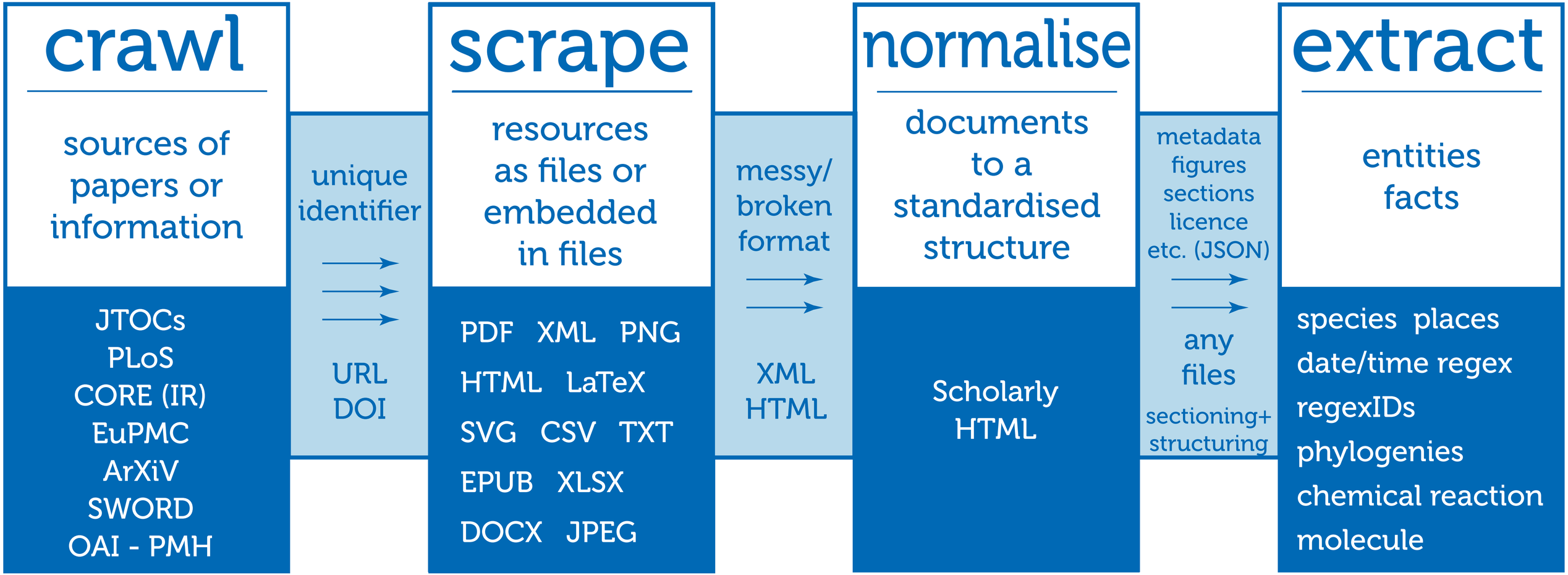
crawl
sources of papers or information
JTOCs
PLoS
CORE (IR)
EuPMC
ArXiV
SWORD
OAI - PMH
unique identifier
URL
DOI
scrape
resources as files or embedded in files
PDF XML PNG
HTML LaTeX
SVG CSV TXT
EPUB XLSX
DOCX JPEG
messy/ broken format
XML
HTML
normalise
documents to a standardised structure
Scholarly HTML
metadata figures sections licence etc. (JSON)
any files
sectioning+ structuring
extract
entities facts
species places
date/time regex
regexIDs
phylogenies
chemical reaction
molecule