Vol.2016-CLE-20 No.3
2016/11/18

MOOC データ二次利用のための加工プロセス

武田俊之[†1]　重田勝介[†2]　森秀樹[†3]

概要：教育・学習履歴データの研究利用においては，学習者の個人情報およびプライバシーの保護が必要である．特にデータ収集時に許諾を明示していない研究において，そのデータを扱う場合には匿名加工などの配慮が必要である．本論文では，ある MOOC コースにおいて得られたデータを，新たな研究課題において利用するためにおこなった，データ加工の検討事項，プロセス，課題について記述する．

キーワード：オンライン・コース，学習履歴データ，プライバシー保護，匿名化

1. はじめに

教育における成果や行動履歴を共有することによって，教育，教材の改善や学習者の行動に関する研究が促進することが期待されている．特に，MOOC (Massive Open Online Course)のように，大人数の受講者がオンライン上で学習をおこなった結果得られるデータは，研究上重要な資源となりうる[1].

しかし，学習行動履歴データの共有には，学習者のプライバシーが侵害される懸念がある．学習履歴によるプライバシー侵害は,差別や将来の不利につながる可能性がある．このような懸念から，児童の学習データリポジトリ構築を目指していた非営利団体 inBloom に対して，プライバシーの観点から反対意見が相次ぎ，2014 年 4 月に活動の停止を発表している．

改正個人情報保護法においては，本人の同意があるか，匿名化加工したデータは共有することが可能である．しかし，教育サービス提供者が学習者からデータの利用許諾を強要することは，望ましくない．許諾時に提供されている情報は，学習者と教育機関の間で非対称である．学習者が実質的に同意する内容を理解していない可能性がある．データの二次利用許諾の影響を，学習者が事前に予測することは困難であろう．

したがって，教育データを二次利用，第三者利用するためには，許諾だけで二次利用するのではなく，匿名化もあわせておこなうことが,現実的な選択肢であろう.しかし，教育研究や教育改善における有用性を維持しながら，個人を特定されないように学習履歴データを匿名化する手法や評価指標は確立されていない．

教育データの提供者，分析者にとっても，プライバシー・リスクのあるデータは利用しづらい．学習履歴データの共有が進展するためには，学習者およびデータ所有者のプライバシーに関する不安を取りのぞくことが必要である．

本論文では,ある MOOC コースのデータについて研究利用許諾をもつ研究者が，許諾を受けていない研究者と共同研究をおこなうために，データの匿名化の手法,プロセス,検討事項について論じる．

2. 匿名化の手法と実践の動向

2.1 用語について

本論文で用いる用語の意味は以下の通りである．

個人データ（パーソナルデータ）は，個人に関する情報であって，氏名等の個人識別情報（識別子; identifier）によって個人を特定できるものである．準識別子（quasi-identifier）は，個人識別情報ではないが，特定の個人を識別することができる情報で，行動履歴，ゲノムなども含まれる可能性がある．

ここで,「特定」とはある個人とデータを結びつけることである．識別（identification）はデータが誰か一人の人と結びついているとわかることである．

仮名化（pseudonymization）は，データから個人を特定しづらいように,識別子を取り除くか,変換することである．仮名化だけでは，プライバシー保護のためには不十分である．

匿名化（匿名加工，非識別化，anonymization / de-identification）とは，個人情報から識別性を取りのぞくことである．匿名化されたデータを非識別化データ，de-identification data）という．改正個人情報保護法では匿名加工情報とも呼ばれる．非識別化データから，特定の個人を識別することを再識別（re-identification）という．

2.2 匿名化の手法と指標

以下の技術を用いて，個人情報の識別性を低減する．

1) マスキング: 属性やその組み合わせを仮名化または削除する．
2) 属性の加工: 属性の一般化（抽象度において上位の値に置きかえる），トップ（ボトム）コーディング（トリミング），値の置換えなどをおこなう．
3) サンプリング（sampling）：データをランダムに抽出す

†1 関西学院大学
Kwansei Gakuin University
†2 北海道大学
Hokkaido University
†3 東京工業大学
Tokyo Institute of Technology

る．抽出の比率が低いほど識別性が低減する．

4) 統計: 統計量を取り，データを縮約する．識別性の評価や低減のために，統計開示制御（statistical disclosure control）や差分プライバシー（differential privacy）が用いられることがある．
5) k-匿名化: 頻度が k 以上となるよう（k-匿名性），準識別子の情報を変換する．

2.3 医療におけるデータの第三者利用の枠組

研究者によるデータ利活用で先行する医療分野では，第三者による研究を促進するために，個人データ匿名化の手法の開発や，枠組の構築の実践が積み重ねられている．オンタリオ州では，州法で認められた研究機関（CHEO）が，医療データのレジストリとプライバシー評価をおこなうために設置されている．CHEO は，以下のプロセスで医療データを取りあつかう．①医療機関がレジストリへデータを登録，②研究者または民間企業が匿名化データを研究プロトコルとデータチェックリストとともにCHEOにリクエスト，③科学審査委員会による研究プロトコルのチェック，データアクセス委員会（DAC）による研究プロトコルとデータチェックリストの審査，④DAC が匿名化方法を研究者と交渉の上決定，⑤研究倫理委員会による提供リスクと倫理面の審査，⑤データ管理者が匿名化方法にしたがってデータを匿名化，⑥安全に匿名化データを研究者，民間企業へ提供．このプロセスは[2]にくわしい．これらの手続きは煩雑であるが，オンタリオ州ではリクエストから 1-2 週間でデータを提供している．

2.4 教育におけるデータ匿名化

(1) HarvardX-MITx データセット

非営利の MOOC プロバイダーHarvardX と MITx は，研究利用のために，匿名化した MOOC コースのデータを公開している．公開されているデータは，edX プラットフォーム上で実施された，HarvardX と MITx の初年度（2 学期）の延べ16 コース分である．その匿名化プロセス[3]および，分析結果は公開されている[4].

コースから得られる生データには，米国の FERPA (Family Educational Rights and Privacy Act)によって保護される項目をはじめとして，個人を特定可能な情報が含まれている．これらの情報は，専門家が推奨するベスト・プラクティス（aggregation, 識別子のランダム化, あいまい化など）によって非識別化された．表 1 は匿名化前後のレコード数の比較である．

HarvardX-MITx データセットは，公開後 1 ヶ月間で，560 回のダウンロードを数えた．

同じ匿名化プロセスによって，Instructure 社が 238 コース，325,000 件のデータを公開している．a

表 1 HarvardX-MITx データセットにおける
匿名化前後のレコード数の比較[3]

INSTITUTION	COURSE	TERM/YEAR	BEFORE	AFTER
HarvardX	CB22x	2013_Spring	43555	30002
HarvardX	CS50x	2012	181410	169621
HarvardX	ER22x	2013_Spring	79750	57406
HarvardX	PH207x	2012_Fall	61170	41592
HarvardX	PH278x	2013_Spring	53335	39602
MITx	14.73x	2013_Spring	39759	27870
MITx	2.01x	2013_Spring	12243	5665
MITx	3.091x	2012_Fall	24493	14215
MITx	3.091x	2013_Spring	12276	6139
MITx	6.002x	2012_Fall	51394	40811
MITx	6.002x	2013_Spring	29050	22235
MITx	6.00x	2012_Fall	84511	66731
MITx	6.00x	2013_Spring	72920	57715
MITx	7.00x	2013_Spring	37997	21009
MITx	8.02x	2013_Spring	41037	31048
MITx	8.MReV	2013_Summer	16787	9477
		TOTAL	**841687**	**641138**

(2) LearnSphere

LearnSpherebは，カーネギーメロン大学が中心となった教育研究者，コミュニティのためのデータレポジトリである. LearnSphere には 550 以上のデータセットが共有されている．

3. MOOC コースとデータの概要

本節では，当該の MOOC コースと，そこで得られたデータについて述べるc.

3.1 コースについて

このオンライン・コースは，MOOC プロバイダーdが制作･配信する 4 週間のコース（スケジュールは図 1 を参照）であり，講師 3 名とティーチング・アシススタント 1 名によって開講された．コースは Open edXeをプラットフォームとしたオープン・アクセスのコースであり，募集開始が 2014 年 4 月，配信開始が 7 月 7 日であった．一般の受講者は 8,118 アカウントであった．

コースは, 週 1 回ごとに 4 週間配信される講義ビデオと，各講義回のセクションごとに付属するクイズ，期中 2 回のレポート，ディスカッション・フォーラム，から構成されている．成績はクイズ（配点: 各週 15 点）とレポート（1 回目 10 点, 2 回目 30 点）の合計 100 点満点で採点される．クイズと第 1 回のレポートはシステムによる自動採点，第 2 回目のレポートは受講者の相互評価による採点であった．オンライン・コースの受講にくわえて，1 回の対面授業をおこなう「反転授業」も同時期に実施された．

a https://dataverse.harvard.edu/dataverse/cn

b http://learnsphere.org/about.html
c コースのコンテンツ，レポート，アンケート結果などの統計は，http://www.daigomi.org/jmooc14-openedu/で公開されている．
d http://gacco.org/
e https://open.edx.org/

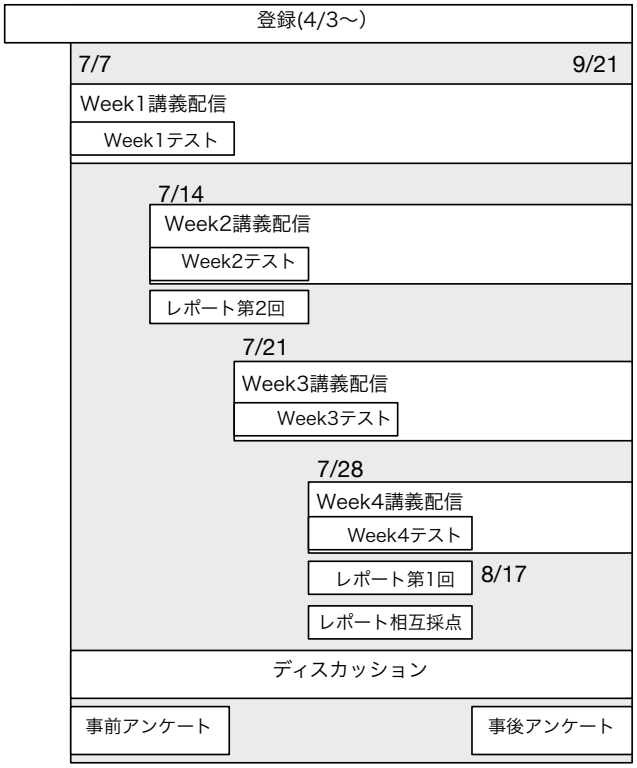

図 1 コースのスケジュール

3.2 学習履歴データの契約上の取り扱い

コース上で得られたデータを，プロバイダーおよび講師が，教育改善および研究のために利用する許諾は，受講者とプロバイダーの契約において得られている．

さらに，このコースの特記事項として，コース中のレポート（成績評価対象）を，氏名削除の上で公開する許諾を得たことが挙げられる．これは，レポート課題の教示において説明された．受講者がレポートの公開を望まない場合は，レポート内にその旨記述することでオプトアウトを可能とした．

3.3 データセットの作成

コースから得られたデータから，分析用データセットを作成した．

データの収集方法は，(1) edX システムの講師画面からダウンロード，(2) コースのページを保存またはクローリング，(3) 運営事務局より入手，のいずれかである．筆者らコースの講師は，講師画面と受講生画面以外の，システム管理者の機能へはアクセスできない．

得られた生データ群に，連結，分割，整理の処理をおこない，データセットを作成した．作成したデータセットは以下の通りである．レコード数などの追加情報を表 1 に記述している．

1) 受講登録データ
コースへの受講登録時に入力した項目等（アカウント名，氏名，性別，生年，教育歴, email, 反転授業参加の有無）．

2) 成績
各クイズ，各レポートの得点，合計点，補正された最終成績．

3) 理解度クイズ
各週各回（5～6 回）のビデオ講義において，それぞれ 1～2 問の理解度クイズを設けている．このデータは，それぞれの問についての，解答，正解，得点から成る．

4) 第 1 回レポート
第 1 回レポートは，オープンエデュケーションのサービスやキーワードを 3 から 5 個収集して，そのタイトル，URL，簡単な説明をするものであった．この内容および提出日がデータ項目である．

5) 第 2 回レポートおよび相互採点結果
第 2 回レポートは，「オープンエデュケーションが広まる世界に生きるある架空の人物のストーリーを想像し，その人生にオープンエデュケーションがどのように関わっていくか」について，800 字程度記述するものであった．内容はルーブリックにしたがって，他の受講者によって採点される．レポートは公開されることが告知された．レポートの公開を望まない場合は，その旨レポートに記入しておくよう教示した．
データ項目は，タイトル，レポート本文，提出日，得点（相互採点結果），被採点人数，である．
他の受講者の第 2 回レポートを採点した結果．

6) ディスカッション・フォーラム
ディスカッション・フォーラムのページをクローリング，パースして，投稿の種類，内容（テキスト），投稿への vote，時刻を収集した．

7) トラッキング・ログ（学習履歴）
Open edX ではユーザーのクリックストリーム等のイベントデータを，トラッキング・ログとして保存しているf．トラッキング．ログは，ブラウザ上のビデオプレイヤーなどの操作と，サーバー上のイベントを記録する JSON ファイルである．
8) 事前アンケート結果
コース開始時に受講者に対しておこなった事前アンケートの結果．
9) 事後アンケート結果
テスト，レポート提出最終日の翌日におこなった，受事後アンケートの結果．
10) コンテンツ
講師が作成したコンテンツ等の講義資料は，コースを構成するデータでもある．本コースにおいては，プレゼンテーション資料，講義映像，講義トランスクリプション，レポート課題，採点基準がコンテンツ・データである．

以上のデータセット（ディスカッションを除く）から，講師とスタッフのデータを取りのぞき，分析用の基礎データセットを作成した．トラッキング・ログとコンテンツ以外のデータはアカウント名と連結している．

3.4 安全管理のための仮名化

安全管理のために，基礎データセットの仮名化をおこない，仮名化データセットを作成した．仮名化の内容は，(1) ユーザーID の仮名への置きかえとその対照表の作成，(2) 個人への到達性の高い符号（ユーザーID，氏名，email, IP アドレス）の削除，の 2 点である．

ユーザーID と仮名の対照表，生データ群，仮名化前の基礎データセットは，仮名化データセットとは別に保存しておく．

4. データ二次利用の方針と匿名化の実施

本件は講師＝研究者のコースであり，データの利用許諾を得ている者が，前節で作成した仮名化データセットを用いる研究は，法的，倫理的，安全管理上において，問題がない．

しかし，データ利用の許諾を得た研究者と，他の研究者を含むチームにおいて，データセットを利用することは，第三者によるデータの二次利用の特殊なケースである．この場合の匿名加工のあり方について，以下のように検討をおこない，匿名化をおこなった．

f トラッキング・ログで記録されるイベント仕様はすべて公開されている．http://edx.readthedocs.io/projects/devdata/en/latest/internal_data_formats/tracking_logs.html

表 2 コースのデータから作成されたデータセット一覧

データの種類	入手形式/入手方法	備考
受講登録データ	CSV / ダウンロード	受講取消等のため，成績データとミスマッチあり．
成績	Excel / 事務局	最終成績および各課題得点．反転授業申込者 60 名．
理解度クイズ	CSV / ダウンロード	第 1 週 12 問，第 2 週 10 問，第 3 週 10 問，第 4 週 12 問
第 1 回レポート	CSV / ダウンロード	
第 2 回レポート 相互採点結果	テキスト CSV / 事務局	CSV ファイルで提供されたものを加工．
ディスカッション・フォーラム	テキスト等 / クローリング	レコード数には講師，スタッフ（6 アカウント）を含む
トラッキング・ログ	JSON / 事務局	
事前アンケート	CSV / ダウンロード	12 項目
事後アンケート	CSV / ダウンロード	30 項目
コンテンツ スライド 講義ビデオ ビデオ書起こし	PDF / ダウンロード MP4/ローカル テキスト / クローリング	

4.1 データの二次利用をおこなう研究の概要

本件において，データセットの二次利用をおこなう研究課題（JSPS 科研費 JP15H02922）について説明する．

この研究課題の目的は，①学習履歴データから教材改善と教育改善に有効なデータを抽出する手法の開発と，②その手法と，指標を可視化するダッシュボードを開発することである．以下の内容がこの研究開発に含まれる．

1) 教材改善に有効なデータを抽出する手法
受講者それぞれの講義ビデオ視聴時間や再生・一時停止等のクリック動作，講義ビデオに付随する知識確認テストの正答率や正答に至るまでの試行回数
2) 教育改善に有効なデータの抽出
受講者の学習時間やログイン時間などの傾向，電子掲示板における議論への参加状況などから，教材・教育の改善で用いられる指標の抽出と算出
3) ダッシュボード開発
4) 受講者のクラスタリング

4.2 データ二次利用の方針

当該研究課題は，教育データを利用する研究として以下の特徴を持つ．

1) データを得たコースの講師かつデータの研究利用許諾を得たメンバーがいる．
2) データの研究利用許諾がない TA を研究者に含む．
3) データの利用許諾がない研究者を含む．

4) コースの受講者であった研究者を含む.

データの利用者に，コースの受講者またはスタッフを含むということは，データから個人を特定・識別するリスク要因であるg. したがって，データそれぞれの識別リスクを考慮した適切な匿名加工をおこなわねばならない.

本研究では，以下の方針でデータを取りあつかうこととした.

1) 充分に安全で実用的な方法によって，匿名加工データを作成する．研究では，原則として匿名加工データを用いる.
2) 各研究者は識別を試みない．一方で，識別可能と考えられる個人データが発見された場合には，匿名化の手順を見直して，再度匿名加工データを作成する．その際，問題のある匿名加工データは破棄する.
3) 利用許諾を得た研究者をメンバーに含む利点を活かし，匿名加工データ分析に用いた分析法を，元のデータに適用することによって，分析の有用性を確保する.
4) データによっては，個人と連結しないことによって，識別別リスクを低減する.
5) 匿名加工データは研究メンバーのみで共有する.

図 2 に，オンライン・コース，各研究，データの流れを示す。

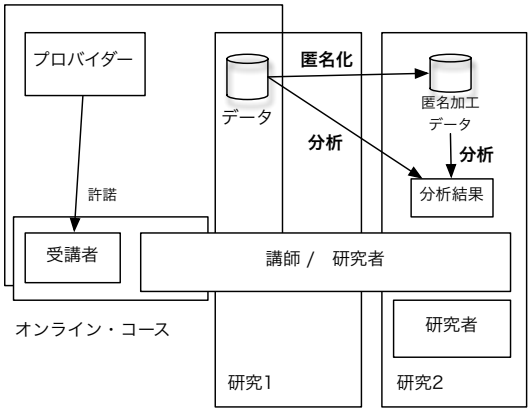

図 2 コース，研究，データの流れ

4.3 匿名加工方法の検討

各データについて，以下のように検討をおこなった.

識別子，準識別子の削除

氏名，email アドレスの識別子および準識別子は，仮名化において，削除されている．生年はすべての受講者が入力しているものではないが k-匿名性を評価する必要がある.

ディスカッション・フォーラム

ディスカッション・フォーラムへの投稿は，受講者全員が読むことが可能である．したがって，投稿内容から投稿者を特定可能である．また，投稿数等の統計量は投稿者を特定または識別可能が高い集計値である．これらと投稿内容や投稿数と仮名化データを連結した場合，個人の特定リスクが高くなる.

第 1 回レポート

第 1 回レポートは客観性の高い内容を求める課題であり，匿名加工の必要はない.

第 2 回レポート

第 2 回レポートは，仮想の個人のライフヒストリーに関わる内容である．受講者によっては，仮想人格に個人データを反映している．仮名とレポートを直接連結することはリスクが高い.

トラッキング・ログ

クリックストリーム等の学習者の操作履歴から，個人を識別しうるかどうかは未知である．しかし，今回のデータは, ビデオプレイヤーの操作と, ブラウジングのみであり，個人を識別性はないと判断した.

アンケート

アンケートの自由記述項目において，個人を識別しうる内容があれば，削除する.

以上の検討にもとづき，以下の匿名加工の実施を計画した.

1) 簡易で効果の高い匿名化の方法としてサンプリングをおこなう．成績データの 1/4 の件数を抽出する.
2) サンプリング前に必要な項目の Ho（2014）と同様に k=5 として，k-匿名性を評価する．k-匿名とみなされない場合は，データのトリミングをおこなう.
3) ディスカッション・フォーラムは，他のデータと連結しない．教材，講義との関連を中心に分析をおこなうこととする．ディスカッションの統計量と，成績，学習行動の関連の分析が必要な場合は，k 匿名化，差分プラバシーなどを用いて匿名化をおこなう.
4) 第 2 回レポートは，個人の属性を反している可能性が高いため，他のデータとは連結しない.

4.4 匿名加工の実施

以上の方針にもとづいて，匿名加工と連結を実施した．表 2 に，匿名加工前後のレコード数と，データに含まれる受講者数を示す.

(1) 仮名化された成績データからランダムに 1/4 件（2030 件）の抽出おこなった.
(2) 個人識別の可能性がある生年の項目について，年齢が 12 歳以下および 80 歳以上になるようトリミングをおこなった．これによって，k=5 の匿名性を実現した.
(3) 各匿名加工データを，仮名によって連結した.

g これは，個々のデータ利用者を信用しないという意味ではなく，リスクの評価であることは言うまでもない.

表 2 匿名加工データのレコード数と含まれる受講者数

データの種類	元データ		匿名加工後	
	レコード数	受講者数	レコード数	受講者数
成績・受講登録データ	8118	8118	2030	2030
理解度クイズ	31674	1941	8084	492
第 1 回レポート	1006	1006	254	254
第 2 回レポート 相互採点結果	901	901	122	122
ディスカッション・フォーラム	1253	250	-	-
トラッキング・ログ	1389218	6881	353761	1712
事前アンケート	2637	2637	649	649
事後アンケート	510	510	122	122

5. まとめと今後の課題

本研究では，オンライン・コースから得られたさまざまなデータについて，その二次利用を目的として，個人の識別可能性を低減するような,匿名加工する方法を探索した．本研究は，公開データの作成ではなく，また，二次利用者に個人データ取り扱いの許諾を得た研究者が含まれるケースであった．その点で，Ho ら[3]のように第三者に提供する公開データの匿名加工とは異なる．

本研究のような教育データの共有を目的とした匿名化について，以下のような技術の研究開発が必要である．

(1) 学習行動履歴からの個人識別性の評価法
(2) レポートのようなテキストデータからの個人識別性の評価法
(3) ディスカッション・フォーラムのように受講生が個人を識別・特定可能なデータと，成績等のデータの連結手法
(4) 複数コースのデータ匿名化，連結手法
(5) 匿名加工支援ツール
(6) 算出・推定されたプロファイルの個人識別性

これ以外にも，受講者からの許諾について，その範囲（内容，非許諾者）の提示方法，教育実践との許諾，オプトアウト手続きの妥当性などが洗練されなければならない．

謝辞 この研究の一部は，JSPS 科研費 JP16K12564 の助成を受けている．オンライン・コースの開講，データの提供について，gacco と JMOOC に感謝する．

参考文献

[1] Reich, Justin. Rebooting MOOC research. Science. 2015, 347.6217 p. 34-35.
[2] El Emam, Khaled, and Luk Arbuckle.. Anonymizing health data: case studies and methods to get you started. O'Reilly Media, Inc., 2013. （邦訳 木村映善, 魔狸訳. データ匿名化手法–ヘルスデータ事例に学ぶ個人情報保護. オーム社. 2015.）
[3] Ho, A. D., Reich, J., Nesterko, S., Seaton, D., Mullaney, T., Waldo, J., and Chuang, I.. Person-Course De-identification Process. 2014. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2381263
[4] Ho, A. D., Reich, J., Nesterko, S., Seaton, D. T., Mullaney, T., Waldo, J. and Chuang, I. arvardX and MITx: The first year of open online courses, fall 2012-summer 2013. 2014, HarvardX and MITx Working Paper No. 1.