

K00U研究支援人材の資質向上WG/研究データ管理・活用WG合同ワークショップ資料（2025）

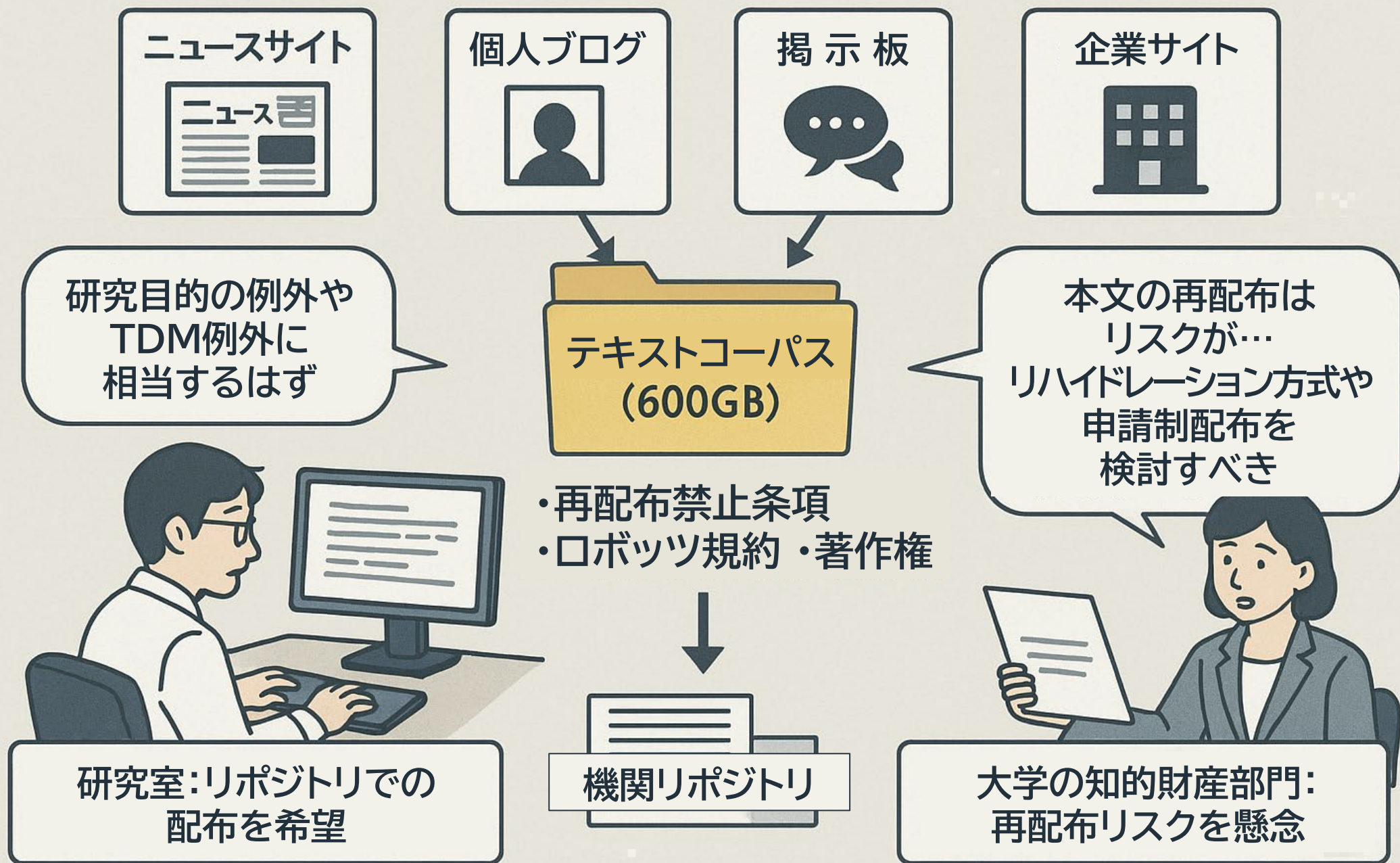
九州大学附属図書館

九州大学データ駆動イノベーション推進本部研究データ管理支援部門

<https://hdl.handle.net/2324/7440576>

出版情報：2025-12-11
バージョン：
権利関係：





公開形態や審査体制をどのように決めるべきか？

Day2-A Case2 Web スクレイピングで作成した日本語コーパスの取り扱い

言語情報学の研究室では、日本語のウェブサイトを数万件収集し、クローラーを用いて自動的に文章を取得し、大規模なテキストデータを構築した。ニュースサイト、個人ブログ、掲示板、企業公式サイトなど多様なページから収集した結果、約 600GB の日本語テキストコーパスが完成した。研究室の教員や学生は、このコーパスを生成 AI の学習や言語研究に活用するだけでなく、国内外の研究者コミュニティと共有するため、大学の機関リポジトリで公開したいと考えている。

しかし、収集したウェブサイトにはそれぞれ利用規約やルールがあり、再配布を禁止するサイトや、robots.txt でクローラー利用を制限するサイトもある。著作権の扱いが明示されている場合もあれば、記載がなく解釈が難しい場合もある。このような条件下で収集したデータを大学からそのまま配布してよいのか、疑問が生じた。

研究室のメンバーは、学術研究目的での利用は法律上の特例(研究目的の例外やテキスト・データマイニング例外)に該当し、コーパスの利用や公開は基本的に許容され则认为している。また、大規模言語モデルの開発や言語研究の発展にはこの種のコーパスが不可欠であり、日本語研究基盤として世界的にも意義があると自負している。一方で、独自判断で公開することへの不安もあり、大学の知的財産部門に相談した。

大学の知的財産部門は、著作権や利用規約に違反する可能性を指摘し、本文を含むデータの公開には懸念を示した。テキストそのものは配布せず、URL と ID のみを記録として共有し、利用者が各自で本文を取得する「リハイドレーション方式」や、利用条件を付けて配布する方法などが提案された。

ちょうどその頃、この日本語コーパスの存在を知った研究者や民間の事業者から、利用希望の問い合わせがあった。研究室ではできるだけ提供して活用してほしいと考えているが、どのようなデータをどのような条件で提供してよいか、判断がつかない。