Błażej Marciniak
Uniwersytet Łódzki; Konsorcjum BBMRI.pl

Dominik Strapagiel
Uniwersytet Łódzki; Konsorcjum BBMRI.pl

Piotr Topolski
Uniwersytet Łódzki

ANONIMIZACJA W DOBIE WIELKICH DANYCH – SYTUACJA BIOBANKÓW W KONTEKŚCIE RODO[1,2]

Skok technologiczny w szeroko rozumianej dziedzinie informatyki, którego byliśmy świadkami na przestrzeni ostatnich lat, oraz spadek cen mocy obliczeniowych (średnio co trzy lata koszt jednostki obliczeniowej zmniejsza się o połowę)[3] umożliwiły przetwarzanie i analizę danych na poziomie, jaki nie był nigdy wcześniej dostępny. Utworzony został nawet termin opisujący zjawisko wielkich danych – *Big Data*. W innych dziedzinach jest podobnie, i tak, na nieznaną dotychczas skalę masowe kolekcjonowanie ludzkiego materiału biologicznego oraz przetwarzanie go do formy cyfrowej, wielokrotne używanie próbek w badaniach oraz łączenie danych dotyczących źródła próbki (człowieka) z danymi na poziomie „krajowych rejestrów zdrowia pacjentów" czy z danymi genetycznymi lub informacji z mediów społecznościowych przyczynia się do wzrostu wiedzy i w przyszłości niewątpliwie przyczyni się do poprawy zdrowia społeczeństwa[4,5]. W tym miejscu należy postawić pytanie, czy sytuacja ta nie niesie ze sobą także zagrożeń.

[1] Artykuł powstał w ramach projektu pt. „Utworzenie sieci biobanków ludzkiego matreiału biologicznego w Polsce w obrębie Infrastruktury Badawczej Biobanków i Zasobów Biomolekularnych BBMRI-ERIC" realizowanego na postawie decyzji nr DIR/WK/2017/01.

[2] Praca częściowo finansowana w ramach projektu „Infrastruktura Badawcza Biobanków i Zasobów Biomolekularnych BBMRI-ERIC" – DIR/WK/2017/01.

[3] G. O'Connor, *Moore's Law Gives Way to Bezos's Law*, GigaOm, April 2014, https://gigaom.com/2014/04/19/moores-law-gives-way-to-bezoss-law/ (dostęp: 3.03.2018 r.).

[4] F. F. Costa, *Big Data in biomedicine*, „Drug Discovery Today" 2014, t. 19, nr 4, s. 433–440.

[5] S. J. Mooney, D. J. Westreich, A. M. El-Sayed, *Epidemiology in the era of Big Data*, „Epidemiology" 2015, t. 26, nr 3, s. 390–394.

Wspomniany rozwój technik i możliwości pozyskiwania oraz analizy danych wymusza jednak zwrócenie uwagi na istniejące rozwiązania prawne. Zwłaszcza w kontekście ochrony danych osobowych. Rozwijające się liczne portale społecznościowe[6], sklepy internetowe, wyszukiwarki, np. google, oraz konta chmur obliczeniowych gromadzą dane i profilują swoich użytkowników, którzy notabene sami nagminnie udostępniają prywatne informacje o sobie (nie tylko w kontekście imion i nazwisk, numerów PESEL, lecz także modeli zachowań, preferencji itp.). Niezależnie od beztroski i łatwości w udostępnianiu swoich danych ludzie są narażeni na ciągłą obserwację niejako bez własnej świadomości, np. przy wykorzystaniu systemów zarządzania relacjami z klientem, programów lojalnościowych[7] itp. Problem ten został zauważony przez administrację Unii Europejskiej, wypracowując nowe regulacje w zakresie ochrony danych osobowych zastępujące dyrektywę 95/46/WE Parlamentu Europejskiego i Rady WE przez rozporządzenie Parlamentu Europejskiego i Rady (UE) 2016/679 z dnia 27 kwietnia 2016 r. w sprawie ochrony osób fizycznych w związku z przetwarzaniem danych osobowych i w sprawie swobodnego przepływu takich danych oraz uchylenia dyrektywy 95/46/WE (ogólne rozporządzenie o ochronie danych)[8] (dalej: RODO). Rozporządzenie to wprowadza znaczące ograniczenia dotyczące profilowania. Jednocześnie, samo profilowanie definiuje jako zautomatyzowane przetwarzanie danych osobowych polegające na ocenie czynników osobowych człowieka w celu analizy lub prognozy aspektów dotyczących pracy tej osoby, jej sytuacji ekonomicznej, stanu zdrowia, osobistych preferencji, zainteresowań, wiarygodności, zachowania, lokalizacji lub przemieszczania się. Definicja ta zawarta jest w art. 4 pkt 4 RODO. W kontekście badań biomedycznych mianem *Big Data* określa się duże pakiety danych, w tym dotyczących stanu zdrowia ochotnika uczestniczącego w badaniach, często w powiązaniu z danymi genetycznymi, informacje o zabiegach i świadczeniach medycznych. Przyjęta definicja ogólna zakłada współwystępowanie 3 V od angielskiego: *volume* – objętość, *velocity* – szybka dostępność i zmienność w czasie oraz *variety* – zróżnicowanie[9]. Obecnie przyjmuje się, że za tym terminem kryją się techniczne i analityczne metody do wyciągania interesujących informacji z kompleksowych i różnorodnych pakietów danych. Za źródła danych szczególnie istotnych dla badań medycznych uznaje się elektroniczne rejestry medyczne, dane wywodzące się z badań klinicznych, dane genomowe i inne dane klasyfikowane do danych omicznych (np. proteomicz-

[6] Zob. http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/ (dostęp: 3.03.2018 r.).

[7] Zob. http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html (dostęp: 3.03.2018 r.).

[8] Dz.Urz. UE L 119 z 4.05.2016, s. 1.

[9] M. May, *Life Science Technologies: big biological impacts from Big Data*, „Science" 2014, t. 344, nr 6189, s. 1298–1300.

nych)[10], dane behawioralne. Jakkolwiek dostęp do tego typu danych może wydawać się pogwałceniem zasad ochrony danych osobowych, jest on niezbędny do prowadzenia nowoczesnych badań naukowych, zwłaszcza z zakresu szeroko pojętej genetyki. Takie stanowisko prezentuje też Wojewódzki Sąd Administracyjny w Warszawie w wyroku z dnia z 14 lutego 2012 r., II SA/Wa 2418/11[11], który oddalił skargę NFZ na decyzję GIODO nakazującą udostępnienie wrażliwych danych osobowych pacjentów leczonych na oddziale kardiologii Śląskiego Uniwersytetu Medycznego. Szpital domagał się ujawnienia informacji o przyjmowanych przez pacjenta świadczeniach medycznych w innych placówkach, jeśli nie był leczony na miejscu. Chodziło o ustalenie skuteczności w czasie zastosowanych metod leczenia, tzw. *follow up*[12]. Warto nadmienić, że ustawa z dnia 29 sierpnia 1997 r. o ochronie danych osobowych[13] w obecnej formie dopuszcza przetwarzanie danych wrażliwych w art. 27 ust. 2 pkt 9, gdy „jest to niezbędne do prowadzenia badań naukowych, w tym do przygotowania rozprawy wymaganej do uzyskania dyplomu ukończenia szkoły wyższej lub stopnia naukowego; publikowanie wyników badań naukowych nie może następować w sposób umożliwiający identyfikację osób, których **dane** zostały przetworzone".

Powszechnie przyjętą metodą stosowaną celem zabezpieczenia danych osobowych jest anonimizacja (ujęcie europejskie, początek XXI wieku), kodowanie (ujęcie amerykańskie, początek XXI wieku)[14]. Definicję anonimizacji zawiera norma ISO 29100:2011, która uznaje, że jest to proces, w wyniku którego informacja identyfikująca daną osobę zostaje nieodwracalnie zmieniona w taki sposób, iż dana osoba (podmiot informacji) nie może być już zidentyfikowany pośrednio lub bezpośrednio ani samodzielnie przez administratora danych osobowych, ani we współpracy z jakimkolwiek innym podmiotem[15]. Jakkolwiek należy pamiętać, że anonimowość w kontekście dostępności danych genomowych jest niemożliwa do osiągnięcia, gdyż teoretycznie do zidentyfikowania osoby wystarczy informacja na temat 30–80 niezależnych względem siebie polimorfizmów pojedynczych nukleotydów w DNA[16]. Staje się to problematyczne zwłaszcza w sytuacji coraz częstszych tzw. ponownych użyć kolekcji materiału biologicznego wraz z dostępną informacją genetyczną do opracowywania kolejnych prac oryginal-

[10] V. Marx, *Biology: the big challenges of Big Data*, „Nature" 2013, t. 498, s. 255–260.

[11] Legalis.

[12] Zob. http://www.rp.pl/artykul/812483-Informacje-wrazliwe-pacjenta-moga-sluzyc-do-badan-naukowych.html (dostęp: 3.03.2018 r.).

[13] Tekst jedn. Dz.U. z 2016 r., poz. 922, z późn. zm.

[14] B. Elger (i in.), *Consent and anonymization in research involving biobanks*, „EMBO reports" 2007, t. 7, nr 7, s. 661–666.

[15] Zob. https://www.iso.org/standard/45123.html (dostęp: 3.03.2018 r.).

[16] Z. Lin, A. B. Owen, R. B. Altman, *Genomic research and human subject privacy*, „Science" 2004, t. 305, nr 5681, s. 183.

nych. Jednakże na mocy definicji danych osobowych zawartej w ustawie o ochronie danych osobowych: „(...) za **dane osobowe** uważa się wszelkie informacje dotyczące zidentyfikowanej lub możliwej do zidentyfikowania osoby fizycznej" (art. 6 ust. 1) i „Informacji nie uważa się za umożliwiającą określenie tożsamości osoby, jeżeli wymagałoby to nadmiernych kosztów, czasu lub działań" (art. 6 ust. 3). Jeszcze obecnie koszt przebadania na tyle szerokiej populacji, by na podstawie genomu ustalić tożsamość przypadkowej osoby, jest tak duży, że pozwala wykluczyć dane genomowe z listy danych osobowych. Problem stanie się realny za kilka lat, gdyż już dziś koszt sekwencjonowania genomu człowieka spadł do około jednego tysiąca dolarów ze stu milionów w 2001 roku[17], a urządzenia służące do sekwencjonowania DNA stają się niewiele większe od *pendrive'a*[18]. Należy też zwrócić uwagę na fakt, że RODO definiuje dane genetyczne i biometryczne jako dane wrażliwe.

Niestety – z punktu widzenia naukowca – RODO zmienia definicję danych osobowych: „«dane osobowe» oznaczają informacje o zidentyfikowanej lub możliwej do zidentyfikowania osobie fizycznej («osobie, której dane dotyczą»); możliwa do zidentyfikowania osoba fizyczna to osoba, którą można bezpośrednio lub pośrednio zidentyfikować, w szczególności na podstawie identyfikatora takiego jak imię i nazwisko, numer identyfikacyjny, dane o lokalizacji, identyfikator internetowy lub jeden bądź kilka szczególnych czynników określających fizyczną, fizjologiczną, genetyczną, psychiczną, ekonomiczną, kulturową lub społeczną tożsamość osoby fizycznej" (art. 4 pkt 1). Co więcej, zgodnie z art. 9 pkt 1 RODO: „Zabrania się przetwarzania danych osobowych ujawniających pochodzenie rasowe lub etniczne, poglądy polityczne, przekonania religijne lub światopoglądowe, przynależność do związków zawodowych oraz przetwarzania danych genetycznych, danych biometrycznych w celu jednoznacznego zidentyfikowania osoby fizycznej lub danych dotyczących zdrowia, seksualności lub orientacji seksualnej tej osoby". Wykorzystanie takich danych do badań naukowych jest dopuszczalne na podstawie art. 89 RODO. Może to przede wszystkim usankcjonować poprawnie skonstruowana zgoda na udział w badaniach naukowych podpisywana przez ochotnika przed przystąpieniem do badań. Taka regulacja stawia niestety pod znakiem zapytania całą ideę otwartych danych, która leży u podstaw wielu programów unijnych w ramach Horyzontu 2020[19], czy też krajowych, jak np. Program Operacyjny Polska Cyfrowa[20]. Należy zatem zmienić zasady dostępu do otwartych danych, np. przez wprowadzenie odpowiednich umów licencyjnych czy komisji naukowych opiekujących się zbiorami

[17] Zob. https://www.genome.gov/sequencingcosts/ (dostęp: 3.03.2018 r.).

[18] Zob. https://nanoporetech.com/products/minion (dostęp: 3.03.2018 r.).

[19] Zob. http://ec.europa.eu/programmes/horizon2020/en (dostęp: 3.03.2018 r.).

[20] Zob. https://cppc.gov.pl/programy/popc-2/ (dostęp: 3.03.2018 r.).

danych – w kontekście danych genomowych, aby nie niszczyć skądinąd słusznej idei. Kroki takie są niezbędne z uwagi na niewystarczające zabezpieczenie danych, jakie daje współcześnie stosowana anonimizacja, lub na błędnie stosowaną anonimizację.

Obecnie (w RODO definicja jest utrzymana) termin „anonimizacja" jest definiowany w regulacjach Unii Europejskiej jako technika, która w sposób nieodwracalny zapobiega identyfikacji. Oceniając sposób, który może być z uzasadnionym prawdopodobieństwem wykorzystany do zidentyfikowania danej osoby, należy wziąć pod uwagę wszelkie obiektywne czynniki, takie jak koszt i czas potrzebne do jej zidentyfikowania, oraz uwzględnić technologię dostępną w momencie przetwarzania danych, jak również postęp technologiczny. Zasady ochrony danych nie powinny więc mieć zastosowania do informacji anonimowych, czyli informacji, które nie wiążą się ze zidentyfikowaną lub możliwą do zidentyfikowania osobą fizyczną, ani do danych osobowych zanonimizowanych w taki sposób, że osób, których dane dotyczą, w ogóle nie można zidentyfikować lub już nie można zidentyfikować. Niniejsze rozporządzenie nie dotyczy więc przetwarzania takich anonimowych informacji, w tym przetwarzania do celów statystycznych lub naukowych[21]. Niestety zapisy te, nie odnoszą się do sytuacji, w której materiał badawczy nie może podlegać pełnej anonimizacji danych. Jak udowodniono za pomocą przypadków opisanych w dalszej części, o pełniej anonimizacji można mówić jedynie w kontekście zagregowanych danych statystycznych. Takie podejście bardzo często oznacza usunięcie kluczowych danych z punktu widzenia badacza. Co więcej, w przypadku badań podłużnych, wieloletnich, gdzie co do zasady konieczne jest powracanie do dawców materiału biologicznego/ochotników/pacjentów i uzyskiwanie aktualizacji danych, w tym np. o charakterze medycznym, pełna anonimizacja na pierwszych etapach biobankowania materiału biologicznego i pochodnych, lub uzupełniających danych przekreślałaby sens istnienia takich kolekcji. Problem ten został zauważony już dawno i podkreślono, że takie praktyki czynią finansowo i praktycznie niemożliwym prowadzenie badań o charakterze prospektywnym[22,23]. W takim przypadku należy zastosować możliwą do odwrócenia przez jak najmniejsze grono osób pseudoanonimizację. Z reguły osoby te muszą dysponować odpowiednim kluczem kodującym. Innym, niezwykle ważnym aspektem anonimizacji, która jest teoretycznie najbezpieczniejszą metodą dla danych ochotnika biorącego udział w badaniach, uniemożliwia mu złożenie oświadczenia woli o chęci wycofania się

[21] RODO, http://eur-lex.europa.eu/legal-content/PL/TXT/PDF/?uri=CONSIL:ST_5418_2016_REV_1&from=PL (dostęp: 3.03.2018 r.).

[22] B. M. Knoppers, M. H. Zawati, E. S. Kirby, *Sampling populations of humans across the world: ELSI issues*, „Annual Review of Genomics and Human Genetics" 2012, t. 13, s. 395–413.

[23] O. Tene, J. Polonetsky, *Privacy in the age of Big Data: a time for big decisions*, „Stanford Law Review Online" 2012, t. 64:63, s. 63–69.

z badań. Technicznie nie ma możliwości zniszczenia próbek materiału biologicznego, jak również danych będących w relacji do niego po procesie anonimizacji.

Rozwiązaniem alternatywnym stosowanym przez biobanki jest pseudoanonimizacja danych osobowych. Technika ta, poprzez stosowanie jedno- lub dwustopniowego kodowania, uważana jest za wystarczająco bezpieczną dla osób powierzających swoje dane i materiał biologiczny do badań naukowych. Zapewniając zbliżony poziom bezpieczeństwa danych, zachowuje jednak możliwość odwrócenia procesu dla ponownej jednoznacznej identyfikacji na potrzeby choćby wycofania zgody na udział w badaniach lub ponownego kontaktu z badanym. Dzięki tej metodzie, możliwe jest również dopasowanie oryginalnych danych do nowego kodu dawcy. Niestety koncepcja ta nie ma wsparcia w istniejących rozwiązaniach prawnych krajów Unii Europejskiej. Do innych metod pseudoanonimizujących można zaliczyć stosowanie skrótów poprzez stosowanie funkcji „hash" (*Hash function*). Technika jest popularna i łatwa do przeprowadzenia. W jej efekcie dane o różnej wielkości są kodowane w zunifikowany sposób, np. Jan Kowalski, Anna Nowak, Marcin Nowakowski zostaną zakodowane „01", „02", „03", niezależnie od prawdziwej długości znaków kodowanych imion i nazwisk[24]. Inną metodą służącą pseudoanonimizacji jest k-anonimizacja zaproponowana przez Laboratorium Fujitsu, która zakłada automatyczne uogólnienia, w sytuacji, gdzie unikalność danych, pomimo zakodowania danych osobowych, może prowadzić do łatwej (relatywnie) identyfikacji osoby biorącej udział w badaniach[25] (tabele: 1, 2, 3, 4, 5).

Tabela 1. Dane wyjściowe

Imię	Nazwisko	Wiek	Płeć	Miejsce ur.	Choroby
Dorin	Owens	33	K	Sopot	Anemia
Franciszek	Dolas	25	M	Warszawa	Amnezja
Albert	Starski	29	M	Łódź	Amnezja
Maksymilian	Paradys	39	M	Warszawa	Anemia

Źródło: Opracowanie własne.

[24] Zob. https://cnpd.public.lu/content/dam/cnpd/fr/publications/groupe-art29/wp216_en.pdf (dostęp: 3.03.2018 r.).

[25] K. Ito, J. Kogure, T. Shimoyama (i in.), *De-identification and Encryption Technologies to Protect Personal Information*, „Fujitsu Scientific & Technical Journal" 2016, t. 52, nr 3, s. 28–36, https://www.fujitsu.com/global/documents/about/resources/publications/fstj/archives/vol52-3/paper05.pdf (dostęp: 3.03.2018 r.).

Tabela 2. Pseudoanonimizacja z wykorzystaniem funkcji haszującej

Hash	Wiek	Płeć	Miejsce ur.	Choroby
G070901	33	K	Sopot	Anemia
Z070902	25	M	Warszawa	Amnezja
N070903	29	M	Łódź	Amnezja
H070904	39	M	Warszawa	Anemia

Źródło: Opracowanie własne.

Mimo zastosowanej pseudoanonimizacji pozostawione dane pozwalają na identyfikację osoby w zbiorze, np. „Płeć" lub „Miejsce urodzenia".

Tabela 3. K-anonimizacja (1)

Wiek	Miejsce ur.	Choroby
30–39	Pow. 100 000	Anemia
20–29	Pow. 100 000	Amnezja
20–29	Pow. 100 000	Amnezja
30–39	Pow. 100 000	Anemia

Źródło: Opracowanie własne.

W tabeli 3 usunięto wszystkie pola pozwalające na jednoznaczną identyfikację osoby oraz „rozmyto" pewne dane – pole „Wiek".

Tabela 4. K-anonimizacja (2)

Wiek	Płeć	Miejsce ur.	Choroby
30–39	n/d	Pow. 100 000	Anemia
20–29	M	Pow. 100 000	Amnezja
20–29	M	Pow. 100 000	Amnezja
30–39	n/d	Pow. 100 000	Anemia

Źródło: Opracowanie własne.

W tabeli 4 przedstawiono k-anonimizację w mniej restrykcyjnej odsłonie – częściowo zachowano dane na temat płci.

Tabela 5. Pseudoanonimizacja z wykorzystaniem obu technik

Hash	Wiek	Płeć	Miejsce ur.	Choroby
G070901	30–39	n/d	Pow. 100 000	Anemia
Z070902	20–29	M	Pow. 100 000	Amnezja
N070903	20–29	M	Pow. 100 000	Amnezja
H070904	30–39	n/d	Pow. 100 000	Anemia

Źródło: Opracowanie własne.

Metoda zawarta w tabeli 5 jest możliwa do zastosowania np. w biobankach. Pozwala ona udostępnić możliwie największą część informacji.

Zastosowanie technik pseudoanonimizujących utrudnia możliwość bezpośredniej identyfikacji osoby, od której pochodzą dane lub materiał biologiczny. Aby dodatkowo zapobiegać reidentyfikacji (bez względu na stosowaną formę anonimizacji czy pseudoanonimizacji), utworzono metodę, w której zasadą jest to, by analizę uruchamiać w miejscu bezpiecznie przechowywanych danych, a nie przekazywać dane do prowadzenia analiz (*taking the analysis to the data*, *not the data to the analysis*). Koncepcja ta została opublikowana pod nazwą *dataSHIELD*[26]. Przy takim podejściu naukowiec prowadzący badania nie ma dostępu do żadnych danych identyfikujących osoby lub też do tzw. danych wrażliwych[27]. Co więcej, wszystkie analizy są przeprowadzane w miejscu przechowywania danych przez personel biobanku, a badacz otrzymuje już przeanalizowane, zagregowane dane statystyczne. Takie zamknięcie danych niewątpliwie podnosi poziom zabezpieczenia danych i choć nie eliminuje niebezpieczeństwa w całości, to wydaje się słusznym kierunkiem rozwoju. Przy obecnym rozwoju techniki jedynie pełne ograniczenie dostępu do danych pozwala zabezpieczyć posiadane dane. W poniżej opisanych przypadkach deanonimizacji danych kluczowym okazał się fakt możliwości porównywania danych z innymi bazami, co w dobie powszechnego profilowania nie jest trudne.

Metody anonimizacyjne i pseudoanonimizacyjne, aby były skuteczne, muszą być stosowane w sposób świadomy i przemyślany. Istnieje kilka udokumentowanych udanych prób deanonimizacji danych, gdzie podstawą sukcesu był brak dogłębnej analizy i pominięcie innych dostępnych źródeł danych.

1. America On Line

W 2006 roku America On Line (AOL) – wyszukiwarka internetowa – opublikowała historię wyszukiwań 650 000 użytkowników. Dane zostały oczywiście

[26] A. Gaye, Y. Marcon, J. Isaeva (i in.), *DataSHIELD: taking the analysis to the data*, *not the data to the analysis*, „International Journal of Epidemiology" 2014, t. 43, nr 6, s. 1929–1944.

[27] S. E. Wallace, A. Gaye, O. Shoush (i in.), *Protecting personal data in epidemiological research: DataSHIELD and UK Law*, „Public Health Genomics" 2014, t. 17, nr 3, s. 149–157.

zanonimizowane oraz usunięte numery IP. Nazwy użytkowników zostały zakodowane z użyciem funkcji haszującej – chodziło o zachowanie możliwości analizowania danych. Na pierwszy rzut oka takie podejście wydawało się prawidłowe. Nic bardziej mylnego. Jak się okazało, poprzez historię wyszukiwania można dojść do danych osobowych użytkownika[28].

2. Netflix

Również w 2006 roku Netflix – portal udostępniający filmy – ogłosił konkurs "Netflix prize": osoba, która stworzy najlepszy algorytm przewidujący ocenę filmów przez użytkownika (bazując na wcześniejszych ocenach) otrzyma 1 000 000 dolarów. W tym celu zostały udostępnione bazy danych zawierające m.in. 100 480 507 ocen filmów wygenerowanych przez 480 189 użytkowników serwisu (około 1/8 ogółu użytkowników w tamtym okresie). Niezbędne było zachowanie unikalności użytkowników, aby analizy mogły być w ogóle możliwe do przeprowadzenia. Dlatego podobnie jak w przypadku AOL identyfikatory użytkowników zostały zakodowane przy użyciu funkcji haszujących. Dodatkowo usunięto wszystkie publicznie dostępne – opublikowane na Netflix – oceny, część dat wystawienia oceny została zmieniona w przedziale +/– 14 dni. Niestety dla użytkowników serwisu takie zabiegi okazały się niewystarczające. Okazało się, że po połączeniu baz danych Netflixa z informacjami z baz danych serwisu IMDB możliwe jest ustalenie tożsamości użytkownika Netflixa. W niektórych przypadkach nawet z 99-procentową pewnością. Co gorsza, na podstawie udostępnionych ocen można próbować określać przekonania religijne, poglądy polityczne czy preferencje seksualne. Badacze z Uniwersytetu w Austin w Texasie zarzucili personelowi zły dobór próby oraz wybór niewłaściwego algorytmu anonimizującego. Błędy te miały ułatwić reidentyfikację danych użytkowników[29].

3. Genome data Identifying Personal Genomes by Surname Inference

Istnieją też przypadki zidentyfikowania ludzi na podstawie genomu, nieważne, czy chodzi o ustalenie dawców spermy[30] czy tych, którzy oddali materiał biologiczny do projektów naukowych[31], np. 1000 Genome project[32]. Zasada jest podobna, należy zidentyfikować charakterystyczne miejsca w genomie, a następnie porównać z istniejącymi bazami danych genealogicznych – tworzonych

[28] Zob. http://www.nytimes.com/2006/08/09/technology/09aol.html (dostęp: 3.03.2018 r.).

[29] Zob. https://arxiv.org/pdf/cs/0610105.pdf (dostęp: 3.03.2018 r.).

[30] Zob. http://www.slate.com/articles/double_x/doublex/2010/02/are_sperm_donors_really_anonymous_anymore.html (dostęp: 3.03.2018 r.).

[31] M. Gymrek, A. L. McGuire, D. Golan (i in.), *Identifying personal genomes by surname inference*, „Science" 2013, t. 339, nr 6117, s. 321–324.

[32] 1000 Genomes Project Consortium, *A map of human genome variation from population-scale sequencing*, „Nature" 2010, t. 467, nr 7319, s. 1061–1073.

np. na potrzeby ludzi szukających swoich przodków[33]. Uzyskane wyniki z baz danych należy przeanalizować pod kątem dat urodzenia, miejsc zamieszkania. W trudniejszych przypadkach uzyskane informacje można skonfrontować z innymi źródłami danych, by móc zidentyfikować daną osobę.

4. Identyfikacja danych gubernatora Williama Welda

W latach 90. XX wieku administracja stanu Massachusetts wykupiła dla swoich pracowników ubezpieczenia zdrowotne. Kilka lat później postanowiono udostępnić dane medyczne osób objętych ubezpieczeniem. Miało to zapewnić transparentność całego procesu i rzetelne wydawanie pieniędzy publicznych. Ówczesny gubernator stanu zapewniał, iż dane są dobrze przygotowane do udostępnienia i należycie zabezpieczone przed ryzykiem reidentyfikacji tożsamości beneficjentów programu. Te słowa zachęciły ludzi do zweryfikowania zapewnień gubernatora, w wyniku czego tożsamość samego Williama Welda została ujawniona. Wystarczyło połączyć opublikowaną bazę świadczeń medycznych z rejestrem wyborczym miasta, z którego pochodził gubernator. W bazie medycznej tylko 6 osób miało tę samą datę urodzenia co gubernator, z czego trzy były kobietami; po uwzględnieniu kodu pocztowego dokumentacja medyczna gubernatora stała się publicznie dostępna[34].

Anonimizacja w dobie *Big Data*, taniej mocy obliczeniowej komputerów i mnogości źródeł danych, nie jest magicznym środkiem pozwalającym na pełną ochronę prywatności osób, których danych dotyczy przetwarzanie. Nie ma też zastosowania w przypadku gromadzenia danych genetycznych, które ze swojej specyfiki identyfikują osoby jednoznacznie, a na drodze do swobodnego dostępu takiej identyfikacji stoją jedynie czynniki ekonomiczne i prawne. Należy rozważyć, czy w ogóle można użyteczne badawczo dane na poziomie pojedynczego człowieka zanonimizować. Ponadto nie można ograniczać dostępu do danych, zwłaszcza w kontekście człowieka – bez nich współczesna nauka zginie. Pomimo stworzenia w przepisach RODO przestrzeni do legalności przetwarzania danych wrażliwych na potrzeby badań naukowych oraz rozwoju medycyny (art. 9 ust. 2 pkt i, j RODO), to już teraz istnieją obawy restrykcyjnych przepisów, które mogą spowodować zamknięcie wielu cennych repozytoriów danych. Aby uniknąć drastycznych skutków dla środowiska naukowego, warto jest zrewidować zasady dostępu do otwartych repozytoriów. Należy wprowadzić umowy określające zasady korzystania z danych, wprowadzić rady naukowe na poziomie repozytorium, które ocenią zasadność dostępu do danych dla określonych badań. Aby

[33] Zob. https://www.ancestry.com/.

[34] P. Ohm, *Broken promises of privacy: Responding to the surprising failure of anonymization*, „UCLA Law Review" 2009, t. 57, nr 6, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450006 (dostęp: 3.03.2018 r.).

zniwelować utrudnienia, jakie niesie ze sobą proponowane rozwiązanie, każde repozytorium powinno udostępniać odpowiedni interfejs wyszukiwania, pozwalając na znalezienie szukanych danych, jednocześnie nie zdradzając ich zawartości do czasu podpisania licencji. Na przykładzie danych genetycznych dla wyszukujących byłaby udostępniona informacja, czy dane miejsce w genomie (SNP) jest zgodne z referencją czy nie, natomiast sama wartość w wynikach wyszukiwania pozostałaby utajniona. Alternatywnie jako standard należałoby wprowadzić koncepcję *dataSHIELD*. Wówczas administrator danych nie traci kontroli nad informacją, którą przetwarza, minimalizowane jest też ryzyko wycieku danych przez ograniczenie ilości ich kopii. W kontekście rodzaju danych przetwarzanych przez instytucje naukowe jest niezwykle istotne, aby administratorzy tych danych utrzymali status zaufania publicznego. Bez spełnienia tego warunku nie ma szans na prowadzenie badań omicznych czy GWAS na szeroką skalę. Stworzenie centralnych jednostek, np. na poziomie województw – biobanków, które będą odpowiedzialne za składowanie materiału biologicznego i przetwarzanie danych, znacząco może przyczynić się do zabezpieczenia interesów zarówno dawców, jak i samych badaczy, czy wręcz całych uczelni, i wydaje się interesującym kierunkiem do rozwoju. Co więcej, podobne rozwiązania są już dyskutowane w świecie naukowym[35].

Współczesna technika i rosnąca świadomość ludzi w zakresie danych osobowych stawia ogromne wyzwania przed badaczami, prawnikami, instytucjami naukowymi, prawodawcami i to od decyzji, które za chwilę zapadną, zależeć będzie kształt przyszłej nauki.

Wydaje się, że słusznym rozwiązaniem jest tworzenie scentralizowanych ośrodków przetwarzania danych wrażliwych, które będą prowadzić szerokie usługi analityczne na podstawie umów i licencji oraz wezmą na siebie pełną odpowiedzialność za skuteczne zabezpieczenie posiadanych danych poprzez stosowanie unormowanych środków bezpieczeństwa oraz zapewnienie rzetelnego kontaktu z osobami, których dane dotyczą.

BIBLIOGRAFIA

Costa F. F., *Big Data in biomedicine*, „Drug Discovery Today" 2014, t. 19, nr 4

Elger B. S. (i in.), *Consent and anonymization in research involving biobanks*, „EMBO reports" 2007, t. 7, nr 7

Gaye A., Marcon Y., Isaeva J. (i in.), *DataSHIELD: taking the analysis to the data*, *not the data to the analysis*, „International Journal of Epidemiology" 2014, t. 43, nr 6

[35] Zob. http://www.nature.com/news/privacy-protections-the-genome-hacker-1.12940 (dostęp: 3.03.2018 r.).

Gymrek M., McGuire A. L., Golan D. (i in.), *Identifying personal genomes by surname inference*, „Science" 2013, t. 339, nr 6117

Ito K., Kogure J., Shimoyama T. (i in.), *De-identification and Encryption Technologies to Protect Personal Information*, „Fujitsu Scientific & Technical Journal" 2016, t. 52, nr 3, https://www.fujitsu.com/global/documents/about/resources/publications/fstj/archives/vol52-3/paper05.pdf (dostęp: 3.03.2018 r.)

Knoppers B. M., Zawati M. H., Kirby E. S., *Sampling populations of humans across the world: ELSI issues*, „Annual Review of Genomics and Human Genetics" 2012, t. 13

Lin Z., Owen A. B., Altman R. B., *Genomic research and human subject privacy*, „Science" 2004, t. 305, nr 5681

Marx V., *Biology: the big challenges of Big Data*, „Nature" 2013, t. 498

May M., *Life Science Technologies: big biological impacts from Big Data*, „Science" 2014, t. 344, nr 6189

Mooney S. J., Westreich D. J., El-Sayed A. M., *Epidemiology in the era of Big Data*, „Epidemiology" 2015, t. 26, nr 3

O'Connor G., *Moore's Law Gives Way to Bezos's Law*, GigaOm, April 2014, https://gigaom.com/2014/04/19/moores-law-gives-way-to-bezoss-law/ (dostęp: 3.03.2018 r.)

Ohm P., *Broken promises of privacy: Responding to the surprising failure of anonymization*, „UCLA Law Review" 2009, t. 57, nr 6, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450006 (dostęp: 3.03.2018 r.)

Tene O., Polonetsky J., *Privacy in the age of Big Data: a time for big decisions*, „Stanford Law Review Online" 2012, t. 64:63, file:///C:/Users/Dell/Downloads/64-SLRO-63%20(1).pdf (dostęp: 3.03.2018 r.)

Wallace S. E., Gaye A., Shoush O. (i in.), *Protecting personal data in epidemiological research: DataSHIELD and UK Law*, „Public Health Genomics" 2014, t. 17, nr 3

1000 Genomes Project Consortium, *A map of human genome variation from population-scale sequencing*, „Nature" 2010, t. 467, nr 7319

RODO http://eur-lex.europa.eu/legal-content/PL/TXT/PDF/?uri=CONSIL:ST_5418_2016_REV_1&from=PL

https://www.ancestry.com/

https://arxiv.org/pdf/cs/0610105.pdf

https://cppc.gov.pl/programy/popc-2/

https://cnpd.public.lu/fr/publications/groupe-art29/wp216_en.pdf

http://ec.europa.eu/programmes/horizon2020/en

https://www.genome.gov/sequencingcosts/

https://www.iso.org/standard/45123.html

https://nanoporetech.com/products/minion

http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

http://www.nytimes.com/2006/08/09/technology/09aol.html

http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/

http://www.nature.com/news/privacy-protections-the-genome-hacker-1.12940

http://www.rp.pl/artykul/812483-Informacje-wrazliwe-pacjenta-moga-sluzyc-do-badan-naukowych.html

http://www.slate.com/articles/double_x/doublex/2010/02/are_sperm_donors_really_anonymous_anymore.html

ANONYMIZING THE BIG DATA – THE SITUATION OF BIOBANKS IN THE GDPR CONTEXT

Summary

Technological development increased abilities of data analysis collecting gathering and obtaining from wide range of sources including national health registries. Aim of, entering into force on May 2018, General Data Protection Regulation (GDPR) is to regulate analysis of wide range of personal data from profiling to healthy issues. New regulations may hamper the implementation of research, or realization of projects under Horizon 2020 or Digital Poland, that based on data sharing. In order not to waste the opportunities and benefits for humanity that modern technology and „open data" can bring, appropriate measures must be taken to protect the privacy of individuals. But then there is next trap. The more data is protected in the context of privacy protection, the less useful they are for scientific purposes. The key point is to find right balance between security and usability.

KEYWORDS

General Data Protection Regulation (GDPR), data anonymization, data generalization, privacy protection, DataSHIELD policy

SŁOWA KLUCZOWE

ogólne rozporządzenie o ochronie danych osobowych, anonimizacja danych, generalizacja danych, ochrona prywatności, koncepcja *DataSHIELD*