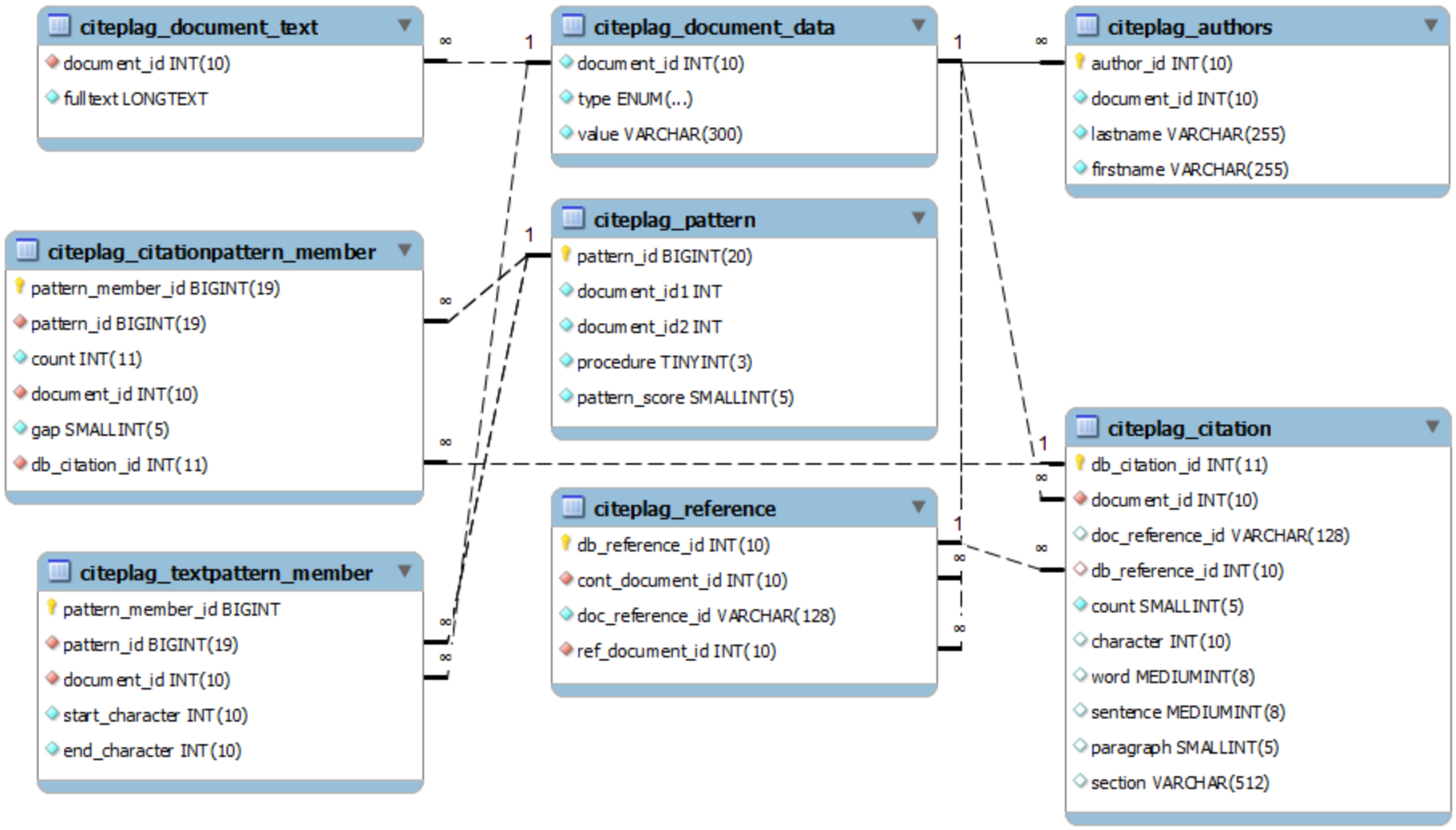

The following list explains the attributes of all tables in the CitePlag database.

citeplag_document_data

- *document_id* → database-internal ID assigned to documents for which the full text is available in the database and to "placeholder" documents representing documents referenced within full texts

- *type* → flag that identifies the type of additional data stored for documents, e.g. title or external document identifiers (PubMed IDs, PMC IDs, DOIs). The ENUM type provides the possibility to add further types, which are not yet considered, in the future.
- *value* → attribute holding the actual data of a certain type, e.g., title

citeplag_document_text

- *document_id* → database-internal ID of the document for which the full text is stored
- *fulltext* → full text of the document

citeplag_author

- *author_id* → database-internal ID for all authors
- *document_id* → ID of the document in which the author appeared. Currently, authors are not disambiguated, i.e., if an author appears in multiple documents, there will be multiple records with the same name in citePlag_authors.
- *last_name*, *first_name* → author name

citeplag_citation

- *db_citation_id* → database-internal ID for all citations
- *document_id* → database-internal ID of the document that contains the citation
- *doc_reference_id* → ID for references used in NXML-documents; is unique only within the NXML-document. Each in-text citation within a NXML-document specifies the ID of the corresponding reference
- *db_reference_id* → unique database-internal ID for references
- *count* → sequential number of a citation within a document's full text
- *character*, *word*, *sentence*, *paragraph*, *section* → positional information of a citation within a document's full text

citeplag_reference

- *db_reference_id* → database-internal ID for references
- *cont_document_id* → document_id of the document that contains the reference

- *doc_reference_id* → ID for references used in NXML-documents; is unique only within the NXML-document
- *ref_document_id* → document_id of the document that is referenced. The referenced document is not necessarily part of the PMC OAS. Therefore, many "placeholder documents" for which no full text is available are contained in the database

citeplag_pattern

- *pattern_id* → database-internal ID for all patterns
- *document_id1, document_id2* → document_ids of the two documents for which the matching pattern has been identified
- *procedure* → ID that denominates the detection algorithm, which was used to identify the pattern, see **Table 1** for short descriptions and IDs of the detection methods.

Table 1: Overview of detection methods and their database-internal IDs

Class	Detection Algorithm	ID
LCCS	LCCS	1
	LCCS distinct	11
GCT	shared citations only	2
	all citations	21
	all citations, matches all shared citations once a match has been found	22
Citation Chunking	one document chunked, only adjacent citations considered, no merge performed	30
	one document chunked, only adjacent citations considered, merge	31
	one document chunked dependent on predecessor, no merge	32
	one document chunked dependent on predecessor, merge	33
	one document chunked dependent on textual proximity, no merge	34

	one document chunked dependent on textual proximity, no merge	35
	both documents chunked, only adjacent citations considered, no merge	40
	both documents chunked, only adjacent citations considered, merge	41
	both documents chunked dependent on predecessor, no merge	42
	both documents chunked dependent on predecessor, merge	43
	both documents chunked dependent on textual proximity, no merge	44
	both documents chunked dependent textual proximity, no merge	45
Encoplot Similarity	Encoplot (global score of document = percentage of similarity)	50
	Encoplot (scores of multiple patterns per document, details on patterns in textpattern_member table)	51
CPA	Basic CPA	60
Bibliographic Coupling	Bibliographic Coupling (score = coupling strength of both documents, no pattern_members)	70
	Bibliographic Coupling / Coupling units (score = total citations of a shared reference, pattern_members: citations that form the coupling)	71
Co-Citation	Co-Citation = number of documents that cite the two documents together	80
Lucene	Lucene MoreLikeThis measure computed on the full text	90

- *pattern_score* → similarity score of the identified pattern. For citation patterns, the score equals the length of the pattern, for character-based patterns see table above.

citeplag_citationpattern_member

- *pattern_member_id* → database-internal ID for all citation_pattern_members
- *pattern_id* → database-internal ID of the pattern formed by the citation_pattern_members

- *document_id* → document_ID of the document that contains the citations. Storing this ID here is redundant, because the citation identified by db_citation_id contains the same information. However, in practice the redundancy saves a join of citeplag_citationpattern_member to citeplag_citation, which significantly improves performance, because citeplag_citation-pattern_member is a very large table (approx. 1.4 bn records).
- *count* → sequential position of the pattern member within the pattern
- gap → number of non-matching citations between two matching citations in a citation pattern
- *db_citation_id* → ID of the citation that represents the pattern member

citeplag_textpattern_member

- *pattern_member_id* → database-internal ID for all text_pattern_members
- *pattern_id* → database-internal ID of the pattern formed by the text_pattern_members
- *document_id* → document_ID of the document that features the text similarity.
- *start_character, end_character* → character count at the start- and ending position of the identified text overlap