Marta Grabowska[★]

Zbiory *Big Data* oraz technologia chmury obliczeniowej dla humanistyki cyfrowej

Wprowadzenie

Po prawie dwóch wiekach epoki przemysłowej, w której maszyny zastępowały głównie siłę mięśni ludzkich, osiągniecia wybitnych matematyków i inżynierów końca XIX i początku XX w., takich jak: David Hilbert (1862–1943), Kurt Goedel (1906–1978), John von Neumann (1903–1957), Charles Babbage (1871–1871) czy Lee De Forest (1873–1961)[1], skierowały wynalazczość w dziedzinie techniki na inne tory, tj. objęły sferę informacji. Pojawienie się komputerów stacjonarnych, a potem mobilnych, coraz większe pojemności pamięci dyskowych, miniaturyzacja, a także zaawansowane oprogramowania komputerowe i rozwój sieci teleinformatycznych przywiodły nas do punktu, w którym posługujemy się ogromnymi ilościami danych, zwanymi *Big Data*, oraz różnorodnymi narzędziami operowania nimi, otwierającymi nowe przestrzenie różnorodnych działań.

Terminem *Big Data* określa się ogromną ilość danych liczonych w terabajtach (1000 GB), petabajtach (1000 TB), exabajtach (1000 PB) i zettabajtach (1000 EX), niemieszczących się już w komputerach o pojemnościach kilkunastu gigabajtów, gromadzonych i przetwarzanych w środowisku nowej technologii, którą jest chmura obliczeniowa (ang. *cloud computing*)[2]. Technologia ta, za pomocą zaawansowanych narzędzi operowania danymi, umożliwia lepszą organizację i zarządzanie wielkimi zbiorami danych i dzięki temu uzyskiwanie nowej wiedzy. Obejmuje ona nowe rozwiązania zarówno w zakresie hardware'u, jak i software'u, w tym też nowoczesnych narzędzi analizy i prezentacji danych.

★ Dr hab. **Marta Grabowska**, prof. UW – Centrum Europejskie Uniwersytetu Warszawskiego, e-mail: mgrabowska@uw.edu.pl.

[1] A. Hodges, *Alan Turing: Enigma*, Wydawnictwo Albatros Andrzej Kuryłowicz S.C., Warszawa 2014.

[2] V. Mayer-Sconberger, K. Cukier, *Big data. Rewolucja, która zmieni nasze myślenie, pracę i życie*, MT Biznes sp. z o.o., Warszawa 2014.

Początki tworzenia wielkich zbiorów danych

Pojęcie *Big Data* wywodzi się z obszaru biznesu, gdzie gromadzone były i są nadal codziennie ogromne ilości danych w takich sektorach, jak np.: bankowość, handel, transport itd. Dane te mają zwykle bardzo różnorodny charakter. Mogą to być tzw. dane ustrukturalizowane (inaczej: uporządkowane czy strukturalne), np. dane o klientach: ich nazwiska, konta bankowe, informacje o transakcjach handlowych, nazwy czy kody kupowanych towarów, ale także dane nieustrukturyzowane, np. pochodzące z korespondencji e-mailowej z klientami czy z wypowiedzi w języku naturalnym umieszczanych np. na portalach społecznościowych, blogach lub w innych miejscach w sieci w postaci tekstu, obrazu czy dźwięku. Wszystkie te dane, napływające jednorazowo czy w sposób ciągły, stanowią wartościowy materiał analityczny. Dodatkowo, po okresach różnego rodzaju ograniczeń dostępu do danych – promowana obecnie polityka tzw. danych otwartych (ang. *open data*[3]) zwiększa jeszcze zasoby, które mogą być przetwarzane w systemach informatycznych.

Początkowo w tradycyjnych bazach biznesowych, tj. w pojedynczych bazach relacyjnych, gromadzono dane ustrukturalizowane, posługując się językiem zapytań SQL (*Structured Query Language*)[4], stworzonym przez firmę IBM w 1960 r. Jednak w przypadku dużych, międzynarodowych korporacji rozwiązania takie nie były wystarczające. W końcu lat 80. XX w., w związku z potrzebą wsparcia procesu zarządzania dużymi przedsiębiorstwami, powstały tzw. hurtownie danych (ang. *data warehouse*)[5], w których gromadzono dane napływające z wielu ośrodków. Zastosowanie w nich lepszych narzędzi wyszukiwawczych wzbogaconych o elementy analityczne, jak np. Online Analytical Processing (OLAP)[6], umożliwiło szybsze przeszukiwanie zagregowanych danych oraz ich analizę według określonych kryteriów. Operacje wykonywane na tych danych, zwane

[3] P. Morin, *Open data structures. An introduction* [e-book] in OPEL (Open Paths to Enriched Learning). Edmonton: AU Press. 2013, http://web.b.ebscohost.com/ehost/ebookviewer/ebook/ZTAwMHh3d19fNjM4OTU2X19BTg2?sid=d8b94be4-7c8e-4f69-9727-d3428c90cdfe@sessionmgr115&vid=3&format=EB&rid=1 (dostęp 7.05.2015).

[4] M.M. David, L. Fesperman, *Advanced standard SQL dynamic structured data modeling and hierarchical processing* [e-book]. Boston, Artech House, 2013, http://web.a.ebscohost.com/ehost/ebookviewer/ebook/ZTAwMHh3d19fNzUzNTk2X19BTg2?sid=96ce560b-b590-47f4-9f8c-18783f9988e6@sessionmgr4002&vid=3&format=EB&rid=1 (dostęp 6.05.2015).

[5] A. Chodkowska-Gyuriks, *Hurtownie danych: teoria i praktyka*, Wydawnictwo Naukowe PWN S.A., Warszawa 2014.

[6] A. Berson, S.J. Smith, *Data warehousing, data mining, and OLAP*, Mc Graw-Hill, New York 2001.

eksploatacją danych (ang. *data mining*)[7], dawały możliwość wychwytywania pewnych trendów, które miały znaczenie w procesie podejmowania decyzji. Operacje mogły być wykonywane na danych historycznych, pokazujących, co się wydarzyło, na danych wpływających na bieżąco, pokazujących zmiany trendów w czasie rzeczywistym, oraz mogły być też dokonywane porównania czy generowane prognozy. Hurtownie danych miały charakter scentralizowany (dane gromadzone w jednym zbiorze) lub składały się z wielu zbiorów, tj. z tzw. minihurtowni, z których każda reprezentowała pewną kategorię danych, tworząc łącznie tzw. ang. *data mart*, co po polsku można określić jako „targowisko danych". Hurtownie danych i stosowane tam narzędzia analityczne dały początek tzw. analityce biznesowej (ang. *business intelligence*)[8]. Ważną rolę odgrywały w nich wbudowane elementy automatycznej reakcji na różne zjawiska, tj. tzw. *feedback* (sprzężenie zwrotne), np. automatyczne generowanie zamówień adresowanych do producentów na brakujące towary za pomocą systemu EDI (*Electronic Data Interchange*)[9], oraz rozbudowane metody wizualizacji danych. Do takich narzędzi analitycznych można też zaliczyć np. systemy ekspertowe[10] z wbudowanymi procedurami wnioskowania czy różne rodzaje tzw. sieci neuronowych[11] (m.in. sieci rekurencyjne[12] stosowane do rozwiązywania problemów optymalizacyjnych czy skojarzeniowych, także samoorganizujące się mapy używane do wybierania i analizowania danych, ich klasyfikacji i wizualizacji), sieci radialne[13] stosowane do rozwiązywania problemów klasyfikacyjnych i prognostycznych. Rodowód hurtowni danych wywodzi się również z firmy IBM. Ich głównym celem było lepsze zarządzanie przedsiębiorstwami i usługami oraz lepsza, bardziej trafna i odpowiadająca indywidualnym zapotrzebowaniom klientów alokacja towarów i usług, co w efekcie dawało obopólne korzyści: zarówno producentom, jak i klientom.

---

[7] Ibidem.

[8] *Business intelligence*, red. J. Suma, Wydawnictwo Naukowe PWN, Warszawa 2013.

[9] M.C. Pparfatt, *What is EDI?: a guide to electronic data interchange*, PNCC Blackwell, Oxford 1992.

[10] P. Jackson, *Introduction to experts systems*, Accisson Wesley Pub.co, Workingham (England) 1986.

[11] S. Osowski, *Sieci neuronowe do przetwarzania informacji*, wyd. 2 popr. i uzup., Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa 2006.

[12] A. Chinea, *Understanding the principles of recursive neural networks. A generative approach to tackle model cpmplexity*, http://arxiv.org/ftp/arxiv/papers/0911/0911.3298.pdf (dostęp 2.05.2015).

[13] L. Rutkowski, *Metodyi i techniki sztucznej inteligencji*, wyd. 2 zmien., Wydawnictwo Naukowe PWN, Warszawa 2012.

Technologia chmury obliczeniowej

Przełomem w organizowaniu i zarządzaniu tego typu danymi, których ilość z biegiem czasu zaczęła wzrastać lawinowo, stała się wspomniana wyżej nowa technologia chmury obliczeniowej (ang. *cloud computing*)[14], której istota polega przede wszystkim na odciążeniu instytucji, firm i organizacji od całego skomplikowanego problemu gromadzenia i zarządzania wielkimi zbiorami danych i oddania go w ręce nowej kategorii wysoce wyspecjalizowanych usług zewnętrznych (ang. *outsourcingowych*) świadczonych przez firmy dysponujące zaawansowanymi technologiami zarówno w zakresie infrastruktury fizycznej, jak i oprogramowania narzędziowego oraz aplikacyjnego. Na usługi te składają się: serwery dysponujące wielkimi pojemnościami pamięci, których odległa lokalizacja, wobec możliwości współczesnych sieci telekomunikacyjnych, praktycznie nie odgrywa roli, rozbudowane oprogramowanie operacyjne i narzędziowe, którego zadaniem jest wszechstronne organizowanie napływających danych i m.in. zapewnienie ich bezpieczeństwa, wreszcie oprogramowania aplikacyjne umożliwiające różnego rodzaju operacje na tych danych, w szczególności dokonywanie ich wszechstronnej analizy oraz prezentacji[15].

W literaturze przedmiotu można znaleźć stanowisko, że technologia chmury obliczeniowej z jednej strony wywodzi się z takich wcześniejszych osiągnięć techniki, jak videotex[16], lecz z drugiej strony, co jest zaskakujące i co podaje V. Mosco, ma swój rodowód w koncepcji centralnie sterowanej gospodarki Związku Radzieckiego. Na początku lat 60. XX w. bowiem Nikita Chruszczow, chcąc zaradzić ciągłym brakom różnych towarów na rynku lub złej ich alokacji, korzystając z ówczesnych osiągnieć naukowych Norberta Wienera, twórcy cybernetyki, tj. nauki o sprawnym zarządzaniu[17], i sformułowanej przez niego zasadzie tzw. sprzężenia zwrotnego, planował stworzenie centrum zarządzania gospodarką wspomaganego przez maszyny liczące (w tamtej epoce były to jeszcze duże komputery lampowe, tzw. *mainframe*). Decyzje miały być podejmowane centralnie

[14] P. Mell, T, Grance, *The NIST definition of cloud computing. Recommendation of the National Institute of Standards and Technology*, Gaithesburg (MD), Computer Security Division, Information Technology Laboratory, National Institute of Standards and Technology, 2011, s. 2, http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf (dostęp 2.05.2015).

[15] V. Mosco, *To the cloud. Big data in a turbulent world*, Paradigm Publishers, Boulder (Colorado), London 2014.

[16] A.F. Alber, *Videotex/teletext: principles and practices*, McGraw Hill Book Company, New York 1986.

[17] N. Wiener, *Cybernetics: or control and communication in the animal and the machine*, M.I.T. Press, New York 1961.

na podstawie z jednej strony napływających drogą telexową informacji z sektora produkcji, a z drugiej – sygnalizowanych braków na rynku. Tak sterowana gospodarka miała być bardziej efektywna. Jak podaje V. Mosco, podobny model – pod nazwą Cybersyn – zaprojektowany został na początku lat 70. w Chile, gdzie upadającą gospodarkę socjalistyczną w okresie prezydentury Salvadora Allende'a próbowno ratować w ten sposób przed katastrofą. Oba kraje, ze względu na perturbacje polityczne, nie zrealizowały jednak do końca swoich projektów[18].

Obecna technologia chmury obliczeniowej rozwinięta w Stanach Zjednoczonych Ameryki i w innych krajach charakteryzujących się wysokim poziomem rozwojem gospodarczego, jak np. Japonia, opiera się na wielu nowych osiągnięciach techniki oraz służy szerszym celom. Przede wszystkim takie rozwiązania, jak: powszechnie stosowane karty magnetyczne i technologia identyfikacji radiowej (*Radio Frequency Identification* – RFID)[19], urządzenia mobilne i technologia zbliżeniowa (*Near Field Communication* – NFC)[20], telekomunikacja bezprzewodowa, miniaturyzacja urządzeń elektronicznych, a także digitalizacja ogromnych zbiorów różnego rodzaju dokumentów, wreszcie sam Internet, otworzyły nowe możliwości gromadzenia i przetwarzania niewyobrażalnych ilości danych już nie tylko w biznesie, ale także w administracji, sektorze zdrowia, transporcie, nauce itd. Za pomocą nowych narzędzi analitycznych stosowanych w tej technologii z danych wyłania się nowa wiedza.

Usługi świadczone przez firmy w ramach chmury obliczeniowej dzieli się zwykle na cztery rodzaje: Kolokacja, IaaS – Infrastructure as a Service, PaaS – Platform as a Service i SaaS – Software as a Service. W pierwszym przypadku użytkownikom udostępnia się jedynie pomieszczenia na serwery, w drugim – dodaje się infrastrukturę informatyczną (np. Elastic Compute Cloud EC2 firmy Amazon)[21], w trzecim – dodaje się system operacyjny z elementami oprogramowania narzędziowego (np. Windows Azure)[22], a w czwartym – udostępnia się również oprogramowanie aplikacyjne, np.

[18] V. Mosco, op.cit., s. 22.

[19] H. Lephamer, *RFID design principles*, Artech House, Inc., Boston, London 2008.

[20] NFC Forum: *NFC Data Exchange Format (NDEF). Technical specification. 2006-07-24*, NFC Forum, Inc., http://www.eet-china.com/ARTICLES/2006AUG/PDF/NFCForum-TS-NDEF.pdf?SOURCES=DOWNLOAD (dostęp 12.03.2014).

[21] Amazon Web Services, *Amazon Elastic Computer Cloud. User Guide for Linux. API version 2014-10-01*, http://www.google.pl/url?sa=t&rct=j&q=&esrc=s&source=web&cd=12&ved=0CG8QFjAL&url=http%3A%2F%2Fawsdocs.s3.amazonaws.com%2FEC2%2Flatest%2Fec2-ug.pdf&ei=jZdLVei2LuTgyQPH_4CoBA&usg=AFQjCNFIg7f8P673xBxButxQueralRAHkg&bvm=bv.92885102,d.bGQ (dostęp 2.05.2015).

[22] Microsoft Azure, http://azure.microsoft.com/pl-pl/ (dostęp 3.05.2015).

systemy poczty elektronicznej (np. Gmail czy Hotmail), a nawet oprogramowanie aplikacyjne może być specjalnie tworzone dla indywidualnego klienta. Ponadto wymienia się również usługi Software+Service, tj. użytkowanie konkretnego oprogramowania firmy, która udostępnia usługi w chmurze (np. Microsoft i jego webowy Microsoft Business Productivity Online Service), CaaS – Communications as a Service – usługodawca zapewnia także platformę telekomunikacyjną oraz IPaaS – świadczenie integracji usług chmurowych[23]. Usługi te mogą mieć charakter publiczny, prywatny, wspólny, tj. społecznościowy albo hybrydowy. Systemy muszą wykazywać się tzw. skalowalnością, tj. możliwością rozbudowy pod każdym względem, szczególnie tak, by sprostać gromadzeniu, analizowaniu i prezentacji napływających danych (*volume of data*), ich różnorodności (*variety of data*) oraz tempu ich napływu (*velocity of data*), a ponadto elastycznością i indywidualizacją usług oraz możliwością przetwarzania danych w zbiorach scentralizowanych, jak i w środowisku rozproszonym. Zagwarantowane muszą być: bezpieczeństwo i ochrona danych oraz ustalony sposób i zakres ich udostępniania[24].

W technologii chmury obliczeniowej upatruje się obecnie źródło przyspieszenia rozwoju gospodarczego i wzrostu nowych miejsc pracy oraz przyrostu PKB w granicach 2%. Unia Europejska wspiera rozwój usług przetwarzania w chmurze obliczeniowej, co wyrażone zostało zarówno w Europejskiej agendzie cyfrowej[25], jak i w Komunikacie Komisji do Parlamentu Europejskiego, Rady, Europejskiego Komitetu Ekonomiczno-Społecznego i Komitetu Regionów w sprawie wykorzystania potencjału chmury obliczeniowej w Europie[26]. Usługodawcy mogą korzystać ze wzorów umów przygotowanych przez Unię Europejską na usługi świadczone w tym zakresie na jej terytorium, jak również dokonuje się porządkowania unijnego prawa w tym obszarze.

[23] P. Szmit, *Cloud computing historia, technologia, perspektywy*, Polska Agencja Rozwoju Przedsiębiorczości, Warszawa 2012, https://www.web.gov.pl/g2/big/2012_06/ebfa211-f1a9f174c7517738f68df2d8b.pdf (dostęp 3.05.2015).

[24] J. Hurwitz, A. Nugent, F. Halper, M. Kaufman, *Big data for dummies*, John Wiley & Sons, Hoboken, NY 2013.

[25] Komisja Europejska. Komunikat Komisji do Parlamentu Europejskiego, Rady, Europejskiego Komitetu Ekonomiczno-Społecznego i Komitetu Regionów *Europejska agenda cyfrowa.* KOM(2010) 245 Bruksela 19.05.2010, wersja ostateczna.

[26] Komisja Europejska. Komunikat Komisji do Parlamentu Europejskiego, Rady, Europejskiego Komitetu Ekonomiczno-Społecznego i Komitetu Regionów w sprawie wykorzystania potencjału chmury obliczeniowej w Europie, Bruksela (2012). KOM(2012) 529, wersja ostateczna.

Zbiory *Big Data* w środowisku chmury obliczeniowej

Pomijając kwestie poziomu infrastruktury fizycznej, na działania w zakresie *Big Data* w środowisku chmury obliczeniowej składają się w zasadzie trzy etapy: 1) pozyskiwanie, gromadzenie i organizowanie danych, 2) analiza danych, 3) raportowanie i wizualizacja rezultatów[27].

Zasadniczą nowością w technologii chmury obliczeniowej w odniesieniu do pierwszego etapu, tj. gromadzenia i porządkowania danych, stało się wprowadzenie nowych, potężnych silników pozyskiwania, gromadzenia i porządkowania danych na przepastnych, wirtualnych obszarach Internetu i różnego rodzaju baz danych. Jednym z najbardziej znanych i publicznie dostępnych (otwartych) tego typu silników jest Hadoop firmy Apache[28], napisany w języku Java, zaprojektowany na podstawie zamkniętego oprogramowania o nazwie BigTable firmy Google[29]. Najistotniejszą cechą Hadoop jest przejęta z BigTable procedura MapReduce[30], umożliwiająca z jednej strony skuteczne rozprowadzenie i zidentyfikowanie problemu (pytania) w rozproszonych systemach informacyjnych (funkcja *map*), z drugiej zaś sprowadzenie z powrotem już wstępnie i odpowiednio uporządkowanych i zagregowanych danych (funkcja *reduce*) jako rezultatu. Procedura ta, paralelnie wykonywana na wielkiej liczbie rozproszonych serwerów zawierających ogromne ilości danych oraz równoległe ich przetwarzanie przez ich rozbicie na mniejsze grupy w celu przyspieszenia całego procesu (Hadoop Distributed File System – HDFS)[31], daje możliwość błyskawicznego przeszukiwania, gromadzenia i porządkowania niewyobrażalnej ilości danych. Oba systemy, tj. Hadoop i BigTable, zapewniają pełną skalowalność. Hadoop został początkowo zaprojektowany dla biznesu prowadzonego przez Yahoo, lecz stał się wkrótce najpotężniejszym narzędziem pozyskiwania danych w technologii chmury obliczeniowej[32].

---

[27] K. Bakshi, *Technologies of big data*, w: *Big data management. Technologies, and applications*, red. Wen-Chen Hu, N. Kaabouch, IGI Global, Hershey (USA) 2014, s. 1–22 (*Information Science Reference*s).

[28] I.K. Saavas, G.N Sofianidon, M-Tachar Kechadi, *Applying the K-Means Algorithm in Big Raw Data Sets with Hadoop and MapReduce*, w: *Big data management. Technologies, and applications*, red. Wen-Chen Hu, N. Kaabouch, IGI Global, Hershey (USA) 2014, s. 23–46.

[29] E. Ciurana, *Developing with Google App Engine*, Apress, Distributed by New York Springer Verlag, Berkeley (CA) 2008.

[30] D. Miner, A. Schock, *MapReduce design patterns*, O'Reilly, Beijing, Koln 2013.

[31] Hadoop, *HDFS Architecture guide*, http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html (dostęp 3.05.2015).

[32] E. Baldeschfieler, *Hadoop at Yahoo*!, http://www.google.pl/url?sa=t&rct=j&q=&esrc=s&source=web&cd=6&ved=0CFMQFjAF&url=http%3A%2F%2Fwww.hadooper.cn%2Fdct%2Fattach%2FY2xiOmNsYjpwZGY6ODE%3D&ei=xX5LVdLpGYupygOGiY

Wspomniany BigTable, zaprojektowany przez firmę Google na podstawie GFS (Google File Systems)[33] – systemu do indeksowania i przetwarzania ogromnych ilości danych tej firmy, umożliwia organizowanie, indeksowanie i przechowywanie wielkich ilości danych „na stołach" składających się z „rzędów" i „kolumn", lecz w porównaniu z możliwościami tradycyjnych baz SQL jest bardziej pojemny, gwarantuje danym większą odporność na zagrożenia, może przechowywać je w środowisku rozproszonym, tj. na wielu różnych serwerach oraz w środowisku wielowymiarowym. Jest jednak oprogramowaniem zamkniętym (prywatnym), choć elementy takiej właśnie organizacji danych są też zawarte w Hadoop).

Dopiero wobec tak zgromadzonych i zorganizowanych danych można zastosować narzędzia drugiego etapu, tj. narzędzia analityczne, w tym także prognostyczne, których istnieje duży wybór w zależności od charakteru danych i oczekiwanych rezultatów. Począwszy od prostych analiz statystycznych, tj. monitoringu danych napływających np. w czasie rzeczywistym w określonych odstępach czasu (np. co kilka sekund), dalej możliwe jest zastosowanie do nich bardziej zaawansowanych narzędzi analitycznych. jak: drzewa klasyfikacyjne, wspomniane wcześniej sieci neuronowe (rekurencyjne), radialne, techniki klastrowe ustalające najbliższe sąsiedztwo określonych danych umożliwiające identyfikację zbliżonych danych według ich stopnia podobieństwa itd. Bardzo istotną rolę w całej konstrukcji projektu odgrywa oprogramowanie pośredniczące[34] między systemem operacyjnym a aplikacjami, tj. tzw. szyny integracyjne (ang. *middleware*) oraz tzw. API's – interfejsy programistyczne aplikacji (*Application Programming Interfaces*), które muszą zagwarantować swobodny przepływ danych między wszystkimi elementami platformy w środowisku rozproszonym, ale także zapewnić ich bezpieczeństwo zgodnie z zadanymi warunkami. Oprogramowania te zorientowane są na świadczone przez system usługi dla klienta (*Service Oriented Architecture* – SOA)[35]. Projektują je zwykle osobne, wyspecjalizowane firmy.

Trzeci etap obejmuje prezentację danych za pomocą określonych narzędzi raportowania i wizualizacji w sposób statyczny lub dynamiczny.

HQDw&usg=AFQjCNGw_wSaqgGr32beq8POCxlne13KpA&bvm=bv.92885102,d.bGQ (dostęp 7.05.2015).

[33] J. Strickland, *How the Google File System works*, http://computer.howstuffworks.com/internet/basics/google-file-system.htm (dostęp 2.05.2015).

[34] D. Serain, *Middleware and enterprise application integration: the architecture of e-business solutions*, Springer Verlag, London 2002.

[35] T. Eri, *Service-Oriented Architecture: concepts, technology, and design* [e-book], NY, Prentice Hall Professional Technical Reference, Upper Saddle River 2005.

Aby cały zamierzony cel został osiągnięty, na poszczególnych poziomach, począwszy od infrastruktury fizycznej aż po prezentację rezultatów, architektura projektu musi być starannie przemyślana i zaprojektowana z uwzględnieniem indywidualnych wymagań klienta i zapewnienia skalowalności całego systemu.

Zbiory *Big Data* i technologia chmury obliczeniowej w humanistyce

Opisane wyżej przetwarzanie danych zapoczątkowane w sferze biznesu i obejmujące dane ustrukturalizowane stopniowo zaczęto uzupełniać przetwarzaniem danych nieustrukturalizowanych, tj. tekstów pochodzących z portali społecznościowych, e-maili klientów, nagrań dźwiękowych, blogów czy innych źródeł internetowych, a później także z zawartości dokumentów, danych graficznych, przestrzennych, bibliotek cyfrowych itd. Do przetwarzania tego typu danych stworzono z kolei narzędzia ogólnie zwane NoSQL (*Not only SQL*)[36].

Systemy NoSQL działają na bazie HBase, rozszerzonej wersji HDFS. HBase w stosunku do HDFS jest rozbudowana horyzontalnie i ma możliwość obsługiwania znacznie większych ilości danych niż HDFS. Baza składa się z jednostki głównej (*Master Node*) i serwerów regionalnych (*Regional Servers*), na których posadowione są kolejne HFiles.

Dla humanistów niewątpliwie najbardziej interesujące jest przetwarzanie zawartości dokumentów znajdujących się w zasobach bibliotek cyfrowych oraz Internetu. Trwający już od lat proces digitalizacji zasobów bibliotecznych osiągający w niektórych krajach wysoki procent całości zbiorów, tworzenie bibliotek cyfrowych, cyfrowych zasobów czasopism naukowych i repozytoriów literatury naukowej, a także rosnące lawinowo zasoby Internetu umożliwiają stosowanie narzędzi z obszaru ICT do ich analizy. Niewyobrażalny przyrost dokumentów podwajający się w coraz krótszym czasie skutkuje tym, że właściwie już żaden człowiek nie jest w stanie przeczytać wszystkich dokumentów, np. literatury naukowej nawet tylko ze swojej dziedziny, jej liczebność bowiem staje się tak duża i różnorodna (choćby ze względu na różnorodność języków), że jest to zadanie niewykonalne. Mówiąc obrazowo, człowiek stara się teraz zaprząc do czytania i analizowania dokumentów maszyny i narzędzia informatyczne wyższej generacji niż to było dotychczas. Za pomocą ekosystemu *Big Data* i *cloud computing* człowiek przenosi się w sferę, w której maszyna wyręczy go w dotychczasowych czynnościach czytania i analizowania

[36] A. Fowler, *NoSQL for dummies*, Jon Wiley & Sons, Inc. Hoboken (NY) 2015.

dokumentów w celu wytworzenia nowej wiedzy. Warunkiem jest jednak elektroniczna wersja dokumentów i zawartych w nich danych, toteż digitalizacja jak największej liczby dokumentów jest etapem wstępnym tych działań, jak i warunkiem *sine qua non* osiągnięcia sukcesu.

Do przetwarzania tekstów w tym ekosystemie stosuje się różne nowe narzędzia zaliczane do tzw. *Natural Language Processing* (NLP)[37]. Nie chodzi tu o znane nam dotychczas systemy wyszukiwania informacji, jak np. systemy słów kluczowych, haseł przedmiotowych, itd., choć posłużenie się nimi może stanowić pierwszy etap naszego działania. Zgodnie z dotychczasową praktyką, po zrealizowaniu wyszukiwania za pomocą np. znanego nam katalogu przedmiotowego czy słów kluczowych i po uzyskaniu dostępu do wyszukanych dokumentów, sami siadaliśmy do pracy i analizowaliśmy te materiały. To na tym etapie wytwarzaliśmy nową wiedzę. W nowym ekosystemie chodzi o takie narzędzia, które zastąpią nas na tym drugim etapie pracy, tj. wybiorą i dokonają analizy dokumentów zgodnie z naszym zapotrzebowaniem i w wyniku tego działania same wytworzą nową wiedzę.

Jeśli wziąć pod uwagę, że narzędzia, takie jak Hadoop, umożliwiają zlokalizowanie, gromadzenie i organizowanie niewyobrażalnie wielkich zbiorów danych, głównie ustrukturalizowanych, których prawdopodobnie nigdy nie zdołalibyśmy zgromadzić i zanalizować sami, to do systemów NoSQL dla sfery tekstów można zaliczyć m.in. Cassandrę[38], platformę rozwiniętą początkowo dla Google, Amazon i Facebooka, w celu wzmocnienia mechanizmów wyszukiwania, a od 2009 r. będącą w gestii Apache i używaną przez IBM, Thompson-Reutersa i Twittera.

Apache Cassandra wzmocniona mechanizmami MapReduce z Hadoop oraz rozszerzonymi mechanizmami BigTable umożliwia lokalizowanie i porządkowanie ogromnych ilości danych, operując Cassandra Query Language[39], pracującym w środowisku rozproszonym (klastrów serwerów). Jest ona wysoce skalowalna oraz zapewnia redundancję uniemożliwiającą unieruchomienie systemu, gdy jeden z elementów (klastrów) okaże się niewydolny. Ponieważ posiada mechanizmy zarówno SQL jak i NoSQL, jest wykorzystywana przez takie firmy, jak np. DdataStax[40], gdzie ustrukturalizowaną informację biznesową uzupełnia się informacjami pochodzącymi z analizy zawartości portali społecznościowych.

---

[37] G.G. Chowchury, *Natural language processing* "Annual Review of Information Science and Technology", vol. 37, no. 1/2003, s. 51–89.

[38] M. Brown, *Learning Apache Cassandra*, Packt Publishing Ltd., Birmingham 2015.

[39] Ibidem.

[40] Datastax, http://www.datastax.com/ (dostęp 2.05.2015).

W ramach NPL stosuje się różne techniki, takie jak: syntaktyczną analizę zdań, morfologiczną i semantyczną analizę wyrazów, ekstrakcję terminów, faktów, wydarzeń, nazw geograficznych, a nawet wyławia się postawy uczuciowe. Stosuje się metody statystyczne, włącza elementy tłumaczenia maszynowego i uczenia się systemów. Można też w efekcie tych działań ustrukturalizować wyniki i na ich podstawie dokonywać porównań, analiz i prognoz. Zestawem różnego rodzaju technik NLP dysponuje np. IBM Watson[41], superkomputer, który, posługując się różnymi algorytmami, jest już w stanie komunikować się z człowiekiem w języku naturalnym. Jest on w stanie zanalizować zawartość 1 mln książek na sekundę, a pierwsze jego praktyczne zastosowanie to analiza ogromnej ilości naukowej literatury medycznej w celu wspomagania i formułowania diagnostyki medycznej raka, szczególnie w sytuacjach natychmiastowej potrzeby informacji w warunkach klinicznych. System operuje ponad 100 różnymi technikami analizy języka naturalnego i w tym języku porozumiewa się także z użytkownikiem. Watson posiada architekturę zarządzania danymi nieustrukturalizowanymi (*Unstructured Information Management Architecture* – UIMA)[42] firmy Apache i elementy silnika wyszukiwawczego Apache Hadoop służącego do działania w wirtualnym środowisku rozproszonym. Często, w celu zrealizowania indywidualnego zapotrzebowania użytkownika/klienta, firmy, a nawet sami użytkownicy dopisują odpowiednie fragmenty oprogramowania w celu wykonania przez system konkretnego zadania.

Jeśli chodzi o zawartość bibliotek cyfrowych i w ogóle różnego typu dokumentów w wersji cyfrowej użytych ewentualnie do tego typu analiz, których liczba nie jest już żadną przeszkodą dla tych systemów, to zasadniczym problemem jest obecnie sprawa praw autorskich. Wobec tego, że legislacja nie zawsze nadąża za rozwojem techniki, dotychczasowe prawo autorskie jeszcze nic nie mówi o tym, czy dokonywanie takich analiz za pomocą tego rodzaju silników na zbiorach *Big Data* w ekosystemie *cloud computing* jest czy nie jest dopuszczalne. Z drugiej strony można jednak przyjąć, że zamiast człowieka korzystającego z dokumentów na prawach bibliotecznych i tworzącego w wyniku tego procesu nową wiedzę, czyni to za niego maszyna. Trudno powiedzieć, czy jest tu jakaś różnica, czy nie. Aby nie złamać prawa autorskiego działanie silnika *Big Data*, na dokumentach musiałoby odbywać się na zasadach bibliotecznych lub być wykonywane na dokumentach ze sfery publicznej. Stąd też niektórzy eksperymentatorzy w świecie bibliotek na razie, wobec braku uregulowań w sferze prawa

[41] *IBM Watson*, http://www.ibm.com/smarterplanet/us/en/ibmwatson/ (dostęp 2.05.2015).

[42] The Apache Software Foundation. *Getting started: Why Apache UIMA*, https://uima.apache.org/doc-uima-why.html (dostęp 2.05.2015).

autorskiego, ostrożnie przeprowadzają lub umożliwiają przeprowadzanie takich działań tylko na wybranych i wydzielonych zbiorach zaliczających się do sfery publicznej. Tak na przykład postąpiła the British Library, która przy współpracy z University College of London przystąpiła do działań eksperymentalnych, używając do tego celu platformy Microsoft Azure[43]. Platforma ta, udostępniająca przede wszystkim usługi PaaS (Platform as a Service), jest w stanie zestawić dowolną architekturę narzędzi w celu wykonania określonego zadania. Inne biblioteki, jak np. amerykańskie, wprowadzają blokady dla każdego indywidualnego dokumentu objętego prawem autorskim, jeśli działanie silnika miałoby odbywać się na innych niż biblioteczne zasadach. W tym miejscu musimy sobie uświadomić, że to właśnie biblioteki będą zapewne tworzyć tego typu centra badawcze, tj. pracownie, gdzie dostępne będą omawiane wyżej narzędzia. Ponieważ jednak najczęściej dostęp do platform i narzędzi *Big Data* w ekosystemie *cloud computing* nie jest wolny od opłat – biblioteki będą pewnie musiały wykupywać licencje na korzystanie z tych narzędzi, które naukowcom/użytkownikom umożliwią prowadzenie tego typu badań. Centra takie, zawiązywane często jako konsorcja bibliotek, będą musiały też być w stałym kontakcie z firmami macierzystymi, które mogą gwarantować w ramach umów licencyjnych dostosowywanie narzędzi do indywidulanych potrzeb badaczy/użytkowników.

Humanistyka, która formalnie definiowana jest jako sfera dotycząca człowieka funkcjonującego w kontekście społecznym, włączając w to jego twórczości[44], jest tu raczej desygnatem szeroko pojętej produkcji treści prezentowanych za pomocą obszernych dokumentów, tj. książek, artykułów z czasopism itd., w których mowa jest o myślach, ideach, pojęciach, koncepcjach, słowach, terminach, języku, wydarzeniach, osobach, gramatykach, tożsamościach, charakterach, lokalizacjach i miejscach – występujących w całkowicie nieustrukturalizowanej formie. Teksty te dają się podzielić na tomy, rozdziały, sekcje, a nawet strony, co mechanizmom, takim jak BigTable, umożliwia wyodrębnienie mniejszych jednostek (np. pojedynczych stron) i błyskawiczne, paralelne, hierarchiczne ich przetwarzanie. Innymi możliwymi wymiarami są np.: język, organizacja myśli w ramach poszczególnych języków, okresy czasu. Wszystko to ułatwia działanie silnikom *Big Data*, które są w stanie błyskawicznie wydobyć wszystkie elementy ich kontekstu, zanalizować je i zaprezentować w do-

[43] J. Baker, *The British Library big data experiment*. *The British Library. Digital scholarship blog*, http://britishlibrary.typepad.co.uk/digital-scholarship/2014/06/the-british-library-big-data-experiment.html (dostęp 2.05.2015).

[44] A.R. Rogers, *The Humanities*, 2nd ed., Libraries unlimited, Inc., Littleton (Colorado) 1979.

godny dla badacza sposób. Warunkiem jednak ich skutecznego działania jest podział analizowanych tekstów na małe porcje, co wymaga dużych pojemności pamięci na procesy indeksowania. Operowanie całymi dokumentami nie zapewni oczekiwanych rezultatów.

Silniki te, jeśli nawet nie zadowolą nas w stu procentach, to są jednak już dziś w stanie zidentyfikować i zanalizować całość interesującego nas oraz dostępnego materiału na dany temat i przynajmniej wskazać kierunek dalszych działań. Najprawdopodobniej w najbliższym czasie żaden naukowiec nie będzie już w stanie prowadzić badań bez zaprzęgnięcia silników tego typu do pracy. W innym przypadku zabraknie mu życia, znajomości języków i czasu na analizę relewantnych źródeł, nie mówiąc o tym, że badania w tym kontekście prowadzone „na piechotę" mogą się okazać niepełne, dublowane i zbaczające z głównego nurtu wobec niemożności uwzględnienia przez badacza ważnych materiałów, których sam nigdy by nie zdobył i nie był w stanie zanalizować. Biblioteki muszą przygotować się do świadczenia tego rodzaju nowych usług, a warunkiem jest całkowita digitalizacja zbiorów. Na marginesie należy dodać, że w literaturze przedmiotu jako *Big Data* w sferze tekstów określa się liczebność zbiorów na minimum 1 mln zdigitalizowanych dokumentów. Do tego ekosystemu nie można zatem zaliczyć np. projektu Guthenberg, który zawiera tylko nieco ponad 45 tys. zdigitalizowanych dokumentów[45].

Przykład

Dobrym przykładem takiej właśnie cyberinfrastruktury jest The HathiTrust Research Centre (HTRC)[46], które to centrum jest partnerstwem dwóch uniwersytetów: Indiana University oraz University of Illinois posiadających łącznie ok. 10 mln zdigitalizowanych dokumentów pełnotekstowych na 2013 r. (w tym ponad 5 mln książek i ponad 274 tys. tytułów wydawnictw ciągłych), co daje łącznie ponad 3,5 mld stron tekstu. Zbiór ten w wersji papierowej zajmuje 125 mil półek, waży ponad 8,5 tys. ton, a w wersji elektronicznej odpowiada temu 473 TB powierzchni dyskowej. Dominują tam dokumenty w zakresie: języka, literatury, historii, socjologii, biznesu i ekonomii. Jest to The HathTtrust Digital Library – biblioteka cyfrowa o rozproszonej architekturze posadowiona w chmu-

[45] Project Gutenberg, *Free e-books – Projekt Gutenberg*, https://www.gutenberg.org/ (dostęp 2.05.2015).

[46] S.T. Kowalczyk [i in.], *Big data at scale for digital humanities; An architecture for The HathiTrust Research Center*, w: *Big data management. Technologies, and applications*, red. W. Ch. Hu, N. Kaabouch, IGI Global, Hershey (USA) 2014, s. 270–294 (*Information Science Reference*).

rze. W ramach tego dziesięciomilionowego zbioru nieco ponad 3 mln to dokumenty ze sfery publicznej[47].

System pracuje na platformie Apache Cassandra oraz, w celu ułatwienia współpracy systemu operacyjnego z aplikacjami, posługuje się oprogramowaniem *middleware* WSO2, również firmy Apache (jest to tzw. szyna integracyjna)[48], gwarantującym właściwe usługi SOA. Do wyłaniania przez badacza wstępnej kolekcji dokumentów, które znajdują się w polu jego zainteresowania, służy otwarte oprogramowanie katalogowe Blacklight[49], umożliwiające przeszukiwanie zbioru zarówno za pomocą metadanych, jak i elementów pochodzących z pełnych tekstów dokumentów. Jest ono oparte na mechanizmach webowej, skalowalnej aplikacji „Ruby and Rails"[50] i współpracuje z oprogramowaniem indeksującym SOLR. Apache SOLR[51], oparty na produkcie firmy Lucene, jest otwartym oprogramowaniem umożliwiającym indeksowanie danych pochodzących zarówno z metadanych o charakterze bibliograficznym, jak i ze zbiorów nieustrukturalizowanych. Za pomocą tych narzędzi tworzona jest wstępna kolekcja przygotowana do analizy. W zakresie metadanych system ma do dyspozycji nie tylko rekordy w formacie MARC (*Machine Readable Cataloguing*), ale także w formacie METS (*Metadata Encoding and Transmission Standard*)[52]. Z kolei do analizy danych i wizualizacji wyników stosuje się różne narzędzia aplikacyjne. Dla humanistyki polecane jest np. stworzone specjalnie w ramach projektu The Software Environment for the Advencement Scholarly Research (SEASR)[53] narzędzie pod nazwą Meandre[54], umożliwiające m.in. ilościową analizę słów w dokumentach oraz inne operacje, w tym także prezentowanie wyników np. w postaci chmury tagów, Simile Timeline[55] do wyławiania i prezentowania dat czy Google map[56] do wyławiania i prezentacji lokalizacji. Oprogramowanie to zostało przygotowane przy pomocy amerykańskiej fundacji wspierającej działalność

[47] Ibidem.

[48] BlueSoft Sp. z o.o., *WSO2 Enterprise Service Bus*, 2015, http://www.bluesoft.net.pl/pl/technologie/wso2-esb/ (dostęp 8.05.2015); S.T. Kowalczyk [i in.], op.cit.

[49] Blacklight, http://projectblacklight.org/ (dostęp 2.05.2015).

[50] B. Burd, *Ruby on Rails for dummies*, Wiley Publishing Inc., Hoboken, NY 2007.

[51] SOLR, http://lucene.apache.org/solr/ (dostęp 3.05.2015).

[52] The Library of Congress, *Metadata Encoding and Transmission Standard*, ostatnio aktualizowane 17.04.2015, http://www.loc.gov/standards/mets/ (dostęp 1.05.2015).

[53] SEASR, http://www.seasr.org/documentation/ (dostęp 3.05.2015).

[54] Meandre, http://www.seasr.org/meandre/ (dostęp 1.05.2014).

[55] *Reference Documentation for Timeline*, http://simile-widgets.org/wiki/Reference_Documentation_for_Timeline (dostęp 4.05.2015).

[56] P. Derbyshrine, A. Derbyshrine, *Getting startED with Google Apps*. [New York], Apress, 2010 (s. 489–515: Google Maps).

bibliotek, tj. The Andrew W. Mellon Foundation[57]. Stosowane są też inne, zaawansowane algorytmy NLP oraz algorytmy do porównywania wyników uzyskanych z różnych przebadanych kolekcji. W projekcie HathiTrust Research Centre odpowiednie algorytmy badacz może wybrać z zestawu dostępnego na ekranie interfejsu systemu. Sam w ten sposób może kierować swoim badaniem, dobierając odpowiednie narzędzia i ukierunkowując system tak, aby uzyskać potrzebne wyniki.

Wyłania się tu wiele problemów związanych z dostępnością, jak i bezpieczeństwem danych, takich jak: wstępna identyfikacja użytkownika, dostęp do dokumentów objętych prawem autorskim, zabezpieczenie dostępu do rezultatów badań wykonanych przez poszczególnych badaczy (stworzonej w ten sposób nowej wiedzy) itd. Niemniej pierwsze próby zrealizowane dotychczas okazują się udane, a w ich wyniku otrzymujemy nową, wartościową wiedzę[58].

Zapewne z czasem systemy tego typu będą ulepszane i optymalizowane. Może doczekamy też chwili, gdy maszyna, po wykonaniu pracy badawczej, napisze za nas książkę! Wtedy to, co nazywany „pracą naukową" w obszarze humanistyki, osiągnie etap całkowitej automatyzacji, a rola człowieka, tak jak w innych zautomatyzowanych dziedzinach, ograniczy się do sterowania tym procesem.

Rozwój humanistyki cyfrowej wspiera też Unia Europejska. Zagadnienie to będzie przedmiotem osobnego opracowania.

Bibliografia

Alber A.F., *Videotex/teletext: principles and practices*, McGraw Hill Book Company, New York 1986.

Amazon Web Services, *Amazon Elastic Computer Cloud. User Guide for Linux. API version 2014-10-01*, http://www.google.pl/url?sa=t&rct=j&q=&esrc=s&source=web&cd=12&ved=0CG8QFjAL&url=http%3A%2F%2Fawsdocs.s3.amazonaws.com%2FEC2%2Flatest%2Fec2-ug.pdf&ei=jZdLVei2LuTgyQPH_4CoBA&usg=AFQjCNFIg7f8P673xBxButxQueralRAHkg&bvm=bv.92885102,d.bGQ.

Baker J., *The British Library big data experiment*. *The British Library. Digital scholarship blog*, http://britishlibrary.typepad.co.uk/digital-scholarship/2014/06/the-british-library-big-data-experiment.html.

Bakshi K., *Technologies of big data*, w: *Big data management. Technologies, and applications*, red. Wen-Chen Hu, N. Kaabouch, IGI Global, Hershey (USA) 2014 (*Information Science Referenc*es).

[57] S.T. Kowalczyk [i in.], op.cit.

[58] Ibidem.

Baldeschfieler E., *Hadoop at Yahoo*! http://www.google.pl/url?sa=t&rct=j&q=&esrc=s&source=web&cd=6&ved=0CFMQFjAF&url=http%3A%2F%2Fwww.hadooper.cn%2Fdct%2Fattach%2FY2xiOmNsYjpwZGY6ODE%3D&ei=xX5LVdLpGYupygOGiYHQDw&usg=AFQjCNGw_wSaqgGr32beq8POCxlne13KpA&bvm=bv.92885102,d.bGQ.

Berson A., Smith S.J., *Data warehousing, data mining, and OLAP*, Mc Graw-Hill New York 2001.

Big data management, technologies, and Applications, red. Wen-Chen Hu, N. Kaabouch, Information Science Reference (IGI Global), Hershey 2014 (Seria: Advences in Data Mining and Database Management).

Blacklight, http://projectblacklight.org/.

BlueSoft Sp. z o.o. *WSO2 Enterprise Service Bus*, http://www.bluesoft.net.pl/pl/technologie/wso2-esb/.

Brown M., *Learning Apache Cassandra*, Packt Publishing Ltd., Birmingham 2015.

Burd B., *Ruby on Rails for dummies*, Wiley Publishing Inc., Hoboken, NY 2007.

Business intelligence, red. J. Suma, Wydawnictwo Naukowe PWN, Warszawa 2013.

Chinea A., *Understanding the principles of recursive neural networks. A generative approach to tackle model cpmplexity*, http://arxiv.org/ftp/arxiv/papers/0911/0911.3298.pdf.

Chodkowska-Gyuriks A., *Hurtownie danych: teoria i praktyka*, Wydawnictwo Naukowe PWN S.A., Warszawa 2014.

Chowchury G.G., *Natural language processing*, Annual Review of Information Science and Technology", vol. 37, no. 1/2003, s. 51–89.

Ciurana E., *Developing with Google App Engine*, Apress, Distributed by New York Springer Verlag, Berkeley (CA) 2008.

Datastax, http://www.datastax.com/.

David M.M., Fesperman L., *Advanced standard SQL dynamic structured data modeling and hierarchical processing* [e-book]. Boston, Artech House, 2013, http://web.a.ebscohost.com/ehost/ebookviewer/ebook/ZTAwMHh3d19fNzUzNTk2X19BTg2?sid=96ce560b-b590-47f4-9f8c-18783f9988e6@sessionmgr4002&vid=3&format=EB&rid=1.

Derbyshrine P., Derbyshrine A., *Getting start ED with Google Apps* [New York], Apress, 2010 (s. 489–515: Google Maps).

Eri T., *Service-Oriented Architecture: concepts, technology, and design* [e-book], Prentice Hall Professional Technical Reference, Upper Saddle River, NY 2005.

Fowler A., *NoSQL for dummies*, Jon Wiley & Sons, Inc. Hoboken (NY) 2015.

Hadoop, *HDFS Architecture guide*, http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html.

Hodges A., *Alan Turing: Enigma*, Wydawnictwo Albatros Andrzej Kuryłowicz S.C., Warszawa 2014.

Hurwitz J., Nugent A., Halper, F. Kaufman M., *Big data for dummies,* John Wiley & Sons. Inc., Hoboken, NY 2013.

IBM Watson, http://www.ibm.com/smarterplanet/us/en/ibmwatson/.

Jackson P., *Introduction to experts systems*, Accisson Wesley Pub.co, Workingham (England) 1986.

Komisja Europejska. Komunikat Komisji do Parlamentu Europejskiego, Rady, Europejskiego Komitetu Ekonomiczno-Społecznego i Komitetu Regionów *Europejska agenda cyfrowa,* KOM(2010) 245 Bruksela 19.05.2010 wersja ostateczna.

Komisja Europejska. Komunikat Komisji do Parlamentu Europejskiego, Rady, Europejskiego Komitetu Ekonomiczno-Społecznego i Komitetu Regionów w sprawie wykorzystania potencjału chmury obliczeniowej w Europie, Bruksela (2012), KOM(2012) 529 wersja ostateczna.

Kowalczyk S.T. [i in.], *Big data at scale for digital humanities; An architecture for The HathiTrust Research Center*, w: *Big data management. Technologies, and applications*, red. Wen-Chen Hu, N. Kaabouch, IGI Global, Hershey (USA) 2014 (*Information Science Reference*), s. 270–294.

Lephamer H., *RFID design principles*, Artech House, Inc. Boston, London 2008.

Mayer-Sconberger V., Cukier K., *Big data. Rewolucja, która zmieni nasze myślenie, pracę i życie*, MT Biznes sp. z o.o., Warszawa 2014.

Meandre, http://www.seasr.org/meandre/.

Mell P., Grance T., *The NIST definition of cloud computing. Recommendation of the National Institute of Standards and Technology*, Computer Security Division, Information Technology Laboratory, National Institute of Standards and Technology, Gaithesburg, (MD) 2011, http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf.

Microsoft Azure, http://azure.microsoft.com/pl-pl/.

Miner D., Schock A., *MapReduce design patterns*, O'Reilly, Beijing; Koln 2013.

Morin P., *Open data structures. An introduction* [e-book] in OPEL (Open Paths to Enriched Learning). Edmonton: AU Press. 2013, http://web.b.ebscohost.com/ehost/ebookviewer/ebook/ZTAwMHh3d19fNjM4OTU2X19BTg2?sid=d8b94be4-7c8e-4f69-9727-d3428c90cdfe@sessionmgr115&vid=3&format=EB&rid=1.

Mosco V., *To the cloud. Big data in a turbulent world*, Paradigm Publishers, Boulder (Colorado), London 2014.

NFC Forum: *NFC Data Exchange Format {NDEF). Technical specification. 2006-07-24*, NFC Forum, Inc., http://www.eet-china.com/ARTICLES/2006AUG/PDF/NFCForum-TS-NDEF.pdf?SOURCES=DOWNLOAD.

Osowski S., *Sieci neuronowe do przetwarzania informacji*, wyd. 2 popr. i uzup., Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa 2006.

Parfatt M.C., *What is EDI?: a guide to electronic data interchange*, PNCC Blackwell, Oxford 1992.

Project Gutenberg, *Free e-books – Projekt Gutenberg*, https://www.gutenberg.org/.

Reference Documentation for Timeline, http://simile-widgets.org/wiki/Reference_Documentation_for_Timeline.

Rogers A.R. *The Humanities*, 2nd ed., Libraries unlimited, Inc., Littleton (Colorado) 1979.

Rutkowski L., *Metody i techniki sztucznej inteligencji*, wyd. 2 zmien., Wydawnictwo Naukowe PWN, Warszawa 2012.

Saavas I.K., Sofianidon G.N, M-Tachar Kechadi, *Applying the K-Means Algorithm in Big Raw Data Sets with Hadoop and MapReduce*, w: *Big data management. Technologies, and applications*, red. Wen-Chen Hu, N. Kaabouch, IGI Global, Hershey (USA) 2014, s. 1–22 (*Information Science Referenc*es).

SEASR, http://www.seasr.org/documentation/.

Serain D., *Middleware and enterprise application integration: the architecture of e-business solutions*, Springer Verlag, London 2002.

SOLR, http://lucene.apache.org/solr/.

Strickland J., *How the Google File System works*, http://computer.howstuffworks.com/internet/basics/google-file-system.htm.

Szmit P., *Cloud computing historia, technologia, perspektywy*, Polska Agencja Rozwoju Przedsiębiorczości, Warszawa 2012, https://www.web.gov.pl/g2/big/2012_06/ebfa211f1a9f174c7517738f68df2d8b.pdf.

The Apache Software Foundation, *Getting started: Why Apache UIMA*, https://uima.apache.org/doc-uima-why.html.

The Library of Congress, *Metadata Encoding and Transmission Standard*, ostatnio aktualizowane 17.04.2015, http://www.loc.gov/standards/mets/.

Wiener N., *Cybernetics: or control and communication in the animal and the machine*, M.I.T. Press, New York 1961.

Słowa kluczowe: humanistyka cyfrowa, technologia chmury obliczeniowej, zbiory *Big Data*

Key words: Digital Humanities, the Cloud Computing Technology, *Big Data* Sets

Big Data Sets and the Cloud Computing Technology for Digital Humanities

Abstract

The main aim of the paper is to draw a concept of sets of *Big Data* which occur in the sphere of digital humanities, to describe new IT tools applied for its analysis and the conditions of its usage in the environment of the *cloud computing* technology. The technological development, which took place in the recent years in the field of information and communications technologies (ICT), especially the change in the way of governing of data in relation to the development of the *cloud computing* technology enables nowadays gathering unimaginable amount of data (*Big Data* sets) being created in various spheres of human and machine world. The definition and the origin of development of such sets of data are presented in the paper as well as the construction of the ecosystem of the *cloud computing* technology in which *Big Data* sets are gathered and processed is described. *Big data* sets are reach sources of new information. Advanced IT analytical tools are applied to investigate its content in order to obtain a new knowledge. In the paper the emphasize is on sets of *Big Data* gathered within the digital humanities area and chosen examples of IT analytical tools currently applied for these purposes are described. Also, the conditions related to the usage of such new information infrastructure in relation to digital humanities are highlited.