REFERENCE

This table contains data directly pertaining to References, which are always parsed out of the "CR" ISI field. Each subsequent reference string contains a series of tokens separated by the string ", " (a comma, then a single space character). Further down below, this article describes the limited set of reference string formats supported, which are defined by what tokens can be parsed out of them and in what positions they can be parsed.

There are various ways to determine what data a token is supposed to specify, depending on what position the token occurred in. For example, if a token that starts with the letter 'V', followed by numeric digits is in the third of four positions, those numeric digits in that token is treated as the reference volume (*REFERENCE_VOLUME*).

This table includes all fields that can possibly be parsed out of any support reference string format, which are:

- *PK*: Automatically generated. Guarantees uniqueness.
- *ANNOTATION*: It is possible for reference strings to contain Sources in various token positions (see below). These Sources are potentially prefixed with what we call annotations; similarly, it is possible for the entire Source token to just be an annotation. The set of annotations that the loader can separate from the actual Source title (if there is one) is:
    - *UNPUB*
    - *IN PRESS*
    - *PREPRINT*
    - *UNPUBLISHED*
    - *CITED INDIRECTLY*
    - *PRIVATE COMMUNICATIO*
    - *UNOPUB* (This was most likely a typo in a dataset we used for testing.)
- *REFERENCE_ARTICLE_NUMBER*: From what we have found, the standard reference string formats can contain an additional token at the end, which is the article number of the document being referenced. This can be used to match References to Documents (as a "is a" relationship).
- *REFERENCE_AUTHOR_FK*: It is possible for reference strings to contain an Author in various token positions (see below). When an Author is specified, it is always treated as the first author. The string specifying the Author is always considered to specify the unsplit abbreviated author name, and it is parsed as such.
- *DIGITAL_OBJECT_IDENTIFIER*: If you examine below, the only supported reference string format that contains 6 tokens has a Digital Object Identifier token in last/sixth position. A valid Digital Object Identifier begins with the string "DOI " and has the subsequent format "**/**" , where *s are wildcards. The actual Java regular expression used to match that subsequent format is: "**./.**".
- *REFERENCE_OTHER_INFORMATION*: For lack of a better name, this is other information that may be specified at the end of any of the formats (similarly to Article Numbers). We determine if other information is specified if the last token starts with one of the following prefixes:
    - *PII*
    - *PMID*
    - *UNSP*
- *REFERENCE_PAGE_NUMBER*: When part of a support format, a page number token starts with "p" or "P", followed by a series of numeric digits. This field contains those numeric digits as an integer.
- *PAPER_FK*: After all references and documents have been parsed, it is possible to match references to documents if they have matching article numbers and/or digital object identifiers. When such matches have been made, a link is made from the Reference and Document that basically states that the Reference is the Document. This field contains the foreign key into the **DOCUMENT** table in such a case.
- *RAW_REFERENCE_STRING*: The exact string as found in the "CR" field.
- *REFERENCE_VOLUME*: When part of a support format, a volume token starts with "v" or "V", followed by a series of numeric digits. This contains those numeric digits as an integer.
- *REFERENCE_SOURCE_FK*: It is possible for reference strings to contain a Source in various token positions (see below). When a Source is specified, it is always assumed that the abbreviated name is provided (what appears in the "J9" field). This field contains the foreign key into the **SOURCE** table in such a case.
- *REFERENCE_WAS_STARRED*: This field is obsolete and will always be empty.

The following reference string formats are supported, organized by number of tokens present:

- 2 tokens:
    - year, source
    - author, source
- 3 tokens:
    - year, source, volume or page number
    - person, source, volume or page number
- 4 tokens:
    - year, source, volume, page number
    - person, year, source, volume or page number
    - person, source, volume, page number
- 5 tokens:
    - person, year, journal, volume, page number
- 6 tokens:
    - person, year, source, volume, page number, digital object identifier

In the various built-in extractions, when References are matched to Documents, they are considered as part of the set of inner documents. When they are not matched to Documents, they are considered as part of the set of outer documents.

See CITED_REFERENCES, DOCUMENT, PERSON, SOURCE, and How Abbreviated Names are Parsed.