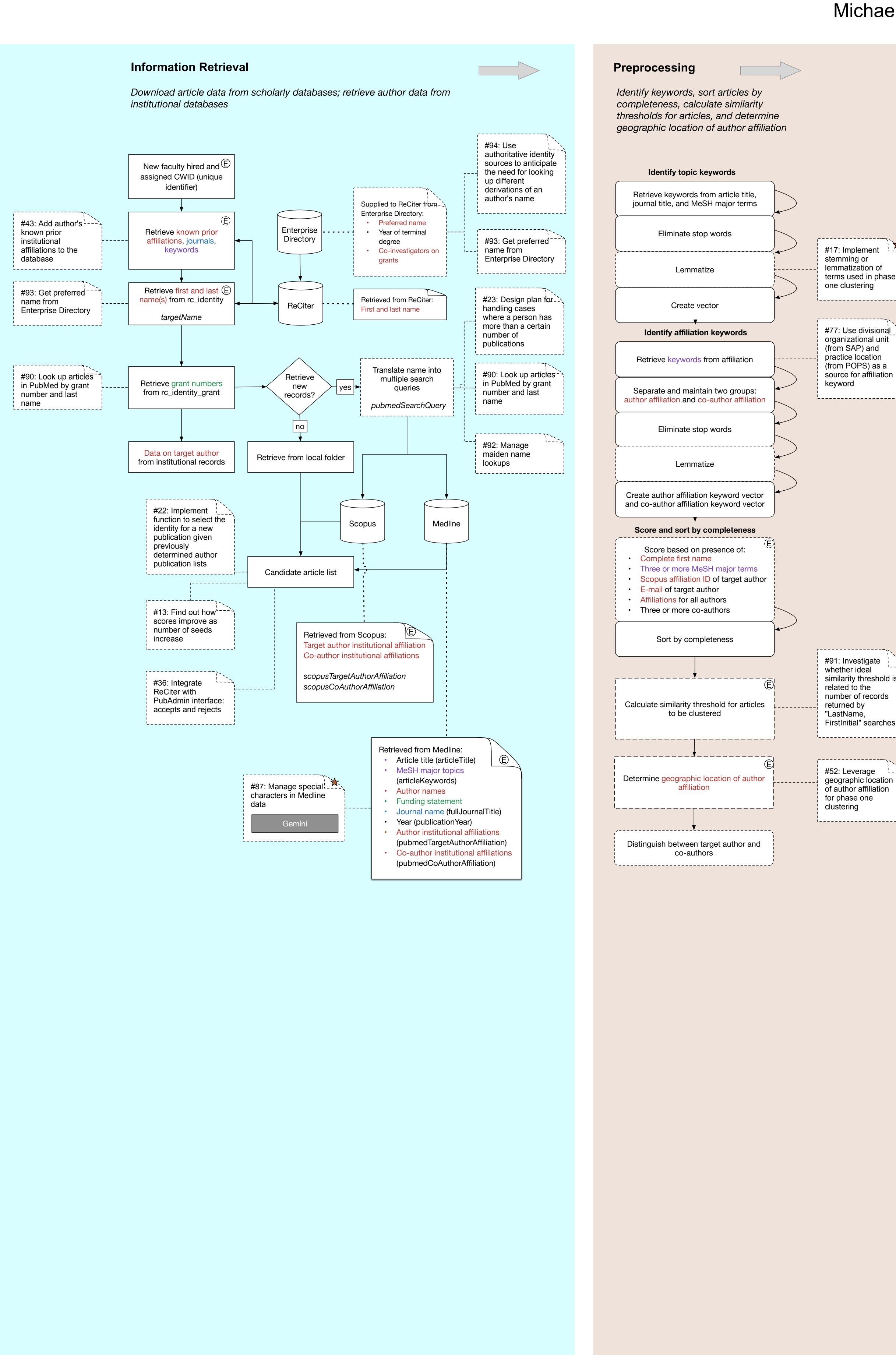
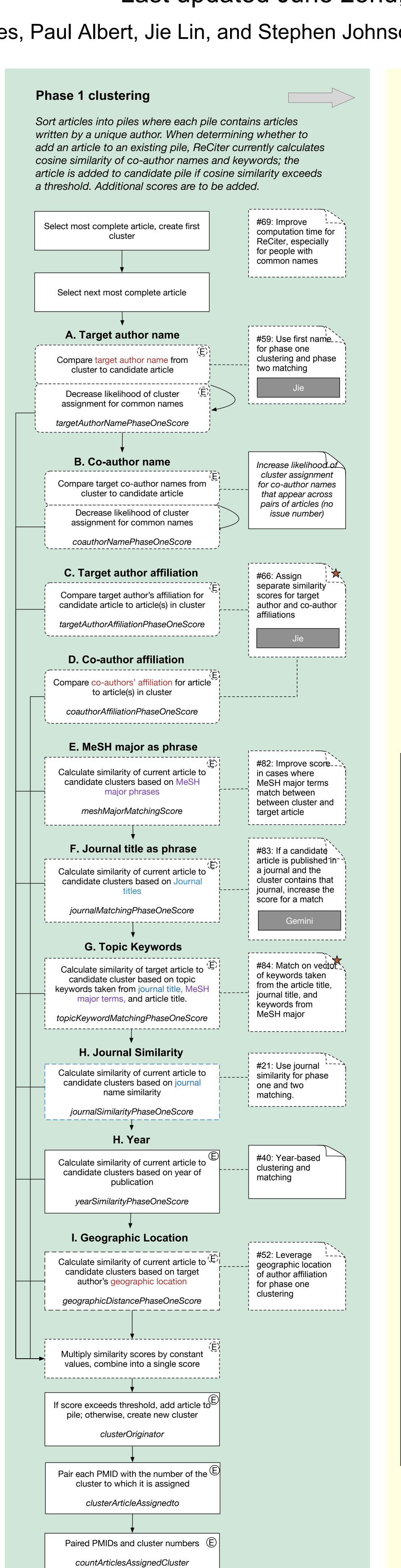
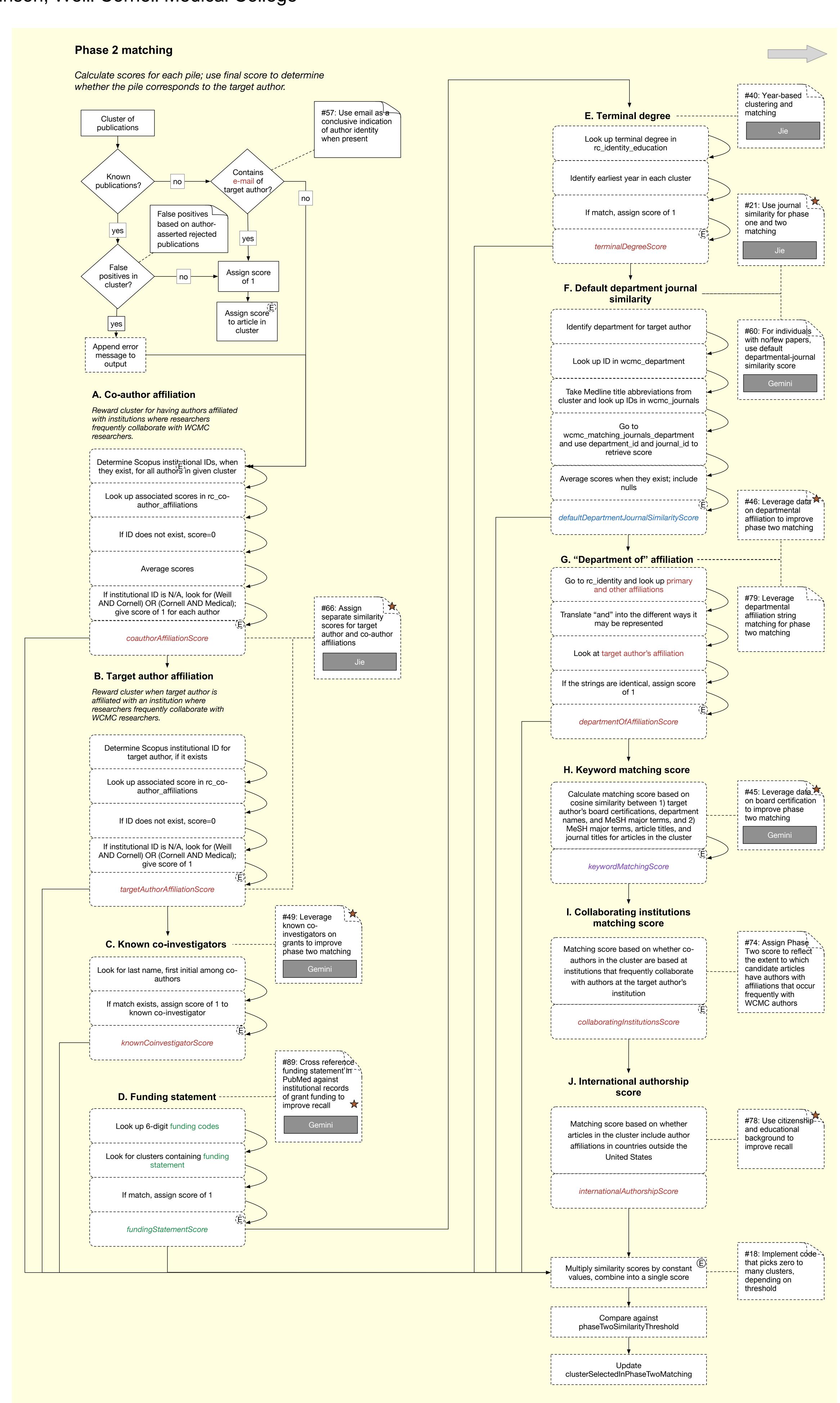
ReCiter Architecture and Data Processing Operations

Last updated June 23nd, 2015

Michael Bales, Paul Albert, Jie Lin, and Stephen Johnson, Weill Cornell Medical College







Data type coloring

Person Grant Journal Keyword

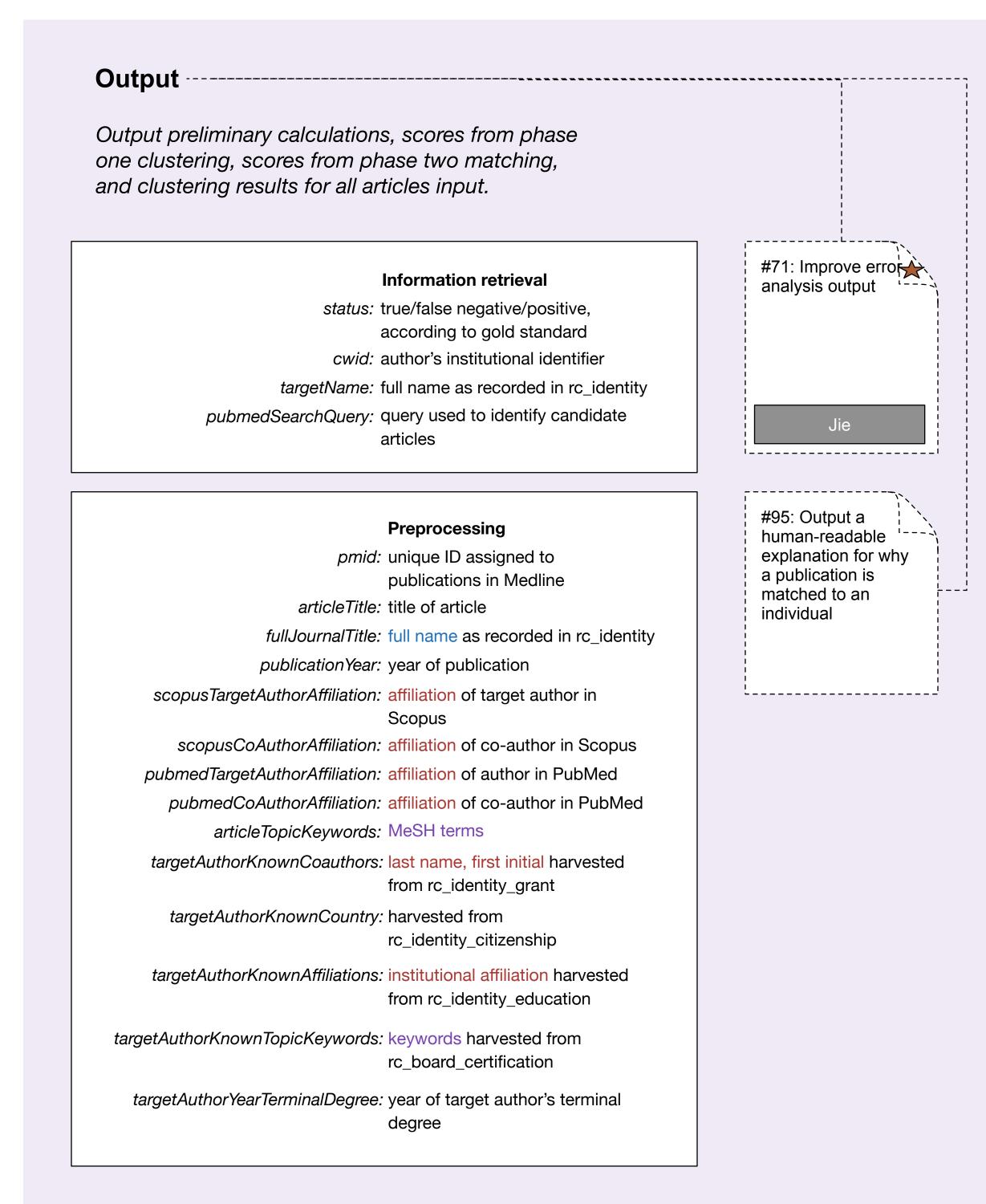
To be completed

Completed

Priority

E Appears in error analysis output

(Ē) To appear in error analysis output



Phase one clustering: clustering results and scores

targetAuthorNamePhaseOneScore: Target author name in cluster versus in candidate article
coauthorNamePhaseOneScore: Co-author names in cluster versus in candidate article
targetAuthorAffiliationPhaseOneScore: Target author's affiliation in article versus cluster
coauthorAffiliationPhaseOneScore: Co-author's affiliation in article versus cluster
meshMajorMatchingScore: Overlap of MeSH major between candidate article and cluster
journalMatchingPhaseOneScore: Overlap of journal titles between candidate article and cluster
topicKeywordMatchingPhaseOneScore: Similarity of topic keywords between cluster and target article
journalSimilarityPhaseOneScore: Journal similarity score between target author and article cluster
yearSimilarityPhaseOneScore: Similarity of current article to candidate clusters based on
geographicDistancePhaseOneScore: Similarity of current article to candidate clusters based on
geographic distance

clusterOriginator: A "*" when an article starts a cluster; else null

clusterArticleAssignedto: Number of the cluster to which the article was assigned countArticlesInAssignedCluster: Number of articles in the cluster to which the article was

Phase two matching: matching scores

coauthorAffiiliationScore: Scores are assigned to co-authors indicating whether or not they are affiliated with institutions where researchers frequently collaborate with WCMC researchers; scores are weighted by institution and averaged across co-authors (issue #47)

targetAuthorAffiiliationScore: Score indicating whether target author is affiliated with institution where researchers frequently collaborate with WCMC researchers; score is weighted by institution (issue #49)

knownCoinvestigatorScore: Score indicating whether co-authors in the cluster match the names of the target author's co-investigators on grants

fundingStatementScore: Score indicating whether any articles in the candidate cluster match one or more of target author's known grant numbers (issue #89)

terminalDegreeScore Score to indicate probability the articles in the cluster belong to the target author, based on year of publication and year of target author's terminal degree (issue #40)

defaultDepartmentJournal Score to indicate probability the articles in the cluster belong SimilarityScore: to the target author, based on journal similarity (issue #21, issue #60)

departmentOfAffiliationScore: Score indicating string match between target author's departmental affiliation and affiliation of target author in Medline

keywordMatchingScore: Matching score based on cosine similarity between 1) target author's board certifications, department names, and MeSH major terms, and 2) MeSH major terms, article titles, and journal titles for articles in the cluster

gInstitutionsScore: Matching score based on whether co-authors in the cluster are based at institutions that frequently collaborate with authors at the target author's institution

international Authorship Score: Matching score based on whether articles in the cluster include author affiliations in countries outside the United

Phase two matching: scoring results

phaseTwoSimilarityThreshold: Similarity threshold in phase two clusterSelectedInPhaseTwoMatching: Whether the cluster was selected in phase two matching