Metody analizy i oceny bezpieczeństwa oraz jakości informacji

Metody analizy i oceny bezpieczeństwa oraz jakości informacji

REDAKCJA NAUKOWA
Wojciech Z. Chmielowski
Dorota Wilk-Kołodziejczyk

Kraków 2012

Rada Wydawnicza Krakowskiej Akademii im. Andrzeja Frycza Modrzewskiego:
Klemens Budzowski, Maria Kapiszewska, Zbigniew Maciąg, Jacek M. Majchrowski

Recenzenci:
prof. dr hab. Jacek Wołoszyn (rozdz. 1–2, 4–9)
prof. dr hab. Paweł Chmielnicki (rozdz. 3)

Redaktor prowadzący: Halina Baszak-Jaroń

Projekt okładki: Joanna Sroka

Korekta: Kamila Zimnicka-Warchoł

ISBN 978-83-7571-249-0

Na zlecenie:

Krakowskiej Akademii
im. Andrzeja Frycza Modrzewskiego
www.ka.edu.pl

Wydawca:
Krakowskie Towarzystwo Edukacyjne sp. z o.o. – Oficyna Wydawnicza AFM,
Kraków 2012

Sprzedaż detaliczną, hurtową i wysyłkową prowadzi:
Księgarnia u Frycza
Kampus Krakowskiej Akademii im. Andrzeja Frycza Modrzewskiego
ul. Gustawa Herlinga-Grudzińskiego 1, 30-705 Kraków
tel./faks: (12) 252 45 93
e-mail: ksiegarnia@kte.pl

Skład: Joanna Sroka

Druk i oprawa: Krakowskie Towarzystwo Edukacyjne sp. z o.o.

Spis treści

Słowo wstępne

Przekazujemy czytelnikowi kolejną monografię wydaną w ramach serii „Informatyka", zawierającą publikacje pracowników związanych z Wydziałem Zarządzania i Komunikacji Społecznej Krakowskiej Akademii im. Andrzeja Frycza Modrzewskiego z obszaru zastosowań informatyki oraz metod numerycznych.

Wśród autorów niniejszej publikacji oprócz pracowników Studium Informatyki KA AFM znajduje się także pracownik naukowy Politechniki Krakowskiej oraz czteroosobowy zespół młodych badaczy, którzy są zatrudnieni w podmiotach gospodarczych i jednostkach finansowo-administracyjnych. Po studiach zainteresowali się oni jednym z tematów (zastosowania praw rozkładu cyfr znaczących w dużych zbiorach danych liczbowych) i prowadzą badania w ramach powołanej przez siebie tzw. „Grupy Benforda". Efektem tej działalności jest już kilka wystąpień na konferencjach naukowych, publikacji, a także portal internetowy www.benford.pl. W planach jest ambitne zadanie wydania kompleksowej monografii, o charakterze encyklopedycznym na tematy związane z prawem Benforda. Druga część niniejszego zbioru zawiera elementy, które po rozwinięciu mogą stanowić przedmiot rozważań w planowanej monografii.

Tematyka monografii koncentruje się na metodach zapewnienia bezpieczeństwa informacji, procedurach wspomagających podnoszenie jakości danych oraz narzędziach zwiększających możliwości pozyskiwania z dostępnych informacji wartościowych i rzetelnych wniosków analitycznych.

Publikacja podzielona jest na trzy części. Pierwsza część zawiera 3 rozdziały, w których podjęto kwestie zapewnienia bezpieczeństwa informacji. Druga część monografii, ujęta w kolejnych trzech rozdziałach, prezentuje wybrane narzędzia i modele organizacji zasobów cyfrowych mających na celu zapewnienie wysokiej jakości, użyteczności oraz wiarygodności zbiorów danych zawartych w repozytoriach. Ostatnia część pracy przedstawia problematykę związaną z prawem Benforda, pozwalającym ocenić stopień rzetelności danych na podstawie analizy rozkładów cyfr w liczbach weryfikowanego zbioru danych.

Monografię rozpoczyna rozdział „Hasło jako podstawowy element bezpieczeństwa systemu informatycznego" pokazujący ciekawą i ważną tematy-

kę haseł dostępu. Omówiono tu m.in. specyfikę haseł w systemach Windows i Linux oraz metody zdobywania i łamania haseł, ze szczególnym uwzględnieniem środowiska programowania równoległego.

W drugim rozdziale „Zastosowanie metod ewolucyjnych w kryptografii – problem faktoryzacji" podjęto próbę udowodnienia przydatności algorytmów ewolucyjnych w procesie faktoryzacji polegającej na znalezieniu liczb, których iloczyn występuje w asymetrycznych procedurach kryptograficznych (np. RSA) i zajmuje stosunkowo dużo czasu.

Trzeci rozdział „Publicznoprawne aspekty bezpieczeństwa danych osobowych w szkołach wyższych w Polsce" dotyka istotnego tematu, jakim jest ochrona danych osobowych. Problematyka pokazana jest od strony formalnoprawnej w sposób jasny i zrozumiały nawet dla osób bez przygotowania prawniczego. Dzięki temu tekst niewątpliwie może być przydatny dla służb informatycznych w uczelniach oraz osób odpowiedzialnych za bezpieczeństwo danych osobowych.

Drugą część monografii rozpoczyna rozdział „Wykorzystanie systemów tablicowych do uporządkowania wiedzy technologicznej". Omówiono w nim metody i narzędzia wykorzystywane w procesie analizy i projektowania baz danych w ramach dużych zbiorów zwanych bazami wiedzy.

Kolejny rozdział drugiej części pracy „Trwała identyfikacja publikacji w repozytoriach cyfrowych – przegląd stosowanych systemów" porusza problematykę konstruowania jednoznacznych i niezmiennych w czasie odwołań do dokumentów cyfrowych zgromadzonych w różnego rodzaju elektronicznych repozytoriach. Jest to bardzo ważne zagadnienie w aspekcie dynamicznego rozwoju technologii internetowych.

Drugą część pracy kończy rozdział „Zabezpieczenie wiarygodności zasobów cyfrowych deponowanych w repozytoriach instytucjonalnych". Wskazano tu na zadania systemu repozytoryjnego oraz trwałej ochrony wiarygodności, autentyczności, integralności i poufności zasobów cyfrowych. Opisano model archiwizacji zasobów cyfrowych OAIS oraz scharakteryzowano podstawowe metody, narzędzia i techniki stosowane w strategii zabezpieczania cyfrowych dokumentów.

Ostatnia część monografii zawiera trzy rozdziały poświęcone wybranym kwestiom związanym z możliwościami tkwiącymi w tzw. prawach Benforda. Prawa te opisują częstości pojawiania się cyfr na określonych miejscach liczb z dużych zbiorów pomiarów. Duże odstępstwa od tych reguł mogą wskazywać na celowe bądź przypadkowe zniekształcenia danych źródłowych i brak możliwości uzyskania wiarygodnych wniosków z analizy opartej na tych danych. W tej części opracowania na uwagę zasługują wnioski wynikające z analizy zbieżności testów i mierników podobieństwa rozkładów częstości, a także przegląd empirycznych praw numerycznych zbliżonych swoim

charakterem do prawa Benforda, np. reguła Pareto, prawa Zipfa, Lotki, ciągi Fibonacciego.

Monografia powinna zainteresować Czytelników zarówno śledzących na bieżąco szybkie zmiany w zakresie metod i narzędzi informatycznych, jak również zajmujących się szeroko pojętymi metodami numerycznymi w naukach ekonomicznych.

Wojciech Z. Chmielowski
Dorota Wilk-Kołodziejczyk

CZĘŚĆ I

Narzędzia wspomagające bezpieczeństwo informacji

Tools supporting information security

STRESZCZENIE:
W części pierwszej poruszono ciekawą i ważną tematykę haseł dostępu, podjęto próbę udowodnienia przydatności algorytmów ewolucyjnych w procesie faktoryzacji, oraz omówiono ochronę danych osobowych w kontekście szkół wyższych.

Słowa kluczowe: hasła, algorytmy ewolucyjne, ochrona danych osobowych.

Cel
Prezentacja metod zabezpiecznia informacji.

Metodyka badań:
Metody: opis pojedynczych przypadków, analiza dokumentacji, modelowanie.

Wynik:
Spójne opracowanie w zakresie tematyki zastosowania hasła jako podstawowego elementu bezpieczeństwa systemu informatycznego oraz zastosowanie metod ewolucyjnych w kryptografii oraz publicznoprawne aspekty bezpieczeństwa danych osobowych w szkołach wyższych w Polsce.

Oryginalność wartość:
Oryginalny układ i połączenie treści.

ABSTRACT:
The first part raised an interesting and important topic of passwords, attempts to prove the usefulness of evolutionary algorithms in the factorization, and said protection of personal data in the context of higher education.

Key words: passwords, evolutionary algorithms, protection of personal data

Presentation of security of information methods

Methods: describing individual cases, analysis of documentation, modeling

Consistent development in the field application password as a key element of system security and the use of evolutionary methods in cryptography and public-safety aspects of personal data in colleges and universities in Poland.

Original layout and combination of content.

1. Hasło jako podstawowy element bezpieczeństwa systemu informatycznego

Joanna Płażek

Wprowadzenie

Główną cechą bezpiecznego systemu informatycznego jest warstwowość systemu zabezpieczeń. Oznacza to, że napastnik musi pokonać wiele poziomów zabezpieczeń, zanim uzyska dostęp do danych użytkownika. Obecnie podstawową metodą uwierzytelniania użytkowników są hasła i to właśnie one są najczęściej atakowanym elementem systemu zabezpieczeń. Haker próbujący ustalić hasło ma generalnie do wyboru dwie strategie. Może próbować przechwycić przesyłane hasło, a następnie użyć narzędzi do jego odszyfrowania (jeśli transmisja była szyfrowana). Może także podjąć próbę odgadnięcia hasła. To właśnie łamanie haseł jest jedną z najstarszych technik wykorzystywanych przez hakerów, a jej skuteczność znacznie się zwiększyła poprzez coraz łatwiejszy dostęp do nowych architektur komputerowych, a co za tym idzie do dużych mocy obliczeniowych, które jeszcze nie tak dawno były dla zwykłego użytkownika w sferze marzeń. Dlatego każdy, kto korzysta z komputera i codziennie loguje się do systemu, podając swój login i hasło, powinien mieć podstawową wiedzę na temat tego, w jaki sposób przechowywana jest większość haseł, jak działają najbardziej popularne programy do ich łamania oraz jakie zagrożenia wynikają z możliwości implementowania tych programów w nowoczesnych, powszechnie dostępnych środowiskach programowania równoległego.

1.1. Hasła w systemach operacyjnych

W systemach wielodostępnych nazwy i hasła użytkowników są przechowywane w plikach systemowych. Niestety w wielu systemach operacyjnych, w tym najpopularniejszych takich jak Unix czy Windows, pliki te są dostępne do czytania dla zarejestrowanych użytkowników. Chociaż zmiany ich zawartości może dokonywać jedynie administrator, to inni mogą je przekopiować i podejmować próbę ich rozszyfrowania na własnym sprzęcie.

Istnieje wiele strategii przechowywania haseł w bazach danych systemów operacyjnych. Najprostszym sposobem jest zapisanie ich w bazie otwartym tekstem. W przypadku, gdy atakujący zdobędzie zawartość takiej bazy danych, hasła użytkowników nie są w żadnym stopniu chronione. Większe bezpieczeństwo daje szyfrowanie haseł. Do tego celu wykorzystuje się różne algorytmy. Obecnie w większości systemów stosuje się tzw. hashowanie haseł. Polega ono na przechowywaniu w bazie danych skrótów haseł użytkowników, czyli haseł zakodowanych przez jednokierunkową funkcję skrótu. Funkcja skrótu przyjmuje argument w postaci ciągu znaków o dowolnej długości i zwraca inny ciąg o ustalonej długości. Funkcje skrótu są jednokierunkowe, tzn. szyfrują hasło, ale nie pozwalają w prosty sposób odtworzyć na podstawie postaci zaszyfrowanej hasła wejściowego. Należy zauważyć, że uznanie funkcji za bezpieczną opiera się zawsze wyłącznie na domniemanej odporności na znane ataki kryptoanalityczne, nie zaś na matematycznych dowodach gwarantujących niemożność złamania hasła.

Hasła w systemach Windows i Linux są przechowywane w postaci jednostronnie zakodowanej, co oznacza, że nie można ich odszyfrować. Logowanie użytkownika polega na zaszyfrowaniu podanego przez niego ciągu znaków i porównaniu go z ciągiem przechowywanym w pliku haseł.

Aby dodatkowo zabezpieczyć hasło, stosuje się technikę „solenia" (ang. *salting*) polegającą na dodawaniu losowego ciągu znaków przed lub po podanym przez użytkownika haśle, a następnie szyfrowaniu złączonego napisu. „Sól" może być uniwersalna dla wszystkich użytkowników lub losowana dla każdego użytkownika osobno. Jeżeli „sól" nie jest przechowywana w tej samej bazie co dane użytkownika, to skutecznie tworzy kolejną warstwę zabezpieczającą system.

Kolejnym zabezpieczeniem przed atakami jest wielokrotne hashowanie. Do hasła dodawana jest „sól", następnie ten ciąg jest szyfrowany, po czym otrzymany wynik jest szyfrowany ponownie. Dla użytkownika opóźnienie wynikające z tej podwójnej operacji nie jest widoczne, ale dla atakującego ma ogromne znaczenie.

Na podobnej zasadzie działa kryptograficzna funkcja bcrypt(), oparta na zmodyfikowanym algorytmie Blowfish. Siła algorytmu tkwi w zastosowaniu „soli" oraz kosztownej obliczeniowo operacji inicjalizacji kluczy algorytmu Blowfish, która została zmodyfikowana tak, aby była jeszcze bardziej kosztowna obliczeniowo. Narzut czasowy na zaszyfrowanie pojedynczego ciągu znaków powoduje, że ataki łamiące hasło są zbyt kosztowne do zastosowania.

Należy jednak pamiętać, że podstawową zasadą bezpieczeństwa naszych zasobów jest używanie przez nas tzw. „silnych" haseł. Przede wszystkim hasło:

- powinno liczyć przynajmniej siedem znaków,
- powinno zawierać duże i małe litery (nie zawsze jest to respektowane, czasami wszystkie litery są zamieniane na duże przed kodowaniem),
- powinno zawierać cyfry,
- może zawierać znaki specjalne,
- może zawierać znaki z górnej połówki tablicy ASCII.

Na pewno nie należy stosować takiej samej nazwy dla loginu i hasła albo tworzyć hasło przez dodanie cyfr do nazwy loginu. Nie należy również stosować popularnych wyrazów z dowolnej dziedziny, nawet jeśli są one stosunkowo długie.

Jednym z najważniejszych czynników ograniczających możliwość złamania hasła jest jego regularna zmiana. Powodzenie bowiem ataku przede wszystkim zależy od czasu uzyskania zabezpieczonego hasła. Jeśli administrator systemu dowie się o włamaniu do bazy haseł i powiadomi o tym użytkowników, to mimo odgadnięcia hasła przez hakera atak nie będzie skuteczny. Administrator również może wymagać odpowiedniej konstrukcji hasła lub narzucić jego zmianę po upływie określonego czasu. W niektórych przypadkach, np. sporadycznego łączenia się z odległymi komputerami, stosuje się ustalenie ważności hasła tylko dla jednego połączenia. Wtedy, za każdym razem przed wylogowaniem się z systemu, hasło musi zostać zmienione.

Hasła w systemie Windows

Wprowadzenie na rynek systemu Windows NT 4.0 dopuściło stosowanie haseł o maksymalnej długości 14 znaków. Hasła krótsze niż 14 znaków są dopełniane znakami ASCII o kodzie 0[1]. System przechowuje dwie wersje każdego hasła użytkownika. Pierwsza wersja nosi nazwę kodu LANMan lub kodu LM. Hasło przed szyfrowaniem (za pomocą algorytmu DES) jest dzielone na połowy. Ostateczny skrót hasła składa się z połączonych skrótów połówek hasła. W efekcie czternastoznakowe hasło użytkownika jest zamieniane na dwa siedmioznakowe hasła, co więcej schemat szyfrowania LANMan ignoruje wielkość liter (wszystkie litery są zamieniane na duże przed kodowaniem), co znacznie skraca czas potrzebny na wykonanie ataku na hasło. Druga wersja jest zwykle nazwana kodem NT. W tym przypadku hasło jest najpierw zapisane w kodzie Unicode, a następnie szyfrowane za pomocą jednokierunkowej funkcji MD4. Natomiast w domenach systemu Windows

[1] *Hack Proofing Your Network. Edycja polska*, praca zbiorowa, Helion, Gliwice 2002.

2000 mogą istnieć już konta z piętnastoznakowymi hasłami. Wyklucza to możliwość ich szyfrowania zgodnie ze schematem LANMan.

Należy pamiętać, że kod LM jest najsłabszą wersją zaszyfrowanego hasła. Dlatego niewątpliwą zaletą systemów Windows XP i Windows 2000 Service Pack 2 jest klucz rejestru umożliwiający usunięcie przechowywanych kluczy LANMan[2]. Można także ustawić powyższy klucz tak, by zakazywał systemowi Windows przechowywania kodów LANMan dla wszystkich haseł zmienianych w przyszłości.

Systemy Windows przechowują hasła w pliku binarnym *Security Accounts Manager* (SAM). Wprawdzie przy próbie jego kopiowania system wyświetla komunikat o błędzie, ale kopie zapasowe pliku SAM są umieszczane najczęściej w katalogu \WINDOWS\repair\ przez program RDISC, który tworzy dyski ratunkowe dla systemu[3]. Standardowo wszyscy użytkownicy mają nadane prawa do tego pliku. Jeśli więc administrator często tworzy dyski awaryjne, to w pliku tym można znaleźć w miarę aktualne dane.

Należy również wspomnieć o mechanizmie wymuszania bezpiecznych haseł wprowadzonym od systemu Windows NT 4.0[4]. Pozwala on administratorowi na ustanawianie reguł tworzenia haseł przez użytkowników. System Windows od tej wersji sprawdza każde nowo utworzone hasło pod kątem zgodności z następującymi regułami:

- hasło nie może być częścią nazwy konta użytkownika,
- hasło musi składać się co najmniej z sześciu znaków,
- hasło musi zawierać znaki z trzech poniższych kategorii:
  - wielkie litery,
  - małe litery,
  - cyfry,
  - znaki niealfanumeryczne.

Aby powyższe reguły zmodyfikować, należy wymienić dołączaną dynamicznie bibliotekę DLL (ang. *Dynamic-Link Library)*, zastępując ją inną zgodną z odpowiednim interfejsem API systemu Windows.

Hasła w systemie Linux

W starszych wersjach systemu Linux dane o użytkownikach i ich hasłach były przechowywane w pliku */etc/passwd*. W nowszych wersjach plik ten podzielono na dwa mniejsze, z których jeden, ogólnie dostępny, zawiera

[2] M. Shema, B.C. Johnson, *Anti Anti-Hacker Tool Kit*. Edycja polska, Helion, Gliwice 2004.
[3] A. Dudek, *Nie tylko wirusy. Hacking, cracking, bezpieczeństwo Internetu*, Helion, Gliwice 2004.
[4] M. Shema, B.C. Johnson, *Anti Anti-Hacker Tool Kit. Edycja polska*, Helion, Gliwice 2004.

wszystkie informacje poza hasłem, a drugi, który czytać może tylko użytkownik uprzywilejowany, zawiera hasła[5]. W większości implementacji ten pierwszy plik zostaje przy nazwie */etc/passwd* i atrybutach 644, a drugi plik posiada nazwę */etc/shadow* oraz atrybuty 600.

W nowym wydaniu plik */etc/passwd* zawiera następujące pola[6]:

- nazwa użytkownika do 8 znaków, ważna jest wielkość liter,
- „x" w miejscu hasła, teraz jest ono przechowywane w pliku */etc/shadow*,
- liczbowo ID użytkownika, które jest przyznawane przez funkcję *adduser()*,
- liczbowo ID grupy,
- pełna nazwa użytkownika do 30 znaków,
- katalog domowy użytkownika,
- powłoka użytkownika.

Plik */etc/shadow* (tylko *root* może go odczytać) zawiera:

- nazwę użytkownika,
- zakodowane hasło,
- ostatnią zmianę hasła,
- liczbę dni, przed upływem których zmiana hasła jest niedozwolona,
- liczbę dni, po upływie których użytkownik musi zmienić hasło,
- informację o tym, na ile dni przed wygaśnięciem ważności hasła użytkownik ma być o tym ostrzeżony,
- liczbę dni, w których konto pozostanie nieaktywne,
- datę wygaśnięcia ważności konta.

Ustawienie wartości *-1* w danym polu oznacza, że jest ono nieaktywne.

Hasło jest szyfrowane za pomocą funkcji systemowej *crypt()*. W starszych systemach wykorzystywała ona algorytm DES, obecnie stosuje się algorytm MD5. Algorytm MD5 potrzebuje do zaszyfrowania hasła około 20 razy więcej czasu procesora niż DES[7]. Dodatkowo algorytm MD5 pozwala na szyfrowanie haseł o dowolnej długości, a DES ograniczał się do haseł co najwyżej 8-znakowych. Oba algorytmy wykorzystują do szyfrowania funkcje jednokierunkowe. Dodatkowo, aby utrudnić złamanie haseł, dopisuje się do hasła losowy ciąg znaków tzw. „sól".

Dodatkowo system Linux zawiera moduł PAM (ang. *Pluggable Authentication Module*) kontrolujący wszystkie odwołania wymagające podania przez użytkownika hasła np. dostęp do usługi telnet, logowanie do konsoli lub zmiana hasła.

---

[5] A. Dudek, *Nie tylko wirusy. Hacking, cracking, bezpieczeństwo Internetu*, Helion, Gliwice 2004.
[6] S. Frampton, *Linux Administration Made Easy*, Iuniverse Inc, December 2000.
[7] B. Toxen, *Bezpieczeństwo w Linuksie. Podręcznik administratora*, Helion, Gliwice 2004.

1.2. Zdobywanie i łamanie haseł

Hasło jest podstawowym elementem bezpieczeństwa pozwalającym zabezpieczyć dane przed nieautoryzowanym dostępem przez osoby postronne. Może ono być przechwycone w różnych miejscach, np.:

- Serwer uwierzytelniania – miejsce, w którym znajdują się dane użytkownika potrzebne do jego identyfikacji i uwierzytelnienia.
- Medium transmisyjne – w celu uzyskania dostępu do zasobów zdalnych hasło zostaje wysłane do serwera uwierzytelnienia, co daje możliwość przechwycenia go.
- Komputer użytkownika – często dla wygody hasło zostaje zapisane w pamięci komputera, aby zaoszczędzić konieczność wpisywania go przy każdorazowym logowaniu. Usługi takie oferują np. przeglądarki internetowe.

W celu zabezpieczenia hasła przed odczytaniem z miejsca przechowywania lub przechwyceniem podczas transmisji stosowane są dwa zabiegi:

- szyfrowanie transmisji – hasło może być przesyłane w postaci ciągów znaków, a ze względu na szyfrowanie całej transmisji danych pozostaje bezpieczne.
- szyfrowanie hasła – obecnie rzadko hasło zostaje zapisane w postaci ciągu znaków. Przed zapisaniem zostaje zaszyfrowane bądź poddane działaniu funkcji hashującej, co czyni proces odtworzenia hasła trudnym i czasochłonnym. Czasem dla zwiększenia efektywności hasło jest hashowane, szyfrowane kilkukrotnie lub wcześniej modyfikowane, tak jak to zostało opisane wyżej.

Jeżeli mimo zabezpieczeń hasło dostanie się w ręce włamywacza, zapewne podejmie on próbę złamania go. Poniżej zostało opisanych kilka strategii łamania hasła.

Metoda słownikowa

Nie jest to metoda gwarantująca złamanie hasła. Przy jej zastosowaniu nie zostają porównane wszystkie możliwe hasła, a jedynie zawarte w dołączonym słowniku[8]. Wykorzystuje ona słabość haseł, gdyż w praktyce większość z nich nie jest hasłami „silnymi", ale ciągami znaków będącymi wyrazami, które można znaleźć w słowniku językowym. Program łamiący hasła metodą słownikową może być bardzo skuteczny, jeśli dodatkowo uwzględni się w nim zwyczaje użytkowników, np. dodawanie cyfr na końcu hasła. Ba-

[8] B. Toxen, *Bezpieczeństwo w Linuksie. Podręcznik administratora*, Helion, Gliwice 2004.

dania wykazują, że istnieją hasła stosowane zdecydowanie częściej niż inne. Hasła takie jak „123456", „qwerty" czy „password" występują tak często, że potrafią stanowić ponad 1% wszystkich haseł. Tym samym program zawierający dobrze dobrany słownik i umiejętnie nim operujący jest w stanie złamać zdecydowanie więcej haseł przy niewielkim nakładzie pracy w porównaniu z podobnym programem łamiącym hasła metodą *bruteforce*.

Metoda bruteforce

Jest to metoda oparta na zasadzie pełnego przeglądu. Teoretycznie pozwala ona złamać każde hasło, należy jednak pamiętać, że w niektórych przypadkach mogłoby to zająć nawet kilkaset lat. Ogólna zasada polega na zdefiniowaniu zbioru X znaków, które mogą wystąpić w haśle, a następnie tworzeniu wszystkich możliwych kombinacji tych znaków o różnej długości. Dla hasła o długości z istnieje X^z możliwości[9]. Ponieważ łamane hasło jest najczęściej przechowywane w zakodowanej postaci, bez znajomości algorytmu, za pomocą którego dokonano kodowania, odgadnięcie hasła jest niemożliwe. Czasami producenci oprogramowania udostępniają informacje o sposobie oraz miejscu zapisu hasła, tym samym dają możliwość rozkodowania go.

Tablice wyszukiwania skrótów

Łamanie haseł metodą bruteforce można znacznie przyśpieszyć, stosując tablice skrótów. Zwykle takie tablice zajmują setki gigabajtów, przez co przy dłuższych hasłach ich przeglądnięcie jest nieefektywne. Bardziej optymalne jest stosowanie tzw. tęczowych tablic (ang. *rainbow tables*). Tęczowa tablica jest bazą skrótów wykorzystywanych w łamaniu haseł zakodowanych jednokierunkową funkcją skrótu. Podstawą jej działania jest tzw. funkcja redukcyjna. Działa ona odwrotnie do funkcji skrótu, ponieważ z hasha tworzy ona hasło w czystym tekście (zawierające tylko określony zestaw znaków, np. tylko małe litery i cyfry). Uzyskane za jej pomocą hasło z hasha nie może być oczywiście hasłem, które dało określony hash (co wynika z własności funkcji skrótu), ale dzięki niej będą tworzone kolejne kombinacje hasła, które znowu zostaną potraktowane funkcją skrótu i porównane z hashem łamanego hasła[10]. Tęczowe tablice można wygenerować samemu, wykorzystując do tego celu między innymi konsolowe narzędzie rtgen z projektu RainbowCrack oraz okienkowy program pod systemy Microsoft o nazwie winrtgen. Można

[9] J. Erickson, *Hacking. Sztuka penetracji*, Helion, Gliwice 2004.
[10] P. Maziarz, *Wykorzystywanie tęczowych tablic do łamania haseł*, „Hakin9" 9/2007 (29).

uzyskać je z serwisów, które udostępniają je za darmo, na przykład http://www.freerainbowtables.com/ lub http://rainbowtables.shmoo.com/ albo kupić poprzez takie serwisy jak np. http://www.rainbowcrack-online.com/.

1.3. Programy do łamania haseł

Łamanie haseł nie odbywa się tylko przez profesjonalnych hakerów. W internecie można znaleźć wiele darmowych aplikacji łamiących hasła zabezpieczone za pomocą różnych algorytmów kryptograficznych. Poniżej zostaną omówione najbardziej popularne.

John the Ripper

Program *John the Ripper* (www.openwall.com/John), czyli Kuba Rozpruwacz, jest jednym z najszybszych, najbardziej uniwersalnych i zapewne najbardziej popularnych dostępnych łamaczy haseł[11]. Może działać w piętnastu różnych systemach operacyjnych i obsługuje wiele dostępnych obecnie procesorów, ze specjalnymi technikami optymalizacji procesorów Pentium i układów RISC włącznie. Obsługuje sześć różnych schematów kodowania haseł stosowanych w różnych odmianach systemów Unix oraz kodowanie Windows LANMan, znane także jako NTLM (używane w systemach Windows NT, 2000 i XP). Do łamania haseł *John the Ripper* wykorzystuje kilka różnych metod, w tym metodę pełnego przeglądu i korzystania z pliku słownikowego. Jedną z jego głównych zalet jest automatyczne zapisywanie stanu przeglądu podczas ataku. Na tej podstawie można wznowić atak w dowolnym czasie i na dowolnym komputerze.

L0phtCrack

Program do łamania haseł w systemach Windows NT i jego następców, Windows 2000 i XP. Wykorzystuje wiele różnych mechanizmów, ale przede wszystkim polega na technice pełnego przeglądu. *L0phtCrack* oprócz odgadywania haseł umożliwia wydobywanie zaszyfrowanych kodów LANMan z dowolnego pliku SAM lokalnego lub zdalnego systemu, a nawet potrafi przechwytywać takie kody przesyłane w sieci[12]. Bardzo często program *L0phtCrack* jest wykorzystywany do zdobywania haseł, a następnie do ich łamania stosuje się opisany wcześniej program *John the Ripper*.

[11] M. Shema, B.C. Johnson, *Anti Anti-Hacker Tool Kit. Edycja polska*, Helion, Gliwice 2004.
[12] *Hack Proofing Your Network. Edycja polska*, praca zbiorowa, Helion, Gliwice 2002.

Crack

Jest bardzo popularnym programem, jednym z pierwszych wykorzystywanych do łamania haseł (http://www.crypticide.com/users/alecm/security/c50-faq.html). Program podejmuje próbę złamania hasła zakodowanego za pomocą tradycyjnej funkcji linuxowej *crypt()* oraz innych systemów kodowania. W tym drugim przypadku konieczne jest dołączenie odpowiednich bibliotek funkcji kodujących. Do wersji programu z 1996 roku dołączono moduł *Crack7*, który rozszerza możliwości łamania haseł o metodę pełnego przeglądu, gdyż do tego momentu dostępna była tylko metoda słownikowa[13].

Ponieważ proces łamania haseł jest długotrwały i kończy się w momencie złamania wszystkich haseł, program *Crack* daje możliwość w dowolnym momencie jego pracy wyświetlenie informacji o złamanych hasłach i napotkanych błędach poprzez uruchomienie pomocniczego programu o nazwie *Reporter*.

RainbowCrack

Tradycyjne programy łamiące hasła metodą pełnego przeglądu sprawdzają wszystkie możliwe kombinacje. Hasła są przekształcane przez funkcję hashującą, a w wyniku tego procesu otrzymane hashe są porównywane z zakodowanym hasłem. Jest to metoda, która pochłania bardzo dużo czasu. W metodzie *RainbowCrack* (http://www.antsight.com/zsl/rainbowcrack) moc obliczeniowa jest skierowana do stworzenia plików zawierających tablice hashy. Po zakończeniu tworzenia tych plików łamanie haseł na ich podstawie może być nawet setki razy szybsze niż w przypadku standardowych łamaczy haseł.

1.4. Łamanie haseł w środowiskach programowania równoległego

Łamanie haseł jest zadaniem wymagającym dużych nakładów obliczeniowych. Do niedawna przyspieszenie tego typu obliczeń uzyskiwano głównie dzięki zastosowaniu coraz nowszych procesorów o ciągle zwiększającej się liczbie wykonywanych cykli na sekundę. F. Alonso w swoim artykule *The Extinction of Password Authentication*[14] przedstawia ewolucję mocy procesorów na przestrzeni lat 1971–2008 i zagrożenia wynikające ze znacznego

[13] Ch. Negus, *Red Hat Linux 9. Biblia*, Helion, Gliwice 2003.
[14] F. Alonso, *The Extinction of Password Authentication*, ISSA Journal, December 2008.

skrócenia czasu łamania „silnych" haseł. Uważa, że uwierzytelnianie kont za pomocą haseł jest w niebezpieczeństwie, a na użytkownikach należy wymóc obowiązek zmiany hasła przynajmniej co dziesięć dni.

Kolejnym krokiem w celu zwiększenia efektywności przetwarzania było zastosowanie procesorów wielordzeniowych, by w końcu łamać hasła przy użyciu maszyn wieloprocesorowych lub multikomputerów. Do najczęściej wykorzystywanych do tego celu środowisk programowania równoległego można zaliczyć MPI (ang. *Message Passing Interface*) oraz OpenMP (ang. *Open Multi Processing*).

Prawdziwą rewolucją stało się wykorzystanie procesorów kart graficznych do obliczeń ogólnego przeznaczenia, czyli GPGPU (ang. *General-Purpose computing on Graphics Processing Units)*. Najczęściej używane środowisko do przetwarzania na kartach graficznych to CUDA (ang. *Compute Unified Device Architecture*) opracowana przez firmę NVIDIA.

Porównanie równoległych implementacji algorytmu łamiącego hasła, realizowanych w różnych środowiskach obliczeń równoległych, daje możliwość odpowiedzi na pytanie, które ze środowisk pozwala na uzyskanie największego przyśpieszenia i jakim nakładem pracy każda z implementacji musi być realizowana.

Poniżej zostanie opisany sposób implementacji dwóch podstawowych metod łamania haseł: *bruteforce* i metody słownikowej, w trzech wymienionych wcześniej środowiskach programowania równoległego.

Implementacja w środowisku MPI

MPI (ang. *Message Passing Interface*) jest standardem interfejsu do przesyłania komunikatów na potrzeby programowania rzeczywistego na maszynach z pamięcią lokalną. Pierwsza implementacja została przedstawiona w maju 1994 roku przez konsorcjum MPI Forum – grupę badaczy z USA i Europy reprezentujących producentów oraz użytkowników maszyn równoległych. MPI nie jest konkretnym pakietem oprogramowania, a formalną specyfikacją interfejsu. Najbardziej znaną implementacją MPI jest MPICH. Interfejs MPI dostarcza funkcje umożliwiające odbieranie oraz wysyłanie komunikatów i synchronizację zadań wykonywanych na różnych komputerach (procesorach). Jego zaletą jest fakt, że program może być wykonywany na maszynach o różnych architekturach. MPI zupełnie rezygnuje z koncepcji pamięci dzielonej. Każdy proces, nawet uruchomiony na tym samym procesorze, posiada własną kopię wszystkich danych. Jedyną drogą komunikacji i wymiany danych między procesami są komunikaty, które służą nie tylko

do synchronizacji i wysyłania informacji kontrolnych, ale głównie do wymiany danych między procesami[15].

Implementacja metody słownikowej w tym środowisku najczęściej polega na podziale słownika na części i przydzieleniu każdej z nich poszczególnym procesorom. Jeżeli korzystamy z kilku słowników, to każdy procesor może operować na innym. Rozsyłane są również kopie plików z zestawem haseł do złamania. Można również rozesłać wszystkim procesorom ten sam słownik, a zbiór łamanych haseł zdekomponować na mniejsze zbiory. Należy pamiętać, że rozsyłanie danych do procesorów jest czasochłonne i opłaca się to robić tylko wtedy, gdy na określonej porcji danych będzie wykonywanych wiele operacji. Programy realizowane przez poszczególne procesy muszą co pewien czas upewniać się, czy dane hasło nie zostało już złamane. W tym celu procesy rozsyłają nieblokujące komunikaty do pozostałych.

Łamanie haseł metodą *bruteforce* polega na sprawdzeniu przez każdy z procesorów zadanego zakresu kombinacji haseł, począwszy od hasła startowego do hasła końcowego. Liczba zakresów najczęściej jest równa liczbie procesorów. Można również wyznaczyć większą liczbę zakresów niż liczba procesorów i wtedy po zbadaniu danego zakresu procesor dostaje następną porcję danych.

Implementacja w środowisku OpenMP

OpenMP (ang. *Open Multi Processing*) to standard służący do tworzenia aplikacji równoległych na komputerach z pamięcią wspólną. Opracowany w latach 90. XX wieku przez największych producentów maszyn równoległych, a następnie przyjęty przez wszystkich producentów oprogramowania. Możliwe jest programowanie w środowiskach obsługujących standard OpenMP w systemach Unix/Linux, a od Visual Studio 2005 również Windows. Głównym filarem standardu są dyrektywy zrównoleglające. Dodatkowymi elementami są zmienne środowiskowe oraz biblioteka procedur, służących głównie do manipulacji tymi zmiennymi. Ponadto procedury pozwalają mierzyć czas oraz identyfikować maszyny[16].

W tej implementacji słownik jest jednokrotnie zapisywany do pamięci komputera i każdy wątek korzysta z tej samej kopii umieszczonej w pamięci dzielonej. W pamięci lokalnej poszczególnych wątków alokowane jest miejsce na kolejne pobierane ze słownika słowa oraz zmienne potrzebne do ich zakodowania. Zrównolegleniu podlega pętla, przy pomocy której kolejne

[15] *Message Passing Interface (MPI) Tutorial*; https://computing.llnl.gov/tutorials/mpi/
[16] *OpenMP Tutorial*; https://computing.llnl.gov/tutorials/openMP/.

słowa ze słownika porównywane są z szukanym hasłem. Jeśli hasło znajduje się wśród porównywanych słów, pętla jest przerywana.

W metodzie *bruteforce* poszczególne wątki będą, podobnie jak w wersji z przesyłaniem komunikatów, sprawdzać zadany zakres kombinacji haseł, począwszy od hasła startowego do hasła końcowego.

Implementacja w środowisku CUDA

CUDA (ang. *Compute Unified Device Architecture*) opracowana przez firmę NVIDIA, równoległa architektura obliczeniowa, która zapewnia radykalny wzrost wydajności obliczeń dzięki wykorzystaniu mocy układów GPU (ang. *Graphics Processing Unit*)[17]. Jest środowiskiem dla ogólnych celów obliczeniowych wykonywanych na kartach graficznych bądź specjalizowanych kartach zbudowanych na ich bazie, czyli dla obliczeń GPGPU (ang. *General-Purpose computing on Graphics Processing Units*). Do niedawna procesory kart graficznych można było wykorzystywać do obliczeń numerycznych tylko za pośrednictwem API (ang. *Application Programming Interface*) dla grafiki komputerowej jak OpenGL czy DirectX. Proces zrównoleglania w tej technologii polega na wykonywaniu jednocześnie przez wiele wątków tych samych partii kodu zwanych jądrami (ang. *kernel*), czyli funkcji opatrzonych kwalifikatorem global. Liczba wątków, którym przypisane jest to samo jądro, jest zdefiniowana przy jego wywołaniu. Wątki tworzą bloki, a te z kolei składają się na gridy[18].

W metodzie słownikowej słownik oraz hasła do złamania zostają wczytane do tablicy w pamięci komputera, po czym w całości skopiowane do pamięci karty graficznej. Każdy wątek wykonuje określoną liczbę sprawdzeń.

W metodzie *bruteforce* każdy wątek wykonuje iteracje dla określonego zakresu haseł wyznaczonego przez odpowiedni przedział znaków.

Podsumowanie

Obecnie dostępne są już w sieci równoległe wersje algorytmów łamiących hasła. Najpopularniejszy program, opisany wyżej *John the Ripper*, jest dostępny w wersjach wykorzystujących zarówno model równoległości na poziomie danych w środowisku OpenMP, jak i przesyłania komunikatów w środowisku MPI. Program w środowisku MPI nosi nazwę

[17] *NVIDIA CUDA, C Programming Guide, Version 3.2*, NVIDIA Corporation, październik 2010.

[18] *NVIDIA CUDA, Reference Manual, Version 3.2 Beta*, sierpień 2010.

Djohn – Distributed John (*http://ktulu.com.ar/djohn*) i jest przeznaczony do zrównoleglonego ataku metodą pełnego przeglądu. Działa tylko pod Linuxem. Serwer dzieli całą przestrzeń haseł na paczki, które podlegają łamaniu. Każda paczka zawiera pierwsze i ostatnie hasło zakresu, jaki ma być testowany. Klient pobiera paczkę z serwera, testuje ją i zwraca otrzymane wyniki.

W literaturze istnieją również przykłady analizy realizacji równoległych implementacji najpopularniejszych algorytmów łamiących hasła. W pracy *Równoległe metody łamania haseł metodą słownikową w środowiskach MPI, OpenMP i CUDA*[19] można znaleźć porównanie implementacji metody słownikowej w trzech różnych środowiskach programowania równoległego. Dla badanego problemu sprzęt, na którym testowane były programy, pozwolił na otrzymanie bardzo dobrych wyników dla programu wykonywanego w środowisku OpenMP. Implementacja w środowisku MPI niestety nie dała tak dobrych wyników i chociaż uzyskała znaczne przyspieszenie w stosunku do wersji sekwencyjnej, to i tak nie może się równać z implementacjami w środowiskach OpenMP czy CUDA. Najciekawszą implementacją jest program wykonywany na procesorze karty graficznej. Co prawda, nie uzyskał on najkrótszego czasu wykonania, ale należy wziąć po uwagę, że procesor komputera, na którym wykonywane były testy, jest procesorem wyższej klasy niż wykorzystana karta graficzna.

W pracy *Równoległe łamanie haseł*[20] zawarto porównanie równoległych implementacji w środowisku MPI algorytmów łamiących hasła za pomocą różnych algorytmów szyfrujących. Wyniki porównano z innymi dostępnymi na rynku aplikacjami.

Wyniki porównujące czasy łamania haseł z wykorzystaniem CPU i środowiska GPGPU dla haseł różnej długości, kodowanych za pomocą różnych algorytmów, można również znaleźć w pracy *GPU-based Password Cracking. On the Security of Password Hashing Schemes regarding Advances in Graphics Processing Units*[21].

Najważniejszym elementem zabezpieczenia zasobów użytkowników jest ochrona ich haseł. Dlatego w systemach operacyjnych należy przede wszystkim chronić bazy danych przechowujące zakodowane hasła użytkowników. Przy obecnie dostępnych nowoczesnych systemach komputerowych wypo-

[19] J. Płażek, M. Podyma, *Równoległe metody łamania haseł metodą słownikową w środowiskach MPI, OpenMP i CUDA*, „Czasopismo Techniczne" 2011.

[20] M. Żak, *Równoległe łamanie haseł*, Politechnika Krakowska, Kraków 2006.

[21] M. Sprengers, *GPU-based Password Cracking. On the Security of Password Hashing Schemes regarding Advances in Graphics Processing Units*, Radboud University Nijmegen 2011.

sażonych w procesory wielordzeniowe lub mocne karty graficzne przystosowane do obliczeń ogólnego przeznaczenia, złamanie posiadanych, zakodowanych wprawdzie, haseł jest tylko kwestią czasu. Dlatego oprócz ochrony plików systemowych z hasłami tak ważne jest częste zmienianie haseł. Dodatkowo, analizując czas łamania haseł, widać, jak ważne jest stosowanie tzw. „silnych" haseł. Nie bez znaczenia jest również wybór algorytmu kryptograficznego do ich hashowania. Wyżej wymienione działania są podstawowymi elementami zabezpieczenia hasła i powinny być stosowane przez każdego użytkownika.

2. Zastosowanie metod ewolucyjnych w kryptografii – problem faktoryzacji

Radosław Bułat

Wprowadzenie

W obecnych czasach ciągły rozwój technologii informatycznych powoduje ich wszechobecność, a jednocześnie coraz łatwiejszą dostępność do informacji przechowywanych lub przesyłanych drogą elektroniczną. Tak duże rozpowszechnienie usług realizowanych drogą cyfrową wymusza ciągłe stosowanie coraz bardziej złożonych obliczeniowo zabezpieczeń przed dostępem do informacji przez osoby niepowołane. W wielkiej ilości przypadków stosowane są asymetryczne metody kodowania, często z wykorzystaniem dużych liczb pierwszych, na których wykonywana jest prosta obliczeniowo operacja mnożenia. Podczas stosowania takich metod, odwrotna dla zastosowanego rozwiązania jest operacja faktoryzacji – odnalezienia liczb, których iloczyn występuje w algorytmach kryptograficznych. Ze względu na swą złożoność czasową i obliczeniową operacja faktoryzacji jest jednym z wyzwań, jakie stawia przed nami współczesna kryptografia – zwłaszcza w epoce, gdy zaszyfrowane informacje o znaczeniu dla bezpieczeństwa publicznego mogą być przesyłane w sieci poza wszelką kontrolą.

Celem opracowania jest zbadanie i ewentualne udowodnienie przydatności metod ewolucyjnych w kryptografii analitycznej. Jednym z wyzwań podczas analizy informacji zaszyfrowanych współczesnymi metodami kryptograficznymi (np. RSA) jest problem faktoryzacji – podziału na czynniki dużych liczb złożonych. Ze względu na trudności analityczne w znajdowaniu takich czynników w grę wchodzić może zastosowanie algorytmów genetycznych, które udowadniają swoją przydatność podczas wielokryterialnej optymalizacji funkcji wielu zmiennych (a do takiej operacji można sprowadzić również faktoryzację). Przeprowadzę zatem testy algorytmu ewolucyjnego przeprowadzającego faktoryzację dla różnych warunków początkowych i parametrów algorytmu, starając się wykazać jego przydatność podczas rozwiązywania tego rodzaju zagadnień matematycznych.

2.1. Algorytm RSA

Algorytm RSA to jeden z najpopularniejszych obecnie algorytmów asymetrycznych (z kluczem prywatnym i publicznym). W podstawowej wersji algorytmu RSA wykorzystywane są dwie liczby pierwsze, których iloczyn poddawany jest dalszym operacjom. Poznanie tych liczb pierwszych gwarantuje rozszyfrowanie wiadomości zakodowanej RSA.

Algorytm postępowania w przypadku generowania klucza:

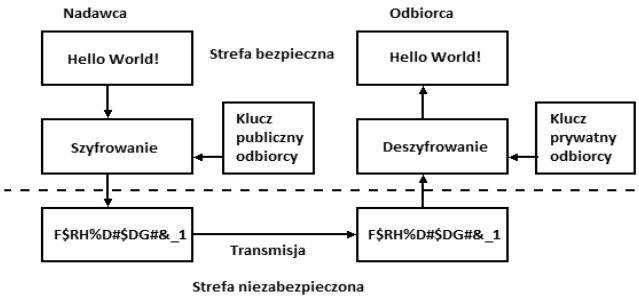
- Wybieramy losowo dwie duże liczby pierwsze p i q.
- Obliczamy ich iloczyn n= pq
- Obliczamy wartość funkcji Eulera dla n: φ(n) = (p-1)(q-1)
- Wybieramy liczbę e (1 < e < φ(n)) względnie pierwszą z φ(n)
- Znajdujemy liczbę d odwrotną do e mod φ(n): d = e-1 mod φ(n)

W tym przypadku kluczem prywatnym staje się para liczb (n,d), zaś kluczem publicznym – (n,e). Klucz publiczny może zostać udostępniony na zewnątrz, klucz prywatny zaś pozostaje tajemnicą użytkownika. Zanim zaszyfrujemy wiadomość, dzielimy ją na bloki m_i o wartości liczbowej nie większej niż n, a następnie każdy z bloków szyfrujemy według wzoru:

$$c_i = m_i^e \quad \text{mod n!} \tag{1}$$

Zaszyfrowana wiadomość będzie się składać z kolejnych bloków c_i. Tak stworzony ciąg znaków deszyfrujemy, odszyfrowując kolejne bloki c_i według wzoru:

$$m_i = c_i^d \quad \text{mod n!} \tag{2}$$



Rysunek 1. Obieg wiadomości.

Źródło: opracowanie własne.

Wiadomości zaszyfrowane kluczem publicznym mogą zostać odczytane jedynie kluczem prywatnym, nie istnieje zaś prosta metoda pozwalająca wyznaczyć klucz prywatny, posiadając wyłącznie klucz publiczny – o ile nie jesteśmy w stanie w prosty czasowo i obliczeniowo sposób sfaktoryzować zawartej w nim liczby n.

2.2. Definicja problemu

Problem, dla którego podjęto próbę znalezienia rozwiązania, zdefiniowany jest następująco. Mając daną całkowitą liczbę złożoną N, znaleźć takie pary liczb całkowitych x,y które pomnożone przez siebie dają N, o własności takiej, że $x^2=y^2$ (mod N) oraz $x \neq \pm y$ (mod N). Ten problem może zastąpić problem znalezienia nietrywialnych (różnych od 1 i N) faktorów N, jako że N dzieli $x^2-y^2=(x+y)(x-y)$, lecz nie jest w stanie podzielić ani (x+y) ani (x-y). Zatem NWD (x-y,N) jest nietrywialnym faktorem N[1].

Opisany problem jest problemem optymalizacji dyskretnej z funkcją minimalizującą

$$f : \{1,...,N-1\} \times \{1,...,N-1\} \rightarrow \{0,...,N-1\},\ f(x,y)=x^2-y^2 (\text{mod } N) \qquad (3)$$

Po uwzględnieniu ograniczenia $x \neq \pm y$ (mod N) możemy zmniejszyć dziedzinę funkcji do:

$$g : \{2,3,...,(N-1)/2\} \times \{2,3,...,(N-1)/2\} \rightarrow \{0,...,N-1\},\ g(x,y)=x^2-y^2 (\text{mod } N) \qquad (2)$$

Minimum tak określonej funkcji pozwoli sfaktoryzować N.

Dla określonej definicji problemu potrzebny jest dobór oraz implementacja algorytmu ewolucyjnego. W omawianym przypadku konieczny jest dobór algorytmu, który jest w stanie podołać zadaniu wielokryterialnej optymalizacji nieulegający zakleszczeniom w ekstremach lokalnych, jak również będący w stanie zapewnić dużą różnorodność przestrzeni rozwiązań.

W przypadku implementacji zrezygnowano z paradygmatu obiektowego na korzyść szybkości działania algorytmu, zatem użytym językiem programowania jest C++. Po implementacji wersji wstępnej algorytmu, zostają przeprowadzone testy faktoryzacji prostych liczb oraz wprowadzone poprawki algorytmiczne.

[1] R.L. Rivest, A. Shamir, L. Adleman, *A method for obtaining digital signatures and public key cryptosystems*, „Communications of the ACM" 1978, t. 21, s. 120–126.

2.3. Zastosowany algorytm

Do rozwiązania problemu użyję podejścia ewolucyjnego z wykorzystaniem algorytmów genetycznych, które udowadniają swoją przydatność podczas optymalizacji funkcji wielu zmiennych[2]. W algorytmie zastosowana jest binarna reprezentacja danych – x oraz y reprezentowane są przez ciągi zerojedynkowe, każda z reprezentacji rozwiązania w puli genetycznej posiada dwa takie ciągi – jeden dla wartości x, drugi dla wartości y, jakie reprezentuje.

1. Losowo inicjowana jest populacja macierzysta p1, zawierająca n osobników. Tymczasowo wszystkie osobniki w populacji początkowej są takie same, być może zostanie to zmienione w kolejnej edycji programu – wpływa to na wyniki podczas stosowania wyłącznie krzyżówek (patrz testy poniżej).
2. Na populacji p1 wykonywane jest n operacji genetycznych, przy czym stosunek liczby mutacji do liczby krzyżówek podawany jest przez użytkownika jako parametr algorytmu. Osobniki potomne tworzą populację p2, o liczebności n.
   - Operator mutacji tworzy osobniki poprzez losową inwersję jednego lub kilku bitów w reprezentacji danych osobnika. Po zamianie jednego z elementów x lub y, drugi jest generowany automatycznie poprzez zastosowanie operacji N div x. lub N div y. Dodatkowym ogranicznikiem jest parametr próbujący stosować mutację do momentu uzyskania osobnika należącego do danej dziedziny – w zależności od ustawienia może on w ogóle nie dopuszczać tworzenia osobników leżących poza dziedziną funkcji lub dopuszczać je, aczkolwiek z bardzo wysokim parametrem fitness.
   - Operator krzyżówki tworzy nowego osobnika poprzez zastąpienie fragmentu łańcucha bitów jednego osobnika macierzystego, łańcuchem bitów innego osobnika.
3. Osobniki z populacji p1 i p2 dobierane są losowo w pary, po jednym osobniku z każdej populacji.
4. Przeprowadzany jest turniej binarny w parach, gdzie preferowany jest osobnik o niższej wartości funkcji minimalizowanej, obliczanej z jego wartości x i y. Osobniki spoza dziedziny funkcji mają przypisywany automatycznie wysoki fitness.

[2] A. Menezes, P. van Oorschot, S. Vanstone, *Handbook of applied cryptography*, CRC Press series on discrete mathematics and its applications, CRC Press, 1996.

5. Zwycięzcy turnieju tworzą nową populację p1 i jeżeli nie został osiągnięty warunek stopu, czyli określona liczba iteracji algorytmu, wykonywany jest powrót do punktu 2.

Ze względu na specyfikę problemu faktoryzacji w kryptografii, N może być tak dobrane, że x oraz y są liczbami pierwszymi lub względnie pierwszymi. W pierwszym przypadku przestrzeń dopuszczalnych argumentów funkcji minimalizowanej sprowadza się do jednego punktu, w przypadku drugim dziedzina funkcji jest większa, jednakże z samego N nie jesteśmy w stanie wywnioskować, z którym z tych przypadków mamy do czynienia.

W związku z tym funkcja mutacji musi dokonywać samooptymalizacji – jeżeli dopuszczałaby tylko osobników należących do dziedziny problemu, to w przypadku gdy x i y są liczbami pierwszymi, całe działanie algorytmu sprowadziłoby się do stosowania operatora mutacji do momentu wylosowania prawidłowego i jedynego osobnika, co byłoby przeszukiwaniem przestrzeni w sposób losowy, nie zaś algorytmem genetycznym. Z drugiej strony w wypadku liczb względnie pierwszych, algorytm taki winien działać bez przeszkód.

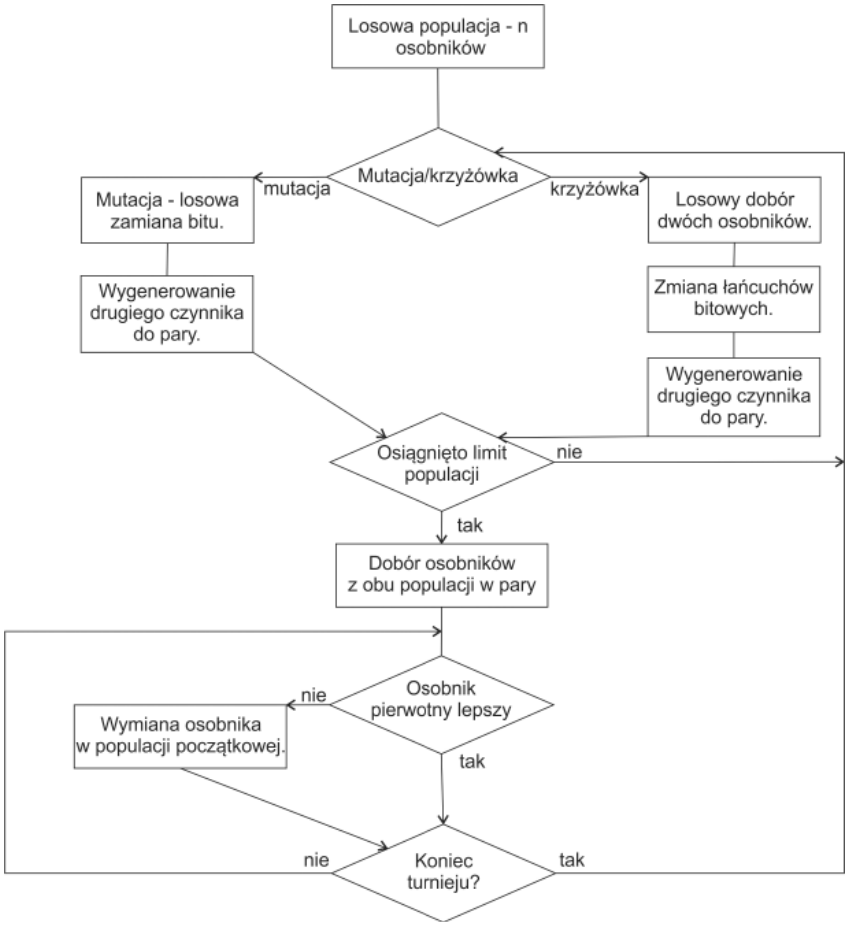
W związku z tym operator mutacji zlicza liczbę wszystkich mutacji przeprowadzonych przez siebie oraz w oddzielnym liczniku mutacje, jakie zaowocowały stworzeniem osobnika spoza dziedziny funkcji. Jeżeli stosunek drugiej z tych wartości do pierwszej przekroczy ustalony przez użytkownika poziom, operator podwaja liczbę prób stosowanych w operatorze mutacji, efektywnie zapewniając, że zostanie stworzony osobnik spełniający warunki. W przeciwnym wypadku zaczyna dopuszczać do puli genetycznej rozwiązania niespełniające warunku x*y=N, aczkolwiek z wyolbrzymioną funkcją fitness, tak aby każde rozwiązanie poprawne było lepsze od nich. W ten sposób wprowadzana jest różnorodność materiału genetycznego. Dodatkowo wspomniana funkcja poprzez swój parametr pozwala zdefiniować, że osobniki niepoprawne akceptujemy zawsze (parametr bardzo bliski 1) lub nigdy (parametr równy 0).

2.4. Testy w zastosowaniu praktycznym

W zaimplementowanym algorytmie mamy możliwość określenia czterech parametrów:

- liczby iteracji algorytmu (warunek stopu),
- wielkość populacji,

- liczby mutacji w jednym przebiegu algorytmu (liczba zastosowanych operatorów jest równa wielkości populacji, parametr określa, ile spośród operacji jest mutacjami, pozostała część to krzyżówki),
- tolerancja – parametr opisany w części „Wprowadzone mechanizmy".



Rysunek 2. Algorytm turnieju
Źródło: opracowanie własne.

*Testy dla wartości N=691*701 (x,y są pierwsze)*

Ilość iteracji=4 Populacja=16 Mutacje=0 Tolerancja – nieistotna, ponieważ nie używamy mutacji.
Algorytm NIE jest w stanie znaleźć poprawnego rozwiązania.

Ilość iteracji=400 Populacja=16 Mutacje=0 Tolerancja – nieistotna, ponieważ nie używamy mutacji.
Algorytm NIE jest w stanie znaleźć poprawnego rozwiązania. Jak widać, w przypadku stosowania operatora krzyżówki nie jesteśmy w stanie znaleźć poprawnego rozwiązania – ze względu na zapętlenie algorytmu na obszarze populacji początkowej – po zakończeniu działania algorytmu populacja ta pozostaje niezmieniona.

Ilość iteracji=4 Populacja=16 Mutacje=16 Tolerancja ≈ 0
Algorytm ZNAJDUJE poprawne rozwiązanie (wszystkie osobniki populacji wynikowej zawierają szukane rozwiązanie), jednakże wykonanie 4 iteracji algorytmu zajmuje około 10 sekund. Spowodowane jest to punktowym rozmiarem dziedziny problemu i sztywnością ograniczeń w mutacji. W takim kształcie algorytm nie różni się specjalnie od losowego przeszukiwania dziedziny problemu.

Ilość iteracji=4 Populacja=16 Mutacje=16 Tolerancja ≈ 1
Algorytm ZNAJDUJE poprawne rozwiązanie w około 50% przypadków (z reguły występuje ono jedno lub dwukrotnie w populacji wynikowej). Czas przeszukiwania wydatnie się zmniejszył (poniżej sekundy).

Ilość iteracji=400 Populacja=16 Mutacje=16 Tolerancja ≈ 1
Algorytm ZNAJDUJE poprawne rozwiązania – przestrzeń rozwiązań po ostatniej iteracji składa się w 80% z poprawnych rozwiązań. Czas wykonania – około 6 sekund. Jak widać, posiłkowanie się rozwiązaniami cząstkowymi znacząco poprawia proces przeszukiwania obszaru rozwiązań.

Ilość iteracji=4 Populacja=16 Mutacje=8 Tolerancja ≈ 0.5
Algorytm ZNAJDUJE poprawne rozwiązania – zajmują one około jednej trzeciej populacji końcowej.
Ilość iteracji=40 Populacja=16 Mutacje=8 Tolerancja ≈ 0.5
Algorytm ZNAJDUJE poprawne rozwiązania – wypełniają one całą popula-

cję końcową. Stabilizacja rozwiązań i zbiegnięcie się algorytmu następuje już około 15 iteracji.

Ilość iteracji=40 Populacja=16 Mutacje=1 Tolerancja ≈ 0.5
Algorytm ZNAJDUJE poprawne rozwiązanie – z reguły jedno lub dwa w populacji końcowej. Widoczny jest przyrost szybkości (operacja krzyżówki sprawdza mniej warunków niż operacja mutacji), a jednocześnie już wprowadzenie jednej mutacji pozwala algorytmowi uwolnić się z pułapki zapętlenia w jednym miejscu.

Jak widać w testach praktycznych (jak na razie dla problemu o niskiej złożoności), algorytm genetyczny jest w stanie zaoferować pewną alternatywę dla przypadków, kiedy metody algebraiczne zawodzą, zaś metody *brute-force* są zbyt kosztowne czasowo. Przy odpowiednim doborze warunków początkowych oraz parametrów przeszukiwania dziedziny problemu jesteśmy w stanie zoptymalizować uzyskanie rozwiązania pod względem prędkości oraz dokładności. Być może właśnie to operatory genetyczne obok komputerów kwantowych staną się kluczem do współczesnej kryptografii? Jest to niewątpliwie innowacyjna, acz trochę niedoceniana ścieżka rozwoju.

Podsumowanie

Natura problemu faktoryzacji ze względu na jego nieprzewidywalność co do możliwej dziedziny (liczby pierwsze tworzą dziedzinę punktową), stwarza trudności w zastosowaniu wyłącznie operatorów genetycznych w niezmodyfikowanej postaci. Jak widać na przeprowadzanych testach, przeszukiwanie wymaga dużej różnorodności materiału genetycznego (algorytm bez stosowania mutacji ulega zakleszczeniu). Dodatkowo, jeżeli dziedzina problemu jest punktowa, o czym wiedzieć *a priori* nie musimy, niemożliwe jest uzyskanie dużej różnorodności materiału doświadczalnego na tak określonej funkcji. Koniecznością staje się wobec tego wprowadzenie do puli rozwiązań niespełniających warunków problemu, lecz dających możliwość przekształcenia w rozwiązania, które takie warunki spełnią, jednocześnie zapewniając, że nie wyprą one z puli rozwiązań poszukiwanych, jeżeli dziedzina zawiera ich wiele. Wprowadzenie do algorytmu jego reaktywnego dostosowania do specyfiki materiału pozwala uniknąć stosowania podejścia *brute-force*, skracając dodatkowo czas obliczeń.

Tak zmodyfikowany algorytm, jak starano się wykazać doświadczalnie, jest w stanie rozwiązać postawiony problem – wszystko zależy od właściwego doboru liczby iteracji oraz współczynnika tolerancji. W przypadku faktoryzacji konieczny jest kompromis, pomiędzy stuprocentową pewnością otrzymania wyniku w potencjalnie długim czasie obliczeń, a mniejszą złożonością czasową, połączoną z potencjalnie mniejszą szansą rozwiązania problemu, przy czym zysk czasowy z reguły jest tak znaczny, iż pozwala na kilkukrotne uruchomienie algorytmu i otrzymanie prawidłowego wyniku, nadal utrzymując przewagę złożoności czasowej. Zastosowanie tego typu reaktywnego mechanizmu w połączeniu z bardziej zaawansowanymi algorytmami może przynieść interesujące rezultaty i warte jest dalszego zgłębiania.

3. Publicznoprawne aspekty bezpieczeństwa danych osobowych w szkołach wyższych w Polsce

Agnieszka Bednarczyk

Wprowadzenie

Zakres zbieranych przez różnorakie instytucje, przedsiębiorstwa i organy państwa danych osobowych nieustannie się zmienia i poszerza. Dane te gromadzone i przetwarzane są w celach handlowych, społecznych, kulturalnych, ale także w celu realizowania zadań przez państwo. W związku z tym istnieje ryzyko przekroczenia uprawnień do zbierania i przetwarzania tych danych zarówno przez podmioty publiczne, jak i prywatne oraz zbyt inwazyjne wkraczanie w sferę prywatności obywateli, co w konsekwencji może prowadzić do naruszenia ich dóbr osobistych i prawa do prywatności. Rozwój praw człowieka do ochrony jego danych osobowych rozpoczął się wraz z erą informatyczną[1]. Komputeryzacja i cyfryzacja zbieranych danych umożliwiły bowiem ich łatwiejsze i szybsze przetwarzanie oraz przechowywanie. Aby zarówno jednostka, której dane dotyczą, jak i podmioty zbierające i przetwarzające te dane posiadały kontrolę nad gromadzonymi informacjami, konieczna w tym celu jest pełna regulacja prawna obligująca te podmioty do wprowadzania procedur zapewniających bezpieczeństwo tworzonych zbiorów danych. Niniejsze opracowanie ma na celu przedstawienie sposobów i wymagań prawnych związanych z ochroną danych osobowych w szkołach wyższych w Polsce.

3.1. Geneza ochrony danych osobowych

Obowiązek zabezpieczania i ochrony danych osobowych wywieść należy z Konwencji nr 108 Rady Europy w dnia 28 stycznia 1981 r. o ochronie osób w związku z automatycznym przetwarzaniem danych osobowych[2]. Postanowienia Konwencji dotyczyły sfery publicznoprawnej, w której działają także

[1] M.T. Tinnefeld, *Ochrona danych osobowych – kamień węgielny budowy Europy,* [w:] *Ochrona danych osobowych*, red. M. Wyrzykowski, Warszawa 1999, s. 35.

[2] Dz.U. z 2003 r., Nr 3, poz. 25 – zwana dalej Konwencją; Konwencja weszła w życie 1 października 1985 r., po ratyfikowaniu jej przez 5 państw. Polska ratyfikowała Konwencję 24 kwietnia 2002 r., a weszła ona w życie 1 września 2002 r.

szkoły wyższe, nie wywołując bezpośrednich skutków prawnych po stronie obywateli państw Konwencji. Zgodnie z art. 1 Konwencji każdy obywatel państw członkowskich Rady Europy może oczekiwać od władz danego państwa ochrony jego praw i wolności, w szczególności prawa do poszanowania sfery osobistej, w związku z automatycznym przetwarzaniem danych osobowych. Zasada minimum, ochrony danych osobowych została wyrażona w art. 4 Konwencji. Ratyfikując Konwencję, każdy kraj będący jej stroną zobowiązuje się do wprowadzenia do prawa wewnętrznego przepisów koniecznych dla realizacji ochrony danych osobowych. Na gruncie przepisów Konwencji za podstawowe zasady ochrony danych osobowych uznano: zasadę adekwatności danych, zasadę związania celem zbierania danych, zasadę odpowiedniego zabezpieczania danych, zasadę respektowania uprawnień informacyjnych osób, których dane dotyczą. Na gruncie niniejszej Konwencji przyjęta została także dyrektywa 95/46/WE w sprawie ochrony osób fizycznych w zakresie przetwarzania danych osobowych oraz swobodnego przepływu tychże danych[3]. Dyrektywa jako akt prawa wtórnego Unii Europejskiej wiąże państwa członkowskie (podobnie jak Konwencja) w odniesieniu do rezultatu, jednak pozostawia organom krajowym swobodę wyboru formy i środków zmierzających do osiągnięcia tego rezultatu[4]. W dyrektywie wyszczególnione zostały podstawowe zasady przetwarzania danych, których przestrzeganie ma zapewnić bezpieczeństwo przetwarzanych danych. Za najważniejszą z zasad uznano wymóg rzetelnego i zgodnego z prawem przetwarzania danych osobowych. Jako kolejne zasady gwarantujące bezpieczeństwo danych i poszanowanie praw osób, których dane dotyczą, uznano: zasadę celowości przetwarzania danych, zasadę adekwatności przetwarzania danych, zasadę poprawności merytorycznej danych, zasadę ograniczania czasowego przetwarzania danych, zasadę poszanowania praw osób fizycznych przy przetwarzaniu ich danych osobowych, zasadę stosowania odpowiednich środków zabezpieczenia danych, zakaz przekazywania danych osobowych poza teren Europejskiego Obszaru Gospodarczego (poza wyjątkami przewidzianymi w dyrektywie)[5]. Wprowadzenie jednolitej ochrony danych osobowych w państwach Unii miało i ma na celu zapewnienie swobodnego przepływu towarów, usług i osób na terenie wspólnoty[6].

[3] Dyrektywa 95/46/WE Parlamentu Europejskiego i Rady z dnia 24 października 1995 r. w sprawie ochrony osób fizycznych w zakresie przetwarzania danych osobowych oraz swobodnego przepływu tychże danych, Dz.Urz. L 281 z dnia 23 listopada 1995 r. – zwana dalej dyrektywą.

[4] S. Biernat, *Prawo Unii Europejskiej a prawo państw członkowskich*, [w:] *Prawo Unii Europejskiej*, red. J. Barcz, Warszawa 2004, s. 212.

[5] Przepisy rozdziału II sekcji I dyrektywy.

[6] Preambuła do dyrektywy 95/46/WE.

W polskim prawie podstaw ochrony danych osobowych należy upatrywać w przepisach Konstytucji RP[7]. W art. 47 Konstytucji zostało zagwarantowane prawo do ochrony życia prywatnego, rodzinnego, czci, dobrego imienia oraz decydowania o swoim życiu osobistym. W art. 51 wyszczególniono natomiast bezpośrednie gwarancje bezpieczeństwa i ochrony danych osobowych, tj. prawo każdej osoby do decydowania o ujawnieniu dotyczących jej informacji, prawo każdej osoby do sprawowania kontroli nad informacjami na swój temat, prawo dostępu do dotyczących tych osób dokumentów i zbiorów danych, prawo do weryfikacji i żądania usunięcia danych osobowych. Konkretyzacją konstytucyjnych uregulowań stała się Ustawa z dnia 29 sierpnia 1997 r. o ochronie danych osobowych. Ustawa ta nie przejęła jednak wprost z dyrektywy rozdziału dotyczącego zasad przetwarzania danych osobowych. Zostało to uregulowane w samym zakresie obowiązywania ustawy. Przepisy ustawy opisują bowiem zasady i tryb postępowania przy przetwarzaniu danych osobowych prawa osób fizycznych, których dane osobowe są lub mogą być przetwarzane w zbiorach oraz zakres działania, zasady organizacji i funkcjonowania Generalnego Inspektora Ochrony Danych Osobowych (GIODO).

Jest wiele powodów, dla których zapewnienie bezpieczeństwa gromadzenia i przetwarzania danych osobowych jest tematyką na tyle istotną, aby regulować ją w sposób ustawowy. Do najważniejszych z nich można zaliczyć w pierwszej kolejności gromadzenie danych w większe i mniejsze zbiory. Zbiory w dzisiejszych czasach praktycznie w całości zapisywane i utrwalane są w formie systemów komputerowych, dzięki którym możliwe jest nie tylko szybkie przeszukiwanie, uaktualnianie i przetwarzanie tych informacji, ale także dużo łatwiejsze ich udostępnianie oraz tworzenie zbiorów o ogromnych rozmiarach. W tak zorganizowanych bazach danych osobowych znacznie wzrasta ryzyko zaistnienia naruszeń. Ze względu na duże możliwości techniczne widoczna staje się waga właściwych rozwiązań prawnych w tym zakresie. Brak uregulowań może prowadzić bowiem do niedozwolonego ingerowania w szeroko rozumianą wolność osobistą i prywatność jednostek, pozbawić ich możliwości kontrolowania i ingerowania w zakres gromadzonych na ich temat danych, a także decydowania, komu mogą zostać udostępnione[8]. Kolejnym powodem konieczności prawidłowego zabezpieczania danych osobowych jest udostępnianie danych i ich przetwarzanie za pomocą sieci komputerowych. Obecne stosunki handlowe, ekonomiczne, polityczne

[7] Konstytucja Rzeczypospolitej Polskiej z 1997 r., Dz.U. Nr 78, poz. 483.

[8] J. Barta, P. Fajgielski, R. Markiewicz, *Ochrona danych osobowych. Komentarz*, wyd. Wolters Kluwer, Kraków 2007, s. 51.

czy społeczne realizowane są w ogromnej mierze za pośrednictwem internetu. Pomimo wielu zalet, które niesie za sobą taki sposób wykorzystywania globalnej sieci, otwiera on również wiele pól, na których może dojść do niepożądanej ingerencji w gromadzone dane z zewnątrz, ich kopiowania, przejmowania i wykorzystywania sprzecznie z celem, dla którego zostały zgromadzone, a przede wszystkim bez zgodny osób, których dotyczą.

Przepisy o ochronie danych osobowych mają zapewnić procedury niezbędne do zabezpieczenia podstawowych praw i wolności, zwłaszcza prawa do prywatności oraz niezbędną bazę dla działania podmiotów prywatnych i publicznych.

3.2. Uprawnienie organów uczelni do przetwarzania danych osobowych

Ustawa o ochronie danych osobowych z zamysłu ustawodawcy w sposób kompleksowy reguluje przetwarzanie danych osobowych zarówno przez podmioty publiczne, jak i podmioty sektora prywatnego[9]. Zakres obowiązywania niniejszej ustawy obejmuje również działania organów szkół wyższych, związanych z gromadzeniem i przetwarzaniem danych studentów w ramach realizacji swoich celów, a także w ramach prowadzonych w uczelniach postępowań administracyjnych. Sama ustawa o szkolnictwie wyższym[10] nie reguluje wprost kwestii związanych z ochroną danych osobowych. W tym zakresie należy się posługiwać bezpośrednio ustawą o ochronie danych osobowych. Aby zapewnić ochronę danych osobowych przetwarzanych przez podmioty takie jak uczelnie[11] na gruncie art. 3 ustawy zobowiązano podmioty publiczne oraz podmioty niepubliczne wykonujące zadania powierzone przez państwo do stosowania przepisów ustawy o ochronie danych osobowych, w przypadku spełnienia łącznie dwóch warunków, tj. zadanie realizowane przez dany podmiot musi należeć do sfery publicznoprawnej, realizacja zadań publicznoprawnych musi wynikać z przepisów ustawowych. W przypadku wykonywania przez podmioty prywatne zadań, które można uznać za zadania publiczne, jednakże niezastrzeżonych ustawowo do kompetencji konkretnych typów podmiotów, tzn. zadań, które mogą być wyko-

[9] § 1 ustawy.

[10] Ustawa z dnia 27 lipca 2005 r. – Prawo o szkolnictwie wyższym, Dz.U. 2005 Nr 164, poz. 1365 z późn. zm.

[11] Zaliczyć do tej kategorii można również choćby przedszkola, niepubliczne zakłady opieki zdrowotnej, szkoły.

nywane przez dowolne podmioty prowadzące działalność gospodarczą, nie przesądza o konieczności stosowania zapisów ustawy[12]. W przypadku uczelni warunek realizowania zadań publicznych zostaje spełniony na mocy zapisów art. 4 ust. 3 Ustawy o szkolnictwie wyższym, zgodnie z którym uczelnie w swojej misji stanowią część narodowego systemu edukacji i nauki, a co za tym idzie: wykonują zadania zastrzeżone dla państwa. Obowiązek wykonywania przez uczelnie zadań powierzonych przez państwo wynika także z art. 18 i 20 tejże ustawy. Co więcej, zadania te nie są dostępne dla wszystkich podmiotów prowadzących działalność gospodarczą, a jedynie dla tych, które jak w przypadku uczelni publicznych zostaną powołane specjalnie w tym celu na mocy aktu ustawowego[13], a w przypadku uczelni niepublicznych ich działalność odbywa się na mocy pozwolenia wydawanego przez Ministra Nauki i Szkolnictwa Wyższego, w drodze decyzji administracyjnej[14].

Danymi osobowymi w szkołach wyższych zarządza rektor, pełniąc w tym zakresie funkcję administratora danych osobowych gromadzonych w uczelni. Uprawnienie rektora w tym zakresie wywieść należy z art. 66 Ustawy o szkolnictwie wyższym, zgodnie z którym rektor kieruje działalnością uczelni i reprezentuje ją na zewnątrz, w połączeniu z art. 7 Ustawy o ochronie danych osobowych. Art. 7 ustawy zawiera definicję instytucji administratora danych osobowych. Administratorem danych osobowych jest organ, jednostka organizacyjna, podmiot lub osoba spełniająca warunki polegające na wykonywaniu zadań publicznych lub w przypadku osoby fizycznej lub prawnej prowadzenia działalności zarobkowej, decyduje o celach i środkach przetwarzania danych.

Obowiązki rektora jako administratora danych to: udostępnianie danych osobom, których dane dotyczą zawartych w zbiorach informacji, w celu ich uaktualniania lub zmiany, obowiązek rejestracyjny zbiorów danych, zabezpieczanie danych, zachowywanie ich poufności, integralności i nienaruszalności. Administrator danych odpowiada za legalność przetwarzania danych i ich bezpieczeństwo. Rektor, gromadząc i przetwarzając dane, nie czyni tego osobiście. Osoby, które wykonują polecenia rektora w zakresie przetwarzania powierzonych uczelni danych osobowych, muszą posiadać w tym zakresie specjalne umocowanie[15]. Upoważnienie wymagane jest nawet w przypadku, gdy do zakresu obowiązków na danym stanowisku należy przetwarzanie danych. Oprócz tych osób w szkołach wyższych odrębne upoważnienia powinni również posiadać:

[12] P. Litwiński, *Ochrona danych osobowych w ogólnym postępowaniu administracyjnym*, Wolters Kluwer, Warszawa 2009, s. 36.

[13] Ustawa o szkolnictwie wyższym, art. 18.

[14] *Ibidem*, art. 20.

[15] Ustawa o ochronie danych osobowych, art. 31.

- kierownicy poszczególnych jednostek organizacyjnych – dziekani i prodziekani w stosunku do przetwarzania danych osobowych studentów na właściwych im wydziałach,
- pracownicy zatrudnieni na stanowiskach księgowych – w zakresie realizowanych zadań,
- pracownicy pionu informatyki, szczególnie administrator systemu informatycznego i pracownik pełniący funkcję administratora bezpieczeństwa informacji,
- uczniowie i praktykanci, jeżeli odbywanie praktyki może się wiązać z przetwarzaniem danych osobowych;
- pracownicy administracyjni, jeżeli w zakresie ich obowiązków leży praca z danymi osobowymi[16].

Formalne upoważnienie do dostępu do danych osobowych przetwarzanych w danej uczelni powinny otrzymać również prorektorzy, zwłaszcza prorektorzy ds. studenckich.

Wyżej wymienione osoby (poza administratorem bezpieczeństwa danych i administratorem systemu informatycznego) nie posiadają przymiotu administratora danych i nie ponoszą bezpośredniej odpowiedzialności za prawidłowość przetwarzania i zabezpieczania danych przed Generalnym Inspektorem Ochrony Danych Osobowych[17]. Upoważnieni pracownicy obowiązani są działać w granicach prawa i polityki bezpieczeństwa wypracowanej w danej uczelni. Ciąży na nich również tajemnica danych oraz sposobów ich zabezpieczania. Pracownicy o wiążącej ich tajemnicy powinni zostać poinformowani w chwili nadania im upoważnienia do przetwarzania danych. Można powiedzieć, że jest to forma tajemnicy zawodowej, zwłaszcza gdy osoby przetwarzające dane czynią to na podstawie umowy o pracę. Obowiązek zachowania danych osobowych w tajemnicy oznacza zakaz ujawniania danych studentów innym osobom. Naruszenie tego obowiązku prowadzi do odpowiedzialności karnej[18].

Na gruncie Ustawy o szkolnictwie wyższym oraz Ustawy o ochronie danych administratorem danych jest nie tylko rektor, ale także administrator bezpieczeństwa informacji (ABI), którego wyznacza administrator ochrony danych osobowych[19]. Obowiązek wyznaczenia ABI dotyczy nie tylko tych podmiotów, które przetwarzają dane w systemach informatycznych, ale tak-

[16] J. Borowicz, *Obowiązek prowadzenia przez pracodawcę dokumentacji osobowej i organizacyjnej z zakresu ochrony danych osobowych. Teza nr 3*, PiZS.2001.3.2, LEX 29032/3.

[17] Ustawa o ochronie danych osobowych, art. 8.

[18] *Ibidem*, art. 51.

[19] Ustawa o ochronie danych osobowych, art. 36 ust. 3

że w przypadku przetwarzania danych ręcznie. Ustawa nie precyzuje jednak szczegółowego zakresu obowiązków ABI. Jego głównym zadaniem jest nadzorowanie środków technicznych i organizacyjnych zapewniających ochronę przetwarzanych danych[20], przede wszystkim zabezpieczanie danych tak, aby nie zostały one udostępnione osobom nieupoważnionym, nie były przetwarzane z naruszeniem ustawy, kontrolowanie kwestii związanych z ich zmianą, uszkodzeniem lub zniszczeniem. Zazwyczaj w uczelniach funkcję ABI pełni pracownik administracyjny uczelni, który zostaje wyposażony w kompetencje nadzorcze i możliwość ingerowania w gromadzone zbiory, a także kompetencje uprawniające do wydawania wiążących wytycznych związanych z przetwarzaniem danych. Wyznaczenie ABI następuje w formie pisemnej. Niejednokrotnie funkcja ABI w uczelniach łączona jest z funkcją ASI, czyli administratora systemu informatycznego. Dzieje się tak ze względu na komputeryzację danych gromadzonych w uczelniach. Ilość danych, które uczelnie obowiązane są gromadzić na temat swoich studentów, wyklucza bowiem możliwość ręcznego przetwarzania tych danych, wymuszając stosowanie rozwiązań systemów informatycznych. Zarządzanie tak zorganizowanym systemem danych, zwłaszcza w kontekście ich ochrony i możliwości szybkiej ingerencji administratora w ich zakres w sytuacji zaistnienia naruszeń przemawia na gruncie szkół wyższych za łączeniem tych funkcji.

3.3. Bezpieczeństwo i ochrona danych osobowych w szkołach wyższych

Definicje

Szkoły wyższe w Polsce gromadzą i przetwarzają dane osobowe studentów i kandydatów na studia w celu wykonywania swoich zadań. Głównym i najważniejszym zadaniem uczelni jest zgodnie z art. 6 Ustawy o szkolnictwie wyższym prowadzenie studiów. W tym właśnie celu gromadzone są dane studentów, które wykorzystywane są później przez organy uczelni do realizowania procesu kształcenia oraz prowadzenia postępowania administracyjnego w sprawach studentów. Przed omówieniem kwestii bezpieczeństwa danych osobowych w uczelniach należy odpowiedzieć na dwa pytania: czym są dane osobowe i zbiory danych.

Definicję danych osobowych zawiera art. 6 o ochronie danych osobowych. Zgodnie z jego treścią za dane osobowe uważa się wszelkie informacje

[20] J. Barta, P. Fajgielski, R. Markiewicz, *Ochrona danych osobowych…*, s. 609.

dotyczące zidentyfikowanej lub możliwej do zidentyfikowania osoby fizycznej. Osobą możliwą do zidentyfikowania jest osoba, której tożsamość można określić bezpośrednio lub pośrednio, w szczególności przez powołanie się na numer identyfikacyjny albo jeden lub kilka specyficznych czynników określających jej cechy fizyczne, fizjologiczne, umysłowe, ekonomiczne, kulturowe lub społeczne. Informacji nie uważa się natomiast za umożliwiającą określenie tożsamości osoby, jeżeli wymagałoby to nadmiernych kosztów, czasu lub działań. Za dane osobowe można uznać na tej podstawie „wszelkie informacje" odnoszące się do każdego aspektu osoby, jej stosunków osobistych i rzeczowych, jej życia zawodowego, prywatnego, wykształcenia, wiedzy czy cech charakteru. Danymi osobowymi są zarówno informacje już rozpowszechnione lub opublikowane (zamieszczone w publikowanych materiałach), jak i w ogóle jeszcze nieujawnione[21]. Aby określoną informację można było zaliczyć do danych osobowych, muszą zostać spełnione dwa warunki: dane muszą dotyczyć osoby fizycznej, a jej tożsamość musi być na ich podstawie możliwa do ustalenia. Zgodnie z tym za dane osobowe nie można uznać danych dotyczących jednostek organizacyjnych (bez względu na to, czy posiadają one osobowość prawną, czy nie) ani szeroko rozumianych organów administracji państwowej. W polskim reżimie prawnym kwestie dotyczące podmiotów innych niż osoby fizyczne regulowane są przez odrębne ustawy. Aby dana informacja mogła zostać uznana za „dotyczącą osoby fizycznej", musi przekazywać informacje, które w jakikolwiek sposób odnoszą się do zidentyfikowanej bądź możliwej do zidentyfikowania osoby fizycznej, a identyfikacja odbywa się na podstawie całokształtu posiadanych informacji[22]. W tym kontekście do danych osobowych można również zaliczyć informację nieprzynależącą do zbioru danych, ale która w połączeniu z informacjami znajdującymi się w zbiorze umożliwia zidentyfikowanie osoby fizycznej[23]. Idąc tym tokiem rozumowania, za dane osobowe należy uznać nie tylko nazwiska czy daty urodzenia, ale także zainteresowania czy kierunek odbywanych studiów. Wyłączenia spośród tej grupy należy dokonać w stosunku do danych dotyczących stosunków majątkowych[24]. Zgodnie z przytoczonym orzeczeniem ujawnienie osobie trzeciej danych dotyczących wysokości wynagrodzenia za pracę pracownika nie stanowi samo w sobie naruszenia prywatności. Jeżeli natomiast taka informacja ujawniała-

[21] *Ibidem*, s. 346.

[22] *Ibidem*, s. 351.

[23] A. Bierć, *Ochrona prawna danych osobowych w sferze działalności gospodarczej w Polsce – aspekty cywilnoprawne*, [w:] *Ochrona danych osobowych w Polsce z perspektywy dziesięciolecia*, red. P. Fajgielski, Lublin 2008, s. 121–122.

[24] Uchwała składu 7 sędziów SN z dnia 16 lipca 1993 r., I PZP 28/93, OSNC 1994, nr 1, poz. 2.

by obowiązki alimentacyjne pracownika, można już mówić o naruszeniu jego praw i tajemnicy danych osobowych. Poza już wymienionymi przykładami danych osobowych należy wskazać numer powszechnego elektronicznego systemu ewidencji ludności (PESEL); numer identyfikacji podatkowej (NIP), numer dokumentu tożsamości (dowodu osobistego oraz paszportu), a także: wygląd zewnętrzny, wzór siatkówki oka (cechy fizyczne); struktura kodu genetycznego, grupa krwi (cechy fizjologiczne); pochodzenie, poglądy polityczne, przekonania religijne lub filozoficzne oraz przynależność wyznaniowa, partyjna lub związkowa (cechy te można zaliczyć do cech umysłowych, kulturowych lub społecznych, w zależności od sposobu interpretacji tych pojęć). Wskazane powyżej czynniki nie wyczerpują otwartego katalogu rodzajów informacji, które mogą być przypisane konkretnej osobie fizycznej[25]. W szkołach wyższych poza danymi wymienionymi powyżej za dane osobowe podlegające ochronie uznaje się dane o: punktach ECTS, kierunkach studiów podjętych przez studenta, dacie skreślenia i przyjęcia na studia, stopniu i formie studiów, studiowaniu na kolejnych kierunkach studiów, rodzaj pobieranych świadczeń pomocy materialnej[26].

Obok pojęcia danych osobowych kluczowa dla stosowania Ustawy o ochronie danych osobowych jest definicja pojęcia „zbioru danych". Pojęcie to jest znaczące w kontekście bezpieczeństwa i ochrony danych, gdyż od tego, czy zestawienie danych jest zbiorem, czy też nie, zależy powstanie obowiązku rejestracji zbioru. W przypadku gdy grupie danych osobowych nie można przypisać cech zbioru, dane takie nie podlegają zgłoszeniu do GIODO[27], który prowadzi rejestr. Za zbiór danych uznaje się posiadający pewną strukturę zestaw danych o charakterze osobowym, dostępnych według określonych kryteriów, bez względu na to, czy zestaw ten jest rozproszony lub podzielony funkcjonalnie[28]. Aby zestaw danych

[25] J. Barta, P. Fajgielski, R. Markiewicz, *Ochrona danych osobowych…*, s. 356.

[26] Art. § 2 Rozporządzenia Ministra Nauki i Szkolnictwa Wyższego w sprawie danych zamieszczanych w ogólnopolskim wykazie studentów z dnia 22 września 2011 r. (Dz.U. Nr 204, poz. 1201).

[27] Art. 40. ustawy o ochronie danych osobowych: „Administrator danych jest obowiązany zgłosić zbiór danych do rejestracji Generalnemu Inspektorowi Danych Osobowych (GIODO), z wyjątkiem przypadków, o których mowa w art. 43 ust. 1. Zgodnie z art. 8 ustawy o ochronie danych GIODO jest organem ochrony danych osobowych. Do jego zadań należy m.in.: kontrola zgodności przetwarzania danych z przepisami o ochronie danych osobowych, wydawanie decyzji administracyjnych i rozpatrywanie skarg w sprawach wykonywania danych osobowych, prowadzenie rejestru zbiorów danych oraz udzielanie informacji o zarejestrowanych zbiorach.

[28] Definicja wyprowadzona na podstawie art. 7 Ustawy o ochronie danych osobowych przez P. Litwińskiego, *Ochrona danych osobowych…*, s. 42.

mógł zostać uznany za zbiór danych, musi on łącznie wykazać następujące cechy: zawierać dane osobowe, posiadać zdefiniowaną własną strukturę, umożliwiać dostęp do danych według określonych kryteriów[29]. Cechą odróżniającą zbiór danych osobowych od innych zbiorów jest możliwość odnalezienia konkretnych informacji o danej osobie bez konieczności przeglądania całej zawartości zbioru. Nie wszystkie bowiem dane gromadzone w uczelniach można zdefiniować jako zbiory. Nie wyłącza to jednak stosowania przez organy uczelni przepisów dotyczących udostępniania i zabezpieczania danych studentów. Od zasady dotyczącej obowiązków rejestracji zbiorów istnieją bowiem wyjątki. Zaliczyć do nich można zbiory obejmujące dane osobowe studentów i pracowników uczelni. Zgodnie z art. 43 ust. 4 Ustawy o ochronie danych osobowych z obowiązku rejestracji zbioru zwolnieni są administratorzy danych przetwarzający dane swoich pracowników lub osób uczących się. Ustawodawca jednak nie przewidział i nie uregulował ani w Ustawie o ochronie danych osobowych, ani w Ustawie o szkolnictwie wyższym kwestii dotyczących danych osobowych kandydatów na studia i absolwentów. Powstaje zatem pytanie, czy takie zbiory podlegają zgłoszeniu, czy też nie. O ile w przypadku zbiorów danych można by pokusić się o interpretację negatywną, o tyle w przypadku zbiorów danych absolwentów sytuacja nie wydaje się już taka prosta. Zbiory danych dotyczące kandydatów mają charakter czasowy, tzn. tworzone są na potrzeby i w czasie trwania procesu rekrutacyjnego na studia, który jest stosunkowo krótki. Status studentów, którzy otrzymują decyzję pozytywną o przyjęciu na studia, zmienia się z kandydata na studenta, a ich dane osobowe zostają umieszczone w zbiorze niepodlegającym zgłoszeniu. Jeżeli natomiast kandydat otrzymuje decyzję negatywną i nie zostaje przyjęty w poczet studentów, jego dane są niszczone i nie występują w żadnym zbiorze. Brak rejestracji zbioru danych kandydatów na studia nie umniejsza ochrony zgromadzonych w nim informacji, gdyż władze uczelni muszą przestrzegać wszelkich reżimów dotyczących ochrony tych danych. Dużo bardziej skomplikowaną kwestią jest istnienie bądź nieistnienie obowiązku rejestracji zbiorów danych absolwentów. Zgodnie z art. 13a Ustawy o szkolnictwie wyższym do zadań uczelni należy monitorowanie karier zawodowych swoich absolwentów. Wiąże się to nierozerwalnie z gromadzeniem w zbiorach danych osobowych absolwentów. Przywołany przepis nie określa czasu, przez jaki uczelnia obowiązana jest zbierać dane na temat karier

[29] M. Sakowska, *Pozycja ustrojowa i zadania Generalnego Inspektora Ochrony Danych Osobowych*, „Przegląd Sejmowy" 2006, nr 2, s. 57.

zawodowych swoich absolwentów. Wiadomo jednak, że okres ten nie może być krótszy niż pięć lat. Wydawałoby się zatem, że zbiory takie podlegają zgłoszeniu, choć nigdzie nie jest to przesądzone. Janusz Barta, Paweł Fajgielski i Ryszard Markiewicz w swoim komentarzu do Ustawy o ochronie danych osobowych twierdzą, że analogicznie jak zbiory danych studentów należy potraktować zbiory danych absolwentów, korzystając z możliwości wyłączenia obowiązku rejestracyjnego[30]. Autorzy niestety bliżej nie uzasadniają swojego poglądu i nie podają przesłanek, które skłoniły ich do przyjęcia takiego stanowiska.

Zabezpieczanie danych osobowych

To, czy uczelnia przetwarza dane w zbiorze, czy też nie, o czym była już mowa wyżej, nie wpływa na ograniczenie zakresu ochrony danych osobowych studentów. Zabezpieczanie danych osobowych jest procesem mającym na celu ograniczenie prawdopodobieństwa i ryzyka wystąpienia naruszeń związanych z gromadzeniem, przetwarzaniem i udostępnianiem danych osobowych. Środki, za pomocą których zabezpieczane są dane osobowe, powinny być stosowane w dwóch etapach. Po pierwsze, przed przystąpieniem do przetwarzania danych, a po drugie, w trakcie przetwarzania danych osobowych. Środki ochrony danych mają za zadanie przeciwdziałać wszelkim zachowaniom ze strony osób trzecich, a nawet działaniu siły wyższej, zmierzającym do nieuprawnionego dostępu do danych. Środki ochrony danych powinny być stosowane nie tylko do zbiorów danych, ale do danych osobowych jako takich. Wymogi określone w art. 36 ustawy o ochronie danych odnoszą się zarówno do danych przetwarzanych w sposób manualny, jak i tych przetwarzanych w systemach informatycznych. Ustawodawca nie podaje, jakie konkretnie środki ma przedsięwziąć rektor w celu zabezpieczenia danych. Można je określić dopiero na podstawie analizy obowiązków, jakie przepisy prawa nakładają na administratora danych.

Zadaniem rektora jest zatem stosowanie skutecznych środków technicznych i organizacyjnych. Ustawodawca nie przesądza również, jakie to mają być środki. Przy stosowaniu zabezpieczeń powinno się też uwzględniać zmieniające się warunki oraz postęp techniczny (informatyczny), co może powodować konieczność zmiany czy modernizowania wprowadzonych wcześniej przez administratora systemów ochrony. Można i należy dopasowywać je do konkretnych okoliczności i warunków przetwarzania danych[31]. Zaznaczyć

[30] J. Barta, P. Fajgielski, R. Markiewicz, *Ochrona danych osobowych…*, s. 640.
[31] *Ibidem*, s. 607.

w tym miejscu należy, że poza zobiektywizowanymi przesłankami podjęcia takich a nie innych rodzajów ochrony danych powinna decydować także kosztochłonność wprowadzanych zabezpieczeń, charakter chronionych danych, szkodę, jaka mogłaby powstać w związku z nieuprawnionym dostępem do danych lub innym ich przetwarzaniem.

Dokumentem opisującym całość środków technicznych i organizacyjnych zapewniających ochronę danych oraz sposoby ich przetwarzania jest dokumentacja przetwarzania danych. Niezbędne elementy, jakie powinna zawierać dokumentacja, a co za tym idzie wskazówki, jakie działania powinny być podjęte w celu zabezpieczenia danych określają przepisy rozporządzenia Ministra Spraw Wewnętrznych i Administracji w sprawie dokumentacji przetwarzania danych osobowych oraz warunków technicznych i organizacyjnych, jakim powinny odpowiadać urządzenia i systemy informatyczne służące do przetwarzania danych osobowych[32]. Zgodnie z § 3 rozporządzenia dokumentację powyższą stanowi: polityka bezpieczeństwa oraz instrukcja zarządzania systemem informatycznym służącym do przetwarzania danych osobowych, które powinny być prowadzone w formie pisemnej. Na politykę bezpieczeństwa składa się:

- wykaz budynków, pomieszczeń lub części pomieszczeń tworzących obszar, w którym przetwarzane są dane osobowe,
- wykaz zbiorów danych osobowych wraz ze wskazaniem programów zastosowanych do przetwarzania tych danych,
- opis struktury zbiorów danych wskazujący zawartość poszczególnych pól informacyjnych i powiązania między nimi,
- sposób przepływu danych pomiędzy poszczególnymi systemami,
- określenie środków technicznych i organizacyjnych niezbędnych do zapewnienia poufności, integralności i rozliczalności przetwarzanych danych[33].

Zacytowany katalog nie stanowi jednak katalogu zamkniętego. Powinny znaleźć się tu wszelkie informacje opisujące sposób i miejsca przetwarzania danych, a także przyjęte w tym zakresie rozwiązania techniczne i inne. Również Generalny Inspektor Ochrony Danych Osobowych posiada uprawnienia do wydawania dokumentów stanowiących pomoc dla administratorów

[32] Rozporządzenie Ministra Spraw Wewnętrznych i Administracji z dnia 29 kwietnia 2004 r. w sprawie dokumentacji przetwarzania danych osobowych oraz warunków technicznych i organizacyjnych, jakim powinny odpowiadać urządzenia i systemy informatyczne służące do przetwarzania danych osobowych, Dz.U. z 2004 r., Nr 100, poz. 1024. – zwane dalej „rozporządzeniem".

[33] § 4 rozporządzenia.

danych osobowych. Wydaje on wytyczne w tym zakresie, w postaci poradników zamieszczanych na stronie internetowej urzędu. GIODO zdefiniował w nich pojęcie „polityka bezpieczeństwa", wskazał cel jej opracowania i wdrożenia oraz poszczególne elementy składowe tego dokumentu w sposób szerszy niż w rozporządzeniu, korzystając z otwartości zamieszczonego w nim katalogu[34]. W wytycznych oparto się m.in. na Polskiej Normie PN-ISO/IEC 17799:2007 Technika Informatyczna.

Zakres informacji, jakie powinny być zawarte w dokumencie polityki bezpieczeństwa zgodnie z Polską Normą PN-ISO/IEC 17799:2007, jest znacznie szerszy, niż przewiduje to rozporządzenie. W wytycznych wskazano, że polityka bezpieczeństwa to „zestaw praw, reguł i praktycznych doświadczeń regulujących sposób zarządzania, ochrony i dystrybucji informacji wrażliwej wewnątrz określonej organizacji". Polityka bezpieczeństwa powinna odnosić się całościowo do problemu zabezpieczenia danych osobowych u administratora danych, tj. zarówno do zabezpieczenia danych przetwarzanych tradycyjnie, jak i danych przetwarzanych w systemach informatycznych. Celem polityki bezpieczeństwa jest wskazanie działań, jakie należy wykonać, oraz ustanowienie zasad i reguł postępowania, które należy stosować, aby właściwie wykonać obowiązki administratora danych w zakresie zabezpieczenia danych osobowych[35]. Zgodnie z wytycznymi zawartymi w Polskiej Normie w zakresie zarządzania bezpieczeństwem systemów informatycznych, dokumentacja powinna zawierać:

a. definicję bezpieczeństwa informacji, jego ogólne cele i zakres oraz znaczenie bezpieczeństwa jako mechanizmu umożliwiającego współużytkowanie informacji;
b. oświadczenie o intencjach kierownictwa, potwierdzające cele i zasady bezpieczeństwa informacji;
c. krótkie wyjaśnienie polityki bezpieczeństwa, zasad, standardów i wymagań zgodności mających szczególne znaczenie dla instytucji, np.:
   - zgodność z prawem i wymaganiami wynikającymi z umów,
   - wymagania dotyczące kształcenia w dziedzinie bezpieczeństwa,
   - zapobieganie i wykrywanie wirusów oraz innego złośliwego oprogramowania,
   - zarządzanie ciągłością działania biznesowego,
   - konsekwencje naruszenia polityki bezpieczeństwa;

[34] http://www.giodo.gov.pl/1520074/j/pl/.
[35] J. Barta, P. Fajgielski, R. Markiewicz, *Ochrona danych osobowych…*, www.lex.pl.

d. definicje ogólnych i szczególnych obowiązków w odniesieniu do zarządzania bezpieczeństwem informacji, w tym zgłaszania przypadków naruszenia bezpieczeństwa;
e. odsyłacze do dokumentacji mogącej uzupełniać politykę, np. bardziej szczegółowych polityk bezpieczeństwa i procedur dla poszczególnych systemów informatycznych lub zasad bezpieczeństwa, których użytkownicy powinni przestrzegać.

Zamieszczenie w polityce bezpieczeństwa wymogów określonych w Polskiej Normie nie jest obowiązkowe i zależy od decyzji administratora danych.

Instrukcja zarządzania systemem informatycznym służącym do przetwarzania danych osobowych jest drugim wymaganym na mocy rozporządzenia dokumentem opisującym system przetwarzania danych osobowych. Dokumentacja ta wymagana jest jednak jedynie w przypadkach administrowania danymi w systemie informatycznym. Administratorzy, którzy przetwarzają dane w sposób tradycyjny, nie muszą jej tworzyć i posiadać. Nie są także obowiązani do powołania ASI (administratora systemu informatycznego). W przepisach nie określono jednak szczegółowo zakresu obowiązków administratora bezpieczeństwa informacji. Ustawodawca ograniczył się tylko do ogólnego stwierdzenia, że administrator bezpieczeństwa informacji ma nadzorować przestrzeganie zasad ochrony, o których mowa w art. 36 ust. 1 Ustawy o ochronie danych osobowych, z czego wynika, że obowiązkiem administratora bezpieczeństwa informacji jest nadzorowanie stosowania środków technicznych i organizacyjnych zapewniających ochronę przetwarzanych danych osobowych. Administrator bezpieczeństwa informacji powinien nadzorować przede wszystkim zabezpieczenie danych przed ich udostępnieniem osobom nieupoważnionym, zabraniem przez osobę nieuprawnioną, przetwarzaniem z naruszeniem ustawy oraz zmianą, utratą, uszkodzeniem lub zniszczeniem[36]. Zgodnie z § 5 rozporządzenia instrukcja zarządzania systemem informatycznym powinna zawierać w szczególności:

- procedury nadawania uprawnień do przetwarzania danych i rejestrowania tych uprawnień w systemie informatycznym oraz wskazanie osoby odpowiedzialnej za te czynności;
- stosowane metody i środki uwierzytelnienia (działań, których celem jest weryfikacja deklarowanej tożsamości podmiotu) oraz procedury związane z ich zarządzaniem i użytkowaniem;
- procedury rozpoczęcia, zawieszenia i zakończenia pracy przeznaczone dla użytkowników systemu;

[36] A. Drozd, *Zabezpieczenie danych osobowych*, Wrocław 2008, s. 65 i n.

- procedury tworzenia kopii zapasowych zbiorów danych oraz programów i narzędzi programowych służących do ich przetwarzania;
- sposób, miejsce i okres przechowywania elektronicznych nośników informacji zawierających dane osobowe, a także kopii zapasowych;
- sposób zabezpieczenia systemu informatycznego przed działalnością oprogramowania, którego celem jest uzyskanie nieuprawnionego dostępu do systemu informatycznego;
- sposób odnotowywania informacji o odbiorcach, którym dane osobowe zostały udostępnione, dacie i zakresie tego udostępnienia, chyba że system informatyczny używany jest do przetwarzania danych zawartych w zbiorach jawnych;
- procedury wykonywania przeglądów i konserwacji systemów oraz nośników informacji służących do przetwarzania danych.

Aby całość dokumentacji opracowanej przez rektora mogła prawidłowo funkcjonować, wymaga nie tylko ogłoszenia jako aktów prawa wewnątrzuczelnianego, ale także musi zostać właściwie wdrożona. Polityka bezpieczeństwa wprowadzana jest bowiem na mocy zarządzenia rektora, co zapewnia jej stosowne miejsce w hierarchii źródeł „prawa uczelnianego". Sam fakt obowiązywania nie oznacza jednak wdrożenia zasad przyjętych celem ochrony danych osobowych. Najistotniejszym elementem wdrożenia jest zaznajomienie z ich treścią osób upoważnionych do przetwarzania danych. Stworzenie i wdrożenie dokumentacji dotyczącej ochrony danych osobowych nie jest czynnością jednorazową[37]. Na administratorze danych osobowych ciąży bowiem obowiązek stałej aktualizacji stworzonej dokumentacji, zarówno pod kątem zmian w przepisach prawa, jak i zamian w sytuacji faktycznego przetwarzania danych osobowych (np. zmiana pomieszczeń, w których przetwarzane są dane, zmiana struktury systemu informatycznego, poprzez jego unowocześnianie, zmiana osób upoważnionych do przetwarzania danych).

Całość opublikowanej i wdrożonej dokumentacji związanej z ochroną danych osobowych w uczelni tworzy swoisty system ochrony danych. Do dokumentacji tej zaliczyć należy także wyznaczenie administratora bezpieczeństwa informacji, upoważnienia do przetwarzania danych, ewidencja osób upoważnionych do przetwarzania danych. Ewidencja osób upoważnionych do przetwarzania danych w praktyce może dzielić się na dwie odrębne ewidencje. Zgodnie z art. 39 ust. 1 Ustawy o ochronie danych w ewidencji powinny się znaleźć następujące informacje: imię i nazwisko osoby upoważ-

[37] A. Gałach, *Instrukcja ochrony danych osobowych w systemie informatycznym*, Gdańsk 2004, s. 53.

nionej; data nadania i ustania oraz zakres upoważnienia do przetwarzania danych; identyfikator, jeżeli dane są przetwarzane w systemie informatycznym. W szkołach wyższych często jednak zdarza się tak, że nie wszystkie osoby przetwarzające dane studentów przetwarzają je w systemie informatycznym. Z tego względu tworzone są dwie odrębne często ewidencje, jedna stanowiąca spis wszystkich osób posiadających upoważnienie do przetwarzania danych oraz druga będąca wykazem osób pracujących na systemach informatycznych. Zawsze jednak osoby posiadające dostęp do systemu informatycznego muszą mieć ogólne upoważnienie do przetwarzania danych. Innymi słowy, każdy pracownik uczelni przetwarzający dane studentów musi posiadać stosowne umocowanie w tym zakresie i zostać ujęty w ewidencji osób upoważnionych do przetwarzania danych, jednak nie każda z tych osób musi posiadać uprawnienie do przetwarzania danych w systemie informatycznym. Co więcej, jeżeli dostęp do danych przetwarzanych w systemie informatycznym mają co najmniej dwie osoby, dla każdej z nich powinien być zarejestrowany odrębny identyfikator, a dostęp do danych powinien być możliwy wyłącznie po wprowadzeniu identyfikatora i dokonaniu uwierzytelnienia.

Ewidencja prowadzona jest zazwyczaj w formie pisemnej. Na gruncie przepisów nie zostało jednak przesądzone, w jakiej formie powinna ona być prowadzona. Wydawałoby się, że uzasadnione jest przyjęcie rozwiązania stosowania formy pisemnej choćby ze względów dowodowych, przy czym w stosunku do osób przetwarzających dane w systemie informatycznym wystarczająca jest forma elektroniczna. Osoby te są już bowiem wymienione w ogólnej ewidencji osób upoważnionych do przetwarzania danych. Poza umocowaniem ze strony ASI muszą mieć pełnomocnictwo nadane przez administratora danych do ich przetwarzania. W konsekwencji tego osoby pracujące w systemie informatycznym już raz zostają wykazane w ewidencji prowadzonej z zachowaniem formy pisemnej.

Ewidencja osób poza funkcją rejestrową pełni także rolę porządkującą. Za pomocą ewidencji administratorzy danych osobowych w prosty sposób weryfikują pracowników posiadających dostęp do danych, choćby pod kątem zachowania przez nich tajemnicy danych.

Podsumowanie

Biorąc pod uwagę całokształt przedstawionych rozważań na temat ochrony danych osobowych w szkołach wyższych, należy podkreślić, że w stosunku do danych przetwarzanych w uczelniach stosuje się ogólne zasady wymienione w Ustawie o ochronie danych. Zgodnie z jej art. 5, ustawodawca przyjmuje zasadę rozstrzygania zbiegu norm na korzyść tych norm, które przewidują wyższy poziom ochrony. Innymi słowy, jeżeli ustawy szczególne regulują kwestie ochrony danych, zapewniając jeszcze bardziej rygorystyczne mechanizmy bezpieczeństwa, stosuje się przepisy tych ustaw. W polskim systemie prawnym istnieje relatywnie wiele przepisów odrębnych ustaw, które odnoszą się do szeroko rozumianego przetwarzania danych, przy czym znaczna ich część została wydana wcześniej niż obowiązująca Ustawa o ochronie danych osobowych. W przypadku szkół wyższych przepisy Ustawy o szkolnictwie wyższym nie regulują kwestii ochrony danych osobowych studentów. Ta swoista ochrona w każdym przypadku (nie tylko w szkołach wyższych) niesie za sobą ochronę własności osób fizycznych. W związku z tym bezpieczeństwo powinno być rozpatrywane w kontekście organizacyjnym, technicznym i prawnym. Interdyscyplinarne podejście do zagadnienia umożliwia zapewnienie wysokiego poziomu jakości wprowadzanych procedur, zwłaszcza że systemy bezpieczeństwa danych powinny być dopasowane do specyfiki działania w uczelniach.

CZĘŚĆ II

Modele organizacji i tworzenia zasobów cyfrowych w kontekście ich jakości oraz wiarygodności

Models the organization and creation of digital resources in terms of their quality and reliability

STRESZCZENIE:
Omówiono metody i narzędzia wykorzystywane w procesie analizy i projektowania baz danych w ramach dużych zbiorów zwanych bazami wiedzy, poruszono problematykę konstruowania jednoznacznych i niezmiennych w czasie odwołań do dokumentów cyfrowych zgromadzonych w różnego rodzaju elektronicznych repozytoriach, wskazano na zadania systemu repozytoriów.

Słowa kluczowe: systemy tablicowe, dokumenty cyfrowe, repozytoria.

Cel:
Prezentacja zagadnień związanych z udostępnianiem danych cyfrowych.

Metodyka badań:
Metody: opisu pojedynczych przypadków, analiza dokumentacji, modelowanie.

Wynik:
Spójne opracowanie w zakresie wybranych metod opracowania cyfrowej dokumentacji.

Oryginalność wartość:
Oryginalny układ i połączenie treści.

ABSTRACT:
This section discusses the methods and tools used in the analysis and design databases within large collections called knowledge bases, raised the issue of constructing a clear and unvarying in time appeals to digital documents stored in various electronic repositories, pointed out the tasks of the system repositories.

Key words: array systems, digital documents, repositories.

Presentation of issues related to the sharing of digital data.

Methods: describing individual cases, analysis of documentation, modeling.

Coherent development of selected developing methods of digital documentation.

Original layout and combination of content.

4. Wykorzystanie systemów tablicowych do uporządkowania wiedzy technologicznej

Dorota Wilk-Kołodziejczyk, Renata Uryga, Agnieszka Smolarek-Grzyb

Wprowadzenie

Ze względu na złożoność problemów przetwarzania informacji zmuszeni jesteśmy często do posługiwania się dużymi zbiorami danych. Niektóre operacje na takich zbiorach oraz przechowywanie danych możliwe są do realizacji dzięki technologii relacyjnych baz danych[1]. Systemy zarządzania relacyjnymi bazami danych są obecnie bardzo ważnymi narzędziami pracy i nieodzownymi elementami nowoczesnych aplikacji[2]. Ułatwiają zarządzanie i przechowywanie dużej ilości danych. W relacyjnych bazach danych powiązania między tablicami są realizowane poprzez atrybuty wspólne. Oznacza to, że atrybut o danej nazwie może występować w kilku relacjach. Jednym z najpopularniejszych systemów zarządzania relacyjną bazą danych używaną głównie w aplikacjach internetowych jest MySQL, który może obsługiwać mające miliony rekordów duże bazy danych. Ponadto cechuje go szybkość, łatwość użycia, bezpieczeństwo, duże zastosowanie[3].

W dalszym ciągu jednak logiczne przetwarzanie danych stanowi bardziej przedmiot badań naukowych niż praktyki inżynierskiej, pomimo że badania modelu logicznego baz danych zostały podjęte stosunkowo wcześnie. Inteligentna analiza danych oraz generacja reguł z przykładów stanowią jedne z najbardziej zaawansowanych kierunków badań.

W rozdziale tym omówiono metody i narzędzia analizy oraz wspomagania projektowania baz danych i baz wiedzy. Do tego wykorzystano wspólny model reprezentacji danych i wiedzy. Oparty jest on na tablicowym schema-

---

[1] P. Beynon-Davies, *Database Systems*, MacMillan Press Ltd. 1996, wyd. polskie: *Systemy baz danych*, Wydawnictwa Naukowo-Techniczne, Warszawa 1998.

[2] Bach M., Kozielski S., *Translacja zapytań do baz danych sformułowanych w języku naturalnym na zapytania w języku SQL*, Konferencja Naukowa, Technologie przetwarzania danych, Wydawnictwo Politechniki Poznańskiej, Poznań 2005.

[3] W. Traczyk, *Jak uczyć się z różnorodnych przykładów, Inżynieria Wiedzy i Systemy Ekspertowe*, red. Z. Bubnicki i A. Grzech, Oficyna Wydawnicza Politechniki Wrocławskiej, t. 1, Wrocław 1997, s. 21–28.

cie relacyjnych baz danych. Bazy te analizowane są pod kątem weryfikacji jakościowych własności teoretycznych, do których należą: zupełność, nadmiarowość, spójność, efektywność reprezentacji, możliwość agregacji, postać rozwinięcia specyfikacji. Celem tej analizy jest dążenie do zapewnienia określonego poziomu jakości baz danych i baz wiedzy.

4.1. Systemy tablicowe

W rozdziale rozważany jest taki model bazowy, który jest wspólny dla reprezentacji danych i wiedzy. W rozważaniach wykorzystano pojedynczą tablicę. Niech $A=\{A_1, A_2,\ldots\ldots A_n\}$ będzie określonym zbiorem własności o dziedzinach odpowiednio $D_1,D_2,\ldots\ldots D_n$, gdzie D_i jest dziedziną atrybutu A_1 dla i=1,2,….,n. Rozważane są dwa rodzaje dziedzin: nieuporządkowane zbiory nazw (nominalne) oraz uporządkowane liniowo (dyskretne). Rozważa się tylko dziedziny skończone. Przy opisie własności obiektów podaje się wartości wszystkich atrybutów lub warunki, które własności te muszą spełniać. Podstawowy zapis faktu mówiącego, że wartość atrybutu A_1 wynosi t, ma postać $A_1=t$, gdzie $t \subseteq D_i$. Zapis ten dopuszcza istotne rozszerzenie w stosunku do klasycznego relacyjnego modelu danych – wartości atrybutów nie muszą być atomiczne, a więc reprezentacja warunków, jakie muszą spełniać poszczególne atrybuty, dopuszcza specyfikację intensjonalną, a wartości te mogą należeć do zbioru lub przedziału.

Ogólny schemat reprezentacji informacji jest wspólny dla danych i wiedzy i ma postać tablicy, której kolumny etykietowane są wybranymi atrybutami, zaś wiersze tablicy odpowiadają opisom kolejnych obiektów lub reguł wnioskowania. Postać takiej tablicy, nazywanej również (atrybutową) tablicą decyzyjną jest w ogólnym przypadku następująca:

A_1	A_2	…	A_j	…	A_n	H
$t_{1,1}$	$t_{1,2}$	…	$t_{1,j}$	…	$t_{1,n}$	h_1
$t_{2,1}$	$t_{2,2}$	…	$t_{2,j}$	…	$t_{2,n}$	h_2
:	:		:		:	:
$t_{i,1}$	$t_{i,2}$	…	$t_{i,j}$	…	$t_{i,n}$	h_2
:	:		:		:	:
$t_{m,1}$	$t_{m,2}$	…	$t_{m,j}$	…	$t_{m,n}$	h_m

Rysunek 3. Schemat tablicowej reprezentacji danych i wiedzy
Źródło: opracowanie własne.

W tablicy opisanych jest *m* reguł. Zakłada się, że są to obiekty jednorodne, tzn. każdy z nich opisywany jest poprzez podanie wartości tego samego zestawu atrybutów. W tablicy mogą być reprezentowane zarówno dane (wartości atrybutów są atomiczne), jak i wzorce danych (wartości atrybutów są wtedy podzbiorami dziedzin), oraz reguły wnioskowania; w przypadku reguł wybrany atrybut H (lub kilka atrybutów) ma charakter konkluzji, a poprzedzające atrybuty definiują prewarunki reguły. W przypadku danych i wzorców danych kolumna etykietowana atrybutem H nie występuje, natomiast informacja zawarta w rekordach tabeli jest deklarowana jako prawdziwa (konkluzje reguł bez prewarunków). Najistotniejsze rozszerzenia w stosunku do relacyjnych baz danych obejmują dopuszczenie nieatomicznych wartości danych (takich jak zbiory czy przedziały) oraz interpretacji rekordów jako reguł wnioskowania.

4.2. Model systemu tablicowego – model logiczny

Jakościowe własności systemów tablicowych definiowane są na poziomie logicznym. Rozważmy pojedynczą tablicę według schematu przedstawionego na rysunku 1. Jeżeli tablica ta reprezentuje informację będącą uogólnieniem bazy danych, to semantyka każdego rekordu definiowana jest formułą logiczną postaci:

$$\varphi_1=[A_1 = t_{i,1}] \wedge [A_2 = t_{i,2}] \wedge \ldots \wedge [A_n = t_{i,n}] \tag{5}$$

Zapis postaci $A_j=t_{i,j}$ oznacza, że formuła φ_i jest prawdziwa dla wszystkich wartości atomicznych należących do zbioru $T_{i,j}$. Tak więc, jeżeli $T_{i,j} = \{d_1,d_2, \ldots,d_k\}$, $A_j = T_{i,j}$ oznacza, że $A_i=d_1 \vee A_i = d_2 \vee \ldots \vee A_i = d_k$. Taki sposób zapisu stanowi skrót ekstensjonalnej reprezentacji, w której pojedynczemu wierszowi tablicy odpowiadałoby k wierszy, takich, że w każdym z nich wartość atrybutu A_i byłaby równa odpowiedniej wartości atomicznej. Tablicy natomiast odpowiada formuła

$$\psi = \varphi_1 \vee \varphi_2 \vee \ldots \vee \varphi_m \tag{6}$$

Jeżeli w tablicy reprezentowane są reguły, to semantyka każdego wiersza definiowana jest formułą postaci

$$\rho_1=[A_1 = t_{i,1}] \wedge [A_2 = t_{i,2}] \wedge \ldots \wedge [A_n = t_{i,n}] \Rightarrow [H = h_i] \tag{7}$$

4.4. Praktyczne zastosowanie systemów tablicowych

W pracy wykorzystano model tablic atrybutowych (decyzyjnych), wspólny dla baz danych i baz wiedzy. Na rysunku pokazano postać tablicy atrybutowej, której kolumny etykietowane są wybranymi atrybutami (A_i), a wiersze odpowiadają opisom kolejnych obiektów (o_j) (w tablicach decyzyjnych są to reguły wnioskowania).

A_1	A_2	A_3	...	A_n	0
$t_{1,1}$	$t_{1,2}$	$t_{1,3}$	...	$t_{1,5}$	0_1
$t_{2,1}$	$t_{2,2}$	$t_{2,3}$	...	$t_{2,5}$	0_2
...	...	...	...	...	
$t_{I,1}$	$t_{I,2}$	$t_{I,3}$	...	$t_{I,5}$	0_I
...	...	...	...	...	
$t_{k,1}$	$t_{k,2}$	$t_{k,3}$	...	$t_{k,5}$	0_k

Rysunek 4. Tablica atrybutowa wad odlewniczych
Źródło: opracowanie własne.

rodzaj uszkodzenia	widoczność	wielkość uszkodzenia	ilość materiału	rozmieszczenie	lokalizacja
wyszczerbienie	widoczne okiem nieuzbrojonym	małe	nadmiar	miejscowe	powierzchnia
odłamanie	dobrze widoczne	nieznaczne	niedomiar	rozproszone	krawędź
zbicie	niewidoczne	wyraźne	nieistotne	skupione	wnętrze
ubicie	niewidoczne okiem nieuzbrojonym			rozległe	naroże
usunięcie części odlewu	trudno dostrzegalny				wystających elementów
odkształcenie	widoczna pod mikroskopem				część
...	....				.....

Rysunek 5. Fragment tabeli zawierającej specyfikacje zbiorów wartości dla poszczególnych atrybutów wad
Źródło: opracowanie własne.

W tej metodzie identyfikacja wady jest dokonywana na podstawie wartości jej atrybutów. Analizując opisy wad zamieszczone w dokumentach źródłowych, sporządzono listę atrybutów wad, które wystąpiły w którymkolwiek z branych pod uwagę systemów, zapisano je w zbiorze A.

A={rodzaj uszkodzenia, wielkość, liczebność, widoczność, kształt, lokalizacja, czas powstania}

Dla każdego z tych atrybutów zdefiniowano zbiory wartości, fragment tablicy ilustruje rysunek 5. Oczywiście nie są to zbiory równoliczne, np. wyspecyfikowano 42 wartości dla atrybutu „rodzaj uszkodzenia", a tylko trzy dla atrybutu „wielkość".

Istotna różnica pomiędzy klasycznymi tablicami decyzyjnymi a stworzonymi tu tablicami atrybutowymi polega na tym, iż jak pokazano na rysunku 4, w naszej tablicy występują miejsca puste. Rysunek ten wskazuje także na logiczny model definiowania nazwy za pomocą atrybutów jako koniunkcja określonych wartości atrybutów.

4.4. Weryfikacja własności jakościowych

Reguła subsumowana

Subsumpcja zachodzi wtedy, gdy pewna reguła jest ogólniejsza od innej. W szczególności, gdy przy słabszych lub identycznych prewarunkach pozwala wyciągnąć te same lub silniejsze wnioski.

Rozważmy dwie reguły zapisane w dwóch wierszach tablicy (rysunek 7):

A_1	A_2	…	A_j	…	A_n	H
$t_{1,1}$	$t_{1,2}$		$t_{1,j}$	…	$t_{1,n}$	h_1
$t_{2,1}$	$t_{2,2}$	…	$t_{2,j}$	…	$t_{2,n}$	h_2

Rysunek 6. Tablica 1

Źródło: opracowanie własne.

Reguła pierwsza pokrywa regułę drugą w przypadku, jeżeli spełnione są następujące warunki: $t_{2,j} \subseteq t_{1,j}$ dla j=1,2,…,n oraz $h_1 \subseteq h_2$. Regułę pierwszą można zastosować w każdym przypadku, w którym można zastosować regułę drugą. Produkowana przez nią konkluzja jest bardziej precyzyjna, więc regułę drugą można usunąć z systemu bez naruszenia potencjalnych możliwości dedukcji.

Przedstawiona definicja pokrywa bardziej szczegółowe przypadki, takie jak identyczność reguł, pokrywanie prewarunków czy brak ograniczeń na wybrane atrybuty w regule bardziej ogólnej.

Własność ta może też być wykrywana w tablicach zawierających wzorce danych. W zależności od interpretacji i zastosowania z tabeli usuwane są szablony bardziej ogólne (jeśli chcemy zachować najbardziej precyzyjny

opis) lub szablony najbardziej szczegółowe (gdy chodzi o znalezienie najbardziej ogólnego pokrycia). Warunek subsumpcji można zapisać w postaci $\varphi_2 \subseteq \varphi_1$ lub logicznie $\varphi_2 \models \varphi_1$.

Redukcja

Usuwanie reguł pozwala na sklejanie przynajmniej dwóch reguł o identycznych konkluzjach. Rozważamy przypadek sklejania dwóch reguł o takich samych konkluzjach, zapisanych w następującej tablicy:

A_1	A_2	…	A_j	…	A_n	H
$t_{1,1}$	$t_{1,2}$		$t_{1,j}$	…	$t_{1,n}$	h_1
$t_{2,1}$	$t_{2,2}$	…	$t_{2,j}$	…	$t_{2,n}$	h_2

Rysunek 7. Tablica 2

Źródło: opracowanie własne.

Jeżeli zachodzą warunki: $t_{2,i} \subseteq t_{1,i}$ dla i=1,2,…,j-1,j+1,…,n, oraz $t_j = t_{1,j} \cup t_{2,j}$, to parę powyższych reguł można zastąpić regułą postaci równoważną parze reguł wyjściowych.

A_1	A_2	…	A_j	…	A_n	H
$t_{1,1}$	$t_{1,2}$		$t_{1,j}$	…	$t_{1,n}$	h_1

Rysunek 8. Tablica 3

Źródło: opracowanie własne.

Operacja sklejania ma na celu redukcję rozmiarów tablicy reguł. W zależności od wyboru reguł maksymalnie zredukowana postać tablicy nie jest podana jednoznacznie. Operacja sklejania może mieć również zastosowanie do tablic zawierających szablony danych. Otrzymana w wyniku sklejania rekordów zredukowana tablica opisuje ten sam zbiór obiektów co tablica wyjściowa, jednak jej rozmiary mogą być znacznie zmniejszone.

Rozbicie

Jest to metoda przeciwna do operacji sklejania. Czasami może być celowe rozbicie reguły na dwie lub więcej równoważnych reguł szczegółowych. W najprostszym przypadku pojedyncza reguła może być sprowadzona do dwóch reguł równoważnych o identycznych konkluzjach, ale przesłankach zdecydowanie bardziej szczegółowych niż w regule wyjściowej.

W przypadku rozbijania reguły przedstawionej w tabeli powyżej na dwie reguły należy wskazać, jak rozbijany jest term t_j. Należy podać specyfikację dwóch zbiorów (rozłącznych lub nie) $t_{1,j}$ oraz $t_{2,j}$ takich, że $t_j = t_{1,j} \cup t_{2,j}$. W tym przypadku powiemy, że rozbicie reguły jest indukowane specyfikacją $t_{1,j}$ oraz $t_{2,j}$.

Determinizm

Występuje wówczas, jeżeli dla żadnej sytuacji decyzyjnej nie jest możliwe równoczesne odpalenie przynajmniej dwóch reguł. Zawsze zostanie odpalona jednoznacznie wybrana reguła lub nie zostanie odpalona żadna z reguł.

Warunkiem determinizmu jest, aby prewarunki reguł opisywały sytuacje rozłączne. Zdarzy się tak, o ile przynajmniej dla jednego atrybutu nie będzie istniała wartość spełniająca warunek cząstkowy związany z tym atrybutem. Wymaganie to można sformułować w ten sposób, że dla pewnego atrybutu A_j musi być spełniony warunek $t_{1,j} \cap t_{2,j} = \varnothing$. Sprawdzenie tego warunku polega więc na stwierdzeniu, że iloczyn zbiorów opisywanych odpowiednimi termami jest zbiorem pustym.

Weryfikacja determinizmu może mieć również miejsce w przypadku tablicy zawierającej szablony danych. Jeżeli dla każdej pary rekordów spełniony jest warunek $t_{1,j} \cap t_{2,j} = \varnothing$ (dla dowolnego j), to poszczególne rekordy opisują zbiory rozłączne.

Zupełność

Baza reguł jest zupełna, jeżeli dla każdej sytuacji decyzyjnej istnieje możliwość odpalenia przynajmniej jednej reguły. W praktyce oznaczałoby to możliwość sklejenia tablicy prewarunków (bez kolumny H).

Istnieje możliwość budowy systemów zupełnych, jednak w praktyce najczęściej pewne sytuacje wejściowe nigdy nie wystąpią.

Niech Ψ będzie tablicą zawierającą szablony danych, które definiują kontekst zastosowania analizowanego tablicowego systemu regułowego reprezentowanego pewną tablicą reguł, taką że Φ jest tablicą prewarunków tych reguł. Logicznie, warunek specyficznej zupełności systemu można zapisać w postaci $\Psi|=\Phi$. Warunek ten będzie spełniony, jeżeli dla każdego wiersza ψ występującego w tablicy Ψ istnieje wiersz (lub sklejenie wierszy) $\varnothing$ występujący w tablicy Φ, taki, że zachodzi warunek pokrywania postaci $\psi \subseteq \varnothing$.

W praktyce może się okazać, że weryfikacja warunku subsumpcji nie jest w pełni możliwa, pomimo że system reguł jest zupełny. Zdarzy się tak w przypadku, gdy pewien rekord tablicy ψ jest tak ogólny, że nie jest pokry-

wany żadnym pojedynczym rekordem tablicy Φ. W takim przypadku należy poszukiwać rozbicia tego rekordu na postacie bardziej szczegółowe wykorzystujące wprost poszczególne elementy dziedzin atrybutów występujących w tabelach. Jednak rozbicie to będzie zazwyczaj zbyt szczegółowe i doprowadzi do powstania zbyt dużej liczby rekordów.

Poprawność

System tablicowy jest potencjalny poprawny, jeżeli jest zupełny w określonym kontekście oraz deterministyczny. Oznacza to, że system taki jest w stanie odpalić dokładnie jedną regułę w każdej sytuacji opisywanej specyfikacją kontekstu działania.

System potencjalnie poprawny nie zawiera reguł pokrywanych przez inne.

Spójność

System tablicowy jest spójny, jeżeli żadne dwie reguły o niepustym przecięciu prewarunków nie prowadzą do sprzecznych konkluzji. W schemacie tablicowej reprezentacji wiedzy nie występuje negacja logiczna, dlatego też dla określenia spójności systemu potrzeba zdefiniować, które wartości atrybutu decyzyjnego wykluczają się nawzajem.

Podsumowanie

W powyższym rozdziale przeprowadzono analizę elementów modelu relacyjnych baz danych. Pokazano równoważny model logiczny tabel o ustalonym schemacie atrybutów. Zwrócono uwagę na niektóre ograniczenia oraz trudności realizacji pewnych operacji w przyjętym modelu. Szczególnie uwzględniono analizę globalnych własności jakościowych. Praktyczna weryfikacja wskazanych własności może odbywać się na poziomie logicznym (w oparciu o mechanizmy takie jak unifikacja czy bd-rezolucja, na poziomie algebraicznym (w oparciu o rachunek podziałów czy metody algebraiczne oraz inne mechanizmy dopasowywania, porównywania i scalania wzorców.

Celem rozważań jest wypracowanie modelu oraz metod analizy i weryfikacji zbiorów danych i baz wiedzy pod kątem wymagań jakościowych. W praktyce inżynierskiej zastosowanie omawianego podejścia może obejmować analizę danych pochodzących z monitorowania procesu (rozpoznawanie sytuacji, diagnostyka, analiza cech jakościowych), analizę danych

ekonomicznych, finansowych oraz medycznych (np. pod kątem spełniania określonych ograniczeń), a także analizę danych pomiarowych, eksperymentalnych czy stanowiących np. specyfikację algorytmów sterowania i procesów decyzyjnych (pod kątem ich poprawności, zupełności, spójności *etc.*). Należy podkreślić, że praca dotyczy zwłaszcza jakościowej, logicznej analizy dużych zbiorów danych zawartych w bazach danych oraz wiedzy reprezentowanej za pomocą systemów regułowych. Celem było zarysowanie obszaru badawczego poprzez naszkicowanie definicji wspomnianych powyżej własności i problemów. W perspektywie celem badań jest wypracowanie elementów podejścia i narzędzi do wspomagania analizy danych oraz analizy i syntezy baz wiedzy spełniających określone wymagania. Będzie to realizowane poprzez opracowanie i przebadanie efektywności i przydatności rozszerzonych metod reprezentacji wiedzy i procedur dla opartego na logice testowania wybranych własności baz danych i baz wiedzy. Rozważany aparat może być też wykorzystany do wspomagania syntezy systemów z bazą wiedzy. Zastosowania praktyczne mogą obejmować dziedziny takie jak analiza baz danych przedsiębiorstw pod kątem spełniania określonych własności, monitorowanie i nadzór procesów, budowa systemów diagnostycznych, klasyfikacyjnych i systemów wspomagania decyzji, budowa systemów wyszukujących i wnioskujących w oparciu o analogię, temporalne bazy danych, systemy dokumentacji *etc*.

5. Trwała identyfikacja publikacji w repozytoriach cyfrowych – przegląd stosowanych systemów

Aneta Januszko-Szakiel

Wprowadzenie

Publikowanie w internecie oraz tworzenie repozytoriów, w których gromadzi się i archiwizuje cyfrowe kolekcje różnorodnych treści, stało się zjawiskiem powszechnym. Istotne atrybuty repozytoriów cyfrowych to przede wszystkim długoterminowa, niekiedy wieczysta archiwizacja oraz jednoznaczna identyfikacja i wyszukiwanie przechowywanych w nich obiektów[1]. Szczególnego znaczenia atrybuty te nabierają w przypadku repozytoriów bibliotecznych, archiwalnych, uczelnianych *etc.*, których obiekty stanowią narodowe dziedzictwo cyfrowe i służą jako zaplecze wiedzy w procesach edukacyjnych, pracach naukowych i badawczych. Ich dostępność oraz czytelność powinna być zagwarantowana pomimo wszelkich technologicznych i organizacyjnych zmian, m.in. poprzez jednoznaczne adresowanie i identyfikowanie. Jeżeli nie zostanie zagwarantowana zarówno dostępność, jak i czytelność publikacji sieciowych, wówczas użyteczność zasobów repozytoriów będzie znacznie ograniczona, np. poprzez brak możliwości cytowania i odwoływania się do treści tych dokumentów.

W przypadku publikacji tradycyjnych system identyfikowania jest powszechnie znany. Polega na przydzielaniu publikacjom znormalizowanych, jednoznacznych i niepowtarzalnych numerów ISBN, ISSN, ISAN bądź ISMN[2].

---

[1] W niniejszym opracowaniu przez pojęcie obiektu w repozytorium cyfrowym należy rozumieć pojedynczy dokument opublikowany w sieci. W zależności od typu repozytorium obiektem może być opublikowana w formie elektronicznej książka, artykuł, rozprawa doktorska, habilitacyjna, baza danych, prezentacja PowerPoint, nagranie wykładu, także inne formy prezentacji treści zapisane w postaci kodu zerojedynkowego. Wraz z terminem *obiekt* zamiennie występują pojęcia: materiał cyfrowy, zasób cyfrowy, dokument cyfrowy, publikacja cyfrowa, obiekt sieciowy.

[2] ISBN – International Standard Book Number (Międzynarodowy Znormalizowany Numer Książki), ISSN – International Standard Serial Number (Międzynarodowy Znormalizowany Numer Wydawnictwa Ciągłego), ISAN – International Standard Audiovisual Number (Międzynarodowy Znormalizowany Numer Utworów Audiowizualnych), ISMN – International Standard Music Number (Międzynarodowy Znormalizowany Numer Druku Muzycznego). Szczegółowe informacje o identyfikatorach dokumentów zamieszcza w swoim serwisie WWW Biblioteka Narodowa: http://www.bn.org.pl/index.php [dostęp: 12.01.2009].

Podobne systemy identyfikacyjne są stosowane dla obiektów sieciowych. W procesach bibliograficznych odesłań i wyszukiwania obiektów sieciowych posługiwanie się tylko obiegowymi adresami internetowymi URL (Uniform Resource Locators) jest niewystarczające, gdyż te zmieniają się zbyt często. Profesjonalne repozytoria cyfrowe, dbające o użyteczność[3] zdeponowanego materiału cyfrowego stosują rozmaite systemy trwałego identyfikowania obiektów.

W niniejszym rozdziale zdefiniowano pojęcie „repozytorium cyfrowe" oraz dokonano przeglądu powszechnie stosowanych systemów trwałej identyfikacji obiektów sieciowych.

5.1. Definicja repozytorium cyfrowego

Z przeglądu definicji dostępnych w piśmiennictwie przedmiotu[4] wynika, że przez pojęcie „repozytorium elektroniczne" tudzież „repozytorium cyfrowe" należy rozumieć organizację ludzi oraz narzędzi lub system złożony z osób oraz przyjętych rozwiązań organizacyjnych i technicznych, powołany w celu zgromadzenia, przechowania oraz zapewnienia długoterminowego dostępu i użyteczności cyfrowego materiału. Działania repozytorium koncentrują się na pracach związanych z przeprowadzeniem cyfrowych dokumentów przez kolejne etapy rozwoju technologicznego, przy użyciu najróżniejszych narzędzi i metod archiwizacji, między innymi migracji oraz emulacji[5]. Docelowo repozytorium ma dostarczyć obecnym oraz przyszłym

[3] Przez pojęcie użyteczności cyfrowych zasobów archiwalnych należy rozumieć m.in. stabilny dostęp do autentycznych i integralnych dokumentów cyfrowych oraz możliwość powoływania się na nie we własnych opracowaniach poprzez stosowanie bibliograficznych odesłań. Źródło: *Attributes of a Trusted Digital Repository: Meeting the Needs of Research Resources. RLG-OCLC Report*, Mountain View, CA, August 2001, [online:] http://www.rlg.org/longterm/attributes01.pdf [dostęp: 20.12.2008], A. Januszko-Szakiel, *Archiwizacja publikacji elektronicznych jako wyzwanie dla bibliotek – zarys problematyki*, „Biuletyn Biblioteki Jagiellońskiej" 2003, s. 216–225.

[4] *Attributes of a Trusted Digital Repository: Meeting the Needs of Research Resources. RLG-OCLC Report*. Mountain View, CA. August 2001, [online:] http://www.rlg.org/longterm/attributes01.pdf [dostęp 20.12.2008]; G. Clavel-Merrin, *The Nedlib List of Terms. Nedlib Report Series 7*, Amsterdam 2000, s. 3; *Kriterienkatalog vertrauenswürdige digitale Langzeitarchive. Version 1. (Entwurf zur öffentlichen Kommentierung). Nestor Materialien 8*. Frankfurt am Main, 2006, [online:] http://edoc.hu-berlin.de/series/nestor-materialien/2006-8/PDF/8.pdf, s. 2 [dostęp: 20.08.2008]; J.M. Reitz, *Dictionary for Library and Information Science*, Westport–London 2004, s. 216.

[5] U.M. Borghoff i in., *Langzeitarchivierung. Methoden zur Erhaltung digitaler Dokumente*, Heidelberg 2003, s. 47–78; A. Januszko-Szakiel, *Rola migracji i emulacji w strategii długoter-*

użytkownikom możliwość odczytu autentycznych, integralnych, wiarygodnych i poufnych dokumentów cyfrowych[6].

W wypowiedziach na temat repozytoriów cyfrowych autorzy często odwołują się do standardu archiwizacji publikacji elektronicznych OAIS, w którym oprócz wymienionych cech uwzględnia się dążenie repozytorium cyfrowego do stałej obserwacji i zabezpieczenia zmieniających się potrzeb docelowej grupy użytkowników, nazywanych niekiedy klientami, tudzież odbiorcami usług repozytorium. Synonimicznie „repozytorium cyfrowe" określane bywa terminem „archiwum cyfrowe"[7]. W dalszej części tekstu terminy „repozytorium" oraz „archiwum cyfrowe" bądź „archiwum elektroniczne" będą stosowane wymiennie.

Model referencyjny repozytoriów cyfrowych OAIS został stworzony przez *Consultative Committee for Space Data Systems* (CCSDS)[8] na potrzeby archiwizacji i wymiany danych elektronicznych, zawierających informacje z badań przestrzeni kosmicznej. W maju 1999 r. zaprezentowana została pierwsza wersja modelu OAIS, a w lutym 2003 r., po licznych poprawkach model OAIS został zaakceptowany przez International Organization for Standarization jako norma postępowania w zakresie długoterminowej archiwizacji danych cyfrowych (ISO 14721:2003).

Pomimo że model OAIS został stworzony głównie z myślą o archiwizacji jednego typu danych elektronicznych, jest on uznawany za uniwersalny model organizowania i funkcjonowania repozytoriów cyfrowych i stosowany do gromadzenia, przechowywania i udostępniania różnych typów dokumentów elektronicznych. OAIS jest wykorzystywany w wielu światowych bibliotekach, archiwach i muzeach, w których realizowane są projekty długoterminowej archiwizacji zbiorów cyfrowych.

Jednym z kluczowych pojęć w modelu referencyjnym OAIS jest pakiet informacyjny – *Information Package*. Składa się on z dwóch komponentów, tj.

---

minowej archiwizacji publikacji elektronicznych, [w:] *Informatyka*, red. M. Pękala, W.Z. Chmielowski, Kraków 2008, s. 121–130.

[6] *Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Nestor Handbuch*, 2008, [online:] http://nestor.sub.unigoettingen.de/handbuch/nestor-handbuch.pdf [dostęp: 20.08.2008]; A. Januszko-Szakiel, *Archiwizacja publikacji elektronicznych jako wyzwanie dla bibliotek – zarys problematyki.* „Biuletyn Biblioteki Jagiellońskiej" 2003, s. 215–220.

[7] J.M. Reitz, *Dictionary for Library and Information Science*, Westport–London 2004, s. 216; *Trusted Digital Repositories: Attributes and Responsibilities. An RLG-OCLC Report*, 2002, [online:] http://www.rlg.org/longterm/repositories.pdf [dostęp: 20.12.2008].

[8] Komitet CCSDS został powołany w 1982 r. Jest organizacją składającą się z przedstawicieli wielu światowych agencji badań przestrzeni kosmicznej i podlega bezpośrednio agencji NASA; szczegółowe informacje zob. http://www.ccsds.org/ [dostęp: 06.01.2009].

kontenera informacyjnego (*Content Information*) oraz informacji dotyczącej przechowywania jego zawartości (*Preservation Description Information* – PDI). PDI to w myśl modelu OAIS wszelkie informacje konieczne do odpowiedniego przechowania informacji treściowej (kontenera informacyjnego). Zalicza się tu cztery typy informacji, określane jako: historia (*Provenance*), powiązania (*Context*), identyfikatory (*Reference*) oraz mechanizmy ochrony danych – *Fixity*.

- *Provenance*, w dosłownym tłumaczeniu „pochodzenie", określa źródło obiektu informacyjnego, wskazuje na podmiot odpowiedzialny za opiekę nad obiektem od momentu jego powstania oraz dostarcza wiedzy na temat historii obiektu.
- *Context* opisuje związek obiektu informacyjnego z innymi obiektami nienależącymi do danego pakietu informacyjnego.
- *Reference* jest odpowiedzialny za dostarczenie identyfikatorów, umożliwiających jednoznaczną identyfikację obiektu informacyjnego. Najogólniej rzecz ujmując, zadaniem identyfikatorów publikacji elektronicznych jest odróżnienie określonej publikacji od innych. W archiwach elektronicznych identyfikatory występują pod nazwą *Digital Object Identifier* (DOI) czy też *Persistent Identifier* (PI).
- *Fixity* to element wprowadzający mechanizmy ochronne, mające na celu zabezpieczenie autentyczności i integralności obiektów informacyjnych przed jakimikolwiek nieudokumentowanymi zmianami.

PDI jest więc zarówno pewnego rodzaju informatorem o pochodzeniu i historii obiektu informacyjnego, jego przynależności oraz powiązaniach z innymi obiektami w archiwum, jak i mechanizmem chroniącym jego integralność i autentyczność.

W celu powiązania obu komponentów pakietu informacyjnego model referencyjny OAIS przewiduje także element w postaci informacji o pakiecie (*Packaging Information*). Jego zadaniem jest identyfikacja poszczególnych składników pakietu informacyjnego.

Elementem niezbędnym w archiwum elektronicznym są wreszcie metadane przechowywanych obiektów (*Information Packages*). W modelu referencyjnym OAIS określane są one terminem *Descriptive Information*. Metadane dostarczają informacji o zawartości pakietu informacyjnego oraz umożliwiają jego odnalezienie w archiwum.

Pakiet informacyjny wraz ze wszystkimi jego elementami składowymi należy traktować jako obiekt archiwizacji w archiwum elektronicznym OAIS.

5.2. Identyfikacja obiektów sieciowych

W celu dotarcia do dokumentów opublikowanych w internecie najczęściej wykorzystuje się adresy URL (Uniform Resource Locators), które umożliwiają wyszukanie dokumentu oraz służą jako identyfikator w procesach cytowania i bibliograficznych odesłań do publikacji internetowych. Mogą być również stosowane w bazach danych, katalogach, indeksach, rejestrach i wszelkich innych typach bibliograficznych wykazów, odsyłających do pełnych tekstów dokumentów internetowych bądź ich metadanych. Jednak zmiana miejsca dokumentu sieciowego powoduje, że zastosowane odesłanie w postaci URL jest nieużyteczne, a więc obiekt cyfrowy przestaje spełniać podstawowe kryterium dostępności.

Należy więc zauważyć, że powszechnie stosowany URL nie powinien być określany mianem identyfikatora, lecz raczej „lokalizatora" obiektu sieciowego, ponieważ wskazuje jedynie lokalizację obiektu, a nie identyfikuje jednoznacznie samego obiektu.

Połowicznym rozwiązaniem jest stosowanie metod zapewniających tak zwaną stabilność okresową obiektów cyfrowych. Do metod tych zalicza się:

- zastosowanie systemu adresowania URL, w którym serwer dynamicznie ustala miejsce zapisu obiektu sieciowego, korzystając z odpowiednich skryptów oraz baz danych zawierających bieżącą lokalizację dokumentów,
- zastosowanie odpowiedniej konfiguracji serwera Web, która umożliwi przekierowanie z nieaktualnego do nowego adresu w formie tzw. *redirects* lub *aliases*,
- przeprowadzanie okresowej kontroli dostępności adresów i powiązanych z nimi obiektów (tzw. URL-Checks) przez administratora i wykonanie uaktualnienia odwołań do dokumentów.

Powyższa metodologia stanowi jednak rozwiązanie krótko- lub średniookresowe. Dzieje się tak z wielu powodów, głównie z racji bardzo prawdopodobnych zmian w metodologii adresowania, wynikających na przykład z technicznych modyfikacji otoczenia systemowego. Za sensowną uznaje się okresową kontrolę URL, jednak tylko przy założeniu tzw. „konsekwentnej pielęgnacji", która oznacza, że w przypadku stwierdzenia, że hiperłącze nie odsyła do pożądanego obiektu, należy ustalić źródło błędu, odszukać właściwy adres do obiektu i nanieść stosowne zmiany we wszelkich wykazach, katalogach, bibliografiach, portalach *etc.*, które do danego obiektu odsyłają. Są to zabiegi pracochłonne. Okresową niedostępność adresów URL mogą też

powodować błędy sieciowe lub niestabilne połączenia z serwerem. Wreszcie dokumenty sieciowe ulegają zmianom w wyniku procesów zachodzących w instytucjach, w których są zlokalizowane i ich identyfikacja oraz adresowanie za pomocą samego URL mogą okazać się zawodne.

W związku z powyższym zachodzi potrzeba zastosowania trwałego mechanizmu archiwizacji obiektów cyfrowych. Zaproponowane rozwiązanie to identyfikatory trwałe (ang. *persistent identifiers*) (PI).

5.3. Systemy trwałej identyfikacji obiektów sieciowych

Identyfikator trwały (PI) to niezmienna (określana też jako stabilna, unikatowa, permanentna) nazwa, którą przyporządkowuje się do obiektu sieciowego jeden raz na cały cykl jego „życia". Zadaniem PI jest jednoznaczna i trwała identyfikacja obiektu sieciowego oraz przynależnych do niego metadanych, niezależnie od miejsca (instytucji), w którym obiekt został zapisany i jest archiwizowany, z uwzględnieniem różnorodnych systemów, ich ograniczeń (granic), zmian oraz w obliczu występowania obiektów cyfrowych w różnych wersjach, postaciach, formach reprezentacji. Na podstawie PI, system obsługi PI powinien umożliwić zlokalizowanie dokumentu i jego odczyt. Obecnie wykorzystywane są głównie trzy systemy PI, tj. PURL, Handle System i URN. Bez względu na wybór zastosowanego systemu ważne jest, aby identyfikatory pozostawały niezmienne. Istotne jest także, aby dany system obsługi PI miał podbudowę instytucjonalną.

5.3.1. PURL – Persistent URL

System Persistent Uniform Resource Locator (PURL) jest rozwinięciem koncepcji URL i funkcjonalnie jest z nim tożsamy. System ten wykorzystuje adresy URL, które zamiast wskazywać na określony obiekt, wskazują na usługę przekierowującą do danego obiektu. Tak więc PURL składa się z adresu serwera usługi przekierowującej oraz identyfikatora obiektu, do którego chcemy uzyskać dostęp. Adresy PURL stosuje się wówczas, gdy przewiduje się częste zmiany położenia poszczególnych obiektów WWW. Pełnią one rolę oficjalnych adresów, pod którymi można znaleźć żądane zasoby, a odpowiednimi przekierowaniami zajmuje się serwer[9]. Baza danych serwera usługi

[9] A. Freedman, *Encyklopedia komputerów*, Gliwice 2004, s. 667.

przekierowań zawiera wszystkie identyfikatory zarejestrowane w danym systemie wraz z przypisanymi im aktualnymi lokalizacjami dokumentu. Można więc powiedzieć, że w systemie tym odróżnia się „identyfikatory" od „lokalizatorów", czyli adresów lokalizacji, w których przechowywane są kopie danego obiektu. W przypadku gdy mamy do czynienia z obiektem sieciowym, lokalizatory mają postać aktualnych URL poszczególnych kopii obiektu.

System PURL został wprowadzony przez Online Computer Library Center (OCLC) w 1995 roku, w ramach inicjatywy „Internet Cataloging Projects", której celem było poprawienie (dookreślenie, uściślenie) adresów internetowych zasobów, wykazywanych w katalogach bibliotecznych.

Składnia adresu PURL wygląda następująco: <Protocol><RA><Name>
Przy czym:

- Protocol to standardowy protokół, np.: http,
- RA to adres serwera usługi przekierowującej do wybranego obiektu,
- Name to nazwa wskazująca na określony obiekt.

Przykład: http://purl.oclc.org/keith/home,
gdzie:

- http – protokół,

purl.oclc.org – adres serwera przekierowującego,
/keith/home – nazwa zasobu.

System ten znalazł zastosowanie m.in. w Bibliotece Kongresu oraz United States Government Printing Office (GPO), eksperymentalnie również w OCLC. Aktualnie system PURL nie jest już rozwijany, natomiast zasady jego działania wykorzystano przy opracowywaniu bardziej kompleksowych systemów, takich jak Handle System i URN.

Najszerzej wykorzystywaną implementacją założeń systemu PURL jest Archival Resource Key – ARK[10]. Stanowi on schemat identyfikacyjny służący do trwałej dostępności cyfrowych obiektów. Identyfikator ARK jest stosowany jako link:

- odsyłający od obiektu cyfrowego do organizacji, do której obiekt przynależy,
- łączący obiekt cyfrowy z jego metadanymi,
- odsyłający do treści obiektu bądź jego kopii.

System ARK znalazł zastosowanie w 15 repozytoriach, m.in. w California Digital Library, Library of Congres, National Library of France.

[10] *Archival Resource Key*, [online:] http://www.cdlib.org/inside/diglib/ark/ [dostęp: 20.12.2008].

Trwałość w tym systemie identyfikacyjnym jest zapewniana przez usługodawcę, a nie składnię nazwy. ARK wskazuje metadane o obiekcie, nie daje gwarancji trwałości identyfikatora, zezwala na integrację innych schematów, a także jego zintegrowanie z innymi schematami.

Składnia ARK jest następująca: http://<NMAH>/ark:/<NAAN>/<Name>

Przy czym:

- NMAH to adres serwera usługi przekierowującej,
- NAAN to identyfikator instytucji nadającej poszczególnym obiektom identyfikatory we własnej przestrzeni nazw,
- Name to nazwa (identyfikator) przydzielona do danego zasobu.

Przykład: http://bnf.fr/ark:/13030/tf5p30086k

5.3.2. Handle-System

Handle-System[11] jest systemem identyfikatorów przypisywanych obiektom cyfrowym niezależnie od ich fizycznego umiejscowienia. Założenia systemu zostały opracowane przez Corporation for National Research Initiatives CNRI[12] i opisane w dokumencie RFC 3650[13].

W dokumencie tym autorzy zdefiniowali m.in. zasadę budowy identyfikatorów, na które składa się prefiks oraz sufiks. Prefiks jest numerycznym kodem, oznaczającym instytucję, która została zarejestrowana w Global Handle Service (instytucji nadzorującej system) jako upoważniona do nadawania obiektom identyfikatorów we własnej przestrzeni nazw. Sufiks identyfikatora jest nazwą (identyfikatorem) danego obiektu, unikatową w przestrzeni nazw danej instytucji i może składać się z dowolnej liczby znaków zgodnych z systemem ASCII.

Składnia Handle-System wygląda następująco: Handle: <HNA> / <HLN>, przy czym:

- HNA – prefiks instytucji nadawany przez Global Handle Service,
- HNL – identyfikator obiektu w przestrzeni nazw danej instytucji.

---

[11] *The Handle System*, [online:] http://www.handle.net/ [dostęp: 20.12.2008].

[12] CNRI – to amerykańska organizacja non profit, założona w 1986 roku, której głównym celem jest wspieranie rozwoju kluczowych technologii przetwarzania i udostępniania wiedzy z użyciem sieci komputerowych. Źródło: Corporation for National Research Initiatives: http://www.cnri.reston.va.us/about_cnri.html [dostęp: 20.12.2009].

[13] S. Sun, L. Lannom, B. Boesch, *Handle System Overview. Request for Comments*: *3650*, CNRI, November 2003, [online:] http://www.ietf.org/rfc/rfc3650.txt [dostęp: 20.12.2009].

Przy rejestracji dany obiekt otrzymuje identyfikator, do którego przypisane są informacje uzupełniające. Handle-System nie narzuca sztywnej struktury metadanych powiązanych z obiektem, więc zarówno rodzaj, jak i zakres tych informacji determinowany jest przez instytucję rejestrującą oraz typ obiektu cyfrowego. Wśród informacji o obiekcie najczęściej znajdują się dane właściciela (autora), opis dokumentu (tytuł, słowa kluczowe) oraz co najmniej jeden wpis pozwalający na dostęp do kopii danego obiektu.

Identyfikatory wraz z powiązanymi metadanymi przechowywane są w centralnej, ogólnodostępnej bazie danych, umożliwiającej szybkie uzyskanie podstawowych informacji na temat określonych obiektów poprzez usługi dostępne w sieciach komputerowych. Funkcje systemu umożliwiają jednostkom rejestrującym dystrybucję, administrację oraz rozwiązywanie (likwidację) identyfikatorów.

Z Handle-System korzysta obecnie wiele instytucji i firm. Przykładem zastosowania Handle-System są m.in. CORDA/ADL i DVIA, czyli systemy Departamentu Obrony Stanów Zjednoczonych, rejestrujące i zarządzające dokumentami związanymi z obronnością Stanów Zjednoczonych. Handle-System jest również użyteczny w projekcie DSpace realizowanym przez MIT, w którego ramach tworzona jest baza danych na temat materiałów edukacyjnych powstających we wszystkich wydziałach i jednostkach tej instytucji. Jeszcze innym projektem stosującym opisywany system jest The National Digital Library. Program, którego założeniem jest digitalizacja i utworzenie bazy danych dzieł zgromadzonych w bibliotekach publicznych i uczelnianych w Stanach Zjednoczonych[14].

Struktura identyfikatorów Handle-System pozwala także na rejestrację tzw. rejestratorów lokalnych, wówczas zarejestrowana instytucja ma możliwość rejestracji instytucji sobie podległych, które dysponują własną przestrzenią nazw dla obiektów cyfrowych. W takim przypadku prefiks identyfikatora składa się z dwóch numerycznych członów oddzielonych kropką (np. 10.1000), przy czym pierwszy człon określa instytucję nadrzędną zarejestrowaną przez Global Handle Service, natomiast drugi jest identyfikatorem lokalnego rejestratora. Drzewiasta struktura Handle-System pozwoliła na powstanie podsystemów identyfikacyjnych, z których najpopularniejszym jest Digital Object Identifier – DOI.

DOI to identyfikator dokumentu elektronicznego, który jest do niego na stałe przypisany i w odróżnieniu od identyfikatora URL nie zależy od fizycz-

[14] Na podstawie informacji dostępnych na stronach Wolnej Encyklopedii – Wikipedia: http://pl.wikipedia.org/wiki/Handle_System [dostęp: 10.01.2009].

nej lokalizacji dokumentu. Zgodnie z definicją proponowaną w *Encyklopedii komputerów*[15] Digital Object Identifier to rozwiązanie pozwalające na przydzielanie dokumentom, publikacjom i wszelkim innym zasobom, dostępnym w internecie, stałych, niezmiennych nazw zamiast adresów URL.

Podstawowym założeniem systemu DOI jest identyfikacja oraz wymiana obiektów cyfrowych. Trwają również prace nad organizacyjnymi oraz technicznymi rozwiązaniami, umożliwiającymi zarządzanie obiektami cyfrowymi oraz powiązanie producentów i dostawców obiektów z użytkownikami[16].

Zarządzaniem systemu zajmuje się Międzynarodowa Fundacja DOI (*International DOI Foundation IDF*), która jest organizacją non profit, finansującą się ze składek członkowskich oraz sprzedaży prefiksów i numerów DOI. Fundacja DOI sprawuje kontrolę nad instytucjami i firmami, które uzyskały prawo do pełnienia roli *DOI Registration Agency* (RA). Podstawowym zadaniem RA jest przydzielanie identyfikatorów wydawcom (Publisher ID) i zapewnienie im infrastruktury umożliwiającej tworzenie identyfikatorów obiektów (Item ID) oraz zarządzanie metadanymi przypisanymi identyfikatorom DOI. Od agencji RA oczekuje się promocji systemu DOI oraz współpracy na rzecz jego rozwoju.

Struktura DOI stanowi od roku 2001 standard ANSI/NISO (Z39.84), a jej komponenty są implementacją założeń Handle-System. System DOI składa się z następujących komponentów: metadane, DOI jako identyfikator trwały (PI) oraz techniczna implementacja Handle-System. Identyfikatory DOI zgodnie z założeniami Handle-System są ciągami znaków ASCII. Składają się przedrostka i końcówki.

Przykład: 10.1000/182,

przy czym:

- 10.1000 to przedrostek, w którym znaki 10 informują, że chodzi o identyfikator DOI,
- 1000 to numer przypisany przez IDF wydawcy (*Publisher ID*),
- natomiast sufiks 182 to końcówka, która jest przypisana do określonego dzieła (*Item ID*).

Publisher ID jest przypisywany wydawcom, którzy zdecydowali się zarejestrować i korzystać z systemu DOI przez agencję, która ma do tego prawo. Item ID jest nadawany przez samego wydawcę, który powinien zagwarantować, że ID będzie unikalne dla każdej wydanej przez niego publikacji. Item ID może, ale nie musi być, numerem katalogowym publikacji pochodzącym

[15] A. Freedman, *Encyklopedia komputerów…*, s. 143.

[16] *The DOI System*, [online:] http://www.doi.org/ [dostęp: 19.12.2008].

z innych systemów rejestrowania, np. ISBN, ISSN. Poprawny sposób podawania odnośników do źródeł wygląda następująco: doi: 10.1000/182.

System DOI jest stosowany m.in. w agencjach praw autorskich, wydawnictwach i bibliotekach. Typowym przykładem zastosowania DOI jest identyfikowanie elektronicznych wersji publikacji naukowych w repozytorium SpringerLink, przy czym identyfikator DOI może otrzymać artykuł, całe czasopismo naukowe, rozdział w książce, plik multimedialny, program komputerowy *etc*.

5.3.3. URN – Uniform Resource Name

Historia systemu URN rozpoczęła się w 1990 roku i ma związek z projektowaniem architektury World Wide Web (WWW). URN został wprowadzony jako ujednolicona forma oznaczania zasobów internetowych. Formy i kierunki rozwoju sieci internet są kontrolowane przez organizację Internet Assigned Numbers Authority (IANA). To właśnie IANA oraz ściśle związana z nią grupa robocza o nazwie Internet Engineering Task Force (IETF) stanowią siłę napędową w rozwoju internetu i *de facto* dyktują standardy, których najbardziej znaną postacią są publikacje pod tytułem Requests for Comments (RFCs). W dokumencie RFC 1737[17] z 1994 roku dość precyzyjnie określono wymagania dotyczące schematu URN, natomiast trzy lata później, w publikacji RFC 2141[18] z 1997 roku zostały wymienione cele rozwoju identyfikatorów trwałych PI.

System URN został świadomie pomyślany jako schemat otwarty, zdolny do integracji z systemami istniejącymi, na przykład z identyfikatorami ISBN albo URL. Od ponad 10 lat URN[19] funkcjonuje jako standard adresowania obiektów w instytucjach objętych obowiązkiem takiego identyfikowania zasobów, aby były one dostępne długotrwale oraz niezależnie od tego, w której instytucji są przechowywane.

System URN cieszy się dużą popularnością. Jest stosowany m.in. w narodowych bibliotekach takich krajów jak Finlandia, Holandia, Austria, Szwajcaria i Wielka Brytania. Istnieje także możliwość integracji identyfikatorów

[17] K. Sollins, L. Masinter, *Functional Requirements for Uniform Resource Names*, [online:] http://www.ietf.org/rfc/rfc1737.txt [dostęp: 20.12.2008].

[18] R. Moats, *URN Syntax. Request for Comments*: 2141, AT&T, May 1997, [online:] http://www.ietf.org/rfc/rfc2141.txt [dostęp: 20.12.2008].

[19] *Uniform Resource Names. A Progress Report*, „D-Lib Magazine", February 1996, [online:] http://www.dlib.org/dlib/february96/02arms.html [dostęp: 20.12.2008].

URN wraz z istniejącymi numerycznymi systemami identyfikacyjnymi dokumentów, np. ISAN, ISSN, ISBN.

Identyfikatory URN składają się z kilku hierarchicznie ułożonych elementów, tj. z *Namespace IDentifier NID* (tzw. identyfikatora przestrzeni nazw) oraz z podporządkowanych mu subelementów (*SNID, NSS*).

Składnia identyfikatora wygląda następująco: urn: <NID> [: SNID] : <NSS> przy czym:

- NID – identyfikator przestrzeni nazw,
- SNID – identyfikator podprzestrzeni nazw (jeśli występuje),
- NSS – unikalny dla danej podprzestrzeni identyfikator zasobu (łańcuch znaków).

Jedną z podprzestrzeni nazw systemu URN jest system NBN – National Bibliographic Number. Został on opracowany w celu wyszczególnienia w bibliografiach narodowych publikacji cyfrowych, na przykład czasopism elektronicznych, rozpraw doktorskich i habilitacyjnych, także innych publikacji, stanowiących narodowe dziedzictwo cyfrowe i podlegających obowiązkowi wieczystej archiwizacji. Koncepcja systemu NBN zrodziła się w ramach popularnych inicjatyw bibliotek narodowych, *Conference of Directors of National Libraries (CDNL)* oraz *Conference of European National Librarians (CENL)*.

NBN jest implementacją założeń systemu URN, w związku z czym składnia jego identyfikatorów wygląda następująco: urn: NBM : <ICC> [:SNS] NBNstring, przy czym:

- ICC to dwuliterowy kod kraju według ISO 3166,
- SNS to podprzestrzeń nazw,
- NBNstring to identyfikator w podanej przestrzeni nazw,

Przykład: urn:NBN:de:kobv:23-2312.

System NBN jest ogólnoświatowym systemem używanym wyłącznie w Bibliotekach Narodowych i wykorzystywanym do jednoznacznej, trwałej identyfikacji zarówno dokumentów cyfrowych, jak i fizycznych. Biblioteki Narodowe przyjmują na siebie obowiązek zarządzania przestrzeniami nazw w obrębie danego kraju.

Podsumowanie

Składowanie i archiwizacja zasobów nauki i kultury w sieci ma sens wówczas, gdy zasoby te w każdej chwili, obecnie i w najbardziej odległej przyszłości, mogą być udostępniane i użytkowane. Instytucje tworzące repozytoria

cyfrowych zasobów decydują się na rozmaite systemy ich trwałego identyfikowania. Zaleca się, aby w procesie decyzyjnym, dotyczącym wyboru systemu identyfikacji instytucje uwzględniły następujące kryteria:

- **Standaryzacja**. Instytucje powinny skłaniać się do stosowania systemów, które zostały zaakceptowane jako standard, najlepiej o światowym zasięgu.
- **Wymagania funkcjonalne**. Wybierane systemy identyfikacyjne powinny charakteryzować się trwałością, jednoznacznością, światowym zasięgiem, niezależnością od miejsca składowania. Identyfikatory trwałe powinny odsyłać równocześnie do wielu kopii jednego obiektu.
- **Elastyczność, skalowalność**. Stosowane systemy powinny być skalowalne oraz zdolne do rozszerzenia o nowe funkcje, bez zaburzenia ich zgodności z przyjętym standardem.
- **Niezależność technologiczna i kompatybilność**. Systemy identyfikacyjne powinny być generyczne, niezależne od protokołów i technologii, a także kompatybilne z funkcjonującymi instalacjami i usługami.
- **Instalacje, polecenia (rekomendacje)**. Przy wyborze systemu należy uwzględnić jego akceptację i popularność w skali międzynarodowej.
- **Koszty oraz trwałość**. Kryterium wyboru systemu powinny być koszty systemu (zarówno wstępne, jak i dalszego utrzymania) oraz jego niezawodność.

Opisane w tym rozdziale systemy trwałej identyfikacji obiektów sieciowych w zasadzie spełniają wszystkie z wymienionych kryteriów i są najczęściej implementowane w profesjonalnych repozytoriach. Należy jednak zaznaczyć, że obok nich istnieje także szereg innych, mniej popularnych rozwiązań: ERRoL – Extensible Repository Resources Lokator, GRI – Grid Resource Identifier, GUUID/UUID – Globally Unique Identifier/Universal Unique Identifier, InfoURI, NLA – National Library of Australian, LSID – Life Science Identifier, POI – PURL-Based Object Identifier, XRI – Extensible Resource Identifier.

6. Zabezpieczenie wiarygodności zasobów cyfrowych deponowanych w repozytoriach instytucjonalnych

Aneta Januszko-Szakiel

Wprowadzenie

Instytucje sektora biznesu, administracji, a także nauki i kultury, w bieżącej działalności wytwarzają i operują rozmaitymi dokumentami w cyfrowej postaci. Dokumenty te tworzą cyfrowe kolekcje i są deponowane w systemach repozytoryjnych, gdzie podlegają krótko- bądź długoterminowej archiwizacji.

System repozytoryjny, określany również jako depozytowy bądź archiwalny, jest definiowany jako organizacja ludzi, narzędzi oraz przyjętych rozwiązań organizacyjnych, technicznych i prawnych, powołany w celu zgromadzenia, bezpiecznego przechowania oraz zapewnienia dostępu i użyteczności zgromadzonych zasobów w określonym czasie[1]. Repozytoria biznesowe, w zależności od rodzaju przechowywanych dokumentów, realizowanych zadań oraz procedur prawnych mają zapewnić dostępność i użyteczność zdeponowanych materiałów w okresie od trzech do pięćdziesięciu lat[2], natomiast systemy repozytoryjne instytucji pamięci powinny gwarantować utrzymanie użyteczności zasobów przez okres stu i więcej lat[3]; w niektórych przypadkach wieczyście[4]. Docelowo system repozytoryj-

[1] Reference Model for an Open Archival Information System (OAIS). Recommendation for Space Data Systems. CCSDS 650.0-B-1. Blue Book, Iss. 1. January 2002. Consultative Committee for Space Data System, Washington D.C., [online:] http://public.ccsds.org/publications/archive/650x0b1.pdf, [dostęp: 10.11.2012]; A. Januszko-Szakiel, Open Archival Information System – standard w zakresie archiwizacji publikacji elektronicznych, „Przegląd Biblioteczny" 2005, nr (73)3, s. 342.

[2] W. Sasin, *Przechowywanie i archiwizowanie dokumentacji przedsiębiorstwa według nowych zasad normatywnych. Poradnik dla wszystkich firm z instrukcją wzorcową. Stan prawny na dzień 1 lutego 2004 r.*, Wydawnictwo „Sigma", Skierniewice 2004; M. Konstankiewicz, *Wykaz ważniejszych resortowych aktów prawnych regulujących zasady postępowania z dokumentacją. Uzupełnienie i kontynuacja (część XVIII)*, „Archiwista Polski" 2005, nr 3(39), s. 49–62; *idem*, *Wykaz ważniejszych aktów prawnych regulujących zasady postępowania z dokumentacją. Uzupełnienie i kontynuacja (część XVIII)*, „Archiwista Polski" 2006, nr 2(42), s. 53–60.

[3] U. M. Borghoff, *Vergleich bestehender Archivierungssysteme*, „Nestor Materialien" 2005, nr 3, [online:] http://files.d-nb.de/nestor/materialien/nestor_mat_03.pdf, [dostęp: 21.11.2012].

[4] Ustawa z dnia 27 czerwca 1997 r. o bibliotekach, Dz.U. z 1997 r., Nr 85, poz. 539.

ny ma zapewnić obecnym i przyszłym użytkownikom możliwość dostępu do kompletnej kolekcji użytecznych dokumentów, w ramach posiadanych przez nich praw dostępu[5]. Użyteczność dokumentów cyfrowych wiąże się z efektywnym korzystaniem z zapisanych w nich treści (informacji). Jest to możliwe, wówczas gdy użytkownik ma pewność, że treści, które czyta, słucha, ogląda, a także, na które powołuje się, są autentyczne i niezafałszowane, to znaczy, że pochodzą od ich autorów i od dnia opublikowania nie uległy zmianie; przedstawiają dokładnie to, co było zamierzeniem ich twórców[6]. Systemy repozytoryjne mają zatem za zadanie zabezpieczać i zagwarantować wiarygodność zdeponowanych zasobów cyfrowych.

Celem bezpiecznego deponowania zasobów cyfrowych jest zapewnienie, że dokumenty i zapisane w nich treści zostały zachowane w niezmienionej postaci. Proces ten polega na zachowaniu niezmienionej substancji dokumentu cyfrowego w postaci kodu zerojedynkowego oraz na zapewnieniu platformy programowo-sprzętowej, która będzie w stanie zdekodować dane cyfrowe i przedstawić je w postaci czytelnej dla użytkownika.

Organizatorzy bezpiecznych i trwałych systemów repozytoryjnych odwołują się do standardu archiwizacji zasobów cyfrowych OAIS[7].

6.1. OAIS – Open Archival Information System

OAIS to referencyjny model organizacji i przebiegu procesu trwałej ochrony obiektów cyfrowych, stworzony przez *Consultative Committee for Space Data Systems* (CCSDS)[8] na potrzeby archiwizacji i wymiany danych cyfrowych, zawierających informacje z badań przestrzeni kosmicznej agencji NASA. W lutym 2003 roku model OAIS został zaakceptowany przez International Organization for Standardization ISO jako norma postępo-

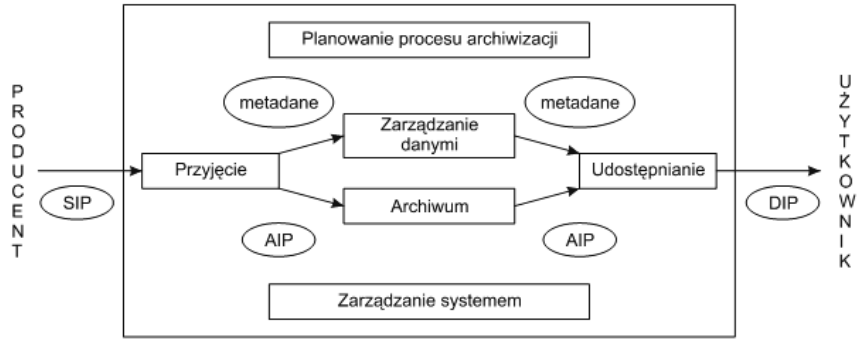
[5] T. Bilski, *Pamięć. Nośniki i systemy przechowywania danych*, Wydawnictwa Naukowo-Techniczne, Warszawa 2008, s. 423–425; *Kriterienkatalog vertrauenswürdige digitale Langzeitarchive. Version 1: Entwurf zur öffentlichen Kommentierung*, „Nestor – Materialien" 2006, nr 8, [online:] http://edoc.hu-berlin.de/series/nestormaterialien/2006-8/PDF/8.pdf [dostęp: 16.11.2012].

[6] W. Coy, *Perspektiven der Langzeitarchivierung multimedialer Objekte*, „Nestor Materialien" 2006, nr 5, [online:] http://files.dnb.de/nestor/materialien/nestor_mat_05.pdf, [dostęp: 10.11.2012].

[7] *Reference Model…*, *op. cit.*

[8] Komitet CCSDS został powołany w 1982 roku. Jest organizacją składającą się z przedstawicieli wielu światowych agencji badań przestrzeni kosmicznej i podlega bezpośrednio agencji NASA. *The Consultative Committee for Space Data System*, Reston VA, USA. CCSDS/AIAA, 2010, [online:] http://www.ccsds.org/, [dostęp: 16.11.2012].

wania w zakresie długoterminowej archiwizacji danych cyfrowych (ISO: nr 14721:2003). Obecnie jest uznawany za uniwersalny model organizowania i funkcjonowania repozytoriów cyfrowych i stosowany do gromadzenia, trwałego i bezpiecznego archiwizowania oraz udostępniania różnych typów dokumentów cyfrowych[9].



Rysunek 9. Funkcjonowanie archiwum OAIS. Na podstawie modelu referencyjnego OAIS

Źródło: *Reference Model…*, *op. cit*.

Repozytorium zasobów cyfrowych, zorganizowane na podstawie modelu OAIS obejmuje sześć funkcjonalnych jednostek oraz drogę dokumentu cyfrowego od producenta do użytkownika. Jednostka „Przyjęcie" odpowiedzialna jest za przyjęcie zgłoszeniowego pakietu informacyjnego SIP (*Submission Information Package* – SIP) oraz za przygotowanie go do umieszczenia oraz administrowania nim w repozytorium. W zakresie jej zadań znajduje się m.in. kontrola kompletności oraz autentyczności zgłoszeniowego pakietu informacyjnego, przekształcenie pakietu SIP w pakiet gotowy do archiwizacji oraz stworzenie do niego metadanych. Następnie archiwizowany pakiet informacyjny AIP (*Archive Information Package* – AIP) przekazywany jest do jednostki zajmującej się archiwizacją, tj. do „Archiwum", a metadane odsyłane do jednostki „Zarządzanie danymi", odpowiedzialnej za zarządzanie zdeponowanymi zasobami.

Kolejną istotną jednostką funkcjonalną systemu OAIS jest „Archiwum", odpowiedzialne za zapis, właściwe przechowywanie pakietów informacyjnych (AIP) oraz możliwość ich odczytu. „Archiwum" odpowiada za długo-

[9] *Reference Model…*, *op. cit*.; *Trusted Digital Repositories. Attributes and Responsibilities. An RLG-OCLC Report* RLG. The Research Libraries Group, Mountain View, CA 2002, [online:] http://www.oclc.org/research/activities/past/rlg/trustedrep/repositories.pdf, [dostęp: 16.11.2012].

terminowe przechowywanie i zapewnienie nienaruszalności pakietów AIP, okresowe przenoszenie danych na media nowszej generacji, migrację do aktualnie stosowanych formatów lub systemów, a w przypadku awarii systemu za ich rekonstrukcję. Na żądanie „Archiwum" przekazuje określony pakiet AIP do jednostki „Udostępnianie".

W archiwum elektronicznym niezbędna jest jednostka „Zarządzanie danymi". Jej zadaniem jest utrzymywanie i udostępnianie szerokiego wachlarza informacji. Przykładami mogą być katalogi i inwentarze, na których podstawie można wyszukać i uzyskać określone zasoby z repozytorium, a także statystyki dotyczące udostępniania zasobów. Do zadań tej jednostki należy także kontrola bezpieczeństwa danych oraz inne procedury związane z ochroną repozytorium narzucane przez OAIS.

Poprawne funkcjonowanie całego repozytorium jest uzależnione od prac jednostki administrującej procesami w nim zachodzącymi. „Zarządzanie systemem" zajmuje się negocjowaniem warunków z deponentami, na których podstawie dokumenty są transferowane do repozytorium, czuwa nad kontrolą zgodności dostarczonych dokumentów ze standardami repozytorium oraz przejmuje odpowiedzialność za utrzymywanie sprawności sprzętu i oprogramowania w repozytorium. Czyni także starania na rzecz rozwoju oraz nadzoru nad standardami, potrzebnymi dla funkcjonowania repozytorium.

Repozytoria zgodne z modelem OAIS starają się zapewnić na przyszłość stabilny dostęp do przechowywanych w nich różnorodnych zasobów cyfrowych. W tym celu w OAIS wyodrębniono jednostkę „Planowanie procesu archiwizacji", która zajmuje się m.in. obserwowaniem rozwoju rynku sprzętu i oprogramowania, testowaniem nowych rozwiązań, kontrolą czy archiwizowane dokumenty dadzą się uruchomić i odczytać. Jednostka ta jest odpowiedzialna za wszelkie decyzje dotyczące strategii postępowania, m.in. częstotliwości odświeżania danych, działań mających na celu dostosowanie rozwiązań do zmieniających się warunków sprzętowo-programowych (emulacja lub migracja) i udostępnienia treści dokumentów w zmienionych warunkach.

Proces udostępniania dokumentów cyfrowych w archiwach OAIS został określony terminem *Access*. W ramach udostępniania zasobów użytkownikom umożliwia się przeglądanie zawartości repozytorium poprzez katalogi online, określenie lokalizacji i dostępności określonych zbiorów. Na zamówienie użytkownika system tworzy i wysyła użytkowe pakiety informacyjne typu DIP (Dissemination Information Package – DIP).

Na podstawie przytoczonego opisu możliwe staje się prześledzenie drogi cyfrowego dokumentu przez repozytorium zgodne z założeniami OAIS.

Deponent, który chce odesłać do repozytorium dokument cyfrowy w celu jego długoterminowego przechowania, powinien nadać dokumentowi właściwą – ustaloną wcześniej z repozytorium – formę oraz dołączyć wszelkie dodatkowe informacje o dokumencie wraz z metadanymi. Dokument wraz z metadanymi przesyłany jest w postaci zgłoszeniowego pakietu informacyjnego SIP do działu przyjęcia, gdzie zostaje „rozpakowany" oraz sprawdzony pod względem kompletności i poprawności wszelkich danych niezbędnych do przyjęcia i długoterminowego zarchiwizowania dokumentu. W systemie repozytoryjnym każdy dokument przyjmuje następnie postać AIP, jest zapisywany na serwerze archiwum i przechowywany w sposób umożliwiający jego długotrwałą, stabilną użyteczność. Wygenerowane metadane deponowanych dokumentów są odsyłane do działu, zajmującego się ich zarządzaniem. Archiwizowany obiekt AIP może być przekształcony w formę DIP (*Dissemination Information Package* – DIP), udostępnianą na żądanie użytkownikowi, w postaci umożliwiającej jego zrozumienie.

Autorzy modelu OAIS zapewniają, że może on być stosowany przy organizacji wszelkich repozytoriów (magazynów, archiwów) danych cyfrowych, ze specjalnym przeznaczeniem dla tych, które mają być odpowiedzialne za długoterminowe przechowywanie zdeponowanych zasobów. Do modelu OAIS odwołują się pracownicy Poznańskiego Centrum Superkomputerowo-Sieciowego, autorzy i producenci Usługi Powszechnej Archiwizacji Platon – U4 oraz dArceo, dedykowanych bezpiecznemu i trwałemu składowaniu cyfrowego zasobu polskiej nauki i kultury[10].

6.2. Atrybuty zasobów repozytoryjnych podlegające trwałej ochronie

Funkcjonujące i planowane repozytoria cyfrowe, zwłaszcza te przechowujące dziedzictwo nauki i kultury, dążą do spełniania wymagań instytucji wiarygodnych, autentycznych, stabilnych i niezawodnych (ang. *trusted digital repository, trustworthy digital repositories,* niem. *vertrauenswürdiges digitales Langzeitarchiv*)[11]. Wiarygodne repozytorium cyfrowe to takie, które

[10] *KMD Krajowy Magazyn Danych*, [online:] http://kmd.pcss.pl/index.html [dostęp: 28.11.2012].

[11] *Trusted Digital Repositories. Attributes and Responsibilities.* An RLG-OCLC Report [online:]http://www.oclc.org/research/activities/past/rlg/trustedrep/repositories.pdf[dostęp: 12.02.2010]; *Kriterienkatalog vertrauenswürdige digitale Langzeitarchive.* Version 1. [Entwurf zur öffentlichen Kommentierung], „Nestor Materialien" 2006, nr 8, [online:] http://edoc.hu-

gwarantuje dostępność przechowywanych i zarządzanych w nim dokumentów obecnie i w odległej przyszłości, przyjmuje odpowiedzialność za przeprowadzanie prac konserwatorskich w imieniu swoich deponentów oraz na rzecz potrzeb obecnych i przyszłych użytkowników. Dostępność (ang. *availibility*) to atrybut ściśle związany z użytecznością dokumentów, czyli możliwością odczytu i interpretacji ich treści w założonym czasie, przez osobę lub instytucję do tego uprawnioną.

Kolejne atrybuty zasobów repozytoryjnych podlegające ochronie to autentyczność i integralność[12]. Integralność danych (ang. *data integrity*), określana także jako spójność, to funkcja bezpieczeństwa polegająca na zagwarantowaniu niezmienności danych, poprzez blokowanie możliwości ich dodania lub usunięcia w nieautoryzowany sposób. Autentyczność (ang. *authenticity*) natomiast to właściwość odpowiadająca za to, że tożsamość podmiotu lub zasobu jest taka jak zadeklarowana. Atrybut ten jest odpowiedzialny za potwierdzenie, że dokument zdeponowany i udostępniany użytkownikom repozytorium jest dokładnie tym samym dokumentem, którego autentyczność potwierdził deponent na etapie zgłoszenia i przyjmowania go do kolekcji repozytoryjnej.

W technice informatycznej i telekomunikacyjnej ochrona integralności zapobiega przypadkowemu zniekształceniu danych podczas odczytu, zapisu, transmisji lub magazynowania. W celu ochrony integralności danych wykonuje się sumy kontrolne i stosuje kody korekcyjne. W bezpieczeństwie teleinformatycznym natomiast ochrona integralności zapobiega celowej modyfikacji danych dokonanej z użyciem zaawansowanych technik, mających na celu ukrycie faktu dokonania zmiany. Wykorzystuje się tutaj techniki kryptograficzne, np. kody MAC odporne na celowe manipulacje. Systemy repozytoryjne, chroniąc integralność danych, dbają o kompletność zdeponowanej kolekcji, prawdziwość dokumentów i gwarantują, że nie zostały one poddane modyfikacji lub manipulacji[13].

W literaturze przedmiotu, obok ochrony dostępności oraz użyteczności dokumentów cyfrowych, wśród celów, którym długoterminowa ochrona ma służyć, wymienia się także poufność. Przez pojęcie poufności należy rozu-

berlin.de/series/nestor-materialien/2006-8/PDF/8.pdf [dostęp: 20.11.2012]; *Trustworthy Repositories Audit & Certification. Criteria and Checklist. Version 1.0. 2007* [online:] http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf [dostęp: 14.11.2012].

[12] *Attributes of a Trusted Digital Repository: Meeting the Needs of Research Resources*. An RLG-OCLC Report. Draft for Public Comment [online:], http://www.worldcat.org/arcviewer/1/OCC/2007/08/08/0000070511/viewer/file2172.pdf, [dostęp: 15.11.2012].

[13] *Społeczeństwo informacyjne*, red. J. Papińska-Kacperek, Wydawnictwo Naukowe PWN, Warszawa 2008, s. 256–257.

mieć stan, w którym dokument nie jest i nie może być ujawniony osobom, podmiotom bądź procesom nieupoważnionym. Zapewnienie poufności dokumentu może wynikać z takich przesłanek jak ochrona prywatności i interesów własnych deponenta, ochrona interesów instytucji archiwizującej, czyli depozytariusza, obowiązujące akty prawne[14].

Zgodnie z modelem OAIS w jednostce *Planowanie procesu archiwizacji* jest opracowywana strategia trwałej ochrony systemu repozytoryjnego. Ochronie podlegają przede wszystkim dane cyfrowe w postaci kodu zerojedynkowego, ale procedurami bezpieczeństwa obejmuje się również algorytmy dotyczące postępowania z danymi zakodowane w metadanych, oprogramowanie, sprzęt, budynki. Strategia ochrony powinna uwzględniać rozmaite rodzaje zagrożeń; błędy ludzkie, awarie sprzętu, oprogramowania i innych części infrastruktury, katastrofy naturalne, destrukcyjne oprogramowania, włamania do systemów[15].

6.3. Metadane zasobów repozytoryjnych

System repozytoryjny powinien zapewnić odpowiedni stopień bezpieczeństwa i nienaruszalności nie tylko zdeponowanych dokumentów, ale także ich metadanych. Metadane opisują dokumenty przechowywane w systemach repozytoryjnych, ich strukturę, atrybuty, ewentualne modyfikacje, wskazują na wytwórcę dokumentu, jego autora, datę powstania, daty transmisji wewnątrz systemu lub na zewnątrz[16]. Umożliwiają wyszukanie, wykorzystanie i administrowanie dokumentem[17]. W strategii ochrony zasobów cyfrowych metadane odgrywają zasadniczą rolę, ponieważ są jedynym sposobem uchwycenia kontekstu archiwizowanych dokumentów[18], czyli wszelkich informacji o pochodzeniu, procesie powstawania, o czasie i celu powstania, o dotychczasowej hi-

[14] T. Bilski, *Pamięć. Nośniki i systemy…*, *op. cit*., s. 423; *Kriterienkatalog Vertrauenswürdige…*, *op. cit.*

[15] *Reference Model…*, *op. cit*.; *Strategie i modele gospodarki elektronicznej*, red. C.M. Olszak, E. Ziemby, Wydawnictwo Naukowe PWN, Warszawa 2007, s. 399–404.

[16] A. Freedman, *Encyklopedia komputerów,* Wydawnictwo Helion, Gliwice 2004, s 456; J. Adamus, *Metadane w archiwizacji dokumentów elektronicznych*, „Zagadnienia Informacji Naukowej" 2009, nr 2(94), s. 14–15.

[17] M. Nahotko, *Biblioteki cyfrowych książek w środowisku akademickim i bibliotekarskim*, [w:] *Cyfrowy świat bibliotek – problemy techniczne, prawne, wdrożeniowe*, Materiały konferencyjne z IX edycji seminarium z cyklu „Archiwizowanie i digitalizacja", 17–18 stycznia 2006 r., Centrum Promocji Informacji, Warszawa, s. 184.

[18] *Ibidem*, s. 31.

storii, warunkach dostępu, sposobach użytkowania i wszelkich powiązaniach dokumentu z innymi komponentami, pozostającymi w repozytorium bądź poza nim. Metadane powinny wspomóc procesy migracji danych przez kolejne generacje sprzętu komputerowego i oprogramowania, umożliwiać rekonstrukcję procesu decyzyjnego dotyczącego prac na dokumentach cyfrowych, dostarczyć rejestr kontroli dokumentu przez cały cykl jego życia[19].

Metadane zwykle ujmowane są w następujące grupy: metadane opisowe, techniczne, administracyjne, strukturalne oraz prawne, niekiedy określane również jako użytkowe[20].

Z punktu widzenia strategii ochrony użyteczności zasobów repozytoryjnych istotną rolę odgrywają metadane techniczne, w których zakodowana jest informacja o okresie przechowywania, platformie sprzętowo-programowej potrzebnej do odczytu i prezentacji treści dokumentu, opis zastosowanego formatu zapisu publikacji wraz z informacjami o dokumentacji tego formatu oraz informacja o zastosowanym nośniku danych. Metadane techniczne określane bywają jako konserwacyjne[21], dlatego że zawierają informacje o planowanych bądź przeprowadzonych dotychczas pracach konserwatorskich na dokumentach. Dodatkowo mogą być uzupełnione o dokumentację sporządzaną podczas takich prac bądź do niej odsyłać. Techniczne metadane są też istotnym czynnikiem umożliwiającym automatyzację określonych prac konserwatorskich na zasobach cyfrowych, np. migracji i emulacji. Mogą też dostarczać dokładny opis zastosowanego identyfikatora trwałego[22] do dokumentu repozytoryjnego.

Kolejny rodzaj metadanych wspomagających procesy ochrony użyteczności zasobów repozytoryjnych to metadane administracyjne. Dotyczą one

---

[19] C. Lupovici, J. Masanès, *Metadata for the Long Term Preservation of Electronic Publications*, Nedlib Report Series 2. The Hague: Koninklijke Bibliotheek, 2000, s. 3–4.

[20] *Metadata, The Tech Terms Komputer Dictionary 2005*, [online:] http://www.techterms.com/definition/metadata [dostęp: 12.11.2012]; A. Januszko-Szakiel, *Dysertacje via Internet. Projekt elektronicznej archiwizacji rozpraw naukowych w Niemieckiej Bibliotece Narodowej*, „Przegląd Biblioteczny" 2006, z. 2, s. 141.

[21] *Ibidem*.

[22] Identyfikator trwały (PI) to niezmienna (określana też jako stabilna, unikalna, permanentna) nazwa, którą przyporządkowuje się do dokumentu cyfrowego przechowywanego w repozytorium jeden raz na cały cykl jego „życia". Zadaniem PI jest jednoznaczna i trwała identyfikacja dokumentu oraz przynależnych do niego metadanych, niezależnie od miejsca (instytucji), w którym dokument został zapisany i jest archiwizowany, z uwzględnieniem różnorodnych systemów, zmian, z uwzględnieniem występowania obiektów w różnych wersjach, postaciach, formach reprezentacji. Zob. *Persistent Identifiers for Cultural Heritage*, Digital Preservation Europe, [online:] http://www.digitalpreservationeurope.eu/publications/briefs/persistent_identifiers.pdf [dostęp: 20.11.2012].

zarządzania dokumentami cyfrowymi w repozytorium. Zawierają między innymi informacje o istniejących wersjach określonego dokumentu, o sporządzanych kopiach. Ważnym elementem tych metadanych są informacje o takich parametrach dokumentów, jak integralność, autentyczność, o wynikach sporządzania sum kontrolnych *etc*. Metadane administracyjne informują również o uprawnieniach poszczególnych pracowników systemu repozytoryjnego do wykonywania określonych czynności na dokumentach.

6.5. Techniczne aspekty ochrony zasobów cyfrowych

Techniczne aspekty zapewnienia ochrony wiarygodności, autentyczności, integralności oraz poufności zasobów cyfrowych można rozpatrywać z punktu widzenia zabezpieczenia fizycznej infrastruktury repozytorium cyfrowego, polityki uprawnień oraz uwierzytelniania użytkowników systemu, a także metod zabezpieczania danych przed niepowołanym odczytem za pomocą metod kryptograficznych.

Podstawowym wymaganiem, które powinien spełniać każdy system repozytoryjny, jest zapewnienie fizycznej nienaruszalności infrastruktury repozytorium i zgromadzonych w nim danych. Zagrożenia fizycznego bezpieczeństwa systemu mogą wynikać zarówno z nieuprawnionych działań ludzi, nieprawidłowej pracy infrastruktury technicznej, jak i działania sił natury. Aby zminimalizować możliwy wpływ zagrożeń fizycznych, konieczne jest odpowiednie zaprojektowanie, zlokalizowanie oraz wyposażenie budynku repozytorium.

Lokalizacja budynku repozytorium powinna być wybrana w taki sposób, aby zminimalizować możliwość jego podtopienia zarówno przez wody opadowe, jak i powierzchniowe, a także zminimalizować możliwość wystąpienia zagrożeń związanych z działalnością ludzi (np. katastrof lotniczych w lokalizacjach położonych blisko korytarzy powietrznych lub w pobliżu lotniska). Dodatkowym zabezpieczeniem szczególnie cennych zasobów powinno być także wykonywanie kopii zapasowej zgromadzonych zasobów, która jest przechowywana w lokalizacji geograficznie oddalonej od repozytorium. Pozwala to na zachowanie zasobu w przypadku zniszczenia repozytorium w wyniku nieprzewidzianych katastrof naturalnych lub np. działań wojennych.

Właściwie zlokalizowany budynek powinien być także odpowiednio wyposażony w systemy zabezpieczające umieszczone w nim urządzenia techniczne.

Jednym z podstawowych zagrożeń podczas pracy urządzeń elektronicznych jest możliwość wystąpienia pożaru, w wyniku którego bardzo prawdopodobne jest uszkodzenie wrażliwych na wysoką temperaturę urządzeń, a w szczególności nośników danych, co pociąga za sobą ich nieodwracalną utratę. W celu zapobiegania rozprzestrzenianiu się ewentualnego pożaru, a także minimalizacji zniszczeń, jakie może powodować, zalecany jest podział budynku na sekcje rozdzielane ścianami ognioodpornymi oraz zastosowanie urządzeń tłumiących ogień. Jednym z najbardziej skutecznych systemów stosowanych w przypadku pomieszczeń z urządzeniami elektronicznymi są urządzenia, które w przypadku wykrycia oznak pożaru wypierają z pomieszczenia powietrze poprzez wtłoczenie mieszaniny gazów o działaniu gaśniczym zgromadzonych w butlach wysokociśnieniowych. Wtłoczenie do pomieszczenia gazów gaśniczych powoduje wyparcie z niego powietrza, dzięki czemu spada stężenie tlenu niezbędnego do podtrzymywania procesu spalania oraz powoduje gwałtowne obniżenie temperatury poprzez rozprężenie gazów gaśniczych. Ze względu na możliwość przebywania w pomieszczeniu ludzi, konieczny jest jednak dobór parametrów gazowego systemu gaśniczego w taki sposób, aby osoby, którym nie udało się ewakuować z pomieszczenia, mogły przeżyć w atmosferze zmodyfikowanej gazami gaśniczymi.

Systemy zabezpieczające powinny także obejmować kontrolę fizycznego dostępu do urządzeń poprzez wykorzystanie ochrony perymetrycznej wnętrza oraz otoczenia budynku, realizowane za pomocą np. czujników geofonicznych, systemów elektromagnetycznej lub impedancyjnej detekcji ruchu, czujników ruchu pracujących w zakresie ultradźwięków lub podczerwieni, a także systemów monitorowania wizyjnego. Pozwala to na minimalizację możliwości celowego działania osób trzecich, których celem mogłoby być zniszczenie infrastruktury lub też ingerencja w zgromadzony zasób[23].

Kolejnym aspektem, który powinien zostać szczegółowo rozważony, jest uwierzytelnianie użytkowników systemu, a więc weryfikacja ich tożsamości. Stosowane obecnie metody można podzielić na trzy grupy: pierwszą z nich stanowią stosowane od lat rozwiązania oparte na identyfikacji użytkownika za pomocą nazwy użytkownika (login) oraz odpowiadającego mu hasła, drugą grupę rozwiązań stanowią biometryczne systemy identyfikacji i uwierzytelniania użytkowników, natomiast trzecią grupę stanowią rozwiązania będące połączeniem elementów obu wyżej wymienionych rodzajów[24].

[23] Systemy ochrony zewnętrznej, [online:] http://www.alkam-security.pl/_cms/view/104/systemy-ochrony-zewnetrznej.html [dostęp: 28.11.2012].

[24] M. Kaeo, *Tworzenie bezpiecznych sieci*, Mikom, Warszawa 2000, *passim*; W. Stallings, *Kryptografia i bezpieczeństwo sieci komputerowych. Matematyka szyfrów i techniki kryptologii*,

Uwierzytelnianie oparte na znajomości identyfikatora (nazwy użytkownika i loginu) i odpowiadającego mu hasła pozwala dowolnej osobie, która zdobędzie te informacje, na dostęp do systemu na prawach przyznanych użytkownikowi, którego tożsamość wykorzystuje. Login i hasło mogą zostać „przejęte" na wiele sposobów, zaczynając od podejrzenia hasła zapisanego przez użytkownika obawiającego się, że je zapomni, poprzez wykorzystanie oprogramowania pozwalającego na generowanie i sprawdzanie poprawności kolejnych kombinacji znaków, po fizyczne wymuszenie ujawnienia danych niezbędnych do uwierzytelnienia.

Drugą grupą metod uwierzytelniania są systemy rozpoznające charakterystyczne cechy osobnicze ludzi – kształt twarzy, wzór tęczówki lub siatkówki oka, brzmienie głosu, geometrię dłoni lub układ linii papilarnych. Projektowane aktualnie systemy biometryczne charakteryzują się coraz wyższą niezawodnością, jednakże niepozwalającą jeszcze na ich zastosowanie jako jedynego źródła uwierzytelniania w systemach, w których wymagany jest bardzo wysoki poziom bezpieczeństwa[25].

W celu podniesienia poziomu bezpieczeństwa systemów uwierzytelniania opartych na metodach biometrycznych mogą być wykorzystywane kombinacje kilku metod biometrycznych, np. skan siatkówki i tęczówki oka wraz z analizą układu linii papilarnych. Zastosowanie równocześnie kilku metod potwierdzenia identyfikacji użytkownika redukuje ryzyko błędnego uwierzytelnienia. Często spotykanym połączeniem jest także wykorzystanie technik biometrycznych jako uzupełnienia klasycznego systemu opartego na loginie i haśle.

Użytkownik uwierzytelniony w systemie może w nim wykonywać działania zgodne z przyznanymi uprawnieniami, odnotowanymi w metadanych technicznych i administracyjnych. Uprawnienia te są zróżnicowane dla danego użytkownika lub grupy użytkowników. Mówi się wówczas o poziomowaniu dostępu do zasobów i operacji, które można na zasobach wykonać (użytkownicy o najniższym poziomie dostępu otrzymują zezwolenie na odczyt ściśle określonych dokumentów, natomiast użytkownicy posiadający status administratora (np. pracownicy działu długoterminowej archiwizacji) posiadają uprawnienia zarówno do odczytu, jak i prac konserwatorskich na danych cyfrowych. Poziom uprawnień przypisanych danemu użytkownikowi może wynikać zarówno z polityki bezpieczeństwa repozytorium, uwarunkowań

Helion, Gliwice 2012, s. 48–49, 408–409; M. Horton, C. Mugge, *Bezpieczeństwo sieci. Notes antyhakera*, Wydawnictwo Translator, Warszawa 2004, s. 124–125.

[25] *Społeczeństwo informacyjne*, red. J. Papińska-Kacperek, Wydawnictwo Naukowe PWN, Warszawa 2008, s. 295.

prawnych, jak i ochrony poufności zasobów. Niezależnie od poziomu uprawnień, wszystkie operacje na danych, wykonywane przez użytkowników wewnętrznych (pracowników repozytorium) oraz użytkowników zewnętrznych (klientów, odbiorców treści deponowanych dokumentów), które mogą wywołać utratę autentyczności dokumentów, powinny być rejestrowane[26].

Właściwe zaplanowanie oraz wdrożenie wspomnianych metod ochrony zgromadzonego zasobu cyfrowego, zarówno przed fizycznym zniszczeniem, jak i nieautoryzowanym fizycznym dostępem, nie gwarantują pełnej ochrony autentyczności, integralności oraz poufności zasobów cyfrowych. Większość repozytoriów cyfrowych świadczy usługi polegające na udostępnianiu treści przechowywanych zasobów za pomocą łącz teleinformatycznych, co powoduje konieczność zapewnienia bezpieczeństwa przesyłanych danych i dostarczenia ich do ściśle określonego odbiorcy. Wykorzystanie łącz teleinformatycznych do udostępniania zasobów powoduje także możliwość podszywania się osób trzecich pod uwierzytelnionego użytkownika, w celu pozyskania poufnych informacji lub też dokonania w repozytorium działań o charakterze destrukcyjnym.

W celu uzyskania wysokiego poziomu bezpieczeństwa na etapie udostępniania dokumentów poprzez łącza teleinformatyczne stosowane są techniki kryptograficzne. Kryptografia jest to dziedzina zajmująca się metodami utajniania informacji poprzez jej szyfrowanie. Dzięki kryptografii można zamienić normalny, zrozumiały tekst lub innego rodzaju wiadomość w taki sposób, że stanie się niezrozumiała dla nieupoważnionego odbiorcy. Właściwy odbiorca wiadomości może po jej otrzymaniu przekształcić ją ponownie do czytelnej postaci. Do niedawna głównymi odbiorcami rozwiązań kryptograficznych były instytucje rządowe, placówki dyplomatyczne oraz wojsko. Rozwój elektronizacji obiegu informacji spowodował, że obszar zastosowań kryptografii znacznie się powiększył. Ze względu na przedmiot szyfrowania można wyróżnić szyfrowanie pojedynczych plików lub systemów plikowych, a także szyfrowanie transmisji[27].

W ostatnich latach postępuje bardzo szybki rozwój kryptografii, która musi sprostać nowym wymaganiom, wynikającym z wcześniej nieznanych

[26] *Reference Model…*, *op. cit*.; *Trusted Digital Repositories. Attributes and Responsibilities*. An RLG-OCLC Report, [online:] http://www.oclc.org/research/activities/past/rlg/trustedrep/repositories.pdf [dostęp: 20.11.2012].

[27] *Zabezpieczenia kryptograficzne*, [online:] http://www.b-skrzypczyk.republika.pl/ ; *Społeczeństwo informacyjne…, op. cit*., s. 300–311 [dostęp: 20.11.2012]; D.E. Denning, *Wojna informacyjna i bezpieczeństwo informacji,* Wydawnictwa Naukowo-Techniczne, Warszawa 2002, s. 326–424.

oczekiwań, dotyczących m.in. bezpieczeństwa transmisji danych w sieciach teleinformatycznych. Jednym z owoców gwałtownego rozwoju kryptografii jest powstanie algorytmu AES (Advanced Encryption Standard), którego założenia zostały zdefiniowane właśnie pod kątem zastosowania w internecie – miał być szybki, wydajny i przede wszystkim bezpieczny. Równocześnie jednak okazało się, że nie wystarczy samo zastosowanie silnego szyfru, należy go jeszcze stosować w bezpieczny sposób. Dopiero prawidłowe zaprojektowanie kompletnego systemu poprawnie wykorzystującego zarówno metody uwierzytelniania użytkowników, jak i protokoły kryptograficzne pozwala na niezawodne i bezpieczne działanie w oparciu o sprawdzone algorytmy szyfrujące. W stosunkowo niedługim okresie, jaki upłynął od cywilnego upowszechnienia kryptografii, ustalono pewien ogólny schemat protokołu służącego do bezpiecznej komunikacji. Dotyczy to zarówno doboru odpowiednich algorytmów zapewniających poufność oraz integralność przekazu, jak i metod projektowania i późniejszego testowania danego protokołu. W przypadku protokołów stosowanych obecnie w internecie standardem stał się otwarty model rozwoju, co oznacza, że stworzony przez grupę ekspertów projekt jest następnie przedstawiany do publicznego wglądu oraz oceny. Na tym etapie wskazywane są najsłabsze punkty projektu, które następnie podlegają korekcie. W ten sposób powstały dwa najszerzej obecnie stosowane protokoły kryptograficzne: SSL v.3 oraz IPSec, używany w sieciach VPN. Praktyka dowiodła, że otwartość procesu projektowania takich protokołów i publiczna ocena dodatnio wpływają na ich bezpieczeństwo, podczas gdy utrzymywanie protokołu w tajemnicy skutkuje prawie zawsze jego złamaniem później, kiedy jest on już szeroko wdrożony i kiedy naprawienie błędów oznacza poważne koszty.

Kryptograficzne zabezpieczenie transmisji danych może być realizowane programowo lub sprzętowo. Programowa realizacja procesu kryptograficznego wymaga stosunkowo dużych mocy obliczeniowych, w związku z czym coraz częściej zastosowanie znajdują specjalistyczne urządzenia zabezpieczające transmisję. Przykładem takiego rozwiązania jest rodzina szyfratorów CompCrypt. Są to urządzenia przeznaczone do zapewnienia wysokiego poziomu bezpieczeństwa informacji niejawnych przesyłanych w sieciach teleinformatycznych. CompCrypt wykorzystują Narodowe Algorytmy Szyfrujące – dzięki czemu mogą być wykorzystywane do ochrony informacji niejawnych – od klauzuli „Zastrzeżone" do „Ściśle tajne". Zastosowany w urządzeniach Narodowy Algorytm Szyfrujący jest faktycznie oparty na zmodyfikowanym algorytmie 3DES (rodzaj i zakres modyfikacji jest tajny).

Powodem zastosowania dodatkowych przekształceń jest prawdopodobnie udaremnienie prób wykorzystania sprzętowych implementacji algorytmu 3DES do szukania kluczy deszyfrujących[28].

Strategia ochrony zasobów cyfrowych składowanych i archiwizowanych w repozytoriach instytucjonalnych, nawet jeśli została opracowana z uwzględnieniem wszelkich możliwych zagrożeń oraz sprawdzonych i skutecznych metod, technik i narzędzi zabezpieczenia wszystkich elementów składowych systemu repozytoryjnego, powinna podlegać okresowej rewizji. Bardzo ważne jest obserwowanie zmian technologicznych (tzw. *technology watch*) oraz zmieniających się oczekiwań klientów repozytorium, czyli deponentów i odbiorców deponowanych treści. W systemach repozytoryjnych obowiązuje zarządzanie zmianami, cykliczne weryfikowanie obowiązujących założeń i w razie potrzeby ich aktualizowanie.

Organizatorzy repozytoriów powinni zastanowić się również nad zawarciem umowy z firmą ubezpieczeniową. Odpowiednio dobrana polisa ubezpieczeniowa może stanowić przydatny element polityki zabezpieczenia systemu repozytoryjnego[29].

Podsumowanie

Zadaniem systemów repozytoryjnych jest zagwarantowanie bezpiecznego przechowania i dostępu do kompletnej kolekcji autentycznych zasobów cyfrowych. W celu zabezpieczenia kompletności i prawdziwości zdeponowanych dokumentów konieczne jest opracowanie strategii ochrony oryginalnego kodu zerojedynkowego oraz pozostałych elementów repozytorium, umożliwiających jego udostępnienie, odczyt i przedstawienie w postaci zrozumiałej dla użytkownika.

Zasoby deponowane w repozytoriach instytucjonalnych zwykle są grupowane w kolekcje z określeniem praw dostępu do nich, zakresu ich użytkowania, przyporządkowania metod ochrony, nadania uprawnień i odpowiedzialności za ochronę bezpieczeństwa[30]. Zaleca się powielanie kolekcji depozytowych. Oprócz podstawowej kolekcji repozytoryjnej tworzone są ich tzw. kopie lustrzane i przechowywane w oddalonym miejscu. Możliwe

[28] *Narodowy Algorytm Szyfrujący*, [online:] http://ipsec.pl/kryptografia/narodowy-algorytm-szyfrujacy-nasz.html, [dostęp: 26.11.2012].

[29] T. Kifner, *Polityka bezpieczeństwa i ochrony informacji*, Helion, Gliwice 1999, s. 119–120.

[30] *Strategie i modele gospodarki elektronicznej*, red. C.M. Olszak, E. Ziemba, Wydawnictwo Naukowe PWN, Warszawa 2007, s. 402–404.

jest przechowywanie kopii kolekcji w kilku miejscach; powstaje wówczas rozbudowany system replikacji geograficznych.

W zależności od przyjętych celów i założeń systemu repozytoryjnego, strategia ochrony dokumentów to połączenie rozmaitych metod, narzędzi i technik. Niektóre z nich to: metadane techniczne i administracyjne, identyfikatory trwałe, sumy kontrolne, systemy kryptograficzne, biometryka, uwierzytelnianie.

CZĘŚĆ III

Prawo Benforda jako procedura weryfikacji jakości zbiorów danych – wybrane problemy

Benford's Law as a procedure to verify the quality of data sets – some problems

STRESZCZENIE:
Ta część monografii poświęcona jest wybranym kwestiom związanym z możliwościami tkwiącymi w tzw. prawach Benforda. Prawa te opisują częstości pojawiania się cyfr na określonych miejscach liczb z dużych zbiorów pomiarów.

Słowa kluczowe: prawo Benforda, zbieżności testów, mierniki podobieństwa.

Cel:
Analiza zbieżności testów i mierników podobieństwa rozkładów częstości, a także przegląd empirycznych praw numerycznych zbliżonych swoim charakterem do prawa Benforda.

Metodyka badań:
Metody: opisu pojedynczych przypadków, analiza dokumentacji, modelowanie.

Wynik:
Spójne opracowanie w zakresie wybranych metod opracowania cyfrowej dokumentacji.

Oryginalność wartość:
Oryginalny układ i połączenie treści.

ABSTRACT:
This part of the monograph is devoted to selected issues related to the possibilities inherent in the so-called. Benford's rights. These laws describe the frequency of occurrence of digits in certain places the number of large sets of measurements.

Key words: Benford's law, convergence tests, measures of similarity.

Convergence analysis of tests and measures the similarity of frequency distributions, as well as an overview of empirical laws numeric close their character to Benford's law.

Methods: describing individual cases, analysis of documentation, modeling.

Coherent development of selected developing methods of digital documentation.

Original layout and combination of content.

7. Geneza, istota i rozwój badań nad prawem Benforda

Marzena Farbaniec, Tadeusz Grabiński, Bartłomiej Zabłocki, Wacław Zając

Wprowadzenie

Na problem rozkładu pierwszych cyfr znaczących w dużych zbiorach danych empirycznych pierwszy zwrócił uwagę amerykański astronom i matematyk Simon Newcomb, który na początku lat 80. XIX wieku zauważył, że strony w tablicach logarytmicznych są bardziej zabrudzone na początku książki niż pod jej koniec (tablice logarytmiczne liczb siedmiocyfrowych zawierają około 200 stron). Oznacza to, że z jakichś powodów użytkownicy tablic logarytmicznych częściej korzystają w obliczeniach z liczb mniejszych niż większych (na początku tablic podawane są logarytmy liczb zaczynających się od 1 000 000 i rosnących do 9 999 999).

Spostrzeżenie to S. Newcomb opublikował w krótkim, dwustronicowym doniesieniu[1], gdzie przytoczył prawdopodobieństwa pojawiania się pierwszej i drugiej cyfry znaczącej, a także stwierdził, że kolejne cyfry znaczące mają rozkład równomierny. Ponadto ustalił, że mantysy logarytmów liczb mają jednakowe prawdopodobieństwa (rozkład równomierny), z czego wynikało, że poszczególne strony tablic antylogarytmów są wykorzystywane z jednakową intensywnością i w odróżnieniu od tablic logarytmów mają jednakowo zabrudzone brzegi. Ponieważ użytkownicy tablic logarytmów nie czytają ich jak książkę beletrystyczną, tylko traktują jako narzędzie ułatwiające obliczenia numeryczne (wczesny kalkulator) i szukają w nich wartości logarytmów nie dla wymyślonych, lecz dla konkretnych liczb, wziętych z rzeczywistych pomiarów, to fakt częstszego korzystania z liczb zaczynających się od mniejszych cyfr nie jest przypadkowy tylko ma charakter ogólnego prawa. Formuła, według której można ustalić prawdopodobieństwo (częstość) pojawiania się pierwszych cyfr znaczących di, ma postać:

$$P(d_i)=log_{10}(1+1/d_i) \text{ dla } d_i=1,2,\ldots,9 \quad (8)$$

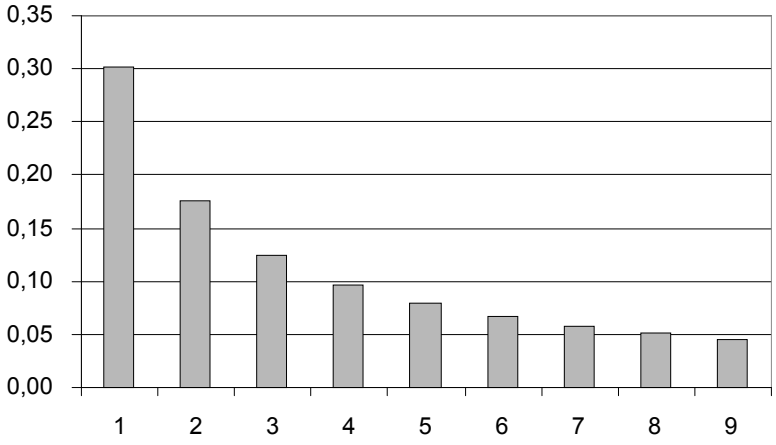
[1] S. Newcomb, *Note on the frequency of use different digits in natural numbers*, „American Journal of Mathematics" 1881, t. 4, s. 39–40.

W tabeli 1 oraz na rysunku 10 podano konkretne wartości tych prawdopodobieństw. Jak się okazuje, wbrew intuicji, która przemawia raczej za rozkładem równomiernym, mamy tu do czynienia z rozkładem szybko malejącym. Co trzecia liczba spotykana w realnych zbiorach danych liczbowych zaczyna się od 1, a tylko jedna na 22 liczby zaczyna się od 9. Inaczej mówiąc, liczb zaczynających się na 1 jest prawie 7x więcej niż liczb zaczynających się na 9.

Tabela 1. Prawdopodobieństwa i częstości pojawiania się pierwszych cyfr znaczących d_i

d(i)	1	2	3	4	5	6	7	8	9
P(d)	0,301	0,176	0,125	0,097	0,079	0,067	0,058	0,051	0,046
%	30,1	17,6	12,5	9,7	7,9	6,7	5,8	5,1	4,6
100/%	3	6	8	10	13	15	17	20	22

Źródło: opracowanie własne.



Rysunek 10. Diagram ilustrujący prawdopodobieństwa pojawiania się pierwszych cyfr znaczących d_i

Źródło: opracowanie własne.

Odkrycie S. Newcomba zostało niezauważone przez 60 lat. W 1938 r. Frank Benford sprawdził empirycznie słuszność formuły Newcomba na wielu empirycznych zbiorach liczb, ale nie wyjaśnił, dlaczego w badanych zbiorach obserwuje się malejące częstości pojawiania się liczb zaczynających się od coraz to większej cyfry. Dopiero w 1995 r. matematyk Theodore Hill

wykazał, na czym polega istota tej prawidłowości oraz podał jej własności i uwarunkowania.

Tak więc nie wiadomo, czy prawo o rozkładzie pierwszych cyfr znaczących powinno być określane prawem Benforda, Newcomba czy Hilla, czy też prawem Newcomba-Benforda lub Benforda-Newcomba-Hilla czy jeszcze inaczej. Wątpliwości budzi również fakt powszechności działania tego prawa, a więc czy w pełni jest tu uzasadnione używanie terminu „prawo" czy też należałoby raczej stosować słabsze określenia, np. prawidłowość, formuła, reguła.

Jako ciekawostkę można podać, że w literaturze funkcjonuje też prawo Benforda dotyczące stopnia zainteresowania sprawami bulwersującymi, mówiące, że „zaciekawienie jest odwrotnie proporcjonalne do ilości realnie dostępnej informacji". Autorem tego prawa jest jednak wspomniany pisarz *science fiction* Gregory Benford i dotyczy ono psychologii oraz socjologii.

7.1. Badania Franka Benforda

Frank Benford, fizyk zatrudniony w laboratorium General Electrics, zainspirowany pracą S. Newcomba, podjął kilkuletnie badania mające na celu sprawdzenie, czy rzeczywiście mamy do czynienia z sytuacją, w której częściej na pierwszym miejscu liczb występują raczej cyfry mniejsze niż większe. W swojej pracy[2] opublikował wyniki analizy ponad 20 tysięcy liczb charakteryzujących 20 różnych zjawisk (m.in. powierzchnia rzek, liczba mieszkańców w miastach USA, dane adresowe osób z Amercian Men of Science, wyniki rozgrywek w baseballu, wskaźniki śmiertelności, liczby znajdujące się w artykułach zamieszczonych w „Readers Digest" itp.). W większości przypadków F. Benford uzyskał podobne wyniki, które doprowadziły go do wniosku o poprawności reguł opisanych przedstawionym wyżej wzorem.

[2] F. Benford, *The Law of Anomalous Numbers*, „Proceedings of the American Philosophical Society" 1938, t. 78, s. 551–572.

Tabela 2. Rozkłady pierwszych cyfr znaczących w zbiorach analizowanych przez F. Benforda, uporządkowane według wartości statystyki chi kwadrat

Nr	Symbol	Nazwa	1	2	3	4	5	6	7	8	9	n	chi kw	p
1	D	Newspapers	30,0	18,0	12,0	10,0	8,0	6,0	6,0	5,0	5,0	100	0,2	1,000
2	F	Pressure	29,6	18,3	12,8	9,8	8,3	6,4	5,7	4,4	4,7	703	1,3	0,996
3	R	Addresses	28,9	19,2	12,6	8,8	8,5	6,4	5,6	5,0	5,0	342	1,3	0,996
4	M	Reader's Digest	33,4	18,5	12,4	7,5	7,1	6,5	5,5	4,9	4,2	308	3,2	0,921
5	G	H,P, Lost	30,0	18,4	11,9	10,8	8,1	7,0	5,1	5,1	3,6	690	3,5	0,899
6	A	Rivers, Area	31,0	16,4	10,7	11,3	7,2	8,6	5,5	4,2	5,1	335	5,0	0,758
7	O	X-Ray Volts	27,9	17,5	14,4	9,0	8,1	7,4	5,1	5,8	4,8	707	5,4	0,714
8	T	Death Rate	27,0	18,6	15,7	9,4	6,7	6,5	7,2	4,8	4,1	418	7,6	0,473
9	Q	Blackbody	31,0	17,3	14,1	8,7	6,6	7,0	5,2	4,7	5,4	1165	9,5	0,302
10	I	Drainage	27,1	23,9	13,8	12,6	8,2	5,0	5,0	2,5	1,9	159	11,1	0,196
11	P	Am, League	32,7	17,6	12,6	9,8	7,4	6,4	4,9	5,6	3,0	1458	14,6	0,067
12	N	Cost Data	32,4	18,8	10,1	10,1	9,8	5,5	4,7	5,5	3,1	741	15,6	0,048
13	J	Atomic Wgt,	47,2	18,7	5,5	4,4	6,6	4,4	3,3	4,4	5,5	91	17,2	0,028
14	L	Design	26,8	14,8	14,3	7,5	8,3	8,4	7,0	7,3	5,6	560	19,2	0,014
15	C	Constants	41,3	14,4	4,8	8,6	10,6	5,8	1,0	2,9	10,6	104	24,4	0,002
16	S	n1,n2…n!	25,3	16,0	12,0	10,0	8,5	8,8	6,8	7,1	5,5	900	25,0	0,002
17	E	Specific Heat	24,0	18,4	16,2	14,6	10,6	4,1	3,2	4,8	4,1	1389	111,2	0,0000
18	B	Population	33,9	20,4	14,2	8,1	7,2	6,2	4,1	3,7	2,2	3259	118,6	0,0000
19	H	Mol, Wgt,	26,7	25,2	15,4	10,8	6,7	5,1	4,1	2,8	3,2	1800	125,8	0,0000
20	K	n-1 ,n1/2	25,7	20,3	9,7	6,8	6,6	6,8	7,2	8,0	8,9	5000	440,8	0,0000
		Benford	30,1	17,6	12,5	9,7	7,9	6,7	5,8	5,1	4,6			
		Średnia	30,6	18,5	12,3	9,4	8	6,4	5,1	4,9	4,8	1011	1,8	0,987
		SUMA	28,9	19,5	12,7	9,1	7,6	6,5	5,4	5,5	5,1	20229	85,4	0,0000

Źródło: opracowanie własne.

W tabeli 2 przytoczono wyniki analiz uzyskane przez F. Benforda. Zbiory danych uporządkowane są według malejących wartości statystyki chi kwadrat oraz odpowiadającemu tej statystyce prawdopodobieństwu *p* polegającemu na odrzuceniu prawdziwej hipotezy o zgodności danego rozkładu empirycznego z rozkładem Benforda. Jak można zauważyć w rozkładzie pierwszych cyfr znaczących cyfra 1 występuje z częstością 30%, następne cyfry pojawiają się coraz rzadziej, aż do cyfry 9, dla której częstość wynosi tylko 4,6%. Rozkład Benforda z dużą dokładnością (R^2=0,999) można aproksymować funkcją potęgową:

$$P(d_i) = 0{,}31331 * d_i^{-0{,}8631} \qquad (d_i = 1{,}2{,}...{,}9) \tag{9}$$

Inne funkcje, np. logarytmiczna, nie oddają tak dobrze przebiegu krzywej Benforda.

W tabeli 3 zestawiono wartości funkcji potęgowej z częstościami rozkładu Benforda. Różnice pomiędzy tymi rozkładami nie są duże – średnia z modułów różnic częstości wynosi 0,3%, natomiast średnia z modułów różnic częstości względnych (w stosunku do częstości w prawie Benforda) – 2%.

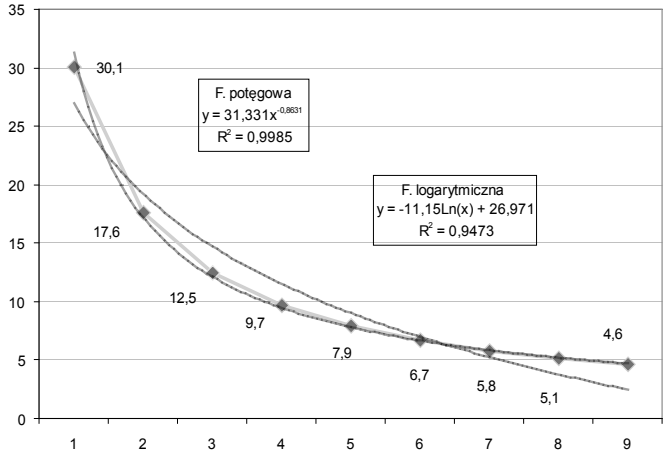
Tabela 3. Funkcja potęgowa P_i=33,33*d_i^-0,863 aproksymująca rozkład Benforda

d	F. potęg.	F. Benf.	P-B	(P-B)/B %
1	31,3	30,1	1,2	4,1
2	17,2	17,6	-0,4	-2,2
3	12,1	12,5	-0,4	-2,8
4	9,5	9,7	-0,2	-2,3
5	7,8	7,9	-0,1	-1,4
6	6,7	6,7	0,0	-0,3
7	5,8	5,8	0,0	0,7
8	5,2	5,1	0,1	1,8
9	4,7	4,6	0,1	2,8
	100,3	100,0	**0,3**	2,0

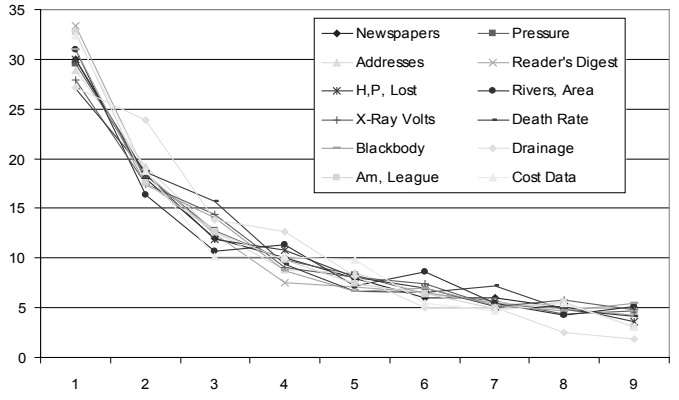
Źródło: opracowanie własne.

Na rysunku 10 podano wykresy częstości pierwszych cyfr znaczących dla 12 zbiorów danych analizowanych przez F. Benforda, w których zaobserwowano największą zgodność rozkładu empirycznego z teoretycznym. Prawdopodobieństwo popełnienia błędu odrzucenia hipotezy o zgodności porównywanych rozkładów kształtuje się w tych przypadkach na poziomie nie większym niż 0,05.

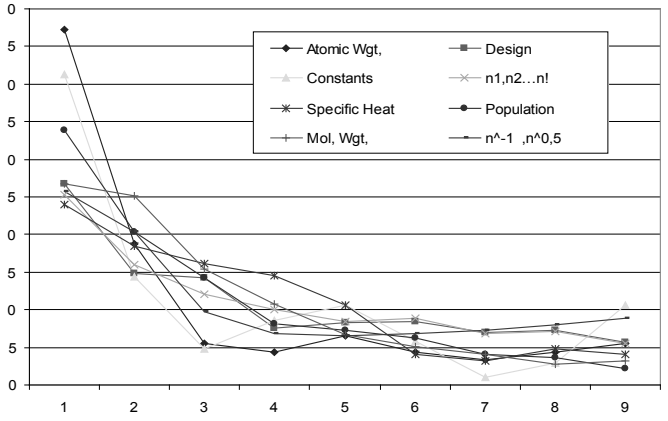
Na rysunku 11 przytoczono wykresy dla pozostałych ośmiu zbiorów danych, w których zgodność z rozkładem Benforda była stosunkowo mniejsza. W przypadku dwóch zbiorów: *(J) masy atomowej* oraz *(L) Dane z projektów* zgodność z prawem Benforda obserwuje się przy ostrzejszym poziomie istotności [>0,01]. Również w przypadku następnych dwóch zbiorów (*C – Stałe* oraz *S – potęgi liczb naturalnych*), obniżając poziom istotności do 0,001 można by przyjąć hipotezę o zgodności rozkładów. Jedynie w czterech ostatnich zbiorach (na 20 analizowanych) trzeba zdecydowanie odrzucić hipotezę o zgodności rozkładów cyfr znaczących z rozkładem Benforda.



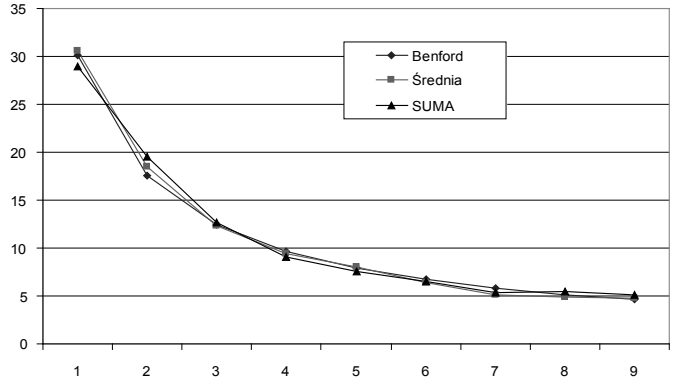
Rysunek 11. Prawo Benforda – rozkład częstości pierwszych cyfr znaczących
Źródło: opracowanie własne.



Rysunek 12. Rozkłady częstości pierwszych cyfr znaczących w 12 zbiorach FB **najbardziej** zgodnych z prawem Benforda
Źródło: opracowanie własne.



Rysunek 13. Rozkłady częstości pierwszych cyfr znaczących w 8 zbiorach FB **najmniej** zgodnych z prawem Benforda
Źródło: opracowanie własne.



Rysunek 14. Rozkłady częstości pierwszych cyfr znaczących według prawa Benforda i dla sumy oraz dla wartości średnich z 20 zbiorów FB
Źródło: opracowanie własne.

Na rysunku 14 przedstawiono rozkłady częstości dla zbioru sumarycznego (dla n=20229), powstałego jako mieszanka wszystkich 20 zbiorów danych (SUMA) oraz rozkład częstości ustalonych jako średnia z częstości dla poszczególnych zbiorów danych (Średnia). Sądząc z wykresu, obydwa te rozkłady charakteryzują się dużą wzajemną zgodnością, jak również zgod-

nością z prawem Benforda. Jednak rzut oka na statystyki chi kwadrat wskazuje, że wniosek o zgodności porównywanych rozkładów uzasadniony jest tylko dla rozkładu uśrednionego (Średnia), natomiast w przypadku rozkładu sumarycznego (SUMA) hipotezę o zgodności z rozkładem Benforda należy odrzucić. Związane to jest z liczebnością zbioru danych n. Liczebności te są bardzo zróżnicowane – liczebność sumaryczna przekracza 20 tysięcy, natomiast liczebność średnia jest 20x mniejsza.

Z podanego przykładu wynika, że wzrokowa, pobieżna ocena podobieństwa wykresów może być zawodna we wnioskowaniu o zgodności rozkładów. Trzeba zwrócić uwagę na wskazania mierników i testów statystycznych służących do obiektywnego ustalania, czy porównywane rozkłady są zbieżne, czy też różnice pomiędzy nimi są zbyt duże, aby je uznać za równorzędne.

Prawo Benforda powiązane jest ściśle z ciągiem Fibonacciego oraz mniej znanym ciągiem Lukasa. Ciągi te określone są rekurencyjnym wzorem:

$$F(i+1) = F(i) + F(\text{i-1}) \tag{10}$$

- ciąg Fibonacciego F(0)=0 F(1)=1 {1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144,…}
- ciąg Lukasa F(0)=2 F(1)=1 {1, 3, 4, 7, 11, 18, 29, 47, 76, 123, 199,…}

Analiza rozkładu 1475 liczb tych ciągów doprowadza do wniosku, że w obydwóch przypadkach rozkłady pierwszych cyfr znaczących są całkowicie zgodne z rozkładem Benforda (por. tabela 4). Na marginesie warto dodać, że więcej liczb niż 1475 w obydwóch ciągach nie udało się uzyskać z uwagi na przekroczenie maksymalnego rzędu dokładności w Excelu. Ostatnie (możliwie największe) wartości w tych ciągach wynoszą odpowiednio:

- dla ciągu Fibonacciego 4,99225460547777000E+307
- dla ciągu Lukasa 1,11630206588347000E+308

Tabela 4. Rozkłady pierwszych cyfr znaczących w ciągach Fibonacciego i Lukasa wraz z rozkładem Benforda dla n=1475

d	Benford	F0=0 F1=1	F0=2 F1=1	F0=F1=2	F0=3 F1=2
1	444	444	444	444	443
2	260	260	260	261	259
3	184	184	184	183	185
4	143	143	143	143	142
5	117	117	117	117	118
6	99	98	98	99	98
7	86	85	85	85	86
8	75	77	77	75	75
9	67	67	67	68	68
	1475	**1475**	**1475**	**1475**	**1474**
		Fibbonacci	**Lukas**		

Źródło: opracowanie własne.

Również dowolne inne ciągi, w których kolejny element jest sumą dwóch poprzednich elementów o dowolnych wartościach, są zgodne z rozkładem Benforda. W tabeli 4 podano rozkłady pierwszych cyfr dla dwóch innych ciągów, w których F(0)=F(1)=2 oraz F(0)=3 i F(1)=1. Różnice w rozkładach wynikają z zaokrągleń i nie przekraczają 1.

7.2. Popularność prawa Benforda na podstawie wskazań Google

Wpisując do wyszukiwarki Google hasła: „Benford", „Frank Benford", „Benford's Law", „Prawo Benforda" uzyskuje się coraz więcej odnośników. W tabeli 5 podano te liczby w ostatnich 4 latach, przy czym pomiarów dokonywano w miesiącach wakacyjnych 2007, 2009 i 2011 roku. Hasła podawano z reguły w dwóch opcjach: bez ograniczeń językowych oraz tylko w języku polskim.

Tabela 5. Liczba wskazań w Google przy hasłach związanych ze słowem „Benford" w latach 2007–2011

Liczba wskazań Google na hasło	2011	2009	2007
„Benford"	3 530 000	814 000	
„Benford" – język polski	580 000	160 000	
%	16,4	19,7	
„Frank Benford"	17 000	14 600	1 800
Frank Benford" – język polski	300	450	13
%	1,8	3,1	0,7
„Benford's Law"	154 000	44 000	39 000
„Benford's Law" – język polski	490		
„Prawo Benforda"	610		
„Prawo Benforda" – język polski	580	690	50

Źródło: opracowanie własne.

(*) Hasło „Benford's Law" w 2005 roku – 10 000 trafień.

Wśród pierwszych trafień na hasło „Benford" można spotkać strony dotyczące maszyn budowlanych produkowanych przez firmę Terex (Benford to marka pełnoobrotowego miniwozidła drogowego oraz walca wibracyjnego), publikacji pisarza *science fiction* Gregory'ego Benforda, autora podręczników języka angielskiego Michaela Benforda, a także witryny zawierające informacje o innych osobach o tym nazwisku, np. Jay Benford, Alec Benford, Mark Benford itp. Wysoko znajdują się witryny poświęcone synowi Harrisona Forda, właścicielowi sieci restauracji w USA – Ben Ford(owi), a także o tym samym nazwisku – leworęcznej gwieździe baseballa. Dwa pierwsze wyniki (w języku angielskim) odnoszą się do haseł w Wikipedii na temat „Frank Benford" oraz „Benford's law". W języku polskim na pierwszym miejscu jest odnośnik do hasła „Frank Benford" w polskiej Wikipedii.

Na hasło „Benford" wyszukiwarka Google podpowiada z reguły następujące terminy:

… online, … law,… gregory centrum galaktyki, …, części, …dumper, … terex.

Witryny dotyczące tematyki prawa Benforda uzyskuje się, podając zawężone hasła, np. „Frank Benford", „Benford's law". W pierwszym przypadku liczba tych odnośników przez ostatnie cztery lata wzrosła na świecie dziesięciokrotnie, z 2 tysięcy w 2007 r. do 17 tysięcy w 2011 r. Jeszcze większy

przyrost trafień nastąpił w języku polskim, z 13 w 2007 roku do 300 w roku 2011. W relacji do zasobów światowych udział witryn w języku polskim nie jest duży i zawiera się w granicach 2–3%. Ciekawy jest fakt spadku odnośników w języku polskim w roku 2011 w stosunku do roku 2009. Dotyczy to zarówno hasła „Frank Benford", jak i hasła „Prawo Benforda".

Na hasło „Benford's Law" wyszukiwarka Google udziela następujących podpowiedzi:

…Excel, …examples, …Wiki, …proof, …of controversy, …explanation, … auditing,

…calculator, …applications, …accounting.

W tabeli 6 podano pięć pierwszych witryn pojawiających się w odpowiedzi na hasło „Benford's law". Kolejność tych witryn jest prawie identyczna w badanych latach. Na pierwszym miejscu jest Wikipedia, następnie portale matematyczne Wolfram MathWorld, MathPages, Intuitor, oraz witryny domowe konsultantów, np. audytora Rexforda Swaina. Zawężając w wyszukiwarce poszukiwania do określonego typu zasobów, na hasło „Benford's law" uzyskuje się (IX 2011) 4700 odnośników do plików w pdf oraz 230 prezentacji w ppt.

Tabela 6. Pięć pierwszych witryn pojawiających się w odpowiedzi na hasło „Benford's law"

Adres witryny	2011	2009	2007
www.en.wikipedia.org/wiki/Benford's_law	1	1	1
www.mathworld.wolfram.com/BenfordsLaw.html	2	2	2
www.rexswain.com/benford.html	3	3	5
www.mathpages.com/HOME/kmath302/kmath302.htm	4	4	4
www.intuitor.com/statistics/Benford's%20Law.html	5	5	3

Źródło: opracowanie własne.

W języku polskim na hasło „Prawo Benforda" na pierwszych pięciu miejscach (tabela 7) podawane są witryny polskiej Wikipedii, blogów na witrynie wordpress.com, strona domowa, portal WSIZ w Rzeszowie oraz portal polskich neuroinformatyków.

Tabela 7. Pięć pierwszych witryn pojawiających się w odpowiedzi na hasło „Prawo Benforda"

Lp	Adresy witryn
1	www.pl.wikipedia.org/wiki/Rozkład_Benforda
2	www.dataminingalapolonaise.wordpress.com/2010/01/06/prawo-benforda/
3	www.ipipan.waw.pl/~ldebowsk/uslugi/index.html
4	portal.wsiz.rzeszow.pl/plik.aspx?id=2618
5	www.neuroinf.pl/Members/danek/swps/2008/Article.2008-05.../getFile

Źródło: opracowanie własne.

7.3. Problematyka rozkładu cyfr znaczących w publikacjach naukowych

W 2007 r. pojawiła się w internecie witryna www.benfordonline.net zawierająca wykaz ważniejszych publikacji z zakresu problematyki dotyczącej prawa Benforda. Jej inicjatorem i głównym realizatorem jest Ted Hill, który wykorzystał zasoby bibliograficzne wcześniej zgromadzone przez M. Nigriniego oraz W. Hurlimana.

Na koniec sierpnia 2011 r. wykaz zawierał 621 pozycji, które można ujmować w porządku chronologicznym, alfabetycznym (według tytułów) oraz według nazwisk autorów. Każda pozycja zawiera obok nazwisk autorów, tytułów, miejsca i roku wydania także informacje dotyczące:

- syntetycznego streszczenia zawartości (autorzy Bibliografii proszą o propozycje w tym zakresie na specjalnym formularzu, zarówno co do prac znajdujących się na witrynie, jak i propozycji uwzględnienia nowych);
- adres witryny, gdzie dana pozycja jest dostępna w pełnej wersji elektronicznej,
- wykaz prac z Bibliografii, które cytują daną pracę,
- wykaz prac z Bibliografii, które cytuje dana praca.

Na podstawie zestawienia prac opublikowanych w latach początkowych 1881–1970 (37 pozycji) oraz w dwóch ostatnich latach 2009–2010 (66 pozycji) wyznaczono rozkład liczby prac opublikowanych w poszczególnych latach okresu 1881–1970. Od opublikowania pierwszej pracy S. Newcomba minęło 130 lat. Przez połowę tego okresu (do 1945 roku) opublikowano 9 prac, tj. 1,5% dotychczasowego dorobku. Druga praca w zakresie omawianej problematyki ukazała się w 1912 roku po ponad 30 latach. W latach

1920–1938 nie pojawiła się na świecie ani jedna praca dotycząca prawa Benforda. W latach 1940–1970 łącznie opublikowano tylko 30 prac, czyli średnio jedną pracę w ciągu roku.

Tabela 8. Liczba ważniejszych prac na temat prawa Benforda opublikowanych w latach 1881–1970

Rok	L. prac	Odstęp czasowy	Skum. l. prac	Rok	L. prac	Odstęp czasowy	Skum. l. prac
1881	1		1	1952	1	2	14
1912	1	31	2	1956	2	4	16
1916	1	4	3	1957	1	1	17
1917	1	1	4	1961	3	4	20
1920	1	3	5	1963	1	2	21
1938	1	18	6	1964	1	1	22
1939	1	1	7	1965	2	1	24
1944	1	5	8	1966	1	1	25
1945	1	1	9	1967	1	1	26
1946	1	1	10	1968	2	1	28
1948	2	2	12	1969	7	1	35
1950	1	2	13	1970	2	1	37

Źródło: opracowanie własne.

Dopiero po roku 1970 liczba publikacji na temat prawa Benforda zaczyna wzrastać, a od roku 1998 następuje gwałtowny ich wzrost. Na rys. 15–17 przytoczono wykresy ilustrujące kształtowanie się liczby publikacji zebranych w omawianej Bibliografii:

– w całym okresie 1881–2009 (rysunek 15),

– w okresie 1970–2009 (rysunek 16)

– w zagregowanych odcinkach czasu (rysunek 17).

Z wykresów tych wynika, że rozwój zainteresowania problematyką prawa Benforda kształtuje się według funkcji wykładniczej.

W tabeli 9 podano wykaz nazwisk autorów największej liczby prac znajdujących się w wykazie www.benfrodonline.net. Zdecydowanymi liderami na tej liście są: M. Nigrini, P. Schatte, T. Hill (po 17–18 prac). Kolejne miejsca zajmują: A. Berger i S. Miller (po 11 prac). Łączna liczba autorów wszystkich 621 prac w Bibliografii wynosi 700. Różnica wynika z faktu występowania prac współautorskich.

Tabela 9. Wykaz autorów o największej liczbie opublikowanych prac na temat prawa Benforda

Lp.	Autor	L. prac
1	Nigrini, MJ	18
2	Schatte, P	18
3	Hill, TP	17
4	Berger, A	11
5	Miller, SJ	11
6	Nagasaka, K	7
7	Turner, PR	7
8	Bhattacharya, S	6
9	Feldstein, A	6

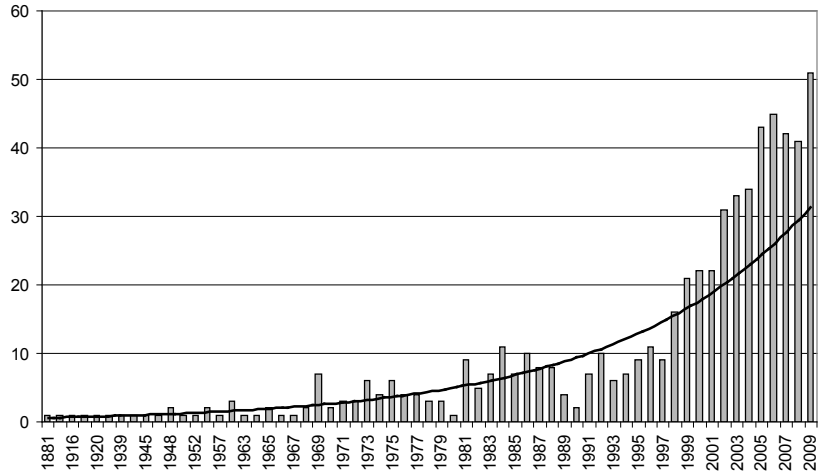
Lp.	Autor	L. prac
10	Baird, JC	5
11	Guan, L	5
12	Lu, F	5
13	Ma, BQ	5
14	Posch, PN	5
15	Shao, L	5
16	Shiue, JS	5
17	Uppuluri, VRR	5

Źródło: opracowanie własne.

Warto zwrócić uwagę, że najczęściej cytowana przez autorów występujących w omawianej Bibliografii jest klasyczna praca F. Benforda z 1938 r. *The law of anomalous numbers*, („Proceedings of the American Philosophical Society" t. 78, s. 551–572). Praca ta jest cytowana przez autorów 240 prac spośród wszystkich 621 prac (to jest 39% ogólnej liczby autorów). Na drugim miejscu jest praca S. Newcomba z 1881 roku *Note on the frequency of use of the different digits in natural numbers* („American Journal of Mathematics", nr 4(1), s. 39–40), którą cytowało 201 autorów innych prac (32%).

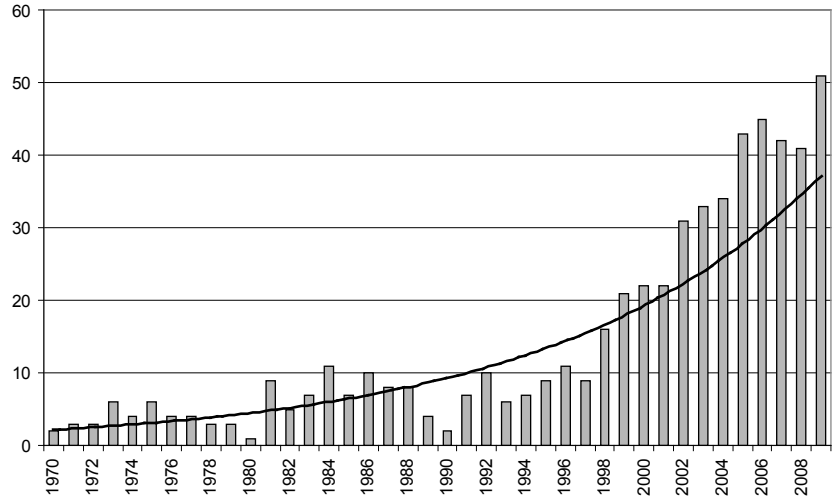
F. Benford w swoim tekście nie zacytował żadnej z pięciu wcześniejszych prac poruszających podobne tematy (ani też nie zacytował żadnej innej publikacji). Co więcej, w następnych latach nie napisał żadnej innej pracy z tego zakresu. Z punktu widzenia historii nauki przypadek ten niewątpliwie zasługuje na uwagę.

Warto tu nadmienić, że S. Newcomb w swojej pracy używa pojęcia liczb naturalnych, natomiast F. Benford stosuje termin: *anomalous number*. Można by z tego wnioskować, że F. Benford uważał, iż jego formuła opisuje raczej anomalie niż sytuacje typowe, podczas gdy prawa naukowe powinny z zasady mieć charakter uniwersalny.



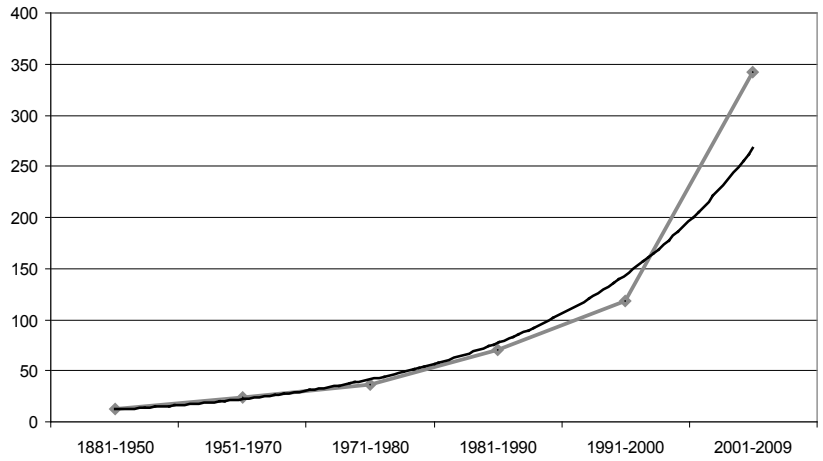
Rysunek 15. Liczba prac dotyczących rozkładu Benforda opublikowanych w latach 1881–2009

Źródło: opracowanie własne.



Rysunek 16. Liczba prac dotyczących rozkładu Benforda opublikowanych w latach 1970–2009

Źródło: opracowanie własne.



Rysunek 17. Liczba opublikowanych prac dotyczących rozkładu Benforda w zagregowanych odcinkach czasowych
Źródło: opracowanie własne.

Również interesujący jest przypadek napisanej w latach 1945–1946 lecz nieopublikowanej pracy G.J. Stiglera *The distribution of leading digits in statistical tabels*. Autor podał w tej pracy istotną modyfikację prawa Benforda, na którą powołują się autorzy innych prac i która stanowi w wielu sytuacjach alternatywę w stosunku do rozkładu Benforda. G. Stigler w latach II wojny światowej pracował w projekcie Manhattan jako matematyk i statystyk. Jako pierwszy (1961) zwrócił uwagę[3] na wartość informacji i stąd uważany jest za inicjatora tej dziedziny ekonomii. W 1982 r. nagrodzony Nagrodą Nobla w zakresie ekonomii jako twórca chicagowskiej szkoły ekonomii, teorii regulacji, teorii *job search.*

7.4. Sylwetki twórców prawa Benforda

Poniżej przytoczono krótkie charakterystyki osób, które szczególnie przyczyniły się do powstania i popularności prawa Benforda. Należy do nich zaliczyć przede wszystkim nieżyjących już twórców prawa – S. Newcomba oraz F. Benforda. Jak się wydaje, do tej listy można dołączyć także T. Hilla oraz M. Nigriniego, którzy od wielu lat prowadzą najbardziej intensywne badania nad prawami rozkładu cyfr znaczących.

[3] G. Stigler, *The Economics of Information*, „Journal of Political Economy" 1961, t. 69.

Simon Newcomb (1835–1909)

- Studiował frenologię, naukę o powiązaniu kształtu czaszek z osobowością.
- W wieku 21 lat zatrudniony jako *computer* (osoba ds. obliczeń) w Nautical Almanac Office, Cambridge, Massachusetts.
- Później dyrektor Nautical Almanac Office.
- Profesor matematyki i astronomii w John Hopkins University.
- Przyczynił się do pomiaru szybkości światła, ustalenia dokładnej orbity Księżyca.
- Stworzył system stałych astronomicznych.
- Założyciel i prezydent: American Astronomical Society, American Mathematical Society, American Association for the Advancement of Science, Philosophical Society of Washington.
- Najbardziej znany amerykański astronom. Otrzymał najwyższe nagrody naukowe w zakresie astronomii w USA, Wielkiej Brytanii, Holandii i Niemczech.
- W Kanadzie przyznawana jest przez Royal Astronomical Society nagroda jego imienia – Simon Newcomb Award.
- Tekst na temat rozkładu pierwszych cyfr znaczących ze wzorem p(d)=log(1+1/d) liczył 2 strony i przez 60 lat nie został przez nikogo dostrzeżony.
- W 1885 r. przyczynił się do powstania prawa Irvinga Fishera (1911) w zakresie ilościowej teorii pieniądza (*Quantity Theory of Money*), wyrażonej wzorem MV=PT, gdzie M – ilość pieniądza w obiegu, V – szybkość obiegu pieniądza (transakcji), P – poziom cen, T – liczba transakcji. Z podanego wzoru wynika, że jeżeli relacja T do V jest stała, to wzrost pieniądza w obiegu powoduje wzrost cen (inflację).

Frank Benford (1883–1948)

- Amerykański inżynier, fizyk, elektrotechnik.
- Po ukończeniu Uniwersytetu w Michigan w 1910 r. pracował w firmie General Electric, najpierw przez 18 lat w Illuminating Engineering Lab, a potem przez 20 lat w Research Lab.
- W 1937 r. skonstruował instrument pomiaru załamywania światła.
- Ekspert pomiarów optycznych, autor 109 dokumentów z zakresu optyki i matematyki.
- Opracował 20 patentów na przyrządy optyczne.

- W 1938 r. opublikował jedyną w swoim dorobku pracę na temat rozkładu cyfr znaczących. Nie wykorzystał w niej informacji wynikających z wcześniejszej pracy S. Newcomba z 1881 r. Publikacja ta nie została zauważona przez 30 lat.

Theodore Preston Hill (ur. 1943)

- Matematyk zatrudniony w George Institute of Technology, specjalizujący się w teorii prawdopodobieństwa, a w szczególności w prawie Benforda.
- Absolwent Akademii West Point (1966), rocznika o największym procencie ofiar wojny w Wietnamie, uczestnik tej wojny.
- Autor prac dotyczących wyboru najlepszego obiektu (problem łowcy posagu, problem sekretarki) ze skończonego zbioru istotnie różnych propozycji, prezentowanych w losowej kolejności (reguła zatrzymywania, *optimal stopping theory*).
- Autor prac związanych z problemem sprawiedliwego podziału *(fair division problem)* oraz strategią *Cut-And-Choose.*
- Twórca serwisu internetowego benfordonline.net, zawierającego bibliografię prac związanych z prawem Benforda, a także dwóch innych portali motherfunctor.org oraz earlyamericanmathbooks.org.

Pełna dokumentacja dorobku naukowego znajduje się na witrynie tphill.net.

Mark Nigrini

- Profesor w St. Michael's College of New Jersey, specjalizacja: rachunkowość finansowa, rachunkowość zarządcza, audyt, zastosowania matematyki, systemy informatyczne, wykrywanie oszustw finansowych.
- Członek American Accounting Association, Institute of Internal Auditors.
- W 1992 r. obronił pracę doktorską (University of Cincinnati) na temat *The Detection of Income Tax Invasion Through an Analysis of Digital Distributions.*
- Analiza zwrotów podatków Billa Clintona w latach 1977–1992.
- Pierwszy z naukowców, który próbuje zarobić pieniądze na znajomości prawa Benforda – płatne seminaria naukowe (>150 $), książki w cenie 32 centy za stronę.
- Zebrał wiele przypadków opisujących oszustwa finansowo-księgowe.
- Wyczerpujące informacje na temat M. Nigriniego i jego prac znaleźć można na witrynie nigrini.com.

W tabeli 10 przedstawiono kalendarium ważniejszych wydarzeń, jakie miały miejsce w zakresie prawa Benforda przed rokiem 2000. Wskazano tu na rolę twórców tego prawa, a także na osiągnięcia kilku innych badaczy, którzy przyczynili się w początkowym okresie do rozwoju, popularyzacji i zastosowań prawa pierwszych cyfr znaczących.

Podsumowanie

Serwis „New Scientist" zaliczył prawo Benforda do jednego z pięciu najbardziej znanych odkryć naukowych o nieodpowiedniej nazwie[4]. Na tej liście (poza prawem Benforda) podano następujące odkrycia.

1. Bakteria salmonella, którą odkrył Theobald Smith – młody pracownik laboratorium kierowanego przez Daniela Salmona.
2. Kometa Halleya, znana już astronomom chińskim w III w., a także J. Keplerowi 75 lat przed tym, jak Halley sformułował hipotezę o okresowości tej komety.
3. Równanie Arrheniusa, opisujące zależność między szybkością reakcji a temperaturą i energią aktywacji, zostało najpierw opisane przez kinetyka holenderskiego van Hoffa, natomiast Arrhenius 5 lat później (powołując się na prace van Hoffa) wyjaśnił istotę tej zależności.
4. Choroba Hansena, czyli trąd. G.A. Hansen wprawdzie pierwszy odkrył bakterie powodujące tę chorobę, ale dopiero jego kolega A. Neissser udowodnił, że bakterie odkryte przez G. Hansena faktycznie powodują trąd.

Na powyższej liście znajduje się też prawo Benforda *(first digit law)* opisujące rozkład częstości występowania **pierwszych cyfr znaczących** (wiodących, *leading digit*) w dużych zbiorach liczb, w miarę możliwości wielocyfrowych, pochodzących z realnych pomiarów oraz dotyczących empirycznych zjawisk i procesów.

[4] http://www.newscientist.com/article/dn14461-five-scientific-discoveries-that-got-the-wrong-name.html? DCMP=ILC-hmts&nsref=news10_head_dn14461.

Tabela 10. Kalendarium ważniejszych osiągnięć w zakresie prawa Benforda przed 2000 rokiem

Rok	Nazwisko	Wyszczególnienie	Publikacja
1881	S. Newcomb	Pierwsza praca na temat rozkładu cyfr znaczących. Niezauważona przez 60 lat.	*Note on the frequency of use of the different digits in natural numbers*, „American Journal of Mathematics" 4(1)/1881, 39–40.
1938	F. Benford	Fundamentalna praca dająca początek teorii i analizie rozkładu cyfr znaczących.	*The law of anomalous numbers*, „Proceedings of the American Philosophical Society" 78/1938, 551–572.
1944	S.A. Goud-smith W.H. Furry	Teza, że prawo Benforda wynika ze sposobu zapisu liczb.	*Significant figure of numbers in statistical tables*, „Nature", 154/1944, p. 800–801.
1945	G.J. Stigler	Modyfikacja formuły wyznaczania pierwszej cyfry znaczącej oparta na innych założeniach niż formuła Benforda.	*The distribution of leading digits in statistical tabels* (praca nieopublikowana).
1948	L.V. Furlan	Prawo Benforda oddaje harmoniczną istotę rzeczywistości (skala logarytmiczna).	*Das Harmoniesgestez der Statistik, Eine Untersuchung uber die metrische Interdependenz der sozialeen Erscheinungen*, Bael, Switzerland, Verlag fur Recht und Geselschaft, XIII/1948.

Rok	Nazwisko	Wyszczególnienie	Publikacja
1961	R.S. Pinkham	Dowód na niezmienniczość skali rozkładu Benforda względem dowolnej operacji arytmetycznej (mnożenia, dzielenia, potęgowania).	*On the Distribution of First Significant Digits*, „Annals of Mathematical Statistics" 32(4)/1961, 1223–1230.
1969	R.A. Raimi	Popularyzacja problematyki prawa Benforda.	*The Peculiar Distribution of First Digits*. „Scientific American" 221(6)/1969, 109–119; *On Distribution of First Significant Figures*, „American Mathematical Monthly" 76(4)/1969, 342–348.
1988	C. Carslaw	Pierwsze zastosowania prawa Benforda do analizy danych finansowych w poszukiwaniu błędów.	*Anomalies in Income Numbers: Evidence of Goal Oriented Behavior*, „The Accounting Review" 63(2)/1988, 321–327.
1993	M.J. Nigrini	Pierwsze próby wykorzystania prawa Benforda w audycie księgowym.	*Can Benford's law be used in forensic accounting?*, The Balance Sheet, VI/1993, 7–8; *Using digital frequencies to detect fraud*, „Fraud Magazine.The White Paper Index" 8(2)/1994, s. 3–6; *A taxpayer compliance application of Benford's law*, „Journal of the American Taxation Association" 18(1)/1996, s. 72–91.

Rok	Nazwisko	Wyszczególnienie	Publikacja
1995	T.P. Hill	Dowód na niezmienniczość podstawy systemu liczbowego, którą to własność rozkład Benforda posiada jako jedyny.	*Base-Invariance Implies Benford's Law*, „Proceedings of the American Mathematical Society" 123(3)/1995, s. 887–895; *The Significant-Digit Phenomenon*, „American Mathematical Monthly" 102(4)/1995, s. 322–327.
1996	T.P. Hill	Prawo Benforda opisuje „rozkład rozkładów": losowo dobrane próby z losowo dobranych rozkładów.	*A Statistical Derivation of the Significant-Digit Law*, „Statistical Science" 10(4)/1996, s. 354–363; *The first digital phenomenon*, „American Scientist" 86/1998, s. 358–363.
1996	E. Leo	Pierwsze zastosowania prawa Benforda do analizy danych giełdowych.	*On the Peculiar Distribution of the US Stock Indexes' Digits*, „American Statistician" 50(4)/1996, s. 311–313;
1997	M.J. Nigrini	Powstaje *digital analysis* – system procedur do analizy własności rozkładów cyfr i ich kombinacji w celu poszukiwania nietypowych wartości	*The use of Benford's Law as an aid in analytical procedures*, „Auditing. A Journal of Practice & Theory" 16(2)/1997, s. 52–67 (M.J. Nigrini, L.J. Mittermaier); *Numerology for Accountants*, „Journal of Accountancy", November 1998, s. 15; *Adding value with digital analysis*, „The Internal Auditor" 56(1)/1999, s. 21–23.

Źródło: opracowanie własne.

8. Empiryczne prawa liczbowe w nauce

Marzena Farbaniec, Tadeusz Grabiński, Bartłomiej Zabłocki, Wacław Zając

Wprowadzenie

Poza prawem rozkładu pierwszych cyfr znaczących odkrytym niezależnie przez S. Newcomba (1881) oraz F. Benforda (1938) w wielu dziedzinach nauki posługujemy się podobnymi regułami. Prawa te na ogół opisują rzeczywistość w dużym przybliżeniu na zasadzie powszechnie znanej prawdy wynikającej z doświadczenia i praktyki i noszą nazwę tzw. reguły kciuka (*rule of thumb*)[1]. Poniżej przytacza się kilka takich reguł głównie z obszaru ekonomii i informatyki.

8.1. Reguły kciuka

Ekonomia

Reguła 72 (niekiedy określana jako reguła 70 lub reguła 69) służy do oceny czasu podwojenia kapitału[2] oprocentowanego na stałym poziomie r(%). Na przykład wyjściowy kapitał oprocentowany w skali rocznej na 8% podwoi się po 72/8=9 latach, przy oprocentowaniu rocznym 6% czas podwojenia kapitału wynosi 72/6=12 lat itd.

Czas podwojenia przy stopie r dany jest formułą T=ln(2)/ln(1+r). W tabeli 11 podano dokładny czas podwojenia T (w latach) przy stopie procentowej r od 1% do 12% (w skali roku) oraz dla różnych podstaw reguły: 72--70-69. W ostatnim wierszu przytoczono średni błąd procentowy modułów różnic pomiędzy faktycznymi okresami podwojenia, a czasami wynikający-

[1] Nazwa pochodzi z czasów, kiedy kciuk był uproszczonym narzędziem i jednostką miary. Na przykład zgodnie z regułą browarnika, jeżeli palec włożony do brzeczki nie parzył, to można było dodawać drożdże. Reguła kapitana statku stosowana była przy nawigacji wzdłuż wybrzeża i polegała na nie zbliżaniu się do linii wybrzeża na odległość kciuka, aby nie wpaść na rafy. Reguła stołu wyrażała się w ułożeniu talerzy od krawędzi stołu na odległość między kciukiem a wskazującym palcem. Według nie potwierdzonych źródeł w prawodawstwie angielskim do końca XIX wieku obowiązywała zasada, że mąż nie mógł bić swojej żony kijem grubszym od swojego kciuka.

[2] Pierwsze wzmianki o tej regule pochodzą z pracy Luca Pacioli (1445–1514) *Summa de Arithmetica* (1494).

mi z poszczególnych reguł. Jak się okazuje, najlepsze wyniki (średni błąd: 1,5%) uzyskuje się, jeżeli za podstawę reguły przyjmie się liczbę 72. Ponadto liczba ta jest bardzo wygodna, gdyż dzieli się bez reszty przez 1,2,3,4,6,8,9,12.

Tabela 11. Czasy podwojenia kapitału wynikające z reguł 72-70-69

%	T	72	70	69
1	69,7	72,0	70,0	69,0
2	35,0	36,0	35,0	34,5
3	23,4	24,0	23,3	23,0
4	17,7	18,0	17,5	17,3
5	14,2	14,4	14,0	13,8
6	11,9	12,0	11,7	11,5
7	10,2	10,3	10,0	9,9
8	9,0	9,0	8,8	8,6
9	8,0	8,0	7,8	7,7
10	7,3	7,2	7,0	6,9
11	6,6	6,5	6,4	6,3
12	6,1	6,0	5,8	5,8
Błąd średni %		**1,5**	**2,2**	3,5

Źródło: opracowanie własne.

Podobne reguły można sformułować na potrojenie kapitału, czterokrotne zwiększenie itd. Odpowiednie podstawy i formuły mają tu postać – dla trzykrotności T=114/r, natomiast dla czterokrotności T=144/r.

Innymi regułami w obszarze ekonomii są:

- Reguła Okuna: każdemu wzrostowi bezrobocia o 1% towarzyszy spadek potencjalnego GDP o 2%.
- Reguła „nafciarska": długoterminowa cena ropy naftowej to 3,5-krotność kosztów poszukiwań i wydobycia (F&D costs).
- Reguła 2 i 3 sigm: 68% danych znajduje się w przedziale ± 2 odchyleń standardowych od średniej, a 95% danych w przedziale ± 3 sigm.

Informatyka

- Reguła Moore'a: moc obliczeniowa komputerów podwaja się co 24 miesiące (przy tym samym koszcie). Dotyczy to liczby tranzystorów w stosunku do powierzchni układu scalonego, mocy obliczeniowej do kosztu, rozmiarów RAM, pojemności dysków twardych, przepustowości sieci itd.

- Reguła Wirtha: oprogramowanie staje się wolniejsze szybciej, niż sprzęt staje się szybszy, np. proces rozruchu komputera z nowoczesnym systemem operacyjnym na nowoczesnym PC trwa coraz dłużej.
- Reguła Gatesa: szybkość oprogramowania maleje o połowę co 18 miesięcy.

Inne reguły kciuka

- Reguła Hellina: bliźnięta rodzą się raz na 89 ciąż, trojaczki raz na 89^2 ciąż, natomiast czworaczki raz na 89^3 ciąż.
- Reguła Cornegie College: na każdą godzinę spędzoną na zajęciach zorganizowanych student powinien przeznaczyć 2–3 godziny pracy własnej.
- Reguła odległości od pioruna: każda sekunda od chwili zobaczenia błyskawicy do momentu usłyszenia grzmotu pomnożona przez 300 metrów.

Wymienione powyżej przykładowe reguły nie są bezpośrednio związane z prawami rozkładu cyfr znaczących. Mają jednak z nimi wspólną cechę, polegającą na empirycznym charakterze i wynikaniu z zaobserwowanych prawidłowości statystycznych, zazwyczaj opartych na prawie wielkich liczb. Poniżej przedstawia się bardziej szczegółowo inne empiryczne prawa „liczbowe", mające już ścisły związek z prawem Newcomba-Benforda oraz twórców tych praw (por. tabela 8).

8.2. Ciągi Fibonacciego

Leonardo Fibonacci (1175–1250)

- Włoski matematyk.
- *Liber Abaci* (1202) – opis systemu pozycyjnego, arytmetyka liczb całkowitych, tablica z zapisem liczb rzymskich i indyjskich.
- Ciąg Fibonacciego przytoczony w pracy *Liber Abaci*, znany wcześniej matematykom hinduskim – Gopala (1135), Hemachandra (1150).
- *Practica geometriae* (1220) – połączenie algebry, geometrii i trygonometrii.
- Sposoby mnożenia liczb tzw. próbą dziesiątkową.
- Rozkład liczb na czynniki pierwsze, cechy podzielności.
- Arytmetyka handlowa oparta na proporcjach.
- Zadania na mieszaninę (ustalenie składników dających stop określonej próby).

- Reguła towarzystwa (podział wielkości proporcjonalnie do części uczestników podziału).
- Reguła poziomów wartości (*figura cata*).

Ciągi Fibonacciego mają silny związek, a w sensie rozkładu pierwszej cyfry znaczącej są nawet tożsame z prawem Benforda. Są to ciągi liczb naturalnych określonych rekurencyjną formułą:

$$F_{i+1} = F_1 + F_{1-2} \tag{11}$$

przy czym zakłada się, że F1=F2=1.

Ze wzoru (11) wynika, że każdy następny wyraz ciągu Fibonacciego jest sumą dwóch poprzednich. Ciąg ten został podany przez Fibonacciego w 1202 r. jako rozwiązanie zadania o rozmnażaniu królików[3].

Ciąg posiada wiele interesujących własności. Poniżej przykładowo podano kilka z nich[4].

- *n*-ty wyraz ciągu wyrażony wzorem Bineta:

$$F_n = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^2 - \left(\frac{1-\sqrt{5}}{2}\right)^2\right] = \frac{\left(1+\sqrt{5}\right)^2 - \left(1-\sqrt{5}\right)^2}{2^n\sqrt{5}} \tag{12}$$

- suma *n* wyrazów ciągu:

$$\sum_{i=1}^{n} F_i = F_{n+2} - 1 \qquad \sum_{i=0}^{n} iF_i = nF_{n+2} - F_{n+3} + 2 \tag{13, 14}$$

- wyraz ciągu jako suma kwadratów wyrazów sąsiednich:

$$F_{2i} = F_{i+1}^2 - F_{i-1}^2 \qquad F_{2i-1} = F_i^2 + F_{i-1}^2 \tag{15, 16}$$

[3] Każda para królików rodzi co miesiąc nową parę młodych królików. Okres rozrodczy królików trwa 2 miesiące. Na początku mamy jedną parę królików [1]. Po miesiącu rodzi się druga para [1] Po dwóch miesiącach rodzą się 2 nowe pary królików [2]. W kolejnym miesiącu rodzą się 3 pary królików (dwie pary z rodziców, i jedna para z „dziadków"), itd. Ciąg Fibonacciego podaje liczbę nowo narodzonych par królików w kolejnych miesiącach. Suma elementów ciągu określa liczebność populacji królików (przy założeniu, że króliki nie umierają).

[4] Więcej informacji na temat ciągów Fibonacciego znaleźć można m.in. w artykułach zamieszczanych na łamach kwartalnika „Fibonacci Quarterly", a także w pracach T. Koshy, *Fibonacci and Lucas Numbers with Applications*, Wiley 2001; L.C. Washington, *Benford's Law for Fibonacci and Lucas Number*, „Fibonacci Quarterly" 1981, t. 19, s. 175–177.

W literaturze można spotkać wiele modyfikacji ciągu Fibonacciego, np.

- Ciąg Fibonacciego, w którym F1=0 oraz F2=1.
- Ciąg Lukasa, w którym F1=2 oraz F2=1.
- Ciąg Tribonacciego, w którym każdy kolejny element powstaje przez zsumowanie trzech poprzedzających go elementów, przy czym F1=0, F2=0, F3=1.
- Ciąg Tetranacciego, w którym każdy kolejny element powstaje przez zsumowanie czterech poprzedzających go elementów, przy czym F1=0, F2=0, F3=0, F4=1.

W tabeli 12 oraz na rysunku 18 przytoczono początkowe wartości ciągów Fibonacciego. Poza ciągami wymienionymi powyżej, przykładowo podano też ciągi dla F1=F2=2 oraz F1=1 i F2=2. Jak można zauważyć, wszystkie te ciągi mają podobną postać i dają się aproksymować funkcją wykładniczą. Na rysunku 18 zamieszczono funkcję wykładniczą dopasowaną do ciągu Fibonacciego F1=F2=1 ze współczynnikiem determinacji R2=0,994.

Ilorazy sąsiednich elementów ciągów Fibonacciego dążą w granicy do tzw. złotej liczby. Liczba ta dana jest wzorem:

$$\varphi = \frac{F_{i+1}}{F_i} \rightarrow \frac{\sqrt{5}+1}{2} = 1{,}61804 \tag{17}$$

i wyznacza tzw. złoty podział (podział harmoniczny, boska proporcja) odcinka. Jest to podział odcinka na dwie części [a;b] takie, że stosunek długości części dłuższej [a] do krótszej [b] jest taki sam jak stosunek długości całego odcinka [a+b] do części dłuższej [a].

$$\frac{a+b}{a} = \frac{a}{b} = \varphi \tag{18}$$

Złota liczba występuje w wielu sytuacjach. Poniżej przytacza się kilka przykładów.

- Proporcje piramidy w Gizie (stosunek wysokości ściany bocznej do połowy wymiaru podstawy), a także proporcje katedry w Mediolanie, Partenonu.
- Proporcje człowieka witruwiańskiego autorstwa Leonarda da Vinci, rzeźby Wenus z Milo. Złote liczby w tym przypadku to stosunek wysokości człowieka do długości dolnej części ciała (od pępka w dół), a także stosunek długości dolnej części ciała do górnej (od pępka w górę).

- Symetria w sposobie ułożenia liści, kwiatów, płatków w roślinach, ziaren w słonecznikach, łusek na szyszce świerkowej itd. (tzw. filotaksja).
- Muszle ślimaków, głowonogów.
- W muzyce skrzypce Stradivariego, utwory Jana Sebastiana Bacha „V Symfonia" Beethovena i wiele innych.

Wartości złotej liczby wyznaczonej dla poszczególnych ciągów Fibonacciego zamieszczono w tabeli 12. Jak można zauważyć dla każdego z tych ciągów ilorazy sąsiednich wyrazów stabilizują się już przy 12–14 wyrazie na poziomie ≈1,61804.

Odwrotność złotej liczby wyznacza tzw. złotą proporcję wykorzystywaną do ustalania poziomów Fibonacciego, m.in. w analizie technicznej do określenia poziomów wsparcia oraz poziomów oporu (punkty zwrotne ruchu cen) na podstawie analizy wykresów cen instrumentów finansowych, indeksów akcji, kontaktów terminowych itp. Poziomy Fibonacciego mogą dotyczyć zarówno pionowej osi cen (miejsca realizacji zysków oraz zleceń obronnych typu stop loss), jak i poziomej osi czasu (okresy pomiędzy kolejnymi ekstremami na wykresie).

Zazwyczaj w analizie technicznej wykorzystuje się kilka poziomów Fibonacciego, które wyznacza się jako graniczne wartości ilorazów wyrazów ciągu Fibonacciego, ale nie wyrazów sąsiednich, lecz wyrazów przesuniętych względem siebie o 2,3,… pozycje:

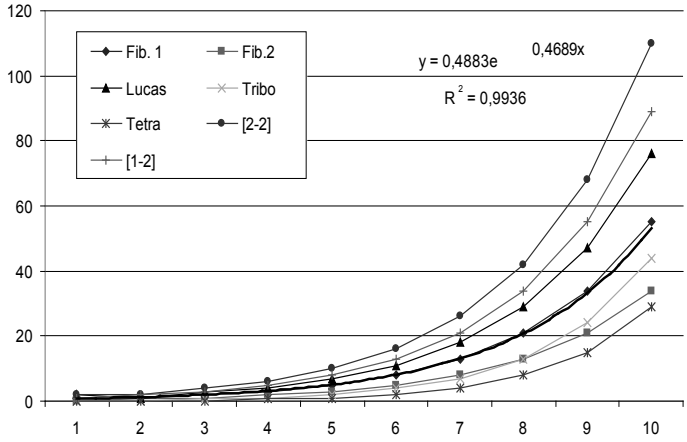
$$P_2 = \frac{F_i}{F_{i+2}};\ P_3 = \frac{F_i}{F_{i+3}};\ P_4 = \frac{F_i}{F_{i+4}};\ P_5 = \frac{F_i}{F_{i+5}} \qquad (19)$$

Tabela 12. Początkowe wyrazy ciągów Fibonacciego

i	Fib. [1;1]	Fib. [0;1]	Lucas	Tribo	Tetra	[2-2]	[1-2]
1	1	0	2	0	0	2	1
2	1	1	1	0	0	2	2
3	2	1	3	1	0	4	3
4	3	2	4	1	1	6	5
5	5	3	7	2	1	10	8
6	8	5	11	4	2	16	13
7	13	8	18	7	4	26	21
8	21	13	29	13	8	42	34
9	34	21	47	24	15	68	55
10	55	34	76	44	29	110	89

i	Fib. [1;1]	Fib. [0;1]	Lucas	Tribo	Tetra	[2-2]	[1-2]
11	89	55	123	81	56	178	144
12	144	89	199	149	108	288	233
13	233	144	322	274	208	466	377
14	377	233	521	504	401	754	610
15	610	377	843	927	773	1220	987
16	987	610	1364	1705	1490	1974	1597
17	1597	987	2207	3136	2872	3194	2584
18	2584	1597	3571	5768	5536	5168	4181
19	4181	2584	5778	10609	10671	8362	6765
20	6765	4181	9349	19513	20569	13530	10946

Źródło: opracowanie własne.



Rysunek 18. Wykresy dziesięciu początkowych wyrazów ciągów Fibonacciego
Źródło: opracowanie własne.

Tabela 13. Ilorazy sąsiednich wyrazów F_{i+1}/F_i ciągów Fibonacciego

i	Fib. [1;1]	Fib. [0;1]	Lucas	Tribo	Tetra	[2-2]	[1-2]
1	1,00000		0,50000			1,00000	2,00000
2	2,00000	1,00000	3,00000			2,00000	1,50000
3	1,50000	2,00000	1,33333	1,00000		1,50000	1,66667
4	1,66667	1,50000	1,75000	2,00000	1,00000	1,66667	1,60000
5	1,60000	1,66667	1,57143	2,00000	2,00000	1,60000	1,62500
6	1,62500	1,60000	1,63636	1,75000	2,00000	1,62500	1,61538
7	1,61538	1,62500	1,61111	1,85714	2,00000	1,61538	1,61905

i	Fib. [1;1]	Fib. [0;1]	Lucas	Tribo	Tetra	[2-2]	[1-2]
8	1,61905	1,61538	1,62069	1,84615	1,87500	1,61905	1,61765
9	1,61765	1,61905	1,61702	1,83333	1,93333	1,61765	1,61818
10	1,61818	1,61765	1,61842	1,84091	1,93103	1,61818	1,61798
11	1,61798	1,61818	1,61789	1,83951	1,92857	1,61798	1,61806
12	1,61806	1,61798	1,61809	1,83893	1,92593	1,61806	1,61803
13	1,61803	1,61806	1,61801	1,83942	1,92788	1,61803	1,61804
14	1,61804	1,61803	1,61804	1,83929	1,92768	1,61804	1,61803
15	1,61803	1,61804	1,61803	1,83927	1,92755	1,61803	1,61803
16	1,61803	1,61803	1,61804	1,83930	1,92752	1,61803	1,61803
17	1,61803	1,61803	1,61803	1,83929	1,92758	1,61803	1,61803
18	1,61803	1,61803	1,61803	1,83929	1,92757	1,61803	1,61803
19	1,61803	1,61803	1,61803	1,83929	1,92756	1,61803	1,61803

Źródło: opracowanie własne.

Tabela 14. Kolejne poziomy Fibonacciego wyznaczone na podstawie ciągu $F_1=F_2=1$

i	Fib. [1;1]	1	2	3	4	5
1	1	1,00000	0,50000	0,33333	0,20000	0,12500
2	1	0,50000	0,33333	0,20000	0,12500	0,07692
3	2	0,66667	0,40000	0,25000	0,15385	0,09524
4	3	0,60000	0,37500	0,23077	0,14286	0,08824
5	5	0,62500	0,38462	0,23810	0,14706	0,09091
6	8	0,61538	0,38095	0,23529	0,14545	0,08989
7	13	0,61905	0,38235	0,23636	0,14607	0,09028
8	21	0,61765	0,38182	0,23596	0,14583	0,09013
9	34	0,61818	0,38202	0,23611	0,14592	0,09019
10	55	0,61798	0,38194	0,23605	0,14589	0,09016
11	89	0,61806	0,38197	0,23607	0,14590	0,09017
12	144	0,61803	0,38196	0,23607	0,14590	0,09017
13	233	0,61804	0,38197	0,23607	0,14590	0,09017
14	377	0,61803	0,38197	0,23607	0,14590	0,09017
15	610	0,61803	0,38197	0,23607	0,14590	0,09017
16	987	0,61803	0,38197	0,23607	0,14590	
17	1597	0,61803	0,38197	0,23607		
18	2584	0,61803	0,38197			
19	4181	0,61803				
	6765	0,61803	0,38197	0,23607	0,14590	0,09017
odwr.		1,61803	2,61803	4,23607	6,85410	11,09017

Źródło: opracowanie własne.

W tabeli 14 podano dla wyjściowego ciągu Fibonacciego F1=F2=1 wartości ilorazów P1, P2 ,…, P5. Jak można zauważyć, ilorazy te stabilizują się już przy 10–11 wyrazie ciągu. Odwrotności tych granicznych wartości można też uzyskać jako kolejne potęgi liczby ψ:

$$P_k = \frac{1}{\varphi^k} \quad (k = \cdots, -3, -2, -1, 0, 1, 2, 3, \ldots) \tag{20}$$

Dokładne wartości poziomów Fibonacciego (w %) dla wybranych wartości parametru k z przedziału [-3;3] zebrano w tabeli 15.

Tabela 15. Poziomy Fibonacciego jako potęgi liczby φ

Wykł. pot. k	ψ = 1,61803	Poziomy Fib. (%)
-3	0,23607	23,6
-2	0,38197	38,2
-1	0,61803	61,8
-0,5	0,78615	78,6
0	1,00000	100,0
0,5	1,27202	127,2
1	1,61803	161,8
2	2,61803	261,8
3	4,23607	423,6

Źródło: opracowanie własne.

8.3. Reguła Pareto (zasada 80/20)

Vilfredo Federico Damasco Pareto (1848–1923)

- Włoski ekonomista i socjolog, współtwórca tzw. lozańskiej szkoły w ekonomii.
- W zakresie ekonomii zajmował się teorią ogólnej równowagi ekonomicznej i podziału dobrobytu oraz zastosowaniami metod matematycznych w ekonomii.
- Twórca pojęcia optymalności Pareta – nie można powiększyć dobrobytu jednostki bez zmniejszenia dobrobytu innej jednostki.
- W zakresie socjologii twórca teorii krążenia elit jako grupy ludzi mających w danej dziedzinie najwyższe osiągnięcia oraz teorii rezyduów (trwałych dyspozycji psychicznych) i derywacji (zmiennych elementów działań człowieka).

V. Pareto na podstawie analizy dochodów ludności we Włoszech (1897) stwierdził, że 80% majątku jest w posiadaniu 20% mieszkańców. W trakcie innych badań nad koncentracją okazało się, że w praktyce bardzo często większość (80%) problemów powodowanych jest przez małą część (20%) możliwych przyczyn. Na przykład 20% klientów generuje 80% zysków, 20% tekstu pozwala zrozumieć 80% zawartych w nim treści, 20% kierowców powoduje 80% wypadków, 20% powierzchni regionu zamieszkuje 80% ludności, 20% życia przynosi 80% szczęścia. Należy podkreślić, że liczby 20–80 są tylko przybliżeniem. Istota prawa polega na tym, że nie jest tak, że aby uzyskać 100% efektów, trzeba koniecznie ponieść 100% nakładów.

Reguła Pareto stanowi uproszczoną metodę diagnozowania zjawisk i planowania przedsięwzięć, ułatwia organizację czasu, pracy w grupie, pozwala ustalać priorytety działań. Jest podstawą metody ABC jako narzędzia zarządzania jakością w TQM. Regułę Pareto wykorzystuje się w badaniach marketingowych i metodach portfelowych w celu analizy potencjału strategicznego przedsiębiorstw.

W bibliotekoznawstwie reguła Pareto jest wykorzystywana w postaci prawa Bradforda. Zgodnie z tym prawem (zwanym prawem rozproszenia) w każdej dziedzinie nauki istnieje ograniczony, nieliczny zestaw najważniejszych czasopism, w których drukowana jest znacząca liczba (około 1/3) wszystkich wartościowych prac z danej dziedziny.

Jeżeli czasopisma uporządkuje się według malejącej liczby artykułów na dany temat, to można wyróżnić grupę czasopism podstawowych oraz kilka grup czasopism z taką samą liczbą artykułów co grupa podstawowa. Liczba czasopism w kolejnych grupach rośnie wtedy geometrycznie [1, n, n2 ,…]. Wystarczy analizować rejestr wypożyczeń, aby zracjonalizować decyzje o prenumeracie czasopism i koszcie ich zakupu. Prawo Bradforda ma związek z listą Impact factor prowadzoną przez Instytut Filadelfijski.

8.4. Prawa Estoupa, Zipfa, Heapsa

George Kingsley Zipf (1902–1950)

- Amerykański lingwista i filolog.
- Praca doktorska (1929) *Relative Frequency as a Determinant of Phonetic Change*, Uniwersytet Harvarda.
- W 1932 r. sformułował twierdzenie dotyczące częstości występowania słów, nie powołując się na znane od sześciu lat zbliżone prawo Lotki, nie dokonując analiz statystycznych ani nie opisując formułowanych zależności za pomocą odpowiednich wzorów matematycznych.

- Obecnie uchodzi za twórcę ilościowej lingwistyki (*zipfian linguistic*) jako składowej informetrii.

Zajmując się analizą częstości występowania słów w różnych językach, G. Zipf doszedł do wniosku, że „**większość słów używana jest rzadko, natomiast liczba słów często wykorzystywanych w tekstach jest stosunkowo nieduża**". Jeżeli uporządkować słowa według częstości ich występowania w tekstach, to częstość *c* wystąpienia r-tego słowa jest proporcjonalna do $1/r^a$ (gdzie *a* to wykładnik potęgowy bliski jedności). Pierwsze (najczęściej występujące w badanym tekście) słowo występuje więc dwa razy częściej niż drugie słowo, trzy razy częściej niż trzecie słowo itd. Warto nadmienić, że nie jest potwierdzona formuła Zipfa dla par słów oraz dla fraz, tzw. N-gramów.

Podobna zasada jak w prawie Zipfa wyrażona jest w postaci **formuły Estoupa-Zipfa**, zgodnie z którą „**iloczyn pozycji danego słowa na liście ich częstości występowania (r) przez częstość (c) jest stały r*c=const i zależy od długości analizowanego tekstu**".

W tabeli 16 podaje się przykłady analizy tekstów zawierających:

- ponad 40 mln wyrazów zaczerpniętych z „Wall Street Journal" (WSJ) z lat 1987–1989,
- ponad 37 mln wyrazów zaczerpniętych z artykułów opublikowanych przez Associated Press (AP) w roku 1989.

W tabeli 16 uwzględniono tylko 10 słów najczęściej występujących w analizowanych tekstach. W przypadku zbioru (korpusu) WSJ analizie poddano zarówno częstości występowania pojedynczych słów, jak i tzw. 2-gramy i 3-gramy. W przypadku bazy AP dostępne były tylko informacje o pojedynczych słowach.

Zgodność list 10 najczęściej wykorzystywanych słów jest duża. W obydwóch zbiorach na obydwu listach znajduje się po 7 słów na tych samych pozycjach: *the, of, to, a, and, in* oraz *for*. Kolejne słowo *that* jest także na obu listach, ale na różnych pozycjach. Listy różnią się tylko dwoma elementami – w zbiorze WSJ występują słowa *that* oraz *is*, natomiast w zbiorze AP – *said* oraz *was*.

Łącznie 10 najczęstszych słów tworzy 19% (w zbiorze WSJ) oraz 23% (w zbiorze AP) ogólnej zawartości analizowanych tekstów. Liczba fraz 2-wyrazowych i 3-wyrazowych jest wyraźnie mniejsza. Liczba 2-gramów stanowi tylko 13% liczby pojedynczych wyrazów, natomiast liczba 3-gramów – tylko 2,6%. Dla uzyskania porównywalności pomiędzy tymi trzema zbiorami, częstości 2-gramów i 3-gramów przeliczono, zakładając, że ich ogólna liczba równa jest liczbie pojedynczych wyrazów w całym zbiorze (N=40 mln).

Dla tak przeliczonych częstości w ujęciu procentowym wyznaczono iloczyny pozycji słów (*r*) przez częstości ich pojawiania się (*c*). Jak można zauważyć, są one zbliżone do stałej wartości *const=0,1* – średnia wartość iloczynów dla zbioru WSJ wynosi 0,074, a dla zbioru AP – 0,088. W ostatnich trzech kolumnach tabeli 16 podano odpowiednie wartości iloczynów *r*c* z dokładnością do trzech, dwóch i jednego miejsca po przecinku. Porównując stopień zgodności iloczynów *r*c* dla różnych zbiorów, obserwuje się, że najmniejsze ich zróżnicowanie (mierzone rozstępem, czyli różnicą pomiędzy maksymalnymi oraz minimalnymi wartościami iloczynów) ma miejsce dla pojedynczych słów (0,058), nieco większe jest dla 2-gramów (0,068) i największe dla 3-gramów (0,076).

Drugi przykład dotyczy tekstu w języku polskim. Analizie poddano utwór Michaiła Bułhakowa *Mistrz i Małgorzata*. Obliczenia wykonano za pomocą programu Hermetic Word Frequency Counter dostępnego na stronie www.hmetic.ch. Firma Hermetic Systems specjalizuje się w programach lingwistycznych, kryptograficznych, matematycznych, aplikacjach internetowych (kalendarze, konwertery).

W analizowanym tekście znajduje się ogółem 73 175 wyrazów, w tym 18 132 wyrazów rzadkich. W tabeli 17 przytoczono listę 40 najczęstszych słów wraz z ich częstością (liczba oraz wskaźnik procentowy udziału w stosunku do ogólnej liczby słów). Natomiast w tabeli 8 zebrano wybrane parametry opisujące własności analizowanego zbioru. Porównując wyniki analizy zbioru w języku polskim ze zbiorami w języku angielskim, można zauważyć duże podobieństwa. Dla przykładu 10 najczęstszych słów zarówno w zbiorze w języku polskim, jak i w zbiorze WSJ tworzy 19,2% całości tekstu. Iloczyny *r*c* w zbiorze w języku polskim kształtują się średnio na poziomie 0,078, natomiast w zbiorze WSJ na podobnym poziomie 0,074.

W zbiorze w języku polskim 40 najczęstszych wyrazów pozwala stworzyć 30% całości tekstu. Te 40 wyrazów to zaledwie 0,2% ogólnej liczby unikalnych wyrazów występujących w całym tekście. Na rysunkach 19–23 przedstawiono wykresy ilustrujące prawo Zipfa na przykładzie zbioru w języku polskim. Pierwsze dwa rysunki przedstawiają zależność w układzie liniowym pomiędzy częstością wyrazów *c* a ich pozycją – rysunek 19 dla pierwszych 100 najczęstszych wyrazów, a rysunek 20 – dla pierwszych 1000 wyrazów (spośród ponad 18 tysięcy). Obydwa wykresy mają postać hiperboli, przy czym uwzględnienie na wykresie dużej liczby wyrazów powoduje tłumienie przebiegu funkcji dla początkowych, najczęstszych wyrazów.

Tabela 16. Przykłady ilustrujące prawo Zipfa

WSJ 1-GRAM	Poz (r)	Częst. c	c (%)	rc (3)	rc (2)	rc (1)		Associated.Press	Poz (r)	Częst. c	c (%)	rc (3)	rc (2)	rc (1)
THE	1	2 057 968	5,145	0,051	0,05	0,1	T	the	1	2 420 778	6,487	0,065	0,06	0,1
OF	2	973 650	2,434	0,049	0,05	0,0	T	of	2	1 045 733	2,802	0,056	0,06	0,1
TO	3	940 525	2,351	0,071	0,07	0,1	T	to	3	968 882	2,596	0,078	0,08	0,1
A	4	853 342	2,133	0,085	0,09	0,1	T	a	4	892 429	2,392	0,096	0,1	0,1
AND	5	825 489	2,064	0,103	0,10	0,1	T	and	5	865 644	2,32	0,116	0,12	0,1
IN	6	711 462	1,779	0,107	0,11	0,1	T	in	6	847 825	2,272	0,136	0,14	0,1
THAT	7	368 012	0,920	0,064	0,06	0,1		said	7	504 593	1,352	0,095	0,09	0,1
FOR	8	362 771	0,907	0,073	0,07	0,1	T	for	8	363 865	0,975	0,078	0,08	0,1
ONE	9	298 646	0,747	0,067	0,07	0,1	X	that	9	347 072	0,93	0,084	0,08	0,1
IS	10	281 190	0,703	0,070	0,07	0,1		was	10	293 027	0,785	0,079	0,08	0,1
	S1	7 673 055	19,2	0,107	=max	0,074				8 549 848	22,9	**0,088**	=średnia	
	N	40 000 000		0,049	=min	0,058								

WSJ 2-GRAM	Poz (r)	Częst. c	c (%)	rc (3)	rc (2)	rc (1)	WSJ 3-GRAM	Poz (r)	Częst. c	c (%)	rc (3)	rc (2)	rc (1)
OF THE	1	217 427	4,140	0,041	0,04	0,0	THE U. S.	1	42 030	4,081	0,041	0,04	0,0
IN THE	2	173 797	3,309	0,066	0,07	0,1	IN NINETEEN EIGHTY	2	27 260	2,647	0,053	0,05	0,1
MILLION DOLLARS	3	110 291	2,100	0,063	0,06	0,1	CENTS A SHARE	3	24 165	2,346	0,070	0,07	0,1
U. S.	4	89 184	1,698	0,068	0,07	0,1	NINETEEN EIGHTY SIX	4	18 233	1,770	0,071	0,07	0,1

WSJ 1-GRAM	Poz (r)	Częst. c	c (%)	rc (3)	rc (2)	rc (1)		Associa-ted.Press	Poz (r)	Częst. c	c (%)	rc (3)	rc (2)	rc (1)
NINETEEN EIGHTY	5	83 799	1,596	0,080	0,08	0,1		NINE-TEEN EIGHTY SEVEN	5	16 786	1,630	0,081	0,08	0,1
FOR THE	6	76 187	1,451	0,087	0,09	0,1		FIVE MILLION DOLLARS	6	15 316	1,487	0,089	0,09	0,1
TO THE	7	72 312	1,377	0,096	0,10	0,1		MILLION DOLLARS OR	7	14 943	1,451	0,102	0,10	0,1
ON THE	8	65 565	1,248	0,100	0,10	0,1		MILLION DOLLARS IN	8	14 517	1,410	0,113	0,11	0,1
ONE HUN-DRED	9	63 838	1,216	0,109	0,11	0,1		IN NEW YORK	9	12 327	1,197	0,108	0,11	0,1
THAT THE	10	55 014	1,048	0,105	0,10	0,1		A YEAR EARLIER	10	11 981	1,163	0,116	0,12	0,1
	S2	1 007 414		0,109	max				S3	197 558		0,116	max	
U2=S2/ S1*100	13,1	5 251 697	U2*N/100	0,041	min	0,068		U3=S3/ S1*100	2,6	1 029 879	U3*N/100	0,041	min	0,076

Źródło: opracowanie własne na podstawie D.B. Paul, J.M. Baker *The Design for the Wall Street Journal – based CSR Corpus*, Proc. ICSLP, 1992, s. 899–902; Le Quan Ha, E.I. Sicilia-Garcia, Ji Ming, F.J. Smith, *Extension of Zipf's Law to Words and Phrases*, The Association for Computational Linguistics, A Digital Archive of Research Papers, www.aclweb.org/anthology/C/C02/C02-1117.pdf.

Tabela 17. Analiza Zipfa utworu Michaiła Bułhakowa *Mistrz i Małgorzata*

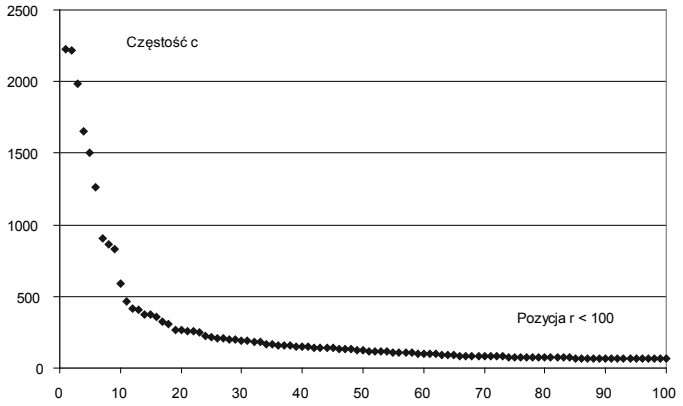
wyraz	ranga r	częst. c	c (%)	rc/100	wyraz	ranga r	częst. c	c (%)	rc/100
się	1	2224	3,040	0,030	go	21	258	0,353	0,074
i	2	2217	3,030	0,061	pan	22	255	0,349	0,077
w	3	1983	2,710	0,081	było	23	252	0,344	0,079
na	4	1649	2,254	0,090	jego	24	223	0,305	0,073
nie	5	1507	2,060	0,103	mu	25	218	0,298	0,074
z	6	1265	1,729	0,104	od	26	204	0,279	0,072
że	7	905	1,237	0,087	tylko	27	204	0,279	0,075
to	8	863	1,180	0,094	jeszcze	28	201	0,275	0,077
do	9	828	1,132	0,102	pod	29	198	0,271	0,078
a	10	586	0,801	0,080	jej	30	194	0,265	0,080
o	11	461	0,630	0,069	przez	31	192	0,262	0,081
ale	12	412	0,563	0,068	tego	32	186	0,254	0,081
jak	13	404	0,552	0,072	ze	33	184	0,251	0,083
co	14	370	0,506	0,071	był	34	168	0,230	0,078
po	15	370	0,506	0,076	kiedy	35	164	0,224	0,078
już	16	354	0,484	0,077	mi	36	158	0,216	0,078
za	17	320	0,437	0,074	sobie	37	156	0,213	0,079
tak	18	304	0,416	0,075	ma	38	154	0,210	0,080
tym	19	267	0,365	0,069	mnie	39	153	0,209	0,082
jest	20	266	0,364	0,073	powiedział	40	153	0,209	0,084

Źródło: opracowanie własne.

Tabela 18. Wyniki analizy Zipfa utworu Michaiła Bułhakowa *Mistrz i Małgorzata*

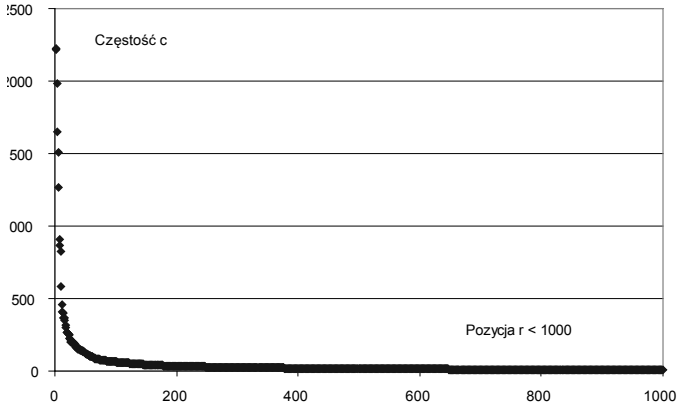
L. wyraz.	Częst. c	c (%)	c (% kum)	słowa (%)	rc
10	14 027	19,17	19,2	0,055	0,083
20	17 555	4,82	24,0	0,110	0,078
30	19 762	3,02	27,0	0,165	0,077
40	21 430	2,28	29,3	0,221	0,078
N	73 163			18 132	

Źródło: opracowanie własne.



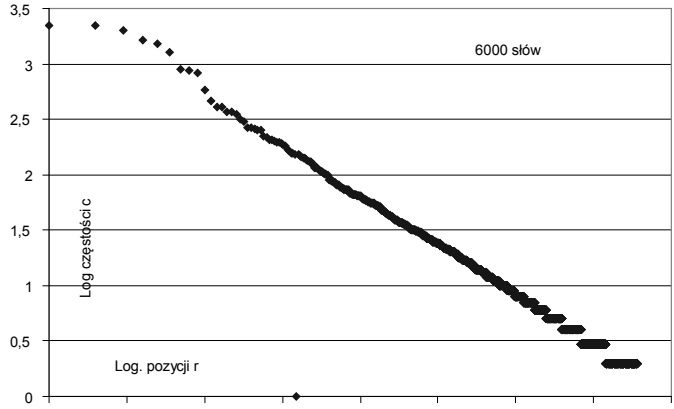
Rysunek 19. Wykres Zipfa typu częstość (pozycja dla 100 najczęstszych wyrazów)

Źródło: opracowanie własne.



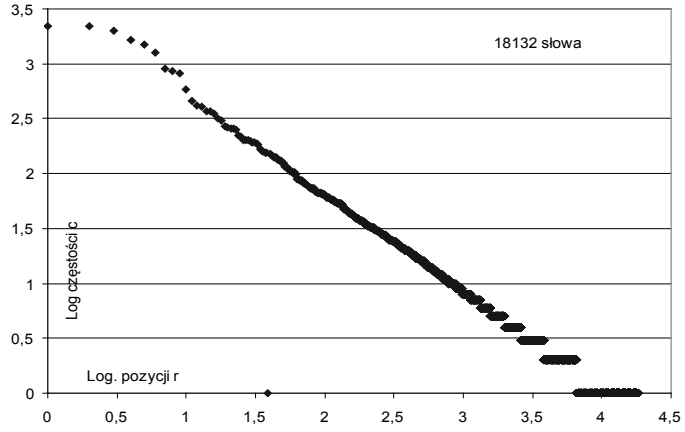
Rysunek 20. Wykres Zipfa typu częstość (pozycja dla 1000 najczęstszych wyrazów)

Źródło: opracowanie własne.



Rysunek 21. Logarytmiczny wykres Zipfa typu częstość (pozycja dla 6000 najczęstszych wyrazów)

Źródło: opracowanie własne.



Rysunek 22. Logarytmiczny wykres Zipfa typu częstość (pozycja dla wszystkich 18 tysięcy wyrazów)

Źródło: opracowanie własne.

Rysunki 21 i 22 przedstawiają tę samą zależność (częstość występowania a pozycja słów), ale nie w układzie liniowym, lecz w logarytmicznym. Podobnie jak poprzednio, uwzględnienie mniejszej liczby słów pozwala dokładniej przeanalizować kształtowanie się wykresu dla słów z „najwyższej półki". Wykresy w układzie podwójnie logarytmicznym przyjmują postać

funkcji prostoliniowej, z zawirowaniami na jej krańcach (to znaczy dla wyrazów znajdujących się na początkowych i końcowych pozycjach).

Prawo Zipfa dotyczy także ludności i powierzchni miast, przychodów i rozmiarów korporacji gospodarczych, rozkładu dochodów osób, rozkładu liczby trzęsień ziemi od najsłabszych do najsilniejszych. Równanie Zipfa pozwala opisać popularność witryn internetowych, dzięki czemu można bardziej wydajnie zaprojektować tabele adresowe serwerów[5].

W pracy z 1941 r. G. Zipf stwierdził[6], że krzywa rozkładów dochodów dla Indonezji wyraźnie odbiega od reguły rozkładu potęgowego, skąd wysnuł tezę, że w tym kraju będą duże napięcia społeczne. Rewolucja w Indonezji rzeczywiście zaczęła się cztery lata później, w 1945 roku.

Prawo Zipfa zakłada, że istnieje ścisły związek pomiędzy rangami obiektów (miasto, firma, osoba) w ich uporządkowaniu ze względu na analizowaną cechę a wielkością tej cechy (*rank-size rule).* Odwrotnością prawa Zipfa jest prawo proporcjonalnego wzrostu Gibrata (1904–1980), zgodnie z którym wielkość firmy i stopień jej wzrostu są niezależne od siebie[7].

Kolejna reguła zbliżona do prawa Zipfa to prawo Heapsa opisujące zależność[8] między rozmiarem tekstu N a liczbą użytych w nim różnych wyrazów V. Relację ta wyraża wzór $V=K^{*}N^{a}$, w którym parametry *K* oraz *a* ustala się empirycznie. Dla języka angielskiego zazwyczaj K ≈ [30;100] natomiast a ≈ [0,4;0,6]. Dla innych języków parametry te przyjmują inne wartości. Reguła Heapsa może być wykorzystana np. w projektowaniu tekstowych baz danych do ustalenia rozmiarów indeksu jako funkcji rozmiaru bazy danych.

---

[5] Inne przykłady zastosowania prawa Zipfa znaleźć można m.in. w pracach A.A. Adamic, B.A. Huberman *Zipf's Law and the Internet*, „Glottometrics", 3/2002, s. 143–150, Y.M. Ioannides, H.G. Overman, *Zipf's law for cities: an empirical examination*, „Regional Science and Urban Economics" 2002; X. Gabaix, *Zipf's Law and the growth of* cities, „American Economic Review", 89/1999, s. 129–132; M. Aida, N. Takahashi, T. Abe, *A proposal of dual Zipfian model fordescribing HTTP access trends and its application to address cache design*. IEICE Transactions on Communications, 81(7)/1998, s. 1475–1485.

[6] G.K. Zipf, *National Unity and Disunity. The Nation as a Bio-Social Organism*, Principia Press, Bloomington Indiana, Princeton Press, 1941.

[7] Więcej informacji na temat G.K. Zipfa i jego osiągnięć znaleźć można w publikacji *To honor G.K. Zipf*, „Glottometrics", 3/2002, m.in. zamieszczonych w tym numerze prac: R. Rousseau, *George Kingsley Zipf: life, ideas, his law and informetrics*, s. 11–18; A. Altman, *Zipfian linquistics*, s. 19–26.

[8] Por. D.C. van Leijenhorst, T.P. van der Weide, *A formal derivation of Heaps' Law*, „Information Science", 170/2005, s. 263–272.

8.5. Prawo produktywności pracowników naukowych Lotki

Alfred James Lotka (1880–1949)

- Amerykański matematyk, chemik, statystyk, biolog, demograf.
- Prezydent Population Association of Americ (1938–1939), American Statistical Association (1942).
- Statystyk w agencji ubezpieczeniowej w Nowym Jorku (1924–1947).
- Współtwórca modelu drapieżnik–ofiara (model Lotki-Volterry). Vito Volterra (1926) podał równanie opisujące populację ryb odławianych w Morzu Adriatyckim, natomiast A. Lotka (1910–1920) równanie opisujące oscylację stężeń substancji w reakcji chemicznej. Model Lotki-Volterry pozwala analizować układy dynamiczne w ekosystemach, w demografii, a także w gospodarce.
- Sformułował równanie łączące strukturę wieku ludności, płodność i umieralność, co w demografii dało początek koncepcji ludności ustabilizowanej.

Prawo produktywności pracowników naukowych sformułowane przez A.J. Lotkę w 1926 roku opisuje zależność między liczbą publikacji (X) a liczbą autorów (Y) mających daną liczbę publikacji (w zadanym okresie czasu i w ustalonym obszarze merytorycznym)[9]. Odpowiedni wzór ma postać $Y=const/X^a$, gdzie *const* oraz *a* to stałe. Parametr $a \approx 2$, natomiast *const* przyjmuje wartość zależną od analizowanej dziedziny i w przybliżeniu równy jest liczbie pracowników, którzy napisali tylko 1 artykuł w badanym obszarze tematycznym.

Inaczej mówiąc, liczba autorów (Y), z których każdy napisał pewną liczbę prac, jest odwrotnie proporcjonalna do kwadratu liczby tych prac (X). Większość autorów (60%) pisze jedną publikację, a bardzo mały odsetek autorów tworzy dużą część publikacji. Dla przykładu, jeżeli przyjmiemy za *const*=200 to liczba autorów, którzy napisali od 1 do 12 artykułów, kształtuje się następująco:

X – liczba artykułów	1	2	3	4	5	6	7	8	9	10	11	12
Y – liczba autorów z liczbą artykułów 1,2,…	200	50	22	13	8	6	4	3	2	2	2	1

[9] A.J. Lotka, *The frequency distribution of scientific productivity*, "Journal of the Washington Academy of Science", 16/1926, s. 317–323.

Podobne zależności mają miejsce na przykład w lotnictwie wojskowym, dla odzwierciedlenia liczby pilotów wojskowych w zależności od liczby zestrzelonych przez nich samolotów.

8.6. Informetria jako wiedza o informacji

Prawa Bradforda, Zipfa, Heapsa, Lotki i ich mutacje są podstawą dyscypliny zwanej **informetrią** (*infometrics*) i będącą częścią informatologii (informologii) jako nauki zajmującej się teoretycznymi i praktycznymi aspektami wiedzy o informacji. Informetria kładzie nacisk na wykorzystanie metod ilościowych i statystycznych do badania praw rządzących informacją.

Innymi działami informatologii[10] są:

- bibliometria, analizująca zjawiska i procesy, w których biorą udział dokumenty (wydzielenie najczęściej cytowanych czasopism, relacje pomiędzy czasopismami krajowymi i zagranicznymi, ocena efektywności działalności bibliotecznej);
- naukometria, zajmująca się ilościową charakterystyką struktury nauki, określeniem dynamiki i kierunków jej rozwoju;
- webometria, badająca dynamikę zmian w środowisku WWW;
- cybermetria, poszerzająca zakres badań webometrii o zasoby elektroniczne;
- infobrokering, tworzy zasady i metody efektywnego pozyskiwania adekwatnych informacji, zwłaszcza w postaci elektronicznej.

[10] Więcej informacji znaleźć można w pracach: M. Dembowska, *Nauka o informacji naukowej (informatologia). Organizacja i problematyka badań w Polsce*, Instytut informacji naukowej, technicznej, ekonomicznej IINTE, Warszawa 1991; J. Ratajewski, *Wybrane problemy metodologiczne informologii nauki (informacji naukowej)*, „Prace Naukowe UŚ", Katowice 1994; B. Sordylowa, *Informacja naukowa w Polsce. Problemy teoretyczne, źródła, organizacja*, Ossolineum, Wrocław 1987; E. Ścibor, *Informacja naukowa w Polsce: tradycja i współczesność*, Olsztyn 1998; *Informacja naukowa: rozwój, metody, organizacja*, red. Z. Żmigrodzki, Wyd. SBP, Warszawa 2006.

9. Metody oceny zgodności rozkładów cyfr znaczących z prawami Benforda

Marzena Farbaniec, Tadeusz Grabiński, Bartłomiej Zabłocki, Wacław Zając

Wprowadzenie

W podręcznikach statystyki można znaleźć wiele metod oceny zgodności rozkładów. Są to: m.in. testy chi kwadrat, Kołmogorowa-Smirnowa, testy istotności oparte na statystyce. Narzędzia te wykorzystywane są do odpowiedzi na pytanie, czy empiryczne rozkłady następujących cyfr:

- F1 – pierwszej cyfry znaczącej,
- F2 – dwóch pierwszych cyfr znaczących,
- F3 – trzech pierwszych cyfr znaczących,
- D2 – dokładnie drugiej cyfry znaczącej,
- D3 – dokładnie trzeciej cyfry znaczącej,
- L1 – ostatniej cyfry,

są zgodne z rozkładami wynikającymi z prawa Benforda.

Poniżej podano wzory i definicje poszczególnych parametrów. We wzorach przyjęto następującą konwencję oznaczeń.

- n – ogólna liczba obserwacji w analizowanym zbiorze;
- k – liczba kombinacji cyfr (w teście F1 – $k=9$, w testach D2, D3, L1 – $k=10$, w teście F2 – $k=90$ natomiast w teście F3 – $k=900$);
- i – subskrypt oznaczający numer kolejny kombinacji cyfr ($i=1,2,\ldots,k$);
- n_i oraz $\hat{n}_i$ $(i=1,2,\ldots,k)$ – liczebności empiryczne i teoretyczne pojawienia się i-tej cyfry (lub i-tej kombinacji cyfr);
- c_i oraz $\hat{c}_i$ $(i=1,2,\ldots,k)$ – częstości empiryczne i teoretyczne dane wzorami:

$$c_i = \frac{n_i}{n}100 \qquad \hat{c}_i = \frac{\hat{n}_i}{n}100 \tag{21}$$

- p_i oraz $\hat{p}_i$ $(i=1,2,\ldots,k)$ – prawdopodobieństwa empiryczne i teoretyczne dane wzorami:

$$p_i = \frac{n_i}{n} \qquad \hat{p}_i = \frac{\hat{n}_i}{n} \tag{22}$$

- f_i oraz $\hat{f}_i$ *(i=1,2,…,k)* – wartości dystrybuanty (kumulanty) empirycznego i teoretycznego rozkładu częstości cyfr dane wzorami:

$$f_i = \sum_{l=1}^{i} p_l \qquad \hat{f}_i = \sum_{l=1}^{i} \hat{p}_l \tag{23}$$

9.1. Test chi kwadrat

Do oceny zgodności rozkładów najczęściej stosowany jest test chi kwadrat, w którym wyznacza się wartość parametru:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} = n\sum_{i=1}^{k} \frac{(p_i - \hat{p}_i)^2}{\hat{p}_i} = \frac{n}{100}\sum_{i=1}^{k} \frac{(c_i - \hat{c}_i)^2}{\hat{c}_i} \tag{24}$$

Statystykę porównuje się z wartością krytyczną testu χ^2 dla założonego poziomu istotności α oraz *k-1* stopni swobody[11]. W tabeli 19 podano wartości krytyczne testu χ^2 dla wybranych poziomów istotności α = 0,1; 0,05, 0,01 i 0,001 oraz dla stopni swobody właściwych dla testów F1–F3, D2, D3 i L1[12].

Tabela 19. Wybrane wartości krytyczne testu χ^2_α

Test	K	α =0,1	α =0,05	α =0,01	α =0,001
F1	9	13,4	15,5	20,1	26,1
D2, D3, L1	10	14,7	16,9	21,7	27,9
F2	90	106,5	112,0	122,9	136,0
F3	900	953,8	969,9	1000,6	1035,8

Źródło: opracowanie własne.

[11] Niektórzy autorzy twierdzą, że test chi kwadrat jest zbyt rygorystyczny, gdyż jego wskazania w dużym stopniu zależą od liczby elementów w analizowanym zbiorze *n*. W analizach związanych z rozkładami Benforda zwykle mamy do czynienia z bardzo dużymi zbiorami i w takich przypadkach nawet niewielkie odchylenia liczebności teoretycznych i empirycznych mogą sugerować istotną niezgodność porównywanych rozkładów. Por. E. Ley, *On the Peculiar Distribution of the U.S. Stock Indexes' Digits*, "The American Statistician", 50/1996, s. 311–313; D.E.A. Giles, *Benford's Law and Naturally Occurring Prices in Certain e-Bay Auctions*, „Applied Economics Letters", 14/2007, s. 157–161.

[12] Wartości te można uzyskać, stosując w Excelu funkcję =ROZKŁAD.CHI.ODWR (α; ss), gdzie ss to liczba stopni swobody (ss=k-1).

Im wyższa jest empiryczna wartość statystyki χ^2, tym bardziej różnią się porównywane rozkłady. Jeżeli $\chi^2 \geq \chi^2_\alpha$, to z prawdopodobieństwem 1-α można twierdzić, że rozkład empiryczny **nie jest zgodny** z regułami prawa Benforda.

W analizach korzysta się także z parametru wskazującego, jaki poziom istotności α odpowiada ustalonej wartości empirycznej testu χ^2. W tabeli 21 podano wartości tych prawdopodobieństw dla statystyk[13] z przedziału, w którym zawierają się typowe wartości poziomów istotności od 0,001 do 0,10. Warto tu zwrócić uwagę, na fakt, że przy 900 stopniach swobody (test F3) oraz statystyce χ^2=890 funkcja w Excelu zwraca błąd LICZBA!

W celu wyjaśnienia tego błędu przeanalizowano ciągi formuł =ROZKŁAD.CHI (chi; ss) dla wartości statystyki χ^2 z przedziału [870;900] ze skokiem co 1, oraz dla liczby stopni swobody z przedziału [870;920] ze skokiem co 10 (tabela 20). Jak się okazuje, błąd zlokalizowany jest na odcinkach o długości 15 jednostek i zaczyna się dla wartości χ^2 o 15 mniejszej niż zadana liczba stopni swobody.

Tabela 20. Wartości prawdopodobieństw w rozkładzie χ^2 dla zadanej wartości testu chi kwadrat [870;900;1] oraz liczbie stopni swobody z przedziału [870;920;10]

chi emp / ss	870	880	890	900	910	920
870	0,484	#LICZBA!	0,669	0,750	0,819	0,875
871	0,475	#LICZBA!	0,661	0,743	0,813	0,870
872	0,465	#LICZBA!	0,652	0,735	0,806	0,864
873	0,455	#LICZBA!	0,643	0,727	0,800	0,859
874	0,446	#LICZBA!	0,634	0,719	0,793	0,854
875	0,437	#LICZBA!	#LICZBA!	0,711	0,786	0,848
876	0,427	#LICZBA!	#LICZBA!	0,702	0,779	0,842
877	0,418	#LICZBA!	#LICZBA!	0,694	0,772	0,836
878	0,409	#LICZBA!	#LICZBA!	0,686	0,764	0,830
879	0,399	0,494	#LICZBA!	0,677	0,757	0,824
880	0,390	0,484	#LICZBA!	0,668	0,749	0,818
881	0,381	0,475	#LICZBA!	0,660	0,741	0,811
882	0,372	0,465	#LICZBA!	0,651	0,734	0,805
883	0,363	0,456	#LICZBA!	0,642	0,726	0,798
884	0,354	0,446	#LICZBA!	#LICZBA!	0,718	0,791
885	0,346	0,437	#LICZBA!	#LICZBA!	0,710	0,784
886	0,337	0,428	#LICZBA!	#LICZBA!	0,701	0,777

[13] W tym przypadku korzysta się z funkcji w Excelu =ROZKŁAD.CHI (stat; ss), gdzie stat=χ^2.

chi emp / ss	870	880	890	900	910	920
887	0,328	0,418	#LICZBA!	#LICZBA!	0,693	0,770
888	0,320	0,409	#LICZBA!	#LICZBA!	0,685	0,763
889	0,311	0,400	0,494	#LICZBA!	0,676	0,756
890	0,303	0,391	0,484	#LICZBA!	0,668	0,748
891	0,295	0,382	0,475	#LICZBA!	0,659	0,740
892	0,287	0,373	0,465	#LICZBA!	0,650	0,733
893	0,279	0,364	0,456	#LICZBA!	#LICZBA!	0,725
894	0,271	0,355	0,447	#LICZBA!	#LICZBA!	0,717
895	0,263	0,346	0,437	#LICZBA!	#LICZBA!	0,709
896	0,256	0,338	0,428	#LICZBA!	#LICZBA!	0,700
897	0,248	0,329	0,419	#LICZBA!	#LICZBA!	0,692
898	0,241	0,321	0,410	#LICZBA!	#LICZBA!	0,684
899	0,233	0,312	0,400	0,494	#LICZBA!	0,675

Źródło: opracowanie własne.

Tabela 21. Wartości prawdopodobieństw w rozkładzie χ^2 dla zadanej wartości testu (chi emp) oraz przy liczbie stopni swobody właściwej dla testów Benforda

Test	F1	D2,D3,L1		F2		F3
chi emp	9	10	**chi emp**	90	**chi emp**	900
4	0,857	0,911	**80**	0,742	**860**	0,821
6	0,647	0,740	**85**	0,600	**875**	0,711
8	0,433	0,534	**90**	0,450	**890**	#LICZBA!
10	0,265	0,350	**95**	0,312	**905**	0,438
12	0,151	0,213	**100**	0,200	**920**	0,306
14	0,082	0,122	**105**	0,118	**935**	0,197
16	0,042	0,067	**110**	0,065	**950**	0,116
18	0,021	0,035	**115**	0,033	**965**	0,062
20	0,010	0,018	**120**	0,016	**980**	0,031
22	0,005	0,009	**125**	0,007	**995**	0,014
24	0,002	0,004	**130**	0,003	**1010**	0,006
26	0,001	0,002	**135**	0,001	**1025**	0,002
28	0,000	0,001	**140**	0,000	**1040**	0,001
30	0,000	0,000	**145**	0,000	**1055**	0,000

Źródło: opracowanie własne.

Dla ustalenia, czy omawiany błąd ma charakter lokalny czy globalny, wyznaczono wartości formuły =ROZKŁAD.CHI (chi; ss) dla parametrów w szerokim przedziale zmienności od 100 do 1800. W tabeli 21 przytoczono fragment tych obliczeń, począwszy od miejsca, gdzie po raz pierwszy stwierdzono obecność tego błędu. Jak można zauważyć, błąd w omawianej formule pojawia się przy parametrach funkcji ROZKŁAD.CHI na poziomie chi=ss=800. Początkowo błąd ten „trwa" krótko, ale stopniowo okres jego obecności systematycznie się wydłuża, co 100 o 12 jednostek. W tabeli 22 podano początkowe i końcowe wartości statystyki χ^2, przy których dla danej liczby stopni swobody zaczyna się wyświetlać błąd LICZBA!. W ostatniej kolumnie tej tabeli znajdują się informacje o długości odcinka liczbowego, na którym błąd ten jest obecny.

Tabela 22. Lokalizacja długości odcinków, na których wyświetlany jest błąd LICZBA! w funkcji ROZKŁAD.CHI

Początek	Koniec/ss	Długość
797	800	3
885	900	15
973	1000	27
1060	1100	40
1148	1200	52
1236	1300	64
1323	1400	77
1411	1500	89
1499	1600	101
1586	1700	114
1674	1800	126

Źródło: opracowanie własne.

Tabela 23. Lokalizacja „usterki" funkcji =ROZKŁAD.CHI w Excelu

Chi/ss	790	800	810	820	830	840	850	860	870
794	0,44	0,54	0,64	0,73	0,80	0,86	0,91	0,94	0,97
796	0,42	0,52	0,62	0,71	0,79	0,85	0,90	0,94	0,96
798	0,40	#LICZBA!	0,60	0,69	0,77	0,84	0,89	0,93	0,96
800	0,38	0,48	0,58	0,68	0,76	0,83	0,88	0,93	0,95
802	0,37	0,46	0,56	0,66	0,74	0,82	0,87	0,92	0,95
804	0,35	0,44	0,54	0,64	0,73	0,80	0,86	0,91	0,94
806	0,33	0,42	#LICZBA!	0,62	0,71	0,79	0,85	0,90	0,94
808	0,31	0,40	#LICZBA!	0,60	0,69	0,77	0,84	0,89	0,93
810	0,29	0,39	0,48	0,58	0,68	0,76	0,83	0,88	0,92
812	0,28	0,37	0,46	0,56	0,66	0,74	0,81	0,87	0,92
814	0,26	0,35	0,44	#LICZBA!	0,64	0,73	0,80	0,86	0,91
816	0,25	0,33	0,42	#LICZBA!	0,62	0,71	0,79	0,85	0,90
818	0,23	0,31	0,41	#LICZBA!	0,60	0,69	0,77	0,84	0,89
820	0,22	0,30	0,39	0,48	0,58	0,67	0,76	0,83	0,88
822	0,20	0,28	0,37	0,46	0,56	0,66	0,74	0,81	0,87
824	0,19	0,26	0,35	0,44	#LICZBA!	0,64	0,72	0,80	0,86
826	0,18	0,25	0,33	0,43	#LICZBA!	0,62	0,71	0,79	0,85
828	0,16	0,23	0,31	0,41	#LICZBA!	0,60	0,69	0,77	0,84
830	0,15	0,22	0,30	0,39	0,48	0,58	0,67	0,76	0,82
832	0,14	0,20	0,28	0,37	0,46	#LICZBA!	0,66	0,74	0,81
834	0,13	0,19	0,26	0,35	0,44	#LICZBA!	0,64	0,72	0,80
836	0,12	0,18	0,25	0,33	0,43	#LICZBA!	0,62	0,71	0,78
838	0,11	0,16	0,23	0,31	0,41	#LICZBA!	0,60	0,69	0,77
840	0,10	0,15	0,22	0,30	0,39	0,48	#LICZBA!	0,67	0,75
842	0,09	0,14	0,20	0,28	0,37	0,46	#LICZBA!	0,65	0,74
844	0,09	0,13	0,19	0,27	0,35	0,45	#LICZBA!	0,64	0,72
846	0,08	0,12	0,18	0,25	0,33	0,43	#LICZBA!	0,62	0,71
848	0,07	0,11	0,17	0,23	0,32	0,41	#LICZBA!	0,60	0,69
850	0,06	0,10	0,15	0,22	0,30	0,39	0,48	#LICZBA!	0,67
852	0,06	0,09	0,14	0,21	0,28	0,37	0,46	#LICZBA!	0,65
854	0,05	0,09	0,13	0,19	0,27	0,35	0,45	#LICZBA!	0,64
856	0,05	0,08	0,12	0,18	0,25	0,33	0,43	#LICZBA!	0,62
858	0,04	0,07	0,11	0,17	0,24	0,32	0,41	#LICZBA!	#LICZBA!
860	0,04	0,07	0,10	0,16	0,22	0,30	0,39	0,48	#LICZBA!
862	0,04	0,06	0,10	0,14	0,21	0,28	0,37	0,46	#LICZBA!
864	0,03	0,05	0,09	0,13	0,19	0,27	0,35	0,45	#LICZBA!
866	0,03	0,05	0,08	0,12	0,18	0,25	0,34	0,43	#LICZBA!
868	0,03	0,05	0,07	0,11	0,17	0,24	0,32	0,41	#LICZBA!
870	0,02	0,04	0,07	0,11	0,16	0,22	0,30	0,39	0,48
872	0,02	0,04	0,06	0,10	0,15	0,21	0,28	0,37	0,46
874	0,02	0,03	0,06	0,09	0,14	0,20	0,27	0,35	0,45
876	0,02	0,03	0,05	0,08	0,13	0,18	0,25	0,34	0,43
878	0,01	0,03	0,05	0,07	0,12	0,17	0,24	0,32	0,41
880	0,01	0,02	0,04	0,07	0,11	0,16	0,22	0,30	0,39
882	0,01	0,02	0,04	0,06	0,10	0,15	0,21	0,29	0,37
884	0,01	0,02	0,03	0,06	0,09	0,14	0,20	0,27	0,35

880	890	900	910	920	930	940	950	960	970
0,98	0,99	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00
0,98	0,99	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00
0,98	0,99	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00
0,97	0,98	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00
0,97	0,98	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00
0,97	0,98	0,99	0,99	1,00	1,00	1,00	1,00	1,00	1,00
0,96	0,98	0,99	0,99	1,00	1,00	1,00	1,00	1,00	1,00
0,96	0,98	0,99	0,99	1,00	1,00	1,00	1,00	1,00	1,00
0,95	0,97	0,98	0,99	1,00	1,00	1,00	1,00	1,00	1,00
0,95	0,97	0,98	0,99	1,00	1,00	1,00	1,00	1,00	1,00
0,94	0,97	0,98	0,99	0,99	1,00	1,00	1,00	1,00	1,00
0,94	0,96	0,98	0,99	0,99	1,00	1,00	1,00	1,00	1,00
0,93	0,96	0,97	0,99	0,99	1,00	1,00	1,00	1,00	1,00
0,92	0,95	0,97	0,98	0,99	1,00	1,00	1,00	1,00	1,00
0,92	0,95	0,97	0,98	0,99	0,99	1,00	1,00	1,00	1,00
0,91	0,94	0,96	0,98	0,99	0,99	1,00	1,00	1,00	1,00
0,90	0,94	0,96	0,98	0,99	0,99	1,00	1,00	1,00	1,00
0,89	0,93	0,96	0,97	0,99	0,99	1,00	1,00	1,00	1,00
0,88	0,92	0,95	0,97	0,98	0,99	1,00	1,00	1,00	1,00
0,87	0,91	0,95	0,97	0,98	0,99	0,99	1,00	1,00	1,00
0,86	0,91	0,94	0,96	0,98	0,99	0,99	1,00	1,00	1,00
0,85	0,90	0,93	0,96	0,98	0,99	0,99	1,00	1,00	1,00
0,84	0,89	0,93	0,95	0,97	0,98	0,99	1,00	1,00	1,00
0,82	0,88	0,92	0,95	0,97	0,98	0,99	1,00	1,00	1,00
0,81	0,87	0,91	0,94	0,97	0,98	0,99	0,99	1,00	1,00
0,80	0,86	0,90	0,94	0,96	0,98	0,99	0,99	1,00	1,00
0,78	0,85	0,90	0,93	0,96	0,98	0,99	0,99	1,00	1,00
0,77	0,83	0,89	0,93	0,95	0,97	0,98	0,99	1,00	1,00
0,75	0,82	0,88	0,92	0,95	0,97	0,98	0,99	0,99	1,00
0,74	0,81	0,87	0,91	0,94	0,97	0,98	0,99	0,99	1,00
0,72	0,80	0,86	0,90	0,94	0,96	0,98	0,99	0,99	1,00
0,70	0,78	0,84	0,89	0,93	0,96	0,97	0,99	0,99	1,00
0,69	0,77	0,83	0,89	0,93	0,95	0,97	0,98	0,99	1,00
0,67	0,75	0,82	0,88	0,92	0,95	0,97	0,98	0,99	0,99
0,65	0,74	0,81	0,87	0,91	0,94	0,97	0,98	0,99	0,99
0,63	0,72	0,79	0,85	0,90	0,94	0,96	0,98	0,99	0,99
0,62	0,70	0,78	0,84	0,89	0,93	0,96	0,97	0,99	0,99
#LICZBA!	0,69	0,77	0,83	0,88	0,92	0,95	0,97	0,98	0,99
#LICZBA!	0,67	0,75	0,82	0,87	0,92	0,95	0,97	0,98	0,99
#LICZBA!	0,65	0,73	0,81	0,86	0,91	0,94	0,96	0,98	0,99
#LICZBA!	0,63	0,72	0,79	0,85	0,90	0,94	0,96	0,98	0,99
#LICZBA!	#LICZBA!	0,70	0,78	0,84	0,89	0,93	0,96	0,97	0,98
#LICZBA!	#LICZBA!	0,69	0,76	0,83	0,88	0,92	0,95	0,97	0,98
0,48	#LICZBA!	0,67	0,75	0,82	0,87	0,92	0,95	0,97	0,98
0,47	#LICZBA!	0,65	0,73	0,80	0,86	0,91	0,94	0,96	0,98
0,45	#LICZBA!	#LICZBA!	0,72	0,79	0,85	0,90	0,93	0,96	0,98

Chi/ss	790	800	810	820	830	840	850	860	870
886	0,01	0,02	0,03	0,05	0,08	0,13	0,18	0,25	0,34
888	0,01	0,02	0,03	0,05	0,08	0,12	0,17	0,24	0,32
890	0,01	0,01	0,02	0,04	0,07	0,11	0,16	0,23	0,30
892	0,01	0,01	0,02	0,04	0,06	0,10	0,15	0,21	0,29
894	0,01	0,01	0,02	0,03	0,06	0,09	0,14	0,20	0,27
896	0,00	0,01	0,02	0,03	0,05	0,08	0,13	0,19	0,26
898	0,00	0,01	0,02	0,03	0,05	0,08	0,12	0,17	0,24
900	0,00	0,01	0,01	0,03	0,04	0,07	0,11	0,16	0,23
902	0,00	0,01	0,01	0,02	0,04	0,06	0,10	0,15	0,21
904	0,00	0,01	0,01	0,02	0,04	0,06	0,09	0,14	0,20
906	0,00	0,00	0,01	0,02	0,03	0,05	0,09	0,13	0,19
908	0,00	0,00	0,01	0,02	0,03	0,05	0,08	0,12	0,17
910	0,00	0,00	0,01	0,01	0,03	0,04	0,07	0,11	0,16
912	0,00	0,00	0,01	0,01	0,02	0,04	0,07	0,10	0,15
914	0,00	0,00	0,01	0,01	0,02	0,04	0,06	0,09	0,14
916	0,00	0,00	0,01	0,01	0,02	0,03	0,05	0,09	0,13
918	0,00	0,00	0,00	0,01	0,02	0,03	0,05	0,08	0,12
920	0,00	0,00	0,00	0,01	0,01	0,03	0,05	0,07	0,11
922	0,00	0,00	0,00	0,01	0,01	0,02	0,04	0,07	0,10
924	0,00	0,00	0,00	0,01	0,01	0,02	0,04	0,06	0,10
926	0,00	0,00	0,00	0,01	0,01	0,02	0,03	0,06	0,09
928	0,00	0,00	0,00	0,00	0,01	0,02	0,03	0,05	0,08
930	0,00	0,00	0,00	0,00	0,01	0,02	0,03	0,05	0,07
932	0,00	0,00	0,00	0,00	0,01	0,01	0,02	0,04	0,07
934	0,00	0,00	0,00	0,00	0,01	0,01	0,02	0,04	0,06
936	0,00	0,00	0,00	0,00	0,01	0,01	0,02	0,03	0,06
938	0,00	0,00	0,00	0,00	0,00	0,01	0,02	0,03	0,05
940	0,00	0,00	0,00	0,00	0,00	0,01	0,02	0,03	0,05
942	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,03	0,04
944	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,02	0,04
946	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,02	0,04
948	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,02	0,03
950	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,02	0,03
952	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,03
954	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,02
956	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,02
958	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,02
960	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,02
962	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01
964	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01
966	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01
968	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01

880	890	900	910	920	930	940	950	960	970
0,43	#LICZBA!	#LICZBA!	0,70	0,78	0,84	0,89	0,93	0,96	0,97
0,41	#LICZBA!	#LICZBA!	0,68	0,76	0,83	0,88	0,92	0,95	0,97
0,39	0,48	#LICZBA!	0,67	0,75	0,82	0,87	0,91	0,95	0,97
0,37	0,47	#LICZBA!	0,65	0,73	0,80	0,86	0,91	0,94	0,96
0,36	0,45	#LICZBA!	#LICZBA!	0,72	0,79	0,85	0,90	0,93	0,96
0,34	0,43	#LICZBA!	#LICZBA!	0,70	0,78	0,84	0,89	0,93	0,95
0,32	0,41	#LICZBA!	#LICZBA!	0,68	0,76	0,83	0,88	0,92	0,95
0,30	0,39	0,48	#LICZBA!	0,67	0,75	0,82	0,87	0,91	0,94
0,29	0,37	0,47	#LICZBA!	#LICZBA!	0,73	0,80	0,86	0,91	0,94
0,27	0,36	0,45	#LICZBA!	#LICZBA!	0,72	0,79	0,85	0,90	0,93
0,26	0,34	0,43	#LICZBA!	#LICZBA!	0,70	0,77	0,84	0,89	0,93
0,24	0,32	0,41	#LICZBA!	#LICZBA!	0,68	0,76	0,83	0,88	0,92
0,23	0,31	0,39	0,48	#LICZBA!	#LICZBA!	0,75	0,81	0,87	0,91
0,21	0,29	0,37	0,47	#LICZBA!	#LICZBA!	0,73	0,80	0,86	0,90
0,20	0,27	0,36	0,45	#LICZBA!	#LICZBA!	0,71	0,79	0,85	0,90
0,19	0,26	0,34	0,43	#LICZBA!	#LICZBA!	0,70	0,77	0,84	0,89
0,18	0,24	0,32	0,41	#LICZBA!	#LICZBA!	0,68	0,76	0,82	0,88
0,16	0,23	0,31	0,39	0,48	#LICZBA!	#LICZBA!	0,74	0,81	0,87
0,15	0,22	0,29	0,37	0,47	#LICZBA!	#LICZBA!	0,73	0,80	0,86
0,14	0,20	0,27	0,36	0,45	#LICZBA!	#LICZBA!	0,71	0,79	0,85
0,13	0,19	0,26	0,34	0,43	#LICZBA!	#LICZBA!	0,70	0,77	0,84
0,12	0,18	0,24	0,32	0,41	#LICZBA!	#LICZBA!	#LICZBA!	0,76	0,82
0,11	0,17	0,23	0,31	0,39	0,48	#LICZBA!	#LICZBA!	0,74	0,81
0,10	0,15	0,22	0,29	0,38	0,47	#LICZBA!	#LICZBA!	0,73	0,80
0,10	0,14	0,20	0,28	0,36	0,45	#LICZBA!	#LICZBA!	0,71	0,79
0,09	0,13	0,19	0,26	0,34	0,43	#LICZBA!	#LICZBA!	0,70	0,77
0,08	0,12	0,18	0,25	0,32	0,41	#LICZBA!	#LICZBA!	#LICZBA!	0,76
0,08	0,11	0,17	0,23	0,31	0,39	0,48	#LICZBA!	#LICZBA!	0,74
0,07	0,11	0,16	0,22	0,29	0,38	0,47	#LICZBA!	#LICZBA!	0,73
0,06	0,10	0,14	0,20	0,28	0,36	0,45	#LICZBA!	#LICZBA!	0,71
0,06	0,09	0,13	0,19	0,26	0,34	0,43	#LICZBA!	#LICZBA!	#LICZBA!
0,05	0,08	0,12	0,18	0,25	0,33	0,41	#LICZBA!	#LICZBA!	#LICZBA!
0,05	0,08	0,12	0,17	0,23	0,31	0,39	0,48	#LICZBA!	#LICZBA!
0,04	0,07	0,11	0,16	0,22	0,29	0,38	0,47	#LICZBA!	#LICZBA!
0,04	0,06	0,10	0,15	0,21	0,28	0,36	0,45	#LICZBA!	#LICZBA!
0,04	0,06	0,09	0,14	0,19	0,26	0,34	0,43	#LICZBA!	#LICZBA!
0,03	0,05	0,08	0,13	0,18	0,25	0,33	0,41	#LICZBA!	#LICZBA!
0,03	0,05	0,08	0,12	0,17	0,23	0,31	0,39	0,48	#LICZBA!
0,03	0,04	0,07	0,11	0,16	0,22	0,29	0,38	0,47	#LICZBA!
0,02	0,04	0,07	0,10	0,15	0,21	0,28	0,36	0,45	#LICZBA!
0,02	0,04	0,06	0,09	0,14	0,19	0,26	0,34	0,43	#LICZBA!
0,02	0,03	0,05	0,09	0,13	0,18	0,25	0,33	0,41	#LICZBA!

9.2. Wpływ zaokrągleń na wyniki testu chi kwadrat

Kolejny problem, jaki pojawia się przy wyznaczaniu wartości mierników dopasowania związany jest z kwestią zaokrągleń. Poniżej przedstawia się wyniki obliczeń wartości testu χ^2 na podstawie danych analizowanych w klasycznej pracy F. Benforda[14]. Rezultaty ujęte są w formie dwóch tabel.

W tabeli 24 wzięto pod uwagę empiryczny rozkład pierwszych cyfr znaczących wynikający z sumy wszystkich 20 zbiorów analizowanych przez F. Benforda. Tabela składa się z 6 modułów (A-F), w których przytoczono wyznaczanie statystyki χ^2 w przypadku, gdy punktem wyjścia były rozkłady procentowe wynikające z rozkładu Benforda (B%) oraz empiryczne rozkłady cyfr znaczących (E%). Na podstawie tych udziałów procentowych wyznaczano teoretyczne $[B_i]$ oraz empiryczne $[E_i]$ liczebności **absolutne** n(i) rozkładów (łączna liczba obserwacji wynosiła tu 20229) oraz poszczególne elementy składowe statystyki $\chi^2 - (B\text{-}E)^2/B$.

W kolejnych modułach tabeli 24 przytoczono rozkłady procentowe w postaci:

(A) wyjściowej (zgodnej z danymi przytoczonymi w pracy F. Benforda);
(B) zaokrąglonej do najbliższej liczby całkowitej przy pomocy funkcji =ZAOKR(x;0);
(C) zaokrąglonej do najbliższej liczby całkowitej (jak w module B) oraz 2 korektach mających na celu doprowadzenie sumy liczby obserwacji do 100%;
(D) zaokrąglonej do najbliższej liczby całkowitej (jak w module B) oraz 2 innych korektach mających na celu doprowadzenie sumy liczby obserwacji do 100%;
(E) zaokrąglonej **w dół** do najbliższej liczby całkowitej za pomocą funkcji =ZAOKR.DO.CAŁK(x);
(F) zaokrąglonej do najbliższej liczby całkowitej (jak w module E) oraz 9 korektach mających na celu doprowadzenie sumy liczby obserwacji do 100%.

Jak wynika z podanego przykładu, zaokrąglanie elementów rozkładu powoduje wyraźne zmiany w statystyce χ^2. „Poprawna, wyjściowa wartość tej statystyki wynosi tu 85,2, natomiast wszystkie pozostałe statystyki są większe (nawet trzykrotnie) i zawierają się w granicach od 117,7 do 279,3.

[14] F. Benford, *The Law of Anomalous Numbers*, "Proceedings of the American Philosophical Society", 78/1938, s. 551–572.

W kolejnym eksperymencie założono, że punktem wyjścia są identyczne rozkłady procentowe jak w poprzednim przykładzie, ale z tą różnicą, że były one wyznaczane na podstawie nie 20 229, lecz tylko 5000 obserwacji. Wyniki analizy (w analogicznym układzie jak poprzednio) zawiera tabela 25. Jak można zauważyć, wartości statystyki χ^2 w każdym przypadku zmniejszyły się, przy czym podobnie jak poprzednio najmniejsze wartości uzyskano dla danych wyjściowych (χ^2. = 2,3), a dla wszystkich pozostałych zbiorów danych statystyki χ^2. kształtowały się na wyższym poziomie, w przedziale od 29,1 do 69,0.

Przy okazji warto zauważyć, że wartości statystyki χ^2 zależą od ogólnej liczby obserwacji. W omawianym przykładzie statystyki te w tabeli 27 są 4x wyższe niż w tabeli 28, gdyż liczba obserwacji w tabeli 27 (20229) jest 4x większa niż w tabeli 28 (5000).

Natomiast wartość krytyczna testu χ^2 nie zależy od liczby obserwacji *n*, lecz od poziomu istotności α oraz liczby przedziałów *k* porównywanych rozkładów, czyli liczby stopni swobody ss=k-1. Oznacza to, że im wyższa jest liczba obserwacji, tym większy musi być stopień podobieństwa rozkładów, aby można było je uznać za identyczne przy danym poziomie istotności.

Tabela 24. Efekt zaokrągleń przy wyznaczaniu statystyki χ^2 na podstawie danych F. Benforda, dla n=20229

(A) Lp	B%	E%	B n(i)	E n(i)	(B-E)^2/B
1	30,1	28,9	6090	5846	9,7
2	17,6	19,5	3562	3945	41,1
3	12,5	12,7	2527	2569	0,7
4	9,7	9,1	1960	1841	7,3
5	7,9	7,5	1602	1517	4,5
6	6,7	6,4	1354	1295	2,6
7	5,8	5,4	1173	1092	5,6
8	5,1	5,5	1035	1113	5,9
9	4,6	5,0	926	1011	8,0
Suma	100,0	100,0	20229	20229	85,2

=ZAOKR(x;0)

(B) Lp	B%	E%	B n(i)	E n(i)	(B-E)^2/B
1	30,0	29,0	6069	5866	6,7
2	18,0	20,0	3641	4046	45,0
3	12,0	13,0	2427	2630	16,9
4	10,0	9,0	2023	1821	20,2
5	8,0	8,0	1618	1618	
6	7,0	6,0	1416	1214	28,9
7	6,0	5,0	1214	1011	33,7
8	5,0	6,0	1011	1214	40,5
9	5,0	5,0	1011	1011	
Suma	101,0	101,0	20431	20431	191,9

=ZAOKR.DO.CAŁK(x)

(E) Lp	B%	E%	B n(i)	E n(i)	(B-E)^2/B
1	30,0	28,0	6069	5664	27,0
2	17,0	19,0	3439	3844	47,6
3	12,0	12,0	2427	2427	
4	9,0	9,0	1821	1821	
5	7,0	7,0	1416	1416	
6	6,0	6,0	1214	1214	
7	5,0	5,0	1011	1011	
8	5,0	5,0	1011	1011	
9	4,0	5,0	809	1011	50,6
Suma	95,0	96,0	19218	19420	125,1

Źródło: opracowanie własne.

(C) Lp	B%	E%	B n(i)	E n(i)	(B-E)^2/B
1	30,0	29,0	6069	5866	6,7
2	18,0	20,0	3641	4046	45,0
3	12,0	13,0	2427	2630	16,9
4	10,0	9,0	2023	1821	20,2
5	8,0	8,0	1618	1618	
6	7,0	6,0	1416	1214	28,9
7	5,0	5,0	1011	1011	
8	5,0	5,0	1011	1011	
9	5,0	5,0	1011	1011	
Suma	100,0	100,0	20229	20229	117,7

=ZAOKR(x;0) + 2 korekty

(D) Lp	B%	E%	B n(i)	E n(i)	(B-E)^2/B
1	30,0	29,0	6069	5866	6,7
2	17,0	20,0	3439	4046	107,1
3	12,0	13,0	2427	2630	16,9
4	10,0	9,0	2023	1821	20,2
5	8,0	7,0	1618	1416	25,3
6	7,0	6,0	1416	1214	28,9
7	6,0	5,0	1214	1011	33,7
8	5,0	6,0	1011	1214	40,5
9	5,0	5,0	1011	1011	
Suma	100,0	100,0	20229	20229	279,3

=ZAOKR.DO.CAŁK(x) + 9 korekt

(F) Lp	B%	E%	B n(i)	E n(i)	(B-E)^2/B
1	30,0	29,0	6069	5866	6,7
2	18,0	20,0	3641	4046	45,0
3	13,0	13,0	2630	2630	
4	10,0	9,0	2023	1821	20,2
5	8,0	8,0	1618	1618	
6	7,0	6,0	1416	1214	28,9
7	5,0	5,0	1011	1011	
8	5,0	5,0	1011	1011	
9	4,0	5,0	809	1011	50,6
Suma	100,0	100,0	20229	20229	151,4

Tabela 25. Efekt zaokrągleń przy wyznaczaniu statystyki χ^2 na podstawie danych F. Benforda, dla n=5000

(A) Lp	B%	E%	B n(i)	E n(i)	(B-E)^2/B
1	30,1	28,9	1505	1445	2,4
2	17,6	19,5	880	975	10,2
3	12,5	12,7	625	635	0,2
4	9,7	9,1	485	455	1,8
5	7,9	7,5	396	375	1,1
6	6,7	6,4	335	320	0,6
7	5,8	5,4	290	270	1,4
8	5,1	5,5	256	275	1,4
9	4,6	5,0	229	250	2,0
Suma	100,0	100,0	5000	5000	21,1

=ZAOKR(x;0)

(B) Lp	B%	E%	B n(i)	E n(i)	(B-E)^2/B
1	30,0	29,0	1500	1450	1,7
2	18,0	20,0	900	1000	11,1
3	12,0	13,0	600	650	4,2
4	10,0	9,0	500	450	5,0
5	8,0	8,0	400	400	0,0
6	7,0	6,0	350	300	7,1
7	6,0	5,0	300	250	8,3
8	5,0	6,0	250	300	10,0
9	5,0	5,0	250	250	0,0
Suma	101,0	101,0	5050	5050	47,4

=ZAOKR.DO.CAŁK(x)

(E) Lp	B%	E%	B n(i)	E n(i)	(B-E)^2/B
1	30,0	28,0	1500	1400	6,7
2	17,0	19,0	850	950	11,8
3	12,0	12,0	600	600	0,0
4	9,0	9,0	450	450	0,0
5	7,0	7,0	350	350	0,0
6	6,0	6,0	300	300	0,0
7	5,0	5,0	250	250	0,0
8	5,0	5,0	250	250	0,0
9	4,0	5,0	200	250	12,5
Suma	95,0	96,0	4750	4800	30,9

Źródło: opracowanie własne.

(C) Lp	B%	E%	B n(i)	E n(i)	(B-E)^2/B
1	30,0	29,0	1500	1450	1,7
2	18,0	20,0	900	1000	11,1
3	12,0	13,0	600	650	4,2
4	10,0	9,0	500	450	5,0
5	8,0	8,0	400	400	0,0
6	7,0	6,0	350	300	7,1
7	5,0	5,0	250	250	0,0
8	5,0	5,0	250	250	0,0
9	5,0	5,0	250	250	0,0
Suma	100,0	100,0	5000	5000	29,1

=ZAOKR(x;0) + 2 korekty

(D) Lp	B%	E%	B n(i)	E n(i)	(B-E)^2/B
1	30,0	29,0	1500	1450	1,7
2	17,0	20,0	850	1000	26,5
3	12,0	13,0	600	650	4,2
4	10,0	9,0	500	450	5,0
5	8,0	7,0	400	350	6,3
6	7,0	6,0	350	300	7,1
7	6,0	5,0	300	250	8,3
8	5,0	6,0	250	300	10,0
9	5,0	5,0	250	250	0,0
Suma	100,0	100,0	5000	5000	69,0

=ZAOKR.DO.CAŁK(x) + 9 korekt

(F) Lp	B%	E%	B n(i)	E n(i)	(B-E)^2/B
1	30,0	29,0	1500	1450	1,7
2	18,0	20,0	900	1000	11,1
3	13,0	13,0	650	650	0,0
4	10,0	9,0	500	450	5,0
5	8,0	8,0	400	400	0,0
6	7,0	6,0	350	300	7,1
7	5,0	5,0	250	250	0,0
8	5,0	5,0	250	250	0,0
9	4,0	5,0	200	250	12,5
Suma	100,0	100,0	5000	5000	37,4

Wzór pozwalający ustalić liczbę obserwacji, dla której można byłoby uznać, że porównywane rozkłady nie różnią się, przy wyjściowej wartości statystyki χ^2 oraz zadanym poziomie istotności α ma postać:

$$n' = \frac{n\chi^2\alpha}{\chi^2} \tag{25}$$

gdzie *n* to faktyczna liczba obserwacji natomiast χ^2_α to wartość krytyczna testu chi kwadrat ustalona przy ss=k-1 stopniach swobody i poziomie istotności α.

Jeżeli w omawianym przykładzie:

- n=20229,
- wartość empiryczna testu chi kwadrat χ^2=85,2
- liczba stopni swobody k-1=9-1=8
- poziom istotności α=0,05,
- wartość krytyczna testu chi kwadrat χ^2_α=15,5

to liczba obserwacji, przy której można byłoby uznać, że porównywane rozkłady nie różnią się od siebie na zadanym poziomie istotności α, wynosi:

$$n' = \frac{20229 * 15{,}5}{85{,}2} = 3680 \tag{26}$$

Zależność pomiędzy liczbą obserwacji a wartością statystyki χ^2 może być niekiedy powodem błędnych wniosków. Poniżej podano przykład, w którym wyznaczono rozkład pierwszych cyfr znaczących dla liczb będących silnią kolejnych liczb naturalnych od 1 do 170.

Dla n=170 wartość 170! wynosi 7,257415615308E+306 i jest to granica dokładności, jaką można uzyskać w Excelu. Na podstawie zbioru wartości tych 170 silni wyznaczono rozkłady pierwszych cyfr znaczących i obliczono wartość testu chi kwadrat, przy czym analizę wykonano w czterech wariantach – dla zbiorów będących wielokrotnością (3-4-10- oraz 20-krotność) zbioru źródłowego.

Jak wynika z tabeli 26 duże zbiory danych (3400 i 1700-elementowe) w świetle testu chi kwadrat należy uznać za niezgodne z rozkładem Benforda. Natomiast w przypadku małych zbiorów (340 i 510-elementowych) można przyjąć hipotezę o zgodności rozkładów pierwszych cyfr znaczących wartości n! z rozkładem Benforda – w pierwszym przypadku na poziomie istotności 0,003, natomiast w drugim na poziomie istotności 0,047.

Tabela 26. Test χ^2 dla wartości n! w zależności od wielkości zbioru danych

d	E	B	(E-B)^2/B	E	B	(E-B)^2/B
1	1080	1024	3,1	540	512	1,6
2	580	599	0,6	290	299	0,3
3	440	425	0,5	220	212	0,3
4	240	329	24,3	120	165	12,2
5	240	269	3,2	120	135	1,6
6	200	228	3,4	100	114	1,7
7	120	197	30,2	60	99	15,1
8	280	174	64,7	140	87	32,4
9	220	156	26,7	110	78	13,3
SUMA	3400	3400	156,7	1700	1700	78,3
	x20	x20	0,000000	x10	x10	0,000000

d	E	B	(E-B)^2/B	E	B	(E-B)^2/B
1	162	154	0,5	108	102	0,3
2	87	90	0,1	58	60	0,1
3	66	64	0,1	44	42	0,1
4	36	49	3,6	24	33	2,4
5	36	40	0,5	24	27	0,3
6	30	34	0,5	20	23	0,3
7	18	30	4,5	12	20	3,0
8	42	26	9,7	28	17	6,5
9	33	23	4,00	22	16	2,7
SUMA	510	510	23,5	340	340	15,7
	x3	x3	0,003	x2	x2	0,047

Źródło: opracowanie własne.

9.3. Pozostałe testy zgodności

Poza testem chi kwadrat w analizach często korzysta się z testu *z*, w którym dla każdej wartości występującej w rozkładzie wyznacza się statystyki:

$$z_i = \frac{p_i - \hat{p}_i}{\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{n}}} \qquad (i = 1,\ldots,k) \tag{27}$$

Jeżeli wartość z_i co do modułu jest większa od wartości krytycznej tego testu, to należy odrzucić hipotezę o zgodności rozkładu empirycznego z roz-

kładem Benforda dla *i*-tej cyfry (lub *i*-tej kombinacji cyfr). Wartość krytyczna tego testu dla α=0,05 wynosi 1,96, natomiast dla α=0,01 odpowiednia wartość krytyczna wynosi 2,58.

Podobnie jak test chi kwadrat, także i test *z* jest zależny od rozmiarów analizowanego zbioru danych. Im większy jest parametr n, tym trudniej jest uzyskać zgodność porównywanych udziałów.

Innym testem służącym do oceny zgodności rozkładów jest test Kołmogorowa-Smirnowa, niezależny od wielkości zbioru n. W wersji oryginalnej tego testu wyznacza się statystykę:

$$KS1 = D\sqrt{\frac{n^2}{2n}} \qquad D = \max_i \left| f_i - \hat{f}_i \right| \qquad (i = 1,...,k) \tag{28}$$

W literaturze proponuje się także inne wersje tego testu. W jednej z nich wyznacza się statystykę:

$$KS2 = D\sqrt{n} \qquad D = \max_i \left| f_i - \hat{f}_i \right| \qquad (i = 1,...,k) \tag{29}$$

Inna modyfikacja testu Kołmogorowa-Smirnowa została zaproponowana przez Kuipera[15]. Uwzględnia ona fakt cykliczności analizowanych rozkładów (*circular distrubution*). Chodzi o to, że różnica pomiędzy liczbami 99,99 a 100,01 jest minimalna, podczas gdy w sensie analizy rozkładu pierwszych cyfr znaczących odpowiadające im liczby 9 i 1 znajdują się na przeciwnych biegunach skali. W związku z powyższym proponuje się wykorzystywać statystykę:

$$KS3 = V_N * [\sqrt{N} + 0{,}155 + 0{,}24N^{-1/2}] \qquad gdzie \qquad (i = 1,...,k)$$
$$V_N = D_N^+ + D_N^- \quad D_N^+ = \sup_i \left[f_i - \hat{f}_i \right] \quad D_N^- = \sup_i \left[\hat{f}_i - f_i \right] \quad N = \frac{n^2}{2n} \tag{30}$$

Wartości krytyczne testów KS1-KS2 wynoszą 1,36 dla α=0,05 oraz 1,63 dla α=0,01. W przypadku zmodyfikowanego testu Kołmogorowa-Smirnowa KS3 wartości krytyczne wynoszą odpowiednio: 1,747 dla α=0,05 oraz 2,001 dla α=0,01.

[15] N.H. Kuiper, *Alternative proof of a theorem of Birnhaum and Pyke*, "Annals of Mathematical Statistics" 30/1959, s. 251–252.

9.4. Mierniki zgodności empirycznych rozkładów cyfr z rozkładami wynikającymi z prawa Benforda

Alternatywnym sposobem pomiaru zgodności dwóch rozkładów są miary ich podobieństwa. Można tu wymienić następujące mierniki.

$$M_1 = \frac{100}{k}\sum_{i=1}^{k}\left|\frac{c_i - \hat{c}_i}{\hat{c}_i}\right| \tag{31}$$

$$M_2 = \frac{1}{k}\sqrt{\sum_{i=1}^{k}(c_i - \hat{c}_i)^2} \tag{32}$$

$$M_3 = \sqrt{\frac{\sum_{i=1}^{k}(c_i - \hat{c}_i)^2}{k}} \tag{33}$$

$$M_4 = \frac{100\sqrt{\sum_{i=1}^{k}(c_i - \hat{c}_i)^2}}{\sqrt{\sum_{i=1}^{k}\hat{c}_i^2}} \tag{34}$$

$$M_5 = \frac{100\sum_{i=1}^{k}|n_i - \hat{n}_i|}{n} \tag{35}$$

Mierniki te są niezależne od wielkości zbioru n i przyjmują tym mniejsze wartości, im bardziej zgodne ze sobą są porównywane rozkłady częstości. Generalnie wskazują one na przeciętną wielkość różnic pomiędzy częstościami faktycznymi a częstościami teoretycznymi w danym teście[16].

[16] Można sformułować analogiczne formuły, w których zamiast częstości c_i występują liczebności n_i- z identyczną interpretacją i wartościami.

Mierniki M2 oraz M3 wskazują na przeciętną różnicę pomiędzy częstościami empirycznymi a teoretycznymi porównywanych rozkładów. W literaturze preferuje się miernik M3, jakkolwiek bardziej naturalny wydaje się miernik M2, gdyż jego interpretacja jest bardziej zbliżona do przeciętnej wielkości różnicy pomiędzy empirycznymi i teoretycznymi częstościami. Miernik M3 jest większy od miernika M2 (w przypadku testów F1, D2, D3, L1 – trzykrotnie większy) i szacuje z dużym nadmiarem wielkość różnic pomiędzy porównywanymi rozkładami.

Mierniki M1 oraz M4 wskazują, jaka jest przeciętna różnica między częstościami empirycznymi a teoretycznymi w relacji do częstości teoretycznych rozkładu. Mierniki M5 pokazuje, jaką część wszystkich obserwacji trzeba by zamienić miejscami, aby rozkłady empiryczne pokryły się z rozkładami teoretycznymi. Wielkości tych trzech mierników (M1, M4, M5) wyrażone są w procentach.

Kolejnym miernikiem zgodności może być współczynnik korelacji linowej *r* pomiędzy empirycznymi i teoretycznymi liczebnościami rozkładu:

$$r = \frac{\sum_{i=1}^{k} \left(n_i - \overline{n}\right)\left(\hat{n}_i - \overline{\hat{n}}\right)}{\sqrt{\sum_{i=1}^{k} \left(n_i - \overline{n}\right)^2 \sum_{i=1}^{k} \left(\hat{n}_i - \overline{\hat{n}}\right)^2}} \tag{36}$$

W powyższym wzorze $\overline{\hat{n}}$ oraz $\overline{\hat{n}}$ to średnie arytmetyczne z empirycznych i teoretycznych liczebności. Identyczne wartości współczynników korelacji uzyskuje się, jeżeli we wzorze (36) przyjmie się nie liczebności rozkładów, ale ich częstości lub prawdopodobieństwa. Współczynnika *r* nie można wyznaczyć dla testu L1, a także D3, gdyż częstości teoretyczne w tych przypadkach są identyczne i nie mają żadnej zmienności.

9.5. Analiza zbieżności mierników zgodności rozkładów

W tabeli 27 podano rozkłady pierwszych cyfr znaczących w zbiorach analizowanych przez F. Benforda wraz z wyznaczonymi na ich podstawie miernikami M1-M5, statystykami $\chi 2$, statystykami testu Kołmogorowa-Smirnowa (KS1, KS2, KS3) oraz parametrami wynikającymi z wartości testu *z*. W tym ostatnim przypadku jest to średnia z modułów statystyk z (ostatnia kolumna tabeli 27) oraz liczba cyfr, dla których statystyki z wska-

zują na istotną rozbieżność pomiędzy porównywanymi częstościami rozkładów. Liczba ta może przyjmować wartości od 0 do 9 i jest ustalana w czterech wariantach, różniących się poziomem istotności, przy którym należy uznać, że porównywane częstości są istotnie różne. Przyjęto następujące progi wartości krytycznych testu *z*:

- z>1,64, dla poziomu istotności α=0,1;
- z>1,96, dla poziomu istotności α=0,05;
- z>2,58, dla poziomu istotności α=0,01;
- z>3,29, dla poziomu istotności α=0,001.

Ponadto wyznaczono współczynniki korelacji pomiędzy empirycznymi a teoretycznymi składowymi rozkładu pierwszych cyfr wraz z prawdopodobieństwem, przy którym należy odrzucić hipotezę o braku istotnego skorelowania. Ze względu na ujemne skorelowania z pozostałymi miernikami tych dwóch ostatnich parametrów zostały one zastąpione dopełnieniami do jedności ich modułów: *1-mod(r) 1-p(r)*. Dzięki tej operacji przy interpretacji wszystkich mierników dopasowania i testów można przyjąć zasadę, że im mniejsze są wartości tych parametrów, tym lepiej dany rozkład empiryczny jest dopasowany do rozkładu Benforda.

Ta zasada nie dotyczy liczebności zbioru danych (pierwsza kolumna tabeli 27). Parametr *n* uwzględniono dla sprawdzenia, czy istnieje związek pomiędzy wielkością zbioru danych a miernikami charakteryzującymi stopień dopasowania rozkładów.

Zbiory uporządkowane są według rosnących wartości testu χ^2, od zbiorów najbardziej do najmniej zgodnych z rozkładem Benforda. Jak można zauważyć, tylko część (12 na 22) zbiorów ma rozkłady cyfr znaczących zgodne w sensie statystyki chi kwadrat z prawem Benforda – są to zbiory, dla których wartość empiryczna statystyki chi kwadrat jest mniejsza od 15,5, przy założeniu 5% poziomu istotności. 5 zbiorów wymienionych na końcu tabeli 30 (jest wśród nich zbiorowość sumaryczna) ewidentnie mają rozkłady niezgodne z prawem Benforda. W przypadku pozostałych 5 zbiorów wymienionych w środku tabeli 27 można przyjąć założenie o zgodności rozkładu pierwszych cyfr znaczących z rozkładem Benforda, pod warunkiem obniżenia poziomu istotności do α= 0,01 (3 zbiory) lub α= 0,001 (2 zbiory).

W tabeli 28 podano rangi zbiorów danych odnoszące się do wartości mierników ich dopasowania do rozkładu Benforda (1 – zbiór najlepiej dopasowany, 22 – zbiór najgorzej dopasowany). Zbiory danych uporządkowano według średniej rangi, ze względu na wszystkie analizowane mierniki (ostatnia kolumna tabeli 28). Stworzenie takiej syntetycznej miary dopasowania zbiorów

ze względu na wszystkie mierniki zawarte w tabeli 27 było niemożliwe z uwagi na różne znaczenie i różny zakres wartości poszczególnych mierników.

Porównując uporządkowanie zbiorów danych z tabeli 27 (według statystyki χ^2) z ich uporządkowaniem w tabeli 28 (ze względu na wszystkie mierniki) można stwierdzić nieznaczne różnice. Sprowadzają się one do przesunięcia kilku zbiorów danych o 1 (rzadziej o dwie) pozycje w górę lub w dół. Tak więc można przyjąć, że test χ^2 jest miarodajnym miernikiem poprawności wnioskowania o podobieństwie rozkładów cyfr znaczących.

W tabeli 29 przytoczono współczynniki korelacji pomiędzy poszczególnymi miernikami dopasowania. W ostatnich wierszach tej tabeli podano liczbę dodatnich oraz ujemnych współczynników korelacji oraz średnią z modułów tych współczynników dla każdego miernika dopasowania. Jak można zauważyć, z wyjątkiem kilku współczynników korelacji pomiędzy liczebnością zbiorów *n* a miernikami M1-M5 i współczynnikiem *1-r* wszystkie pozostałe współczynniki korelacji są dodatnie. Nawet te ujemne współczynniki korelacji są niewielkie i statystycznie nieistotne (dla α=005 oraz przy n=20 wartość krytyczna modułu współczynnika korelacji wynosi 0,47). Oznacza to, że wszystkie mierniki dopasowania mają identyczne zasady interpretacji – im większe przyjmują wartości, tym dany rozkład jest mniej podobny do rozkładu Benforda.

Warto tu zwrócić uwagę na dużą zgodność wskazań mierników w ramach ich grup definicyjnych. I tak mierniki M2-M3-M4 dają identyczne oceny stopnia zgodności (przy różnych wartościach mierników). Wynika stąd, że w analizach wystarcza skorzystać z dowolnego miernika z tej grupy. Podobne rezultaty (r=0,98) jak mierniki M2-4 daje miernik M, a nieco bardziej odmienne (r=0,92) – miernik M4.

Również identyczne wnioski uzyskuje się w przypadku statystyk Kołmogorowa-Smirnowa KS1-KS2. Zgodność wskazań trzeciego testu KS3 z dwoma poprzednimi jest też wysoka (r=0,95).

Miary oparte na statystyce *z* cechuje duży, jakkolwiek mniejszy w poprzednich przypadkach stopień zgodności wskazań. Interesujący jest współczynnik korelacji pomiędzy testem χ^2 a poziomem istotności odpowiadającym tej statystyce. Jest on zadziwiająco niski (r=0,45). Można nawet powiedzieć, że te dwie miary są ze sobą nieskorelowane.

Biorąc pod uwagę sumaryczny stopień skorelowania danego miernika z pozostałymi (ostatni wiersz tabeli 29), najbardziej diagnostyczne w sensie zgodności wskazań są statystyki Kołmogorowa-Smirnowa, zwłaszcza statystyka KS3, a w następnej kolejności mierniki oparte na statystyce *z*.

W tabeli 30 przedstawiono wyniki podziału współczynników korelacji na cztery kategorie ze względu na ich poziom. Kryterium podziału stanowiły wartości krytyczne wynikające Studenta dla różnych poziomów istotności[17]:

- Kategoria 3 – poziom wysoki r>0,8, α1= α2=0,00001
- Kategoria 2 – poziom średni (0,5<r<0,8) α1=0,009 α2=0,019
- Kategoria 1 – poziom niski (0,35<r<0,5) α1=0,056 α2=0,112
- Kategoria 0 – brak korelacji r<0,35

W ostatniej kolumnie tabeli podano średnie wartości rang przypisanych poszczególnym kategoriom. Stanowią one syntetyczną miarę zgodności danego miernika z pozostałymi. Kolumny i wiersze macierzy współczynników uporządkowane zostały według malejących wartości tej średniej. Jak się okazuje, wśród mierników dających wskazania najbardziej podobne do wskazań innych mierników należą dwa mierniki oparte na statystyce *z (z>1,64 oraz z>1,96)* oraz statystyka Kołmogorowa-Smirnowa KS3. Z drugiej strony znajdują się mierniki o najmniejszej zgodności z pozostałymi *–r(p), M2, M3, M4*, a także statystyka χ^2.

Podobną analizę można przeprowadzić, wykorzystując nie średnie wartości rang, lecz rangi danego typu, np. tylko kategorii „3". W ostatnich wierszach tabeli 30 podano liczbę mierników zaliczonych do poszczególnych kategorii. Najwięcej kategorii „3" jest przypisanych do pięciu mierników: *KS1-KS2-KS3, z sred* oraz *z>2,58*. Jest to więc nieco inny zestaw mierników, niż wynikało to z ich uporządkowania według średnich wartości rang.

Tabela 31 ma analogiczną konstrukcję jak tabela 30, z tym że zawiera informacje uzyskane nie na podstawie macierzy współczynników korelacji liniowej, lecz macierzy współczynników korelacji rang Spearmana, obrazujących stopień zgodności pomiędzy poszczególnymi miernikami opisującymi podobieństwo rozkładów cyfr znaczących w zbiorach Benforda z prawem Benforda.

Przyjęto nieco wyższe wartości graniczne określające poszczególne kategorie współczynników. Zamiast [0,35-0,5-0,8] są to wartości [0,5-0,75-0,9]. Wynikało to z wyższego poziomu wartości współczynników korelacji Spearmana niż współczynników korelacji liniowej Pearsona. Pomimo podwyższonych wartości granicznych średni poziom zgodności wszystkich mierników

[17] Podano tu wartości krytyczne testu Studenta zarówno dla jednostronnego α1, jak i dwustronnego α2 obszaru krytycznego Z uwagi na dodatniość wszystkich współczynników korelacji bardziej właściwy wydaje się jednostronny obszar krytyczny. Wartości rozkładu Studenta wyznaczane są przy pomocy funkcji Excela ROZKŁAD.T (t; ss; 1) gdzie $t=[r/(1-r^2)^{1/2}]*ss^{1/2}$ oraz ss=n-2, natomiast *n* to liczba obserwacji (w tym przypadku n=22).

zawartych w macierzy (prawy dolny narożnik tabel) w przypadku współczynników Spearmana i tak jest wyższy (1,81) niż dla współczynników Pearsona (1,67).

Analiza parametrów zawartych w tabeli 31 tylko częściowo potwierdza rezultaty analizy uzyskane poprzednio. Do „zgodnych" mierników nadal zalicza się statystyki KS1-KS2-KS3, ale także statystykę χ^2, która poprzednio była zaliczona do grupy mierników dających odmienne wskazania niż pozostałe mierniki. Wśród parametrów o niskiej zgodności analiza współczynników Spearmana nie wykazuje mierników M2-M4, które to mierniki wynikały z analizy współczynników Pearsona.

Reasumując, problem oceny diagnostyczności testów i mierników charakteryzujących podobieństwo rozkładów nie jest prosty i wymaga dużej rozwagi oraz dalszych badań. Badania te powinny mieć charakter symulacyjny i opierać się na danych generowanych z kontrolowanym stopniem podobieństwa rozkładów.

9.6. Klasyfikacja mierników podobieństwa rozkładów taksonometryczną metodą Czekanowskiego

Macierz mierników podobieństwa rozkładów (tabela 27), którą w ogólnej postaci można zapisać:

$$[X_{ij}] \quad (i = 1,2 \ldots , p; \qquad j = 1,2, \ldots q) \tag{37}$$

gdzie *p* to liczba analizowanych zbiorów danych, natomiast *q* – liczba mierników podobieństwa (w przypadku analizy zbiorów Benforda p=20, q=18) może stanowić punkt wyjścia analizy taksonometrycznej[18]. Jej celem jest klasyfikacja zbioru obiektów (dane Benforda lub mierniki podobieństwa) na bardziej jednorodne podzbiory. Chodzi o taki podział, aby elementy danego podzbioru były jak najbardziej podobne do siebie z punktu widzenia opisujących je charakterystyk, a jednocześnie jak najmniej podobne do obiektów tworzących inne podzbiory.

Klasyfikacja pozwala wprowadzić porządek do analizowanych zjawisk, a tym samym ułatwia i sprzyja wyciąganiu poprawnych wniosków. Dla

[18] Przegląd metod taksonometrii znaleźć można m.in. w pracach: T. Grabiński, *Metody taksonometrii*, Akademia Ekonomiczna w Krakowie, Kraków 1991; *idem*, *Analiza taksonometryczna krajów Europy w ujęciu regionów*, Akademia Ekonomiczna w Krakowie, Kraków 2003.

przykładu, jeżeli przeprowadzimy klasyfikację 20 mierników podobieństwa i okaże się, że dają się one podzielić np. na 3 jednorodne podzbiory, to w analizie można nie uwzględniać wszystkich 20 mierników, a tylko 3, będące reprezentantami grup, do których te mierniki należą.

Klasyfikacji mogą podlegać obiekty znajdujące się w wierszach macierzy danych, które opisane są charakterystykami (cechami) ujętymi w kolumnach macierzy (16), względnie cechy traktowane jako punkty w wielowymiarowej przestrzeni obiektów (tzw. zadanie dualne). W literaturze znanych jest wiele metod klasyfikacji. W niniejszej pracy wykorzystano najstarszą metodę taksonometryczną opracowaną przez polskiego antropologa Jana Czekanowskiego. Program obliczeniowy dostępny jest na witrynie P. Jaskulskiego Archeo-Data poświęconej problematyce wykorzystania komputerów w antropologii i archeologii[19].

Rezultaty analizy taksonometrycznej przedstawiono w postaci tzw. diagramów Czekanowskiego: rysunki 23–26 (klasyfikacja zbiorów Benforda w przestrzeni miar zgodności rozkładów) oraz rysunki 27–30 (klasyfikacja miar zgodności rozkładów w przestrzeni zbiorów Benforda). W diagramach Czekanowskiego symbolami graficznymi o coraz to większym stopniu zaczernienia przedstawia się poziomy miar podobieństwa klasyfikowanych obiektów (zbiorów Benforda, miar zgodności rozkładów). Im bardziej zaczerniony jest element diagramu, tym bardziej podobne do siebie są obiekty przyporządkowane temu elementowi.

[19] http://eskimo73.republika.pl/maczek.html oraz eskimo73.republika.pl/download/manual_30_pl.pdf.

Tabela 27. Miary dopasowania dla 20 zbiorów analizowanych przez F. Benforda

Symbol	Nazwa	n	1-r	r(p)	M5	M1	M2	M3	M4	chi	1-chi-(p)	KS1	KS2	KS3	z>1,64	z>1,96	z>2,58	z>3,29	z sred
D	Czasopisma	100	0,001	0,000	2,8	4,0	0,12	0,37	2,7	0,16	0,000	0,04	0,05	0,06					0,11
F	Ciśnienie	703	0,001	0,000	3,2	4,1	0,14	0,42	3,1	1,27	0,004	0,18	0,26	0,28					0,33
R	Adresy	342	0,005	0,000	5,4	5,5	0,26	0,78	5,7	1,30	0,004	0,16	0,22	0,22					0,35
	Średnia	1011	0,001	0,000	3,2	3,9	0,15	0,44	3,3	1,75	0,012	0,31	0,44	0,34					0,38
M	Dane Reader's Digest	308	0,004	0,000	8,4	7,8	0,46	1,39	10,3	3,23	0,081	0,52	0,73	0,53					0,49
G	Wiatr	690	0,003	0,000	4,8	6,8	0,22	0,65	4,8	3,46	0,098	0,31	0,44	0,34					0,51
A	Dorzecza rzek	335	0,012	0,000	9,9	12,5	0,41	1,22	9,0	4,96	0,238	0,27	0,39	0,39					0,70
O	Promienie X	707	0,007	0,000	7,4	8,2	0,36	1,08	8,0	5,43	0,289	0,43	0,61	0,44					0,70
T	Śmiertelność	418	0,022	0,000	11,2	11,5	0,55	1,66	12,2	7,55	0,522	0,45	0,63	0,61	1	1			0,80
Q	Prom. ciała czarnego	1165	0,006	0,000	7,3	9,5	0,30	0,91	6,7	9,52	0,700	0,53	0,75	0,73	2				0,97
I	Drenaż	159	0,058	0,000	21,6	26,5	0,97	2,91	21,5	11,14	0,806	0,69	0,98	0,98	1	1			1,02
P	Amer. liga baseball	1458	0,002	0,000	6,6	9,0	0,36	1,09	8,0	14,60	0,932	0,76	1,07	0,76	2	2	1		0,98
N	Koszty	741	0,011	0,000	12,3	15,3	0,51	1,53	11,3	15,60	0,952	0,67	0,95	0,68	3	1			1,27
J	Masa atomowa	91	0,048	0,000	38,2	33,5	2,18	6,54	48,2	17,25	0,972	1,23	1,74	1,33	3	2	1	1	1,18
L	Dane z projektów	560	0,021	0,000	16,6	20,0	0,68	2,03	15,0	19,21	0,986	1,09	1,54	1,10	4	1			1,46
C	Stałe	104	0,095	0,001	39,8	48,1	1,82	5,46	40,3	24,44	0,998	0,81	1,14	1,27	4	4	1		1,50
S	n^1, n^2, ... n!	900	0,003	0,000	13,8	16,4	0,67	2,02	14,9	24,99	0,998	1,46	2,07	1,48	3	3	2		1,52
	Suma	20229	0,005	0,000	5,7	6,2	0,27	0,82	6,0	84,10	1,000	1,19	1,69	2,03	7	7	4	2	2,87
E	Ciepło	1389	0,093	0,001	24,2	26,6	1,09	3,28	24,2	111,21	1,000	1,61	2,27	3,21	6	6	6	6	3,24
B	Populacja	3259	0,004	0,000	16,6	20,4	0,69	2,08	15,3	118,63	1,000	3,35	4,73	3,36	7	7	7	5	3,54
H	Masa cząsteczkowa	1800	0,068	0,000	23,2	25,9	1,07	3,22	23,7	125,76	1,000	2,46	3,48	3,50	8	7	7	3	3,54
K	n^(-1), n^(0,5)	5000	0,066	0,000	22,8	30,6	0,95	2,86	21,1	440,76	1,000	4,36	6,16	4,37	8	8	8	8	6,28
α	Wart. teoret. chi kw.	0,05	15,507	0,01	20,09	0,001	26,124												

Źródło: opracowanie własne na podstawie pracy: F. Benford, *The Law of Anomalous Numbers*, „Proceedings of the Amercian Philosophical Society", 78/1938, s. 551–572.

Tabela 28. Rangi zbiorów Benforda według mierników zgodności rozkładów cyfr znaczących

Lp.	Symbol	Nazwa	n	1-r	r(p)	M5	M1	M2	M3	M4	chi	1-ch(p)	KS1	KS2	KS3	z>1,64	z>1,96	z>2,58	z>3,29	z sred	R
1	D	Czasopisma	2	2	2	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1,2
2	F	Ciśnienie	11	3	3	3	3	2	2	2	2	2	3	3	3	1	1	1	1	2	2,7
3		Średnia	15	1	1	2	1	3	3	3	4	4	5	5	5	1	1	1	1	4	3,3
4	R	Adresy	7	9	9	5	4	5	5	5	3	3	2	2	2	1	1	1	1	3	3,8
5	G	Wiatr	10	6	6	4	6	4	4	4	6	6	6	6	4	1	1	1	1	6	4,6
6	M	Reader's Digest	5	7	7	10	7	11	11	11	5	5	9	9	8	1	1	1	1	5	6,3
7	A	Dorzecza rzek	6	14	14	11	12	10	10	10	7	7	4	4	6	1	1	1	1	7	7,0
8	O	Promienie X	12	12	12	9	8	8	8	8	8	8	7	7	7	1	1	1	1	8	7,0
9	Q	Prom. ciała czarnego	16	11	11	8	10	7	7	7	10	10	10	10	11	11	1	1	1	10	8,4
10	T	Śmiertelność	8	16	16	12	11	13	13	13	9	9	8	8	9	9	10	1	1	9	9,7
11	P	Amer. liga baseball	18	4	4	7	9	9	9	9	12	12	13	13	12	11	14	14	1	11	10,1
12	N	Koszty	13	13	13	13	13	12	12	12	13	13	11	11	10	13	10	1	1	14	11,0
13	I	Drenaż	4	18	18	17	18	18	18	18	11	11	12	12	13	9	10	1	1	12	12,3
14	L	Dane z projektów	9	15	15	16	15	15	15	15	15	15	15	15	14	16	10	1	1	15	12,9
15	S	n^1, n^2, ... n!	14	5	5	14	14	14	14	14	17	17	18	18	17	13	16	17	1	17	13,6
16		SUMA	22	10	10	6	5	6	6	6	18	18	16	16	18	19	19	18	18	18	13,8
17	J	Masa atomowa	1	17	17	21	21	22	22	22	14	14	17	17	16	13	14	14	17	13	16,2
18	C	Stałe	3	22	22	22	22	21	21	21	16	16	14	14	15	16	17	14	1	16	16,3
19	B	Populacja	20	8	8	15	16	16	16	16	20	19	21	21	20	19	19	20	20	21	17,5
20	E	Ciepło	17	21	21	20	19	20	20	20	19	19	19	19	19	18	18	19	21	19	19,3
21	H	Masa cząsteczkowa	19	20	20	19	17	19	19	19	21	19	20	20	21	21	19	20	19	20	19,6
22	K	n^(-1), n^(0,5)	21	19	19	18	20	17	17	17	22	19	22	22	22	21	22	22	22	22	20,2

Źródło: opracowanie własne.

Tabela 29. Współczynniki korelacji liniowej pomiędzy miernikami zgodności

	n	1-r	r(p)	M5	M1	M2	M3	M4	chi	1-chi(p)	KS1	KS2	KS3	z>1,64	z>1,96	z>2,58	z>3,29	z sred
n	1,00	-0,08	-0,07	-0,14	-0,13	-0,14	-0,14	-0,14	0,32	0,30	0,27	0,27	0,35	0,51	0,55	0,40	0,31	0,43
1-r	-0,08	1,00	0,89	0,85	0,87	0,79	0,79	0,79	0,44	0,53	0,42	0,42	0,58	0,53	0,54	0,48	0,50	0,53
r(p)	-0,07	0,89	1,00	0,70	0,74	0,62	0,62	0,62	0,28	0,40	0,25	0,25	0,45	0,43	0,47	0,39	0,42	0,38
M5	-0,14	0,85	0,70	1,00	0,98	0,98	0,98	0,98	0,31	0,66	0,45	0,45	0,51	0,52	0,48	0,37	0,36	0,43
M1	-0,13	0,87	0,74	0,98	1,00	0,92	0,92	0,92	0,40	0,69	0,51	0,51	0,56	0,56	0,53	0,42	0,41	0,51
M2	-0,14	0,79	0,62	0,98	0,92	1,00	1,00	1,00	0,23	0,61	0,38	0,38	0,44	0,44	0,41	0,30	0,29	0,35
M3	-0,14	0,79	0,62	0,98	0,92	1,00	1,00	1,00	0,23	0,61	0,38	0,38	0,44	0,44	0,41	0,30	0,29	0,35
M4	-0,14	0,79	0,62	0,98	0,92	1,00	1,00	1,00	0,23	0,61	0,38	0,38	0,44	0,44	0,41	0,30	0,29	0,35
chi	0,32	0,44	0,28	0,31	0,40	0,23	0,23	0,23	1,00	0,43	0,89	0,89	0,85	0,73	0,76	0,82	0,89	0,93
1-chi(p)	0,30	0,53	0,40	0,66	0,69	0,61	0,61	0,61	0,43	1,00	0,63	0,63	0,68	0,81	0,71	0,57	0,46	0,65
KS1	0,27	0,42	0,25	0,45	0,51	0,38	0,38	0,38	0,89	0,63	1,00	1,00	0,95	0,86	0,86	0,91	0,88	0,95
KS2	0,27	0,42	0,25	0,45	0,51	0,38	0,38	0,38	0,89	0,63	1,00	1,00	0,95	0,86	0,86	0,91	0,88	0,95
KS3	0,35	0,58	0,45	0,51	0,56	0,44	0,44	0,44	0,85	0,68	0,95	0,95	1,00	0,93	0,95	0,97	0,92	0,97
z>1,64	0,51	0,53	0,43	0,52	0,56	0,44	0,44	0,44	0,73	0,81	0,86	0,86	0,93	1,00	0,96	0,90	0,79	0,91
z>1,96	0,55	0,54	0,47	0,48	0,53	0,41	0,41	0,41	0,76	0,71	0,86	0,86	0,95	0,96	1,00	0,95	0,84	0,92
z>2,58	0,40	0,48	0,39	0,37	0,42	0,30	0,30	0,30	0,82	0,57	0,91	0,91	0,97	0,90	0,95	1,00	0,93	0,94
z>3,29	0,31	0,50	0,42	0,36	0,41	0,29	0,29	0,29	0,89	0,46	0,88	0,88	0,92	0,79	0,84	0,93	1,00	0,93
z sred	0,43	0,53	0,38	0,43	0,51	0,35	0,35	0,35	0,93	0,65	0,95	0,95	0,97	0,91	0,92	0,94	0,93	1,00
>0	10	16	16	16	16	16	16	16	17	17	17	17	17	17	17	17	17	17
<0	7	1	1	1	1	1	1	1										
średnia	0,31	0,61	0,50	0,62	0,64	0,57	0,57	0,57	0,59	0,61	0,67	0,67	0,72	0,70	0,70	0,66	0,63	0,69

Źródło: opracowanie własne.

Tabela 30. Kategorie współczynników korelacji liniowej *r* między miernikami zgodności rozkładów cyfr znaczących ze względu na ich poziom [0<0,35]; [1 <0,35;0,5]; [2 <0,5;0,8]; [3 >0,8]

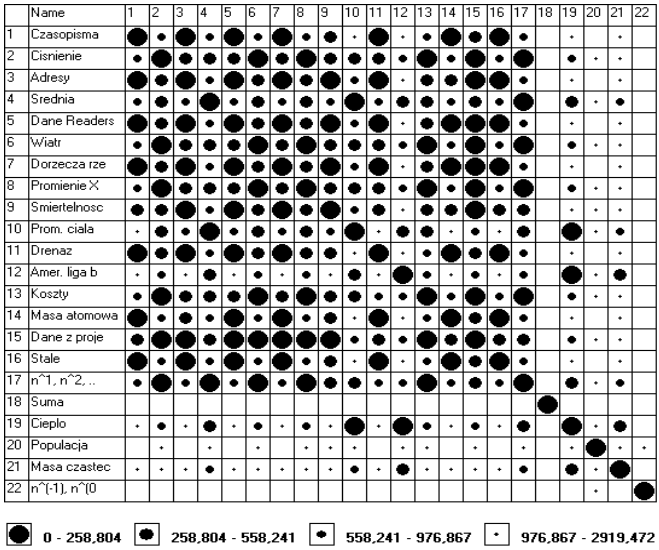
	z>1,64	KS3	z>1,96	M1	KS1	KS2	z sred	1-r	z>2,58	1-chi(p)	M5	z>3,29	chi	M2	M3	M4	r(p)	n	Średnia
z>1,64	3	3	3	2	3	3	3	2	3	3	2	2	2	1	1	1	1	2	2,18
KS3	3	3	3	2	3	3	3	2	3	2	2	3	3	1	1	1	1		2,12
z>1,96	3	3	3	2	3	3	3	2	3	2	1	3	2	1	1	1	1	2	2,12
M1	2	2	2	3	2	2	2	3	1	2	3	1	1	3	3	3	2		2,00
KS1	3	3	3	2	3	3	3	1	3	2	1	3	3	1	1	1			1,94
KS2	3	3	3	2	3	3	3	1	3	2	1	3	3	1	1	1			1,94
z sred	3	3	3	2	3	3	3	2	3	2	1	3	3				1	1	1,94
1-r	2	2	2	3	1	1	2	3	1	2	3	2	1	2	2	2	3		1,82
z>2,58	3	3	3	1	3	3	3	1	3	2	1	3	3				1	1	1,82
1-chi(p)	3	2	2	2	2	2	2	2	2	3	2	1	1	2	2	2	1		1,76
M5	2	2	1	3	1	1	1	3	1	2	3	1		3	3	3	2		1,71
z>3,29	2	3	3	1	3	3	3	2	3	1	1	3	3				1		1,71
chi	2	3	2	1	3	3	3	1	3	1		3	3						1,47
M2	1	1	1	3	1	1		2		2	3			3	3	3	2		1,35
M3	1	1	1	3	1	1		2		2	3			3	3	3	2		1,35
M4	1	1	1	3	1	1		2		2	3			3	3	3	2		1,35
r(p)	1	1	1	2			1	3	1	1	2	1		2	2	2	3		1,18
n	2		2				1		1									3	0,35
L.3>0,8	7	8	7	5	8	8	8	3	8	1	5	7	6	4	4	4	1		91
L.2<0,8	6	4	5	8	2	2	3	9	1	12	4	2	2	3	3	3	5	2	73
L.1<0,5	4	4	5	3	5	5	3	4	6	4	6	4	3	5	5	5	8	2	78
L.0<0,35		1		1	2	2	3	1	2		2	4	6	5	5	5	3	13	52
Średnia	2,18	2,12	2,12	2,00	1,94	1,94	1,94	1,82	1,82	1,76	1,71	1,71	1,47	1,35	1,35	1,35	1,18	0,35	1,67

Źródło: opracowanie własne.

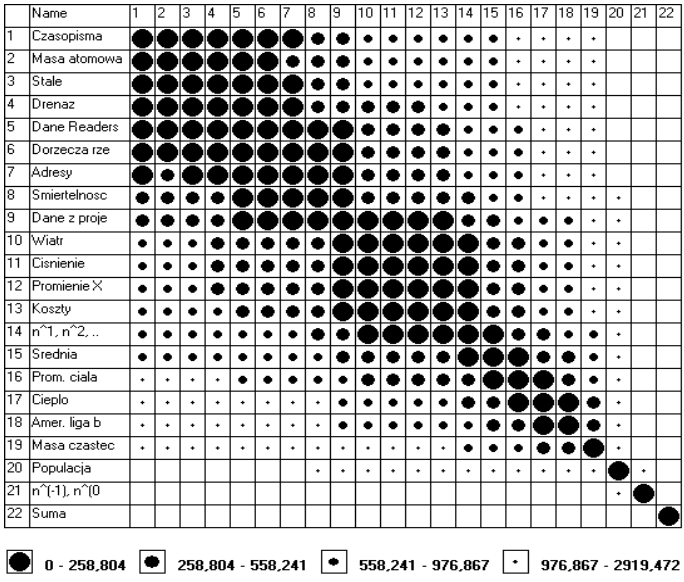
Tabela 31. Kategorie współczynników korelacji rang Spearmana między miernikami zgodności rozkładów cyfr znaczących ze względu na ich poziom [0 <0,5]; [1 <0,5;0,75]; [2 <0,75;0,9]; [3 >0,9]

	KS1	KS2	KS3	chi	1-chi-(p)	z sred	R	z>1,64	z>1,96	M2	M3	M4	M5	M1	z>2,58	1-r	r(p)	z>3,29	n	Średnia
KS1	3	3	3	3	3	3	3	3	3	2	2	2	2	2	2	1	1	2	1	2,28
KS2	3	3	3	3	3	3	3	3	3	2	2	2	2	2	2	1	1	2	1	2,28
KS3	3	3	3	3	3	3	3	3	3	2	2	2	2	2	2	1	1	2	1	2,28
chi	3	3	3	3	3	3	3	3	3	2	2	2	2	2	2	1	1	1	1	2,22
1-chi(p)	3	3	3	3	3	3	3	3	3	2	2	2	2	2	2	1	1	1	1	2,22
z sred	3	3	3	3	3	3	3	3	3	2	2	2	2	2	2	1	1	1	1	2,22
R	3	3	3	3	3	3	3	3	3	2	2	2	2	2	2	1	1	2		2,22
z>1,64	3	3	3	3	3	3	3	3	3	1	1	1	1	1	2	1	1	1	1	1,94
z>1,96	3	3	3	3	3	3	3	3	3	1	1	1	1	1	3	1	1	1		1,94
M2	2	2	2	2	2	2	2	1	1	3	3	3	3	3	1	2	2	1		1,89
M3	2	2	2	2	2	2	2	1	1	3	3	3	3	3	1	2	2	1		1,89
M4	2	2	2	2	2	2	2	1	1	3	3	3	3	3	1	2	2	1		1,89
M5	2	2	2	2	2	2	2	1	1	3	3	3	3	3	1	2	2			1,83
M1	2	2	2	2	2	2	2	1	1	3	3	3	3	3	1	2	2			1,83
z>2,58	2	2	2	2	2	2	2	2	3	1	1	1	1	1	3			2	1	1,50
1-r	1	1	1	1	1	1	1	1	1	2	2	2	2	2		3	3			1,22
r(p)	1	1	1	1	1	1	1	1	1	2	2	2	2	2		3	3			1,22
z>3,29	2	2	2	1	1	1	2	1	1	1	1	1			2			3	1	1,06
n	1	1	1	1	1	1		1							1			1	3	0,50
L.3>0,9	8	8	8	8	8	8	8	8	9	4	4	4	4	4	1	1	1			93
L.2<0,9	7	7	7	6	6	6	7	1		9	9	9	9	9	9	5	5	5		113
L.1<0,75	3	3	3	4	4	4	2	9	8	4	4	4	3	3	6	9	9	9	9	97
L.0<0,5	1	1	1	1	1	1	2	1	2	2	2	2	3	3	3	4	4	5	10	46
Średnia	2,28	2,28	2,28	2,22	2,22	2,22	2,22	1,94	1,94	1,89	1,89	1,89	1,83	1,83	1,50	1,22	1,22	1,06	0,50	1,81

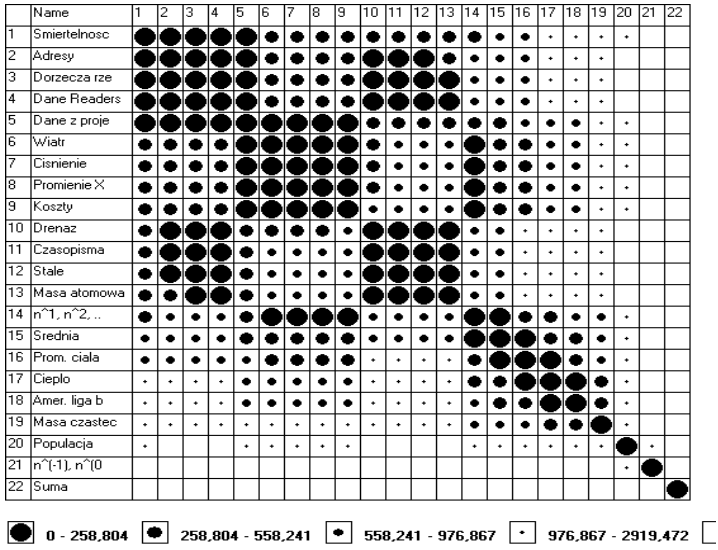
Źródło: opracowanie własne.



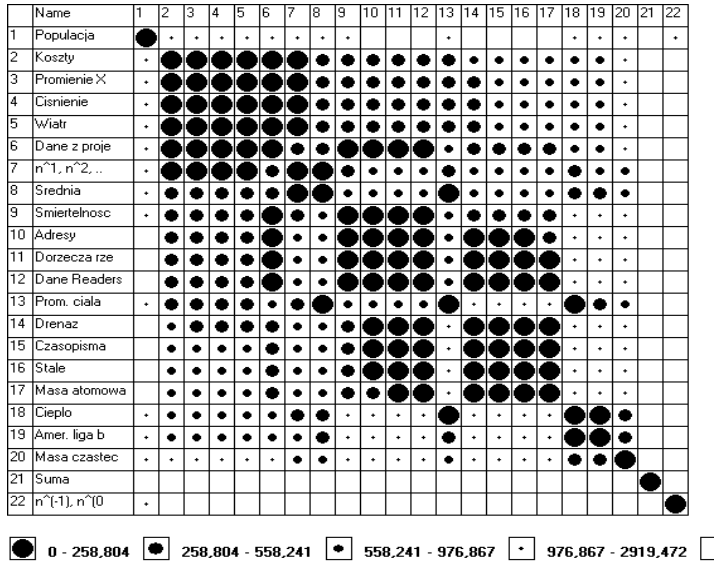
Rysunek 23. Nieuporządkowany diagram Czekanowskiego zbiorów Benforda w przestrzeni mierników zgodności rozkładów cyfr znaczących



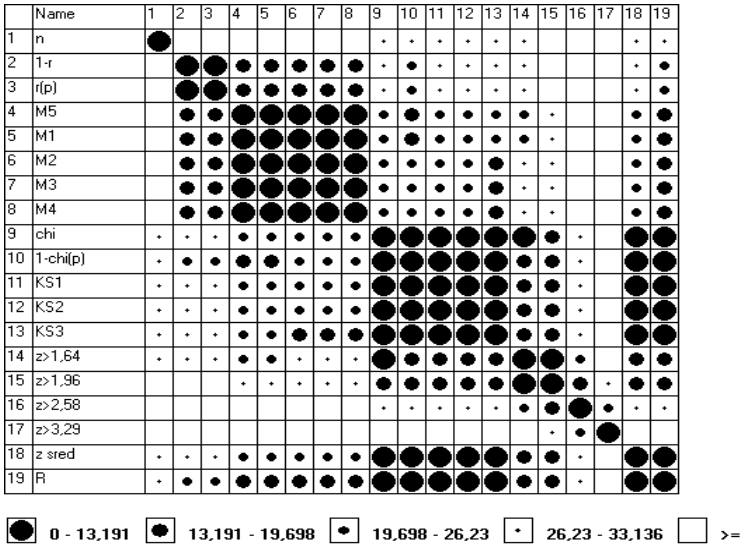
Rysunek 24. Uporządkowany diagram Czekanowskiego zbiorów Benforda w przestrzeni mierników zgodności rozkładów cyfr znaczących (metoda A)



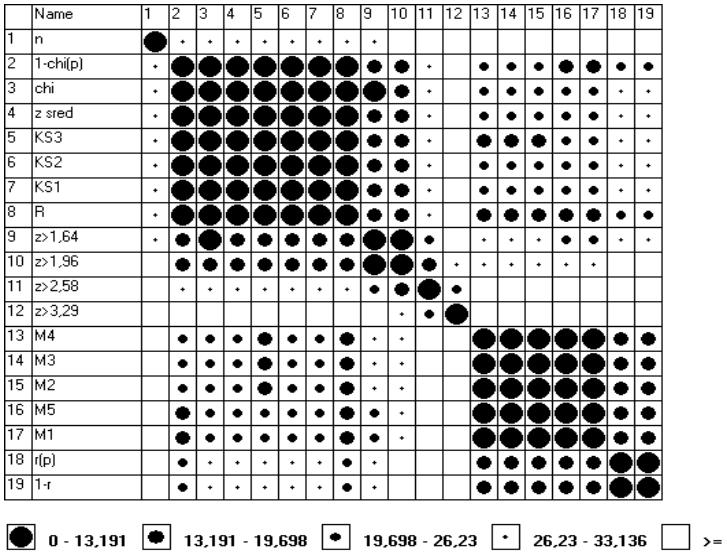
Rysunek 25. Uporządkowany diagram Czekanowskiego zbiorów Benforda w przestrzeni mierników zgodności rozkładów cyfr znaczących (metoda B)



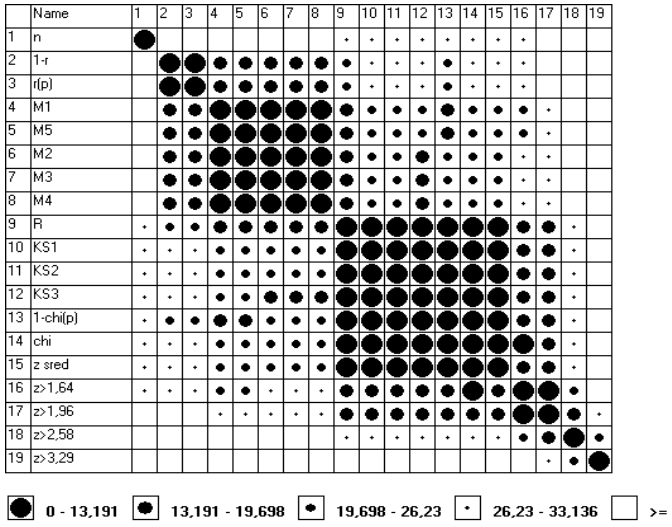
Rysunek 26. Uporządkowany diagram Czekanowskiego zbiorów Benforda w przestrzeni mierników zgodności rozkładów cyfr znaczących (metoda C)



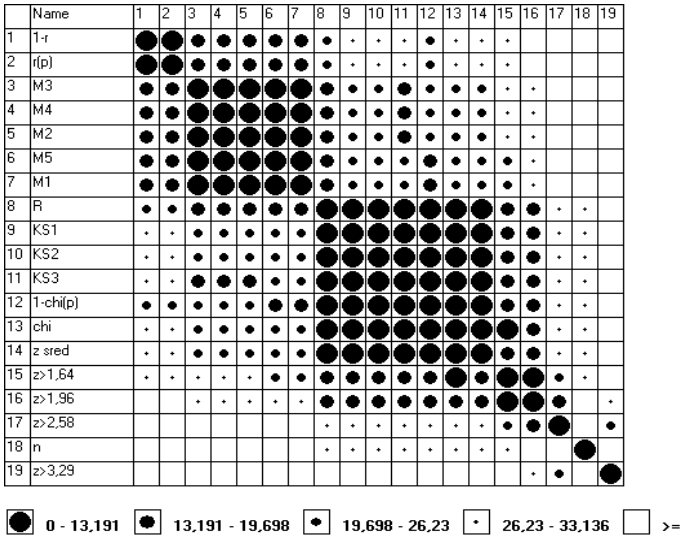
Rysunek 27. Nieuporządkowany diagram Czekanowskiego mierników zgodności rozkładów cyfr znaczących w przestrzeni zbiorów Benforda



Rysunek 28. Uporządkowany diagram Czekanowskiego mierników zgodności rozkładów cyfr znaczących w przestrzeni zbiorów Benforda (metoda A)



Rysunek 29. Uporządkowany diagram Czekanowskiego mierników zgodności rozkładów cyfr znaczących w przestrzeni zbiorów Benforda (metoda B)



Rysunek 30. Uporządkowany diagram Czekanowskiego mierników zgodności rozkładów cyfr znaczących w przestrzeni zbiorów Benforda (metoda C)

Podsumowanie

Istota metody Czekanowskiego[20] polega na takim uporządkowaniu zbioru klasyfikowanych obiektów, aby wzdłuż głównej przekątnej diagramu znalazły się elementy jak najbardziej zaczernione, oraz im dalej od głównej przekątnej – tym te elementy powinny być mniej zaczernione.

Na rysunku 23 przedstawiono nieuporządkowany diagram Czekanowskiego dla zbiorów Benforda, w którym przyjęto 5 klas podobieństwa. Trudno w tym diagramie zauważyć jakikolwiek porządek. Układ i konfiguracja elementów diagramu są chaotyczne.

Rysunki 24–26 zawierają uporządkowane diagramy Czekanowskiego, za pomocą różnych algorytmów (A-B-C) ujętych w programie MaCzek. Analiza tych diagramów prowadzi do wniosku, że wśród 20 zbiorów Benforda można wyróżnić 4 podgrupy zbiorów o podobnych wartościach mierników zgodności oraz trzy zbiory jednoelementowe (populacja, potęga liczb naturalnych, zbiór sumaryczny), dla których miary zgodności kształtują się inaczej niż dla pozostałych zbiorów Benforda.

Podobną analizę wykonano dla zbioru miar zgodności rozkładów (rysunki 27–30). Z uzyskanych diagramów Czekanowskiego wynika, że w rozpatrywanym zbiorze miar zgodności można wyróżnić następujące bardziej jednorodne podzbiory:

a. mierniki M1-M5,
b. współczynniki *1-r* oraz *r(p)*,
c. statystyki *KS1, KS2, KS3, z sred* oraz statystyki *chi* i *1-ch(p)*,
d. statystyki *z>1,64* oraz *z>1,96*,
e. trzy zbiory jednoelementowe, zawierające mierniki *z>2,58; z>3,29* oraz *n*.

Ponadto daje się zauważyć podobieństwo podzbiorów (a) i (b) oraz (c) i (d). Otrzymane wyniki zebrano w tabeli 32. Z przeprowadzonej analizy można wyciągnąć wniosek, że w praktyce należy przy ocenie zgodności rozkładów cyfr znaczących uwzględnić nie 18, ale co najwyżej 6 mierników – po jednym mierniku z każdej z grup od (a) do (d) oraz mierniki tworzące grupy (e) oraz (f).

Alternatywnym rozwiązaniem jest uwzględnienie 3 mierników. Jeden powinien reprezentować podzbiór {a; b}, drugi podzbiór {c; d}, natomiast trzeci miernik podzbiory {e; f}.

[20] Por. np. T. Grabiński, *Propozycje w zakresie porządkowania diagramu Jana Czekanowskiego*, [w:] *Studia z zakresu metod ilościowych w ekonomii, demografii i socjologii*, "Prace Komisji Socjologicznej PAN, o. Kraków", nr 40, Wrocław–Warszawa–Kraków–Gdańsk 1977.

Tabela 32. Wyniki klasyfikacji miar zgodności rozkładów

(a)	M2	M3	M4	M5	**M1**		(b)	**1-r**	r(p)
(c)	KS1	KS2	**KS3**	chi	1-chi(p)	z sred	(d)	**z>1,64**	z>1,96
(e)	z>2,58	(f)	z>3,29		n				

Źródło: opracowanie własne.

Odrębną kwestią jest wybór reprezentanta poszczególnych podzbiorów. Można tu posłużyć się kategoryzacją miar zgodności (por. tab. 30–31). Biorąc pod uwagę te parametry w wariancie oszczędnym „najlepszymi" (w sensie reprezentatywności całego zbioru miar) miarami zgodności rozkładów mogą być: miernik *M1*, statystyka *KS3* oraz *z>2,58*. W wariancie poszerzonym można by dodatkowo uwzględnić miary *1-r* oraz *z>1,64.*

Warto zauważyć, że wśród wskazanych miar nie ma statystyki χ^2. Biorąc jednak pod uwagę jej popularność, w praktyce warto też uwzględniać i ten parametr.

Bibliografia

1. Alonso F., *The Extinction of Password Authentication*, ISSA Journal, December 2008.
2. *Archival Resource Key*, [online:] http://www.cdlib.org/inside/diglib/ark/ [dostęp: 20.12.2008].
3. *Attributes of a Trusted Digital Repository: Meeting the Needs of Research Resources. RLG-OCLC Report*. Mountain View, CA. August 2001, [online:] http://www.rlg.org/longterm/attributes01.pdf [dostęp 20.12.2008].
4. Bach M., Kozielski S., *Translacja zapytań do baz danych sformułowanych w języku naturalnym na zapytania w języku SQL*, Konferencja Naukowa, Technologie przetwarzania danych, Wydawnictwo Politechniki Poznańskiej, Poznań 2005.
5. Barta J., Fajgielski P., Markiewicz R., *Ochrona danych osobowych. Komentarz*, wyd. Wolters Kluwer, Kraków 2007.
6. Bench-Capon T.J.M. et al., *Validation, Verification and Integrity Issues in Expert and Database Systems*, „Expert Update" Spring 1999, t. 2, nr 1, s. 31–35.
7. Beynon-Davies P., *Database Systems*, MacMillan Press Ltd. 1996, wyd. polskie: *Systemy baz danych*, Wydawnictwa Naukowo-Techniczne, Warszawa 1998.
8. Bierć A., *Ochrona prawna danych osobowych w sferze działalności gospodarczej w Polsce – aspekty cywilnoprawne,* [w:] *Ochrona danych osobowych w Polsce z perspektywy dziesięciolecia*, red. P. Fajgielski, Lublin 2008.
9. Biernat S., *Prawo Unii Europejskiej a prawo państw członkowskich*, [w:] *Prawo Unii Europejskiej,* red. J. Barcz, Warszawa 2004.
10. Borghoff U.M., i in., *Langzeitarchivierung. Methoden zur Erhaltung digitaler Dokumente*, Heidelberg 2003.
11. Borowicz J., *Obowiązek prowadzenia przez pracodawcę dokumentacji osobowej i organizacyjnej z zakresu ochrony danych osobowych. Teza nr 3*, PiZS.2001.3.2, LEX 29032/3.
12. Clavel-Merrin G., *The Nedlib List of Terms. Nedlib Report Series 7*, Amsterdam 2000.
13. Coenen F. *et al.*, *Validation and verification of knowledge based systems: Report on EUROVAV'99*, „Knowledge Engineering Review" (submitted for publication).
14. Consultative Committee for Space Data System, *Reference Model for an Open Archival Information System (OAIS). Draft Recommendation for Space Data System Standards. CCSDS 650.0-R-1. Red Book*, Newport Beach, Ca.,

1999, [online:] http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/p2/CCSDS-650.0-R-1.pdf [dostęp: 20.12.2009].

15. Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System (OAIS). Recommendation for Space Data System Standards. CCSDS 650.0-B-1. Blue Book*, 2002, [online:] http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf [dostęp: 20.12.2009].
16. Cormen T.H., *Wprowadzenie do algorytmów*, wyd. 7, Wydawnictwa Naukowo-Techniczne, Warszawa 2005.
17. Drozd A., *Zabezpieczenie danych osobowych,* Wrocław 2008.
18. Dudek A., *Nie tylko wirusy. Hacking, cracking, bezpieczeństwo Internetu*, Helion, Gliwice 2004.
19. *Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Nestor Handbuch*, 2008, [online:] http://nestor.sub.unigoettingen.de/handbuch/nestor-handbuch.pdf [dostęp: 20.08.2008].
20. Erickson J., *Hacking. Sztuka penetracji*, Helion, Gliwice 2004.
21. Famili A., Shen W.-M., Weber R., Simoudis E., *Data preprocessing and intelligent data analysis*, „Intelligent Data Analysis" 1996, t. 1.
22. Frampton S., *Linux Administration Made Easy*, Iuniverse Inc, December 2000.
23. Freedman A., *Encyklopedia komputerów*, Gliwice 2004.
24. Gałach A., *Instrukcja ochrony danych osobowych w systemie informatycznym,* Gdańsk 2004.
25. Gallaire H., Minker J. (red.), *Logic and Databases*, Plenum Press, 1978.
26. Gatnar E., *Symboliczne metody klasyfikacji danych*, Wydawnictwo Naukowe PWN, Warszawa 1998.
27. *Hack Proofing Your Network. Edycja polska*, praca zbiorowa, Helion, Gliwice 2002.
28. Internet Assigned Numbers Authority IANA, http://www.iana.org/ [dostęp: 20.12.2008].
29. Januszko-Szakiel A., *Archiwizacja publikacji elektronicznych jako wyzwanie dla bibliotek – zarys problematyki.* „Biuletyn Biblioteki Jagiellońskiej" 2003, s. 216–225.
30. Januszko-Szakiel A., *Open Archival Information System – standard w zakresie archiwizacji publikacji elektronicznych*. „Przegląd Biblioteczny" 2005, z. 3(73), s. 341.
31. Januszko-Szakiel A., *Rola migracji i emulacji w strategii długoterminowej archiwizacji publikacji elektronicznych*, [w:] *Informatyka*, red. M. Pękala, W.Z. Chmielowski, Kraków 2008.

32. *Kriterienkatalog vertrauenswürdige digitale Langzeitarchive. Version 1. [Entwurf zur öffentlichen Kommentierung]. Nestor Materialien 8*, Frankfurt am Main, 2006, [online:] http://edoc.hu-berlin.de/series/nestor-materialien/2006-8/PDF/8.pdf [dostęp: 20.08.2008].
33. Laskari E.C., Meletiou G.C., Stamatiou Y.C., Vrahatis M.N., *Evolutionary Computation Based Cryptanalysis: A First Study*, „Nonlinear Analysis: Theory, Methods and Applications" 2005, t. 63.
34. Laskari E.C., Meletiou G.C., Vrahatis M.N., *Problems of Cryptography as Discrete Optimization Tasks*, „Nonlinear Analysis: Theory, Methods and Applications" 2005, t. 63.
35. Ligęza A., *Completeness verification of rule-based control systems*, "Systems Analysis – Modelling – Simulation" (SAMS Int. Journal), t. 24, 1996, s. 211–220.
36. Ligęza A., Fuster Parra P., *Towards Logical Analysis of Rule-Based Systems. Quality and Reliability Verification*, [w:] *Proceedings of the 13th European Meeting on Cybernetics and Systems Research EMCSR'96*, t. 2, Vienna 1996, s. 1211–1216 (Extended version in: LAAS du CNRS Report, No. 96185 Toulouse 1996).
37. Ligęza A., *Intelligent data and knowledge analysis and verification: towards a taxonomy of specific problems*, [19], 1999, s. 313–325.
38. Ligęza A., *Towards logical analysis of tabular rule-based systems*, Extended version submitted to „International Journal of Intelligent Systems" 1998, t. 20, s. 30–35.
39. Litwiński P., *Ochrona danych osobowych w ogólnym postępowaniu administracyjnym,* wyd. Wolters Kluwer, Warszawa 2009.
40. Liu B., Ku L.-P., Hsu W., *Discovering interesting holes in data*, [w:] *Proceedings of IJCAI'87*, t. 2, Nagoya 1997, s. 930–935.
41. Maziarz P., *Wykorzystywanie tęczowych tablic do łamania haseł*, „Hakin9", nr 9/2007 (29).
42. Menezes A., Oorschot P. van, S. Vanstone, *Handbook of applied cryptography*, CRC Press series on discrete mathematics and its applications, CRC Press, 1996.
43. *Message Passing Interface (MPI) Tutorial*; https://computing.llnl.gov/tutorials/mpi.
44. Negus Ch., *Red Hat Linux 9. Biblia*, Helion, Gliwice 2003.
45. Niederliński A., *Regułowe systemy ekspertowe*, Gliwice 2000.
46. *OpenMP Tutorial*; https://computing.llnl.gov/tutorials/openMP.

47. Pawlak Z., *Rough Sets. Theoretical Aspects of Reasoning about Data, Dordrecht*, Kluwer Academic Publishers, 1991.
48. *Persistent Identifier. Eindeutige Bezeichner für digitale Inhalte*, [online:] http://www.persistent-identifier.de/?link=201 [dostęp: 8.01.2009].
49. Płażek J., Podyma M., *Równoległe metody łamania haseł metodą słownikową w środowiskach MPI, OpenMP i CUDA*, „Czasopismo Techniczne" 2011.
50. Popławski Ł., Bułat R., *The utilization of genetic computations in the planning of sustainable territorial budget*, [w:] *Local development – chosen factors of sustainable development of Poland*, Szczecin 2010.
51. Quirchmayr G., Schweighofer E., Bench-Capon T.J.M. (red.), *Database and Expert Systems Applications. 9th International Conference DEXA'98* (Lecture Notes in Computer Science, 1460) Springer, Vienna 1998.
52. Reitz J.M., *Dictionary for Library and Information Science*, Westport, London 2004.
53. Repozytorium SpringerLink: http://springerlink.metapress.com/home/main.mpx [dostęp: 20.12.2008].
54. Rivest R.L., Shamir A., Adleman L., A *method for obtaining digital signatures and public key cryptosystems*, „Communications of the ACM" 1978, t. 21, s. 120–126.
55. Sakowska M., *Pozycja ustrojowa i zadania Generalnego Inspektora Ochrony Danych Osobowych*, „Przegląd Sejmowy" 2006, nr 2.
56. Schneier B., *Applied Cryptography*, Wiley, New York 1996.
57. Schöning W.Ch., *Der Uniform Resource Name (URN),* [w:] *Eine kleine Enzyklopädie der digitalen Langzeitarchivierung . Nestor Handbuch*, 2008, [online:] http://nestor.sub.unigoettingen.de/handbuch/nestor-handbuch.pdf [dostęp: 20.08.2008].
58. Schroeder K., *Persistent Identifier (PI) – ein Überblick*, [w:] *Eine kleine Enzyklopädie der digitalen Langzeitarchivierung Nestor Handbuch,* 2008, [online:] http://nestor.sub.uni-goettingen.de/handbuch/nestor-handbuch.pdf [dostęp: 20.08.2008].
59. Shema M., B.C. Johnson, *Anti Anti-Hacker Tool Kit. Edycja polska*, Helion, Gliwice 2004.
60. Sollins K., Masinter L., *Functional Requirements for Uniform Resource Names*, [online:] http://www.ietf.org/rfc/rfc1737.txt [dostęp: 20.12.2008].
61. Sprengers M., *GPU-based Password Cracking. On the Security of Password Hashing Schemes regarding Advances in Graphics Processing Units*, Radboud University Nijmegen 2011.

62. Storn R., Price K., *Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces*, „Journal Global Optimization" 1997, t. 11.
63. Sun S., Lannom L., Boesch B., *Handle System Overview. Request for Comments*: *3650*, CNRI, November 2003, [online:] http://www.ietf.org/rfc/rfc3650.txt [dostęp: 20.12.2009].
64. Tinnefeld M.T., *Ochrona danych osobowych – kamień węgielny budowy Europy*, [w:] *Ochrona danych osobowych*, red. M. Wyrzykowski, Warszawa 1999.
65. Toxen B., *Bezpieczeństwo w Linuksie. Podręcznik administratora*, Helion, Gliwice 2004.
66. Traczyk W., *Jak uczyć się z różnorodnych przykładów,* [w:] *Inżynieria wiedzy i systemy ekspertowe*, red. Z. Bubnicki i A. Grzech, Oficyna Wydawnicza Politechniki Wrocławskiej, t. 1, Wrocław 1997, s. 21–28.
67. *Trusted Digital Repositories: Attributes and Responsibilities. An RLG-OCLC Report*, 2002, [online:] http://www.rlg.org/longterm/repositories.pdf [dostęp: 20.12.2008].
68. *Uniform Resource Names. A Progress Report*, „D-Lib Magazine", February 1996, [online:] http://www.dlib.org/dlib/february96/02arms.html [dostęp: 20.12.2008].
69. Wasiołka B., Wasiołka S., *Optymalizacja systemów zarządzania relacyjnymi bazami danych na przykładzie MySQL*, Konferencja Naukowa „Technologie przetwarzania danych", Wydawnictwo Politechniki Poznańskiej, Poznań 2005.
70. Węsierski Ł.N., *Podstawy logiki i wnioskowania*, Oficyna Wydawnicza Politechniki Rzeszowskiej, Rzeszów 2004.
71. Żak M., *Równoległe łamanie haseł*, Politechnika Krakowska, Kraków 2006.

Akty prawne

1. Dyrektywa 95/46/WE Parlamentu Europejskiego i Rady z dnia 24 października 1995 r. w sprawie ochrony osób fizycznych w zakresie przetwarzania danych osobowych oraz swobodnego przepływu tychże danych, Dz.Urz. L 281 z dnia 23 listopada 1995 r.
2. Konstytucja Rzeczypospolitej Polskiej z 1997 r., Dz.U. Nr 78, poz. 483.
3. Konwencja nr 108 rady Europy w dnia 28 stycznia 1981 r. o ochronie osób w związku z automatycznym przetwarzaniem danych osobowych, Dz.U. z 2003 r., Nr 3, poz. 25

4. Moats R., *URN Syntax*. *Request for Comments*: 2141, AT&T, May 1997, [online:] http://www.ietf.org/rfc/rfc2141.txt [dostęp: 20.12.2008].
5. *NVIDIA CUDA, C Programming Guide, Version 3.2,* NVIDIA Corporation, październik 2010.
6. *NVIDIA CUDA, Reference Manual, Version 3.2 Beta,* sierpień 2010.
7. Rozporządzenie Ministra Nauki i Szkolnictwa Wyższego w sprawie danych zamieszczanych w ogólnopolskim wykazie studentów z dnia 22 września 2011 r., Dz.U. Nr 204, poz. 1201.
8. Rozporządzenie Ministra Spraw Wewnętrznych i Administracji z dnia 29 kwietnia 2004 r. w sprawie dokumentacji przetwarzania danych osobowych oraz warunków technicznych i organizacyjnych, jakim powinny odpowiadać urządzenia u systemy informatyczne służące do przetwarzania danych osobowych, Dz.U. z 2004 r., Nr 100, poz.1024.
9. *The DOI System*, [online:] http://www.doi.org/ [dostęp: 19.12.2008].
10. *The Handle System*, [online:] http://www.handle.net/ [dostęp: 20.12.2008].
11. Uchwała składu 7 sędziów SN z dnia 16 lipca 1993 r., I PZP 28/93, OSNC 1994, nr 1, poz. 2.
12. Ustawa z dnia 27 lipca 2005 r. – Prawo o szkolnictwie wyższym, Dz.U. 2005 Nr 164, poz. 1365 z późn. zm.
13. Ustawa z dnia 29 sierpnia 1997 r. o ochronie danych osobowych. Tekst jednolity Dz. z 2002 r., Nr 101, poz. 936 z późn. zm.

Spis rysunków

Spis tabel

Noty o autorach

- Agnieszka Bednarczyk
  Wydział Prawa, Administracji i Stosunków Międzynarodowych,
  Krakowska Akademia im. Andrzeja Frycza Modrzewskiego,
  e-mail: bednarczyk@afm.edu.pl

- Radosław Bułat
  Politechnika Krakowska

- Marzena Farbaniec
  Wydział Zarządzania i Komunikacji Społecznej,
  Krakowska Akademia im. Andrzeja Frycza Modrzewskiego

- Tadeusz Grabiński
  Wydział Zarządzania i Komunikacji Społecznej,
  Krakowska Akademia im. Andrzeja Frycza Modrzewskiego,
  e-mail: tg@uek.krakow.pl

- Aneta Januszko-Szakiel
  Wydział Psychologii i Nauk Humanistycznych,
  Krakowska Akademia im. Andrzeja Frycza Modrzewskiego,
  e-mail: ajanuszko-szakiel@afm.edu.pl

- Joanna Płażek
  Wydział Zarządzania i Komunikacji Społecznej,
  Krakowska Akademia im. Andrzeja Frycza Modrzewskiego,
  e-mail: joannaplazek@gmail.com

- Agnieszka Smolarek-Grzyb
  Wydział Zarządzania i Komunikacji Społecznej,
  Krakowska Akademia im. Andrzeja Frycza Modrzewskiego,
  e-mail: agasmolarek@gmail.com

▪ Renata Uryga
Wydział Zarządzania i Komunikacji Społecznej,
Krakowska Akademia im. Andrzeja Frycza Modrzewskiego,
e-mail: ruryga@afm.edu.pl

▪ Dorota Wilk-Kołodziejczyk
Wydział Zarządzania i Komunikacji Społecznej,
Krakowska Akademia im. Andrzeja Frycza Modrzewskiego,
e-mail: wilk-kolodziejczyk@afm.edu.pl

▪ Bartłomiej Zabłocki
Wydział Zarządzania i Komunikacji Społecznej,
Krakowska Akademia im. Andrzeja Frycza Modrzewskiego

▪ Wacław Zając
Wydział Zarządzania i Komunikacji Społecznej,
Krakowska Akademia im. Andrzeja Frycza Modrzewskiego