

ウェブ情報をデータベースに結びつける

相澤 彰子 Akiko Aizawa 高久 雅生 Masao Takaku 大山 敬三 Keizo Oyama

Bibliographic reference resolver

引用文献やウェブ上の文献リストをデータベースに対応づけるための「書誌リンクエンジン」

- 数千万件規模のレコードを検索対象として、オンラインで同定結果を出力
- フォーマットを持たないテキストをデータベースのレコードに対応づける
- 入力テキスト全体ではなく、一部に照合するレコードが出力される
- 出力結果として得られるのは、類似レコードのランキングではなく、同一であると判定されたレコード



自動認識した文字列
OCR認識誤りを含む引用文字列
文献データベースに登録されていることをあらかじめ人手で確認した評価用サンプル200件

既存検索サービスの出力

参照先を同定

既存の検索エンジン

そのまま検索窓に入力すると
⇒11件/200件しかヒットしない

i-Linkage

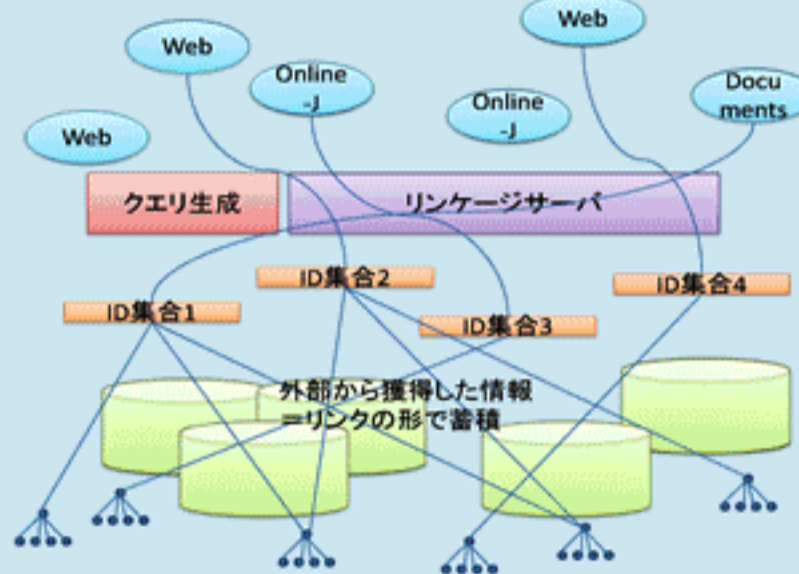
そのまま検索窓に入力すると
⇒196件/200件ヒット、正規化ルールを入れると全件ヒット



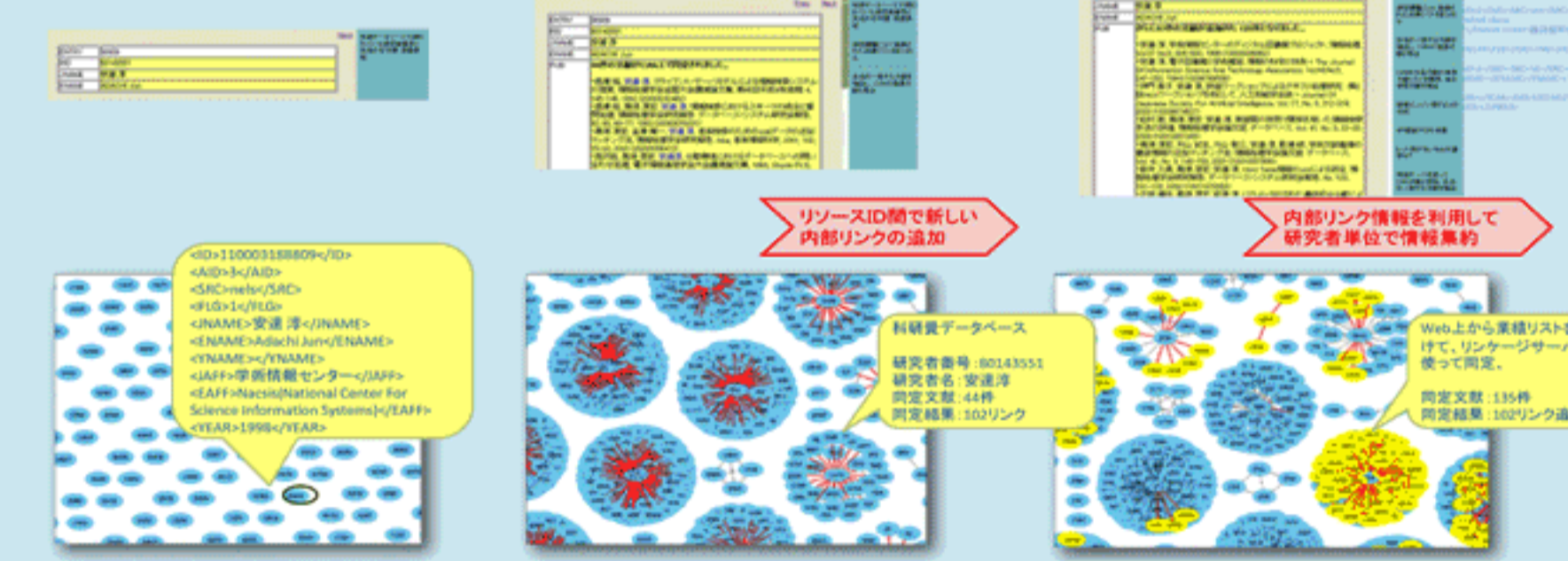
<http://i-linkage.nii.ac.jp/cinii/>

他のデータベースやウェブ上の情報による同定結果を利用して、データベース上の情報を再構築

Academic linkage



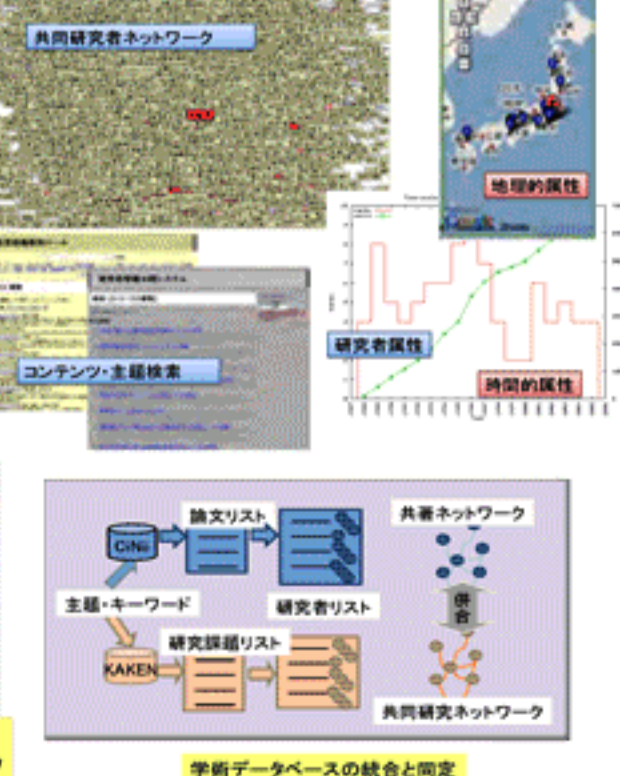
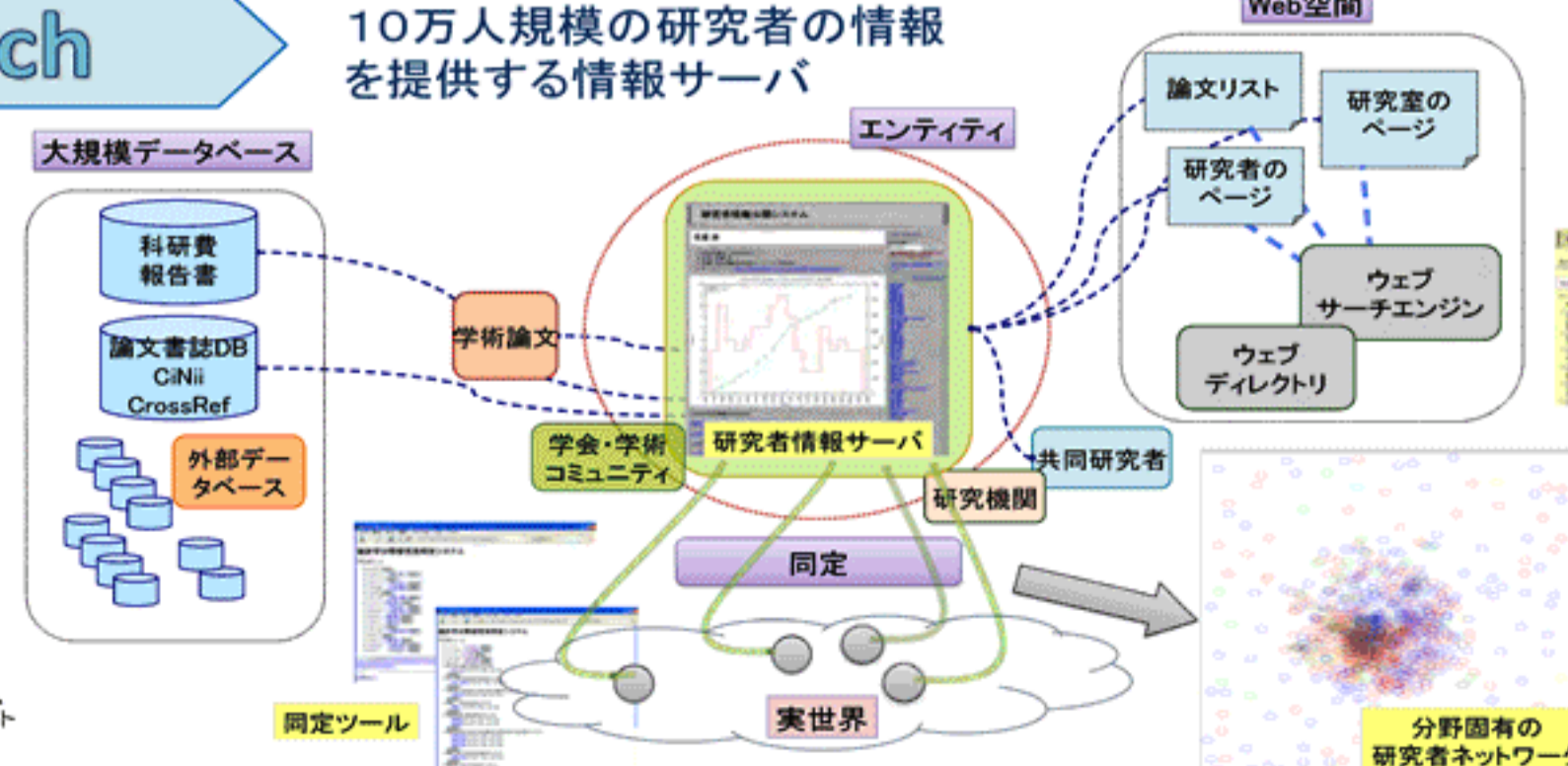
- 任意の文書 (ex. 報告書、論文、研究室のWebページ)
- リンクエンジンサーバで文書をリソースIDの集合に変換
- 既存のリソースからのクエリ自動生成
- 収集した関連URLのウェブ上のリンク先をたどる
- 大規模計算基盤 Intrigger で同定処理を並列実行



RexSearch

10万人規模の研究者の情報を提供する情報サーバ

- 複数情報源の統合 (Web, 外部DBなど)
 - 共同研究者の地理分布
 - 研究活動履歴
- 人名同定ツールの開発
 - 複数学名名簿との照合、可視化
 - 1000名程度の同定データセット構築
- 研究者の各種属性を切り口としたエキスパートサーチ
 - 複数情報源の統合・同定
 - 多種属性情報とネットワーク情報の相補的活用



異なるソースからの情報を結びつけ、新しい価値を発見するためのプラットフォーム構築を目指しています。