ZESZYTY NAUKOWE UNIWERSYTETU SZCZECIŃSKIEGO

NR 733 STUDIA INFORMATICA NR 30 2012

JACEK UNOLD[*]
Wyższa Szkoła Bankowa we Wrocławiu

SPOSOBY EFEKTYWNEGO ZARZĄDZANIA DANYMI W DRUGIEJ GENERACJI WWW

Wprowadzenie

W obszarze problematyki dotyczącej danych analizowanych w kontekście funkcjonowania cyberprzestrzeni widać wyraźne przenikanie się różnorodnych nurtów i wzorców charakterystycznych dla Web 2.0 z każdym z trzech etapów przetwarzania informacji w WWW, to jest: identyfikacją potrzeb informacyjnych, pozyskiwaniem danych oraz ich wykorzystaniem. Już samo pojęcie gospodarki opartej na danych zasadza się między innymi na otwartych standardach danych (ang. *open data*), efekcie sieciowym oraz koncepcji oprogramowania jako usługi. Z kolei udostępnianie danych dla obszarów niszowych w cyberprzestrzeni wymaga atomizacji (granulacji) informacji i danych, czyli uczynienia informacji dostępnej w małych porcjach. Jest to jednocześnie dyskontowanie efektu *długiego ogona*, zjawiska będącego jednym z aspektów pozyskiwania informacji. Wynikające z granulacji możliwości w zakresie remiksu treści informacyjnych (ang. *mash-up*), to jedna z kategorii wykorzystania danych (trzeci etap przetwarzania). Ta obserwacja dotyczy większości kategorii charakteryzujących Web 2.0, to znaczy można je analizować w każdym z trzech aspektów przetwarzania informacji.

[*] jacek.unold@wsb.wroclaw.pl.

W artykule przedstawiono zarys strategii efektywnego zarządzania danymi w cyberprzestrzeni doby Web 2.0. Zidentyfikowano i scharakteryzowano podstawowe zasady efektywnego wykorzystania unikalnych danych oraz przedstawiono istotę tworzenia tak zwanej architektury współuczestnictwa i związane z tym możliwości w zakresie integracji danych. Praktyczne aplikacje zastosowane przez Amazon.com i Flickr.com pozwoliły na zilustrowanie oddziaływania zasady *Dane jako następny Intel w środku*, a pięć zidentyfikowanych strategii pozwoliło na odniesienie się do problematyki prawa własności danych w cyberprzestrzeni.

1. Podstawowe zasady wykorzystania unikalnych danych

Choć historia rozwoju koncepcji Web 2.0 jest wciąż krótka, a opublikowana dotąd literatura niezbyt bogata, jednak umożliwiają one identyfikację kilku podstawowych zasad, jakie powinny pomóc w tworzeniu i zarządzaniu unikalnymi bazami danych. Na przykład P.R. Jackson i T. Musser dają następujące wskazówki[1]:

1. Należy starać się pozyskać unikalne, trudne do zdobycia lub odtworzenia dane.
2. Należy podnosić wartość podstawowych zbiorów danych.
3. Użytkownicy powinni mieć możliwość kontrolowania swoich własnych danych.
4. Nie wszystkie prawa powinny być zastrzeżone.
5. Należy określić strategię zbiorów danych.
6. Należy być właścicielem indeksu, nazwy bądź formatu.
7. Należy projektować dane dla ponownego wykorzystania.
8. Nie ma potrzeby, by samemu zarządzać dostępem do danych.

Ad. 1. Przykładem jest NAVTEQ, czołowy światowy dostarczyciel cyfrowych danych kartograficznych, operujący w 58 krajach. Na bazie tych codziennie uaktualnianych danych stworzono tak znane aplikacje, jak: Google Maps, Microsoft Virtual Earth czy AOL MapQuest.

[1] P.R. Jackson, *Web 2.0 Knowledge Technologies and the Enterprise: Smarter, Lighter, Cheaper*, Chandos Publishing, Oxford 2010; J. Musser, T. O'Reilly, *Web 2.0: Principles and Best Practices*, O'Reilly Media, Sebastopol 2007.

Ad. 2. Wartość podstawowych zbiorów danych można zwiększać przez opinie i recenzje użytkowników, udostępnianie historii dokonanych zakupów, możliwość tworzenia tagów i oceny. Przykładem jest Amazon.com, który wzbogacił podstawowy katalog oferowanych książek przez dodanie wielu z tych opcji. Jednym z najważniejszych efektów takich zabiegów jest uzyskanie zagregowanych wzorców zachowań konsumenckich, co pozwala uzyskać decydującą przewagę konkurencyjną, jak w przypadku Amazon.com.

Ad. 3. Użytkownik powinien mieć możliwość ściągania danych z oferowanej aplikacji, a także, co najważniejsze, wycofania własnych danych. Podobnie jak zamknięte aplikacje programowe ustąpiły w końcu idei *open source*, coraz większego znaczenia nabiera koncepcja tzw. otwartych danych (ang. *open data*). W najpełniejszym wymiarze tendencji ta jest widoczna w tworzeniu formatów dla dystrybucji otwartych danych, takich jak RSS czy GeoRSS, a także mikroformatów oraz otwartych interfejsów programowania.

Ad. 4. Zbyt rygorystyczne przestrzeganie praw własności intelektualnej nie jest wskazane w sytuacji, kiedy zasadnicze korzyści są uzyskiwane z oddziaływania efektu sieciowego, *wirusowości* i zbiorowej adopcji. Jednym z rozwiązań jest zezwolenie na dzielenie się daną własnością, na przykład oprogramowaniem czy danymi, przy oferowaniu autorskich licencji samym twórcom (przypadki Creative Commons czy General Public License).

Ad. 5. Łańcuch wartości w przypadku danych znajdujących się w cyberprzestrzeni jest zazwyczaj bardzo rozbudowany. Na przykład niewielu jest właścicieli pełnych zbiorów danych kartograficznych (wspomniany wcześniej NAVTEQ). Inni uczestnicy cyberrynku przetwarzają te dane (na przykład deCarta), a jeszcze inni dostarczają je użytkownikowi końcowemu (na przykład Google w aplikacji Google Maps lub Google Earth). Najczęściej granice między poszczególnymi uczestnikami rynku są bardzo płynne, ponadto pojawia się wiele podmiotów oferujących kompilacje danych kartograficznych z innymi kategoriami danych, na przykład z rynku nieruchomości, tworząc tzw. *mashupy*. Stąd tak istotne jest określenie własnego miejsca w tym łańcuchu i próba uzyskania tam przewagi strategicznej.

Ad. 6. Do osiągnięcia sukcesu w cyberprzestrzeni nie zawsze konieczne jest posiadanie zbiorów unikalnych i trudno dostępnych danych. Równie dobrym sposobem jest uzyskanie wiodącej pozycji w zakresie znajdowania, dostępu, rangowania bądź formatowania danych. Najlepszym przykładem jest

tu indeks wyszukiwarki Google – Google nie jest właścicielem rangowanych stron, posiada jednak najlepszy indeks do tych stron. Podobnie Technorati posiada indeks najlepszych blogów, Amazon ranguje produkty, a Alexa – witryny internetowe.

Ad. 7. Ponieważ dane stają się równie istotne jak funkcje, możliwość ponownego ich wykorzystania staje się wręcz tak ważna, jak w przypadku samego oprogramowania. W literaturze anglojęzycznej używa się określenia *projektowanie danych* (ang. *data design*), które jest pewnym skrótem myślowym. Chodzi o szerszą koncepcję tworzenia takich mechanizmów dostępu do danych, ich prezentacji, formatowania i licencjonowania, by posiadane zbiory uczynić adresowalnymi, dającymi się wyszukiwać i gotowymi na syndykację (RSS).

Ad. 8. Wprawdzie koszty wykorzystania pasma i przechowywania danych stale ulegają istotnemu obniżaniu, ale samo zarządzanie pokaźnymi zbiorami danych może wciąż stanowić duży problem. W szczególności dotyczy to bardzo dużych zbiorów danych, które są bazą plików audio, foto i wideo. Na przykład aż trzy z pierwszej dziesiątki najpopularniejszych witryn udostępniających wymianę zdjęć, to znaczy Photobucket.com, ImageShack i Slide.com, uzyskują aż połowę ruchu na swoich witrynach z MySpace.com – portalu społecznościowego, który zapewnia *outsourcing* w tej dziedzinie[2].

2. Integracja danych a wykorzystanie architektury współuczestnictwa

Dynamiczny rozwój cyberprzestrzeni w ostatnich latach, odzwierciedlany przez coraz wyraźniejszą wszechobecność Internetu i upowszechnienie urządzeń komputerowych, to także coraz większa dynamika w zakresie pozyskiwania i integracji danych oraz zarządzania nimi. Drugi z ośmiu zasadniczych wzorców charakteryzujących cyberprzestrzeń ery Web 2.0 brzmi *Dane jako następny Intel w środku*, a zarządzanie bazą danych zostało uznane za jedną z podstawowych kompetencji (ang. *core competences*) Web 2.0[3]. Chodzi tu między innymi o wykorzystanie unikalnych, trudnych do od-

[2] *Photobucket Leads Photo-sharing Sites* , „Hitwise" 21 czerwca 2006 r.

[3] T. O'Reilly, *What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software*, http://oreilly.com/lpt/a/6228, 11.03.2011.

tworzenia źródeł danych w oparciu o przekonanie, że dane stają się równie istotne jak funkcje. Innym aspektem jest też takie projektowanie zbiorów danych, aby można je było łatwo wykorzystać ponownie, i to wielokrotnie. Jednym z większych wyzwań na obecnym etapie rozwoju Internetu jest samo zagadnienie integracji. Chodzi o integrację danych dotyczących tego samego obszaru tematycznego, ale pochodzących z różnych źródeł, dostarczanych w rozmaitych formatach itd. Na przykład jedna baza danych może rejestrować wiek konsumenta, a inna – datę urodzenia tego samego konsumenta, jedna – zarobek roczny, inna – zarobki w rozbiciu miesięcznym itd. Nawet jeżeli dane te dotyczą tych samych informacji, są one inaczej oznaczane i formatowane, a to może sprawiać podstawowe problemy w późniejszym zarządzaniu nimi. Integracja danych ma więc bezpośredni związek z możliwością pozyskiwania relewantnych informacji.

Niezależnie od tego typu wyzwań i trudności, w cyberprzestrzeni można zaobserwować prawdziwy wyścig w pozyskiwaniu pewnych podstawowych klas danych: lokalizacji, tożsamości, kalendarza ważnych wydarzeń, identyfikatorów produktów, nazw. Warunkiem realizacji takiej strategii stało się stworzenie udanej architektury współuczestnictwa, co jest jednym z wiodących konceptów Web 2.0. Tworzone w ten sposób bazy danych są oparte na treściach wygenerowanych i stale ulepszanych przez samych użytkowników. Jednakże dane, jakie krążą w sieci, i na podstawie których użytkownicy otrzymują interesujące ich informacje, są wprowadzane do cyberprzestrzeni nie tylko przez samych użytkowników – blogujących, piszących w Wiki, dzielących się recenzjami i rekomendacjami. Wiele dzisiejszych aplikacji generuje dane w sposób automatyczny, na przykład aplikacje stosowane w handlu elektronicznym czy w telefonii cyfrowej. Dane są wówczas gromadzone w sposób celowy, na przykład przez wyszukiwarki internetowe. Gromadzenie danych może też być efektem ubocznym określonej działalności w cyberprzestrzeni. Można zatem podzielić dane na te, generowane przez użytkownika i generowane automatycznie.

Dane generowane automatycznie są często redundantne, a ich ilość jest tak wielka, że do ich przekształcania w użyteczne informacje potrzebne są specjalne techniki, jak: magazynowanie danych, ich integracja czy drążenie. Wprawdzie magazynowanie i klasyfikacja danych nie są zjawiskami nowymi (książka telefoniczna została wprowadzona do użytku pod koniec XIX wie-

ku), jednak decydująca jest tutaj forma danych. Elektroniczny format danych pozwala na ich integrację z najróżniejszymi aplikacjami. Dane stają się w ten sposób podstawą funkcjonowania wielu nieznanych wcześniej usług. W rezultacie, wprowadzane standardy i rozwiązania coraz efektywniej przekształcają całe klasy danych w bardzo konkretne podsystemy internetowych systemów operacyjnych (ang. *Internet Operating Systems*) czy webowych systemów operacyjnych WebOS (ang. *Web Operating Systems*)[4].

Pozostając przy przykładzie telefonii, można powiedzieć, że dane elektroniczne pozwalają na odwrócenie procesu kojarzenia nazwiska z numerem telefonu. Tradycyjny sposób pozwalał na odczytanie numeru, jeśli miało się uprzednio podane nazwisko. Nowy sposób zwany *odwrotną książką telefoniczną* (ang. *inverse phone directory*) pozwala na natychmiastową identyfikacje nazwiska i adresu na podstawie numeru. Przykładem witryny specjalizującej się w *odwrotnym* wyszukiwaniu abonentów na podstawie numeru stacjonarnego lub komórkowego jest Phone Search Central.

Jest to jednocześnie przykład odzwierciedlający kluczową rolę danych w pierwszym etapie przetwarzania informacji, związanym z potrzebami informacyjnymi, a ściślej – z kreowaniem potrzeb informacyjnych. Często są to bowiem potrzeby wcześniej nieuświadamiane przez samych użytkowników.

3. Wybrane aplikacje praktyczne – Amazon i Flickr

Omówione niżej przykłady dotyczą działalności dwóch sztandarowych przedsięwzięć ery Web 2.0, to jest: Amazon.com i Flickr.com. Obie firmy zastosowały z powodzeniem strategię *Dane jako następny Intel w środku*. Amazon wprowadził własny wewnętrzny system identyfikacji swoich produktów – ASIN (Amazon Standard Identification Number). Dla książek posiadających ISBN numer ASIN jest równoznaczny z kodem ISBN. Książkom bez ISBN i wszystkim innym produktom również przypisuje się ASIN. W ten sposób Amazon stworzył unikalną bazę danych wraz z efektywnym systemem zarządzania nią.

Z kolei witryna Flickr.com jest typowym przykładem bazy danych tworzonej całkowicie przez użytkowników. Zaczęto od podstawowych danych,

[4] Zob. m.in. M. Allen, *Palm webOS*, O'Reilly Media, 2009.

jakimi są pliki zdjęć, a następnie zaproszono użytkowników do współpracy w zamieszczaniu nie tylko podstawowych danych, ale też metadanych, takich jak: tagi, notatki, profile, grupy, dyskusje itd. Efekt sieciowy dokonał reszty, a rezultatem jest trudna do odtworzenia i niezwykle cenna baza danych. Główna strona Flickr.com (http://www.flickr.com) prezentuje cztery ikony w dolnym lewym rogu, umożliwiające użytkownikowi aktywny udział w tworzeniu architektury partycypacji, rozbudowie bazy danych i życiu tej społeczności. Ikony zachęcają do wykonania następujących czynności:

– podziel się i bądź w kontakcie (*Share & stay in touch*);
– załaduj fotografię i organizuj zbiór (*Upload & organize*);
– przytnij, dopasuj, edytuj (*Crop, fix, edit*);
– przeglądaj (*Explore*).

W ten sposób Flickr realizuje jedną z naczelnych strategii Web 2.0, a jednocześnie w aktywny sposób kreuje nowe, nieznane wcześniej potrzeby informacyjne użytkowników cyberprzestrzeni.

Te dwa przykłady wskazują też na przesunięcie znaczenia z funkcji właśnie na dane, a to z kolei jest wynikiem przesunięcia z indywidualnego komputera stacjonarnego w stronę ogólnodostępnych usług *on-line*. Funkcje, jak na przykład komputerowe redagowanie tekstu charakterystyczne dla ery desktopu, mają dziś mniejsze znaczenie niż dostęp do danych, na przykład: baz danych wyszukiwarek czy aukcji internetowych, baz wiedzy, encyklopedii, baz z odwzorowaniami geograficznymi czy, jak w przypadku Flickra, zbiorów plików multimedialnych (w tym przypadku fotografii). Dopiero na podstawie czy raczej na bazie tych cennych danych kreowane są, częściowo przez samych użytkowników, a częściowo automatycznie, odpowiednie funkcje. Na przykład automatyczne systemy rekomendacyjne analizują dane transakcyjne dotyczące każdej sprzedaży, wykorzystują też uwagi użytkowników (np. oceny), ścieżki kliknięć, a następnie konstruują klasyfikacje i profile konsumentów na podstawie zidentyfikowanych preferencji. Rekomendacje będą następnie sugerowały zakup innego produktu.

4. Strategie a problematyka prawa własności

Na podstawie dostępnej literatury[5], jak również badań własnych można zidentyfikować pięć rodzajów najefektywniejszych strategii wykorzystania danych w cyberprzestrzeni, a mianowicie:

– strategie kreacji, takie jak posiadanie kosztownych, trudnych do zebrania zbiorów danych lub zbudowanie takiej bazy danych w oparciu o efekt sieciowy;
– strategie kontroli, które wykorzystują dedykowane formaty plików lub specjalne mechanizmy dostępu, na przykład archiwa lub katalogi;
– strategie struktury bazujące na klasach danych, co dostarcza kontekstu dla wielu innych usług, takich jak na przykład identyfikowanie miejsca, tożsamości, czasu itd.;
– strategie dostępu ułatwiające uzyskanie dostępu do trudno dostępnych danych;
– strategie infrastruktury zapewniające infrastrukturę dla pozyskiwania i magazynowania danych.

W tym kontekście pojawia się pojęcie prawa własności danych (ang. *data ownership*). D. Loshin uważa, że *prawo własności danych odnosi się zarówno do posiadania informacji, jak i do odpowiedzialności za tę informację. Sam fakt posiadania oznacza władzę i możliwość sprawowania kontroli. Kontrola nad informacją to nie tylko możliwość jej pozyskania, wytworzenia, modyfikacji, promowania, wykorzystania, odsprzedania i usunięcia, lecz także prawo do nadania tych przywilejów odnośnie informacji innym*[6]. Należałoby uzupełnić tę wypowiedź: to także możliwość kreowania nowych potrzeb informacyjnych, często nieznanych wcześniej użytkownikom cyberprzestrzeni.

Podsumowanie

Na koniec podjęto próbę sformułowania najważniejszych praktycznych zaleceń dotyczących optymalizacji wykorzystania unikalnych danych we współczesnej cyberprzestrzeni. Wskazania te są następujące:

– należy tworzyć zindywidualizowane formaty plików;

[5] Zob. m.in. Musser J., *Web 2.0: Principles and Best*...; A. Shuen, *Web 2.0: A Strategy Guide*, O'ReillyMedia, 2009.

[6] D. Loshin, *Master Data Management*, Morgan Kaufmann, 2008, s. 58.

- należy rozwijać specjalne sposoby uzyskiwania dostępu do danych, na przykład z archiwów;
- należy wykorzystywać dane do tworzenia nowych usług, na przykład dane katalogowe, indeksy produktów, identyfikatory cyfrowe itp.;
- należy rozwijać sposoby uzyskiwania dostępu do danych trudnych do zlokalizowania w cyberprzestrzeni;
- należy rozwijać infrastrukturę do gromadzenia danych i udostępnianie tej usługi innym użytkownikom;
- działania, operacje na danych powinny być zasadniczym wymogiem kompetencyjnym;
- należy upraszczać dostęp do danych;
- oprócz tradycyjnych stron internetowych należy wykorzystywać też inne sposoby udostępniania danych *on-line*;
- należy gromadzić informacje dotyczące zachowań użytkowników;
- należy maksymalizować tak zwaną *wirusową adopcję*;
- osoby trzecie powinny mieć możliwość pracy na istniejącym zbiorze danych;
- osoby trzecie powinny mieć możliwość wzbogacania istniejącego zbioru o własne dane;
- osoby trzecie powinny mieć możliwość agregowania danych;
- należy dywersyfikować kategorie danych dostarczane użytkownikom (osobom trzecim);
- należy udostępnić osobom trzecim mechanizmy do eksportu własnych danych;
- dane powinny być oferowane w różnych dodatkowych formatach dla zwiększenia zakresu oddziaływania (XML, RSS, mikroformaty itd.);
- należy identyfikować specyficzne, niszowe, wysokospecjalizowane rynki, które mogłyby być zainteresowane danymi rozpowszechnianymi przez syndykację;
- należy maksymalizować liczbę danych dostarczanych przez osoby trzecie;
- należy dążyć do jak największej granulacji oferowanych informacji i plików danych.

Literatura

Allen M., *Palm webOS*, O'Reilly Media, 2009.
Jackson P.R., *Web 2.0 Knowledge Technologies and the Enterprise: Smarter, Lighter, Cheaper*, Chandos Publishing, Oxford 2010.
Loshin D., *Master Data Management*, Morgan Kaufmann, 2008.
Musser J., *Web 2.0: Principles and Best Practices*, O'Reilly Media, 2007.
O'Reilly T., *What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software*, http://oreilly.com/lpt/a/6228, 11.03.2011.
Photobucket Leads Photo-sharing Sites, „ Hitwise", 21 czerwca 2006 r.
Shuen A., *Web 2.0: A Strategy Guide*, O'ReillyMedia, 2009.

WAYS OF EFFECTIVE DATA MANAGEMENT IN THE SECOND GENERATION OF WWW

Summary

This article presents an outline of a strategy of an effective management of data in cyberspace in the era of Web 2.0. The basic rules for the utilization of hard--to-get and unique data are presented and the architecture of participation, one of the leading catchphrases of Web 2.0 is discussed. This concept allows for the integration of cyberspace data. The practical applications introduced by Amazon.com and Flickr.com helped illustrate the pattern of "Data as the Next Intel Inside", another leading idea of the second incarnation of the WWW. Last but not least, the five identified strategies allowed to refer to the issue of data ownership in contemporary cyberspace.

Keywords: Web 2.0, data management strategy, architecture of participation, data integration, unique data.

Translated by Jacek Unold