

# crawl

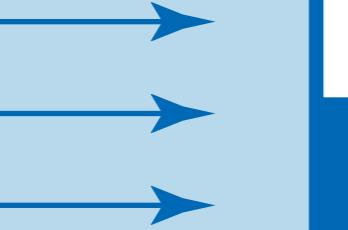
sources of  
papers or  
information

JTOCs  
PLoS  
CORE (IR)  
EuPMC  
ArXiV  
SWORD  
OAI - PMH

# scrape

resources  
as files or  
embedded  
in files

unique  
identifier



URL  
DOI

PDF XML PNG  
HTML LaTeX  
SVG CSV TXT  
EPUB XLSX  
DOCX JPEG

# normalise

documents  
to a  
standardised  
structure

messy/  
broken  
format



Scholarly  
HTML

metadata  
figures  
sections  
licence  
etc. (JSON)



species places  
date/time regex  
regexIDs  
phylogenies  
chemical reaction  
molecule

# extract

entities  
facts