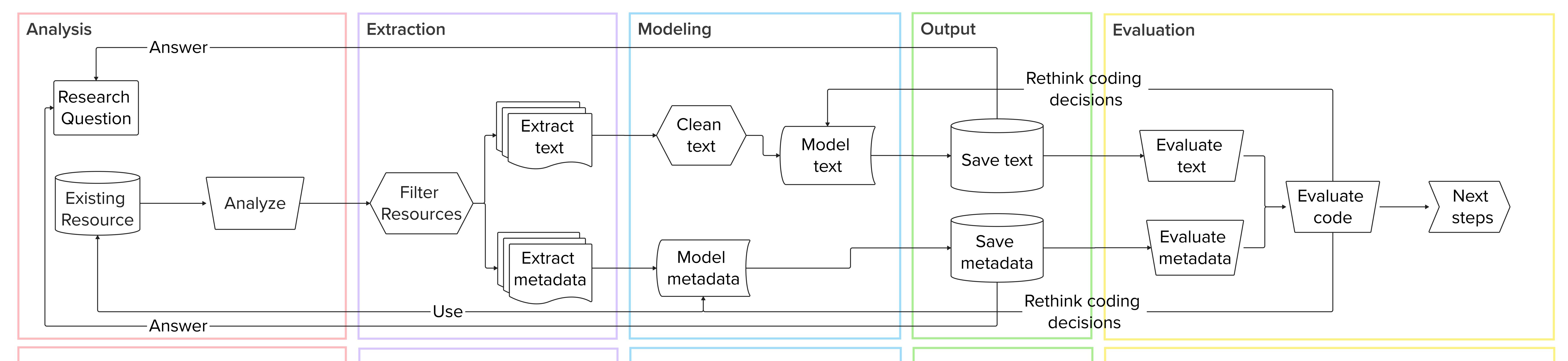
Pipeline for the GartenlaubeExtractor



Is serialization a factor in increasing the suspense level of texts?

wikisource, Die Datenlaube

existing metadata information, text quality, API structure

filter poems and already collected texts

fetch every article of the *Gartenlaube* journals on *wikisource*

fetch metadata of every article of the *Gartenlaube* journals on wikisource

split text into episodes

remove image captions, add footnote counter

use information from wikisource,
the extracted text, and external
resources (name lists, Neuer
Deutscher Novellenschatz) to
collect the metadata needed for
the corpus

store text and episode files in output directory

append metadata to
the existing
metadata table
(Bibliographie.csv)

inspect and correct extracted text and metadata

specify ethical concerns (bias, harmful content)

ensure reusability (user experience, provide environment file, add documentation) and efficiency (reduce complexity)

specify missing metadata (e.g. pseudonyms of the authors) and future work (e.g. using *wikisource* to extract the pseudonyms)