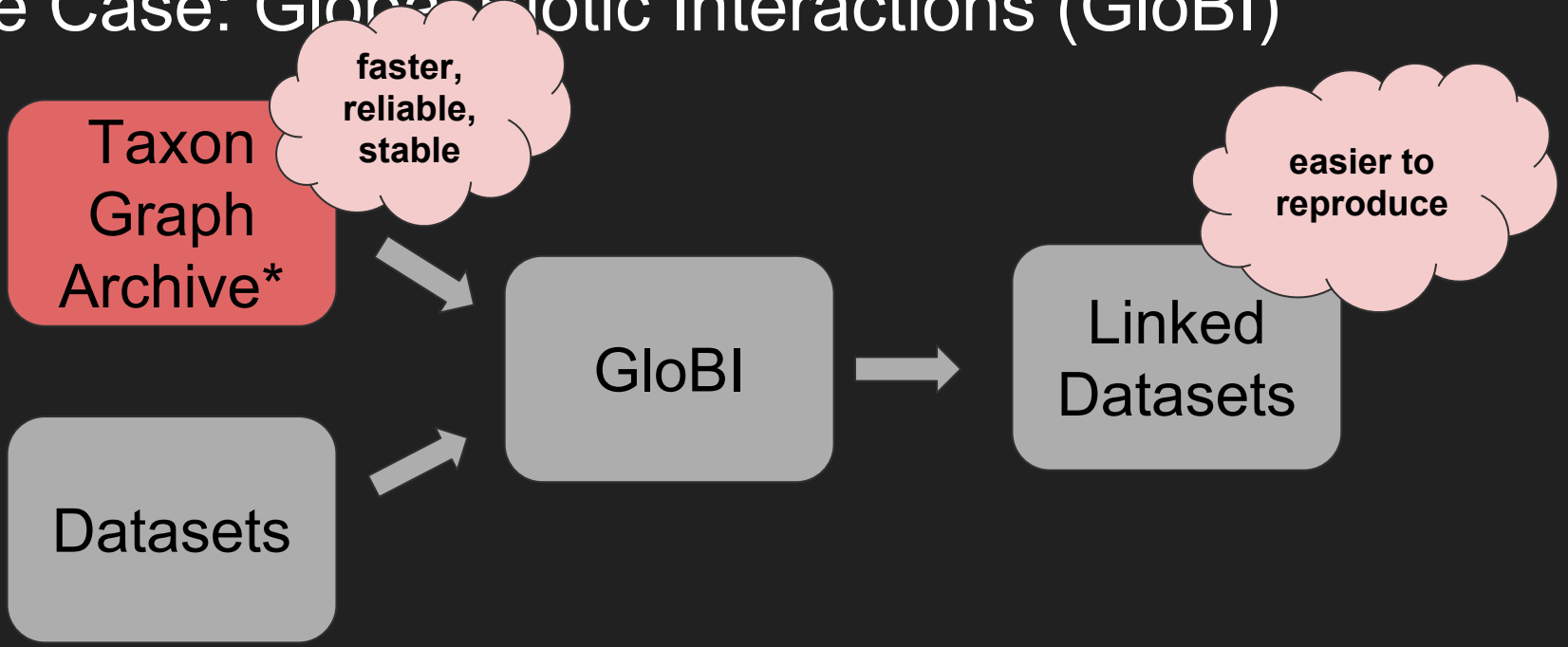
Use Case: Global Biotic Interactions (GloBI)



Use Case: Global Biotic Interactions (GloBI)



Use Case: Global Biotic Interactions (GloBI)



* Poelen, Jorrit H. (2018). Global Biotic Interactions: Taxon Graph (Version 0.3.2) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.1250572 . Also see https://github.com/globalbioticinteractions/taxon-graph-builder .

Acknowledgments

Thanks to **Marie-Angélique Laporte, Austin Meier, Carl Boettiger, Chris Mungall, Pankaj Jaiswal** and many others.

With support from **Planteome** (http://planteome.org) via Oregon State University, Corvallis and **Encyclopedia of Life** (http://eol.org) via Smithsonian.

Computing facilities provided by **Heitlinger Lab**, Humboldt State University, Berlin and **GUODA** via Advanced Computing and Information Systems (**ACIS**) Laboratory, University of Florida, Gainesville.

Discussion Prompts

Reflection on how to promote, reuse and improve frugal tools in biodiversity research.

On what **action do you spend most of your time** when mobilizing digital biodiversity data? And if you removed a bottleneck, how?

What are the **input** and **output** of this action?

What **tools** do you use to perform this action?

What are the computational building blocks in biodiversity research?