ブログコーパスの研究目的利用ガイドライン

第 1.0 版

国立情報学研究所 企画型共同研究
「大規模テキストコーパス整備における
個人情報等取り扱いの検討」
共同研究グループ

代表者　大山敬三
国立情報学研究所 コンテンツ科学研究系

2010 年 3 月 8 日

目次

まえがき

ブログを始めとするいわゆる消費者生成メディア（Consumer Generated Media。以下，CGM）の大規模テキストコーパス（以下，コーパス）は自然言語処理や情報検索などの研究分野においては不可欠の研究資源となっているばかりでなく，多くの研究者が共通のコーパスを用いることにより技術の客観的評価が容易になることにより研究の一層の発展が期待できる。しかし，コーパス整備のためには質の高いデータを大量に収集する必要があり多大なコストがかかる。そこで，例えばブログのホスティングなどを提供しているインターネットサービスプロバイダ（以下，ISP）等からデータの提供を受けて整備を進めることが有効である。一方，ISP 等ではサービス利用者や一般社会の信頼を獲得・維持することが重要視されており，とりわけ個人情報や反社会的情報等の取り扱いについては細心の注意が払われているところである。そこで，データ提供についての理解を ISP 等から得るためには，このような ISP 等における状況への理解と個人情報等に関する意識を研究コミュニティが共有し，コーパス利用上の規定を研究コミュニティが自主的に定め従うことにより，ISP 等から信頼を得ることが重要となる。

本ガイドラインは，このような認識に基づき，研究コミュニティの活動の一助とすべく，CGM のコーパス利用者が心得ておくべき諸事項を示すことを目的として，特にブログコーパスを中心的な対象として，国立情報学研究所共同研究「大規模テキストコーパス整備における個人情報等取り扱いの検討」により関連分野の研究者有志が策定したものである。

なお，本ガイドラインの考え方は，ブログ以外の CGM（例えば Q&A サイト）のコーパスにも通ずるところが多いので，本ガイドラインはそれらの利用者にも大いに参考になるものと期待される。

共同研究グループ名簿

国立情報学研究所 企画型共同研究（平成 20，21 年度）
「大規模テキストコーパス整備における個人情報等取り扱いの検討」
共同研究者

氏　名	所　属　・　職　名
大山 敬三	国立情報学研究所・コンテンツ科学研究系・教授（代表者）
東倉 洋一	国立情報学研究所・副所長・教授
安達 淳	国立情報学研究所・コンテンツ科学研究系・教授
大須賀 智子	国立情報学研究所・技術補佐員
喜連川 優	東京大学・生産技術研究所・教授
辻井 潤一	東京大学・大学院情報理工学系研究科・教授
石塚 満	東京大学・大学院情報理工学系研究科・教授
木戸 冬子	東京大学・大学院情報理工学系研究科・社会連携担当
黒橋 禎夫	京都大学・大学院情報学研究科・教授
奥村 学	東京工業大学・精密工学研究所・教授
前川 喜久雄	国立国語研究所・言語資源研究系・教授
山崎 誠	国立国語研究所・言語資源研究系・准教授（平成 21 年度）
堀下 剛司	ヤフー株式会社・ソーシャルネット事業部・企画部・リーダー（平成 20 年度）
堀野 亜紀	ヤフー株式会社・ソーシャルネット事業部・企画部（平成 20 年度）
寺岡 宏彰	ヤフー株式会社・メディア事業統括本部・メディアサービス本部ソーシャル企画部・チームリーダー（平成 21 年度）
東保 知子	ヤフー株式会社・メディア事業統括本部・メディアサービス本部ソーシャル企画部（平成 21 年度）

1. はじめに

1.1. ガイドラインの目的

本ガイドラインでは，ブログ記事を収集したコーパス（以下，コーパス）の利用者である研究者が心得ておくべき諸事項とその考え方を示す。

ブログ記事を収集したコーパス（以下，コーパス）では，元となる記事（原データ）に個人情報や反社会的情報などが含まれている場合があり，コーパスの使用によって法制上，倫理上，あるいは経済上の問題が発生する可能性がある。このため，コーパスを研究者に提供するに当たっては，原データを収集して提供する者（原データ提供者），原データからコーパスを作成する者（以下，コーパス作成者），及びコーパスを研究者に提供する者（以下，コーパス提供者。三者を併せて「コーパス提供者等」という）は，予め投稿者から研究目的での利用許諾を得たり，個人情報の除去等の可能な限りの処理をしたり，利用者を一定の範囲に限定したりするなど，データの準備と提供には十分な注意を払っている。しかし，これらによっても問題が発生する可能性をゼロにすることはできず，もし仮に一部のコーパス利用者の行為によりこのような問題が引き起こされた場合，その影響の範囲は当該コーパスの提供者等や利用者にとどまらない。

将来にわたってより多くの研究者にコーパスを提供可能とし，さらにコーパスの一層の拡充を実現していくためには，コーパス提供者等（及びその可能性のある者）からの一層の理解と信頼を得てゆくことが重要である。それには，コーパスの全ての利用者が，適切で節度あるコーパス利用の実績を積み上げることが必要である。

そこで本ガイドラインでは，上記のような問題の発生を防ぐとともにコーパス提供者等の理解と信頼を得るために，コーパス利用者が守るべき心得を定める。コーパス利用者は，著作権法や個人情報保護法などの諸法令，学会や関係機関の倫理規程，及びコーパスごとの契約や利用規程等を遵守することはもとより，本ガイドラインをその主旨を十分に理解して遵守しなければならない。

1.2. ガイドラインの対象

本ガイドラインは，言語学や自然言語処理，情報検索など（以下，自然言語処理等）の研究分野における，主に表現媒体としてのテキストを研究対象とする学術研究を対象とする。

コーパスは上記のような研究分野ばかりでなく，そのコンテンツによっては社会心理学や政治学などのさまざまな研究分野においても有用な研究資源となる。しかしこのような研究分野が研究対象とするものは，表現媒体としてのテキストというよりは，テキストに表現された個人やグループの心理や意見などであり，自然言語処理等の研究分野と同列に扱うことはできない。このため，自然言語処理等以外の研究分野は本ガイドラインの対象には含めない。

また，出版や製品の研究開発などの商用を目的とした利用については，より厳密な利用

条件等を定める必要があるため，本ガイドラインの対象には含めない。

1.3. 用語の定義

(1) コーパス提供者
コーパス利用者に対してコーパスを提供する者。なお，コーパスを構成するデータに何らかの権利が存在する場合は，権利者からコーパスの提供の許諾を得ているものとする。

(2) コーパス利用者
コーパス提供者からコーパスを利用することを許可された者。なお，本ガイドラインではコーパス提供者はコーパス利用者を継続的に把握しているものと想定している。

(3) 二次的公開
コーパス利用者が，論文，口頭発表，デモンストレーション等において，コーパスに含まれるデータまたはその一部を，原文のまま，または要約や言い替えを行って，コーパス利用者以外の者に対して示すこと。

(4) データ
コーパスを構成するひとまとまりのテキストの単位であり，「文書」と呼ばれることもある。ブログや新聞では記事，ウェブではページに相当する。

2. 利用上の心得

2.1. 管理上の注意事項

(1) データの保管
コーパスは情報セキュリティ対策の施されたサーバ等に格納し，コーパス利用者以外の者がデータにアクセスできないように適切な管理がなされなければならない。また，バックアップのメディア等についても十分注意して保管しなければならない。

(2) 利用者の管理
複数のコーパス利用者が共同でコーパスを利用する場合は，管理責任者を定め，必要に応じてコーパス利用者の名簿を作成しコーパスの利用状況を把握しなければならない。

2.2. 二次的公開が不適切なデータとその取り扱い

コーパスにはさまざまなデータが含まれている可能性があり，特に，ブログなどインターネット上のコンテンツに基づくコーパスの場合は，内容によるデータの選別などは行われないのが普通である。また，インターネットで公開されているデータに基づいてコーパスが作成されている場合であっても，コーパス作成後にさまざまな理由でデータが削除されることがある点に注意する必要がある。

このため，コーパスには，二次的公開を行うことが不適切な情報や表現が含まれる可能

性があり，以下に示すような取り扱いが必要となる。

(1) 個人情報，プライバシー侵害のおそれがある情報
一般的な個人情報，事件事故の被害者，犯罪の容疑者，裁判の被告，受刑者などについては，第三者が当事者を特定できる情報を公開してはならない。また，例えば当事者や関係者が知らない病名や戸籍などの情報を第三者が記載しているような場合では，当事者やその関係者が，当事者を特定または推定できることが問題となり得るため，このような情報を含むデータの公開は慎重に行うべきである。

(2) 誹謗中傷，脅迫，人権を侵害するおそれがある表現等
誹謗中傷や脅迫，あるいは身体障害，精神障害，出身，身分，職業等で不当に人権を侵害する表現などについては，これらを含むデータを公開してはならない。

(3) デマや詐欺，有害情報，公序良俗に反する表現，スパムデータ
デマや詐欺，性的・暴力的・差別的表現，犯罪や自殺の教唆，著作権を侵害する表現，スパムデータなどについては，学術的な知見の公表として行う場合には公開することは差し支えないが，公開媒体や読者層等を考慮して適切な換言や説明を行うことが必要である。ただし，2.3 節に該当する場合はそれに従うこと。なお，現在，多くの学術論文がインターネット上で一般に公開され，その全文が検索できるものも少なくないこと，また全ての学術論文が将来は同様に公開される可能性が高いことから，潜在的には全てのインターネットユーザが読者となり得ることに注意を要する。

2.3. 削除情報等の取り扱い

コーパス提供者等が公開すべきでないと判断した情報や表現などについては，予め削除や伏せ字処理など（以下，削除処理等）を行った状態でコーパスとして提供されることがある。これらは，仮に機械的処理や他の情報源の参照などにより技術的に復元などが可能であっても，コーパス利用者はそのような行為を行ってはならない。

また，技術的な不完全性のために，本来なら削除処理等が行われるべきものがそのまま残っているような場合も，コーパス提供者等の意図に即したコーパスの利用を心掛けなければならない。例えば，研究発表においてそのような部分を取り立てて紹介するような行為を行ってはならない。

2.4. データ削除依頼への対応

コーパス提供者がコーパス利用者にコーパスを提供した後に，何らかの理由によりコーパス提供者が一部のデータの削除を依頼することがある。このような場合，コーパス利用者は当該データとその複製を削除すること。また，当該データを処理することにより得ら

れた情報についても可能な限り削除すること。

2.5. 対応するインターネットデータの取り扱い

コーパスに含まれるデータの元となっている実データがインターネット上のサービスで公開されている場合がある。コーパス利用者がそのようなサービスやデータを利用することは問題ないが，コーパスは，インターネット上のサービスとは異なったリスクに関する判断基準に基づいて提供されていることを理解し，以下の事項に注意しなければならない。

(1) 対応付け及び評価の禁止

コーパスに含まれるデータは大量一括処理の対象となるため，原データの著者が想定していなかった情報が取得される可能性がある。このため，コーパス提供者が明示的に許可している場合を除き，コーパス中のデータや著者とインターネット上のデータや著者との対応付けを行ったり，また，対応のとれないデータや著者の検出や分析等を行ったりしてはならない。

(2) コーパスに準じた実データの取り扱い

コーパスとは独立に，対応するインターネット上の実データをクローリングして利用する場合においても，コーパス提供者等のコーパスを提供する目的や設定した利用条件等の主旨を十分に理解し，コーパスを利用する場合に準じて実データの取り扱いを行うように心がけなければならない。

3. まとめ

本ガイドラインはブログ記事を収集したコーパスを研究目的で利用する研究者に求められる一般的な心得をまとめたものである。個々のコーパスの提供に当たっては，それぞれのコーパス提供者等は，本ガイドラインを参考にしつつ，コーパスごとの実状と関係者の意向を十分に検討した上で，より具体的な規程等を定めることが期待される。また，研究者は，本ガイドラインとともに，提供を受けるコーパスについて定められた規程等を遵守することが求められる。本ガイドラインが，研究者による節度ある適正なコーパスの利用に貢献し，研究コミュニティの活動の発展の一助となることを期待する。