EBIB Biuletyn EBIB, nr 5 (188)/2019,
Statystyki i analizy biblioteczne – nowe metody i narzędzia
Artykuły

Aleksander Radwański
Stowarzyszenie EBIB
aleksander.radwanski@ebib.pl

Statystyka dłubana

Streszczenie: Statystyka oparta na formularzach jest mało precyzyjna i zawiera niespójne dane. Poprawę sytuacji może zapewnić opracowanie standardów i protokołów pozwalających na bezpośrednie pobieranie danych z systemów bibliotecznych, przez oprogramowanie GUS.

Słowa kluczowe: GUS, statystyka biblioteczna

There are three kinds of lies: lies, damned lies, and statistics[1].

Czy nie jest ironią losu, że to popularne zdanie z motta o kłamstwie ma fałszywe autorstwo? Dotarcie do rzeczywistości jest żmudne i wymaga wielkiej rzetelności. Czy statystyki Głównego Urzędu Statystycznego opisują rzeczywistość, czy tylko podtrzymują mit o przydatności skwapliwie zbieranych danych?

Nieufność do statystyk ma w moim przypadku dwa źródła. W czasach słusznie minionych (młodszym czytelnikom wyjaśniam: do roku 1980) głównym celem statystyki było wykreślenie rosnącej krzywej. Wtedy narodziła się tradycja interpretowania statystyk tak, żeby rosło. A jeśli spada, to trzeba tak policzyć, żeby chociaż stało. Druga przyczyna nieufności to obserwacje, jak w praktyce wypełniane są formularze statystyczne. Nie mam wątpliwości, że wielu bibliotekarzy bardzo się stara wypełnić je rzetelnie, ale nie ma żadnych przeszkód, by wpisać tam zupełnie zmyślone wartości. Czasem po prostu wpisywane jest to, co się konkretnym bibliotekarzom wydaje, bo nie są w stanie uzyskać precyzyjnej informacji.

Główny Urząd Statystyczny w swoich metodach nawiązuje do średniowiecznej praktyki benedyktyńskiej, gdzie kilkunastotysięczna rzesza cierpliwych kopistów wypełnia formularze, które w założeniu mają zawierać skategoryzowane, jednolite dane, ale w praktyce nie zawierają. Tak jak kopistom zdarzało się, że przebiegający kot wywrócił im kałamarz lub ich artystyczne predylekcje skłaniały ich do bardziej kwiecistej ornamentyki, tak też współcześni bibliotekarze wypełniają rubryki z różnym poziomem zrozumienia i w różnych warunkach, dysponując różnym materiałem źródłowym. Jedyny ukłon GUS-u w stronę nowoczesności to przeniesienie papierowego formularza na formę elektroniczną. Dzięki i za to, ale nie jest to żadne unowocześnienie metodologii, która pozostaje niezmiennie średniowieczna.

[1] *Istnieją trzy rodzaje kłamstw: kłamstwa, wierutne kłamstwa i statystyka.* Zdanie z motta sformułował XIX-wieczny brytyjski premier Benjamin Disraeli, ale znane jest z przekazu Marka Twaina, któremu mylnie przypisuje się to powiedzenie.

Idea, że poprzez skategoryzowanie informacji na formularzu uzyskamy jednolite dane, ma kilka słabości. Pierwszą jest precyzja kategoryzacji. Biblioteki są bardzo zróżnicowanymi placówkami obsługującymi różne środowiska. Dlatego niektóre kategorie mają dość szeroki zakres interpretacyjny lub są wręcz nieadekwatne dla praktyki danej biblioteki. Druga to „czynnik ludzki", czyli wspomniana już wiedza, zdolności i motywacje osób wypełniających formularze. Zebrane w ten sposób dane już na samym początku mogą być obarczone wieloma błędami. Nie ma zatem znaczenia, jak bardzo zaawansowane narzędzia zostaną zastosowane do analizy tych danych, bo zasada GIGO (Garbage In, Garbage Out) działa nadal. Nie chcę przez to powiedzieć, że jestem w stanie wymyśleć lepszy formularz, bardziej precyzyjną kategoryzację czy lepszy sposób poinstruowania bibliotekarzy, jak wypełnić formularz poprawnie. Chcę powiedzieć, że formularze są passé i zawsze będą dawały błędne wyniki, szczególnie jeśli wypełnia je tak ogromna rzesza ludzi. Formularze są wygodne tylko dla GUS-u, który może nadal karmić się złudzeniem, że otrzymuje w ten sposób przybliżone odzwierciedlenie rzeczywistości.

Jest kwestią sporną, czy ta liczebność wypełniających sprzyja niwelacji potencjalnych błędów, czy je pogłębia. Zapewne są na ten temat badania przeprowadzone w Stanach Zjednoczonych w latach 60. lub jakaś teoria sprzed 30 lat, z którą nie zamierzam polemizować. Własna intuicja podpowiada mi, że raczej pogłębia. I nie dotyczy to jedynie takich danych, które są podatne na interpretacje. Popularne przekonanie, że liczby nie kłamią, wcale nie jest takie racjonalne, jak się wydaje. Weźmy jako przykład dwie najbardziej oczywiste liczby: liczbę egzemplarzy książek i liczbę wypożyczeń książek. Obie liczby można łatwo i precyzyjnie ustalić. Wydają się obiektywne. No bo albo ma się ileś egzemplarzy albo nie. Albo się je wypożyczyło, albo nie. Jak to się jednak odnosi do rzeczywistości? Wysmakowana biblioteka o doborowym księgozbiorze i biblioteka mająca 30% przestarzałej makulatury, 50% popularnego badziewia i 20% wartościowej literatury mogą mieć przybliżone wartości w obu kategoriach. Jeśli uzależnimy od tych wskaźników jakiekolwiek wnioski, np. którą z nich warto dofinansować i w jakiej kwocie albo jaka obsada personalna jest właściwa dla każdej z nich, to konsekwencje będą oczywiste. Makulatura i badziewie wygra. Na szczęście istnieją poza statystyką również inne kryteria oceny, ale urzędnicy domagający się „obiektywnych" wskaźników, mogą je brać pod uwagę lub nie.

Relacja pomiędzy liczbą egzemplarzy a liczbą wypożyczeń może być symptomem zupełnie różnych stanów rzeczy. Nie wykazuje na przykład, jaka część księgozbioru jest zwyczajnie martwa. Statystykę wypożyczeń może „robić" kilka popularnych kryminałów lub zekranizowanych romansów, których inne biblioteki nie gromadzą. Długość okresu wypożyczenia również może znacząco wpłynąć na końcową statystykę wypożyczeń. Jeśli popularną książkę można wypożyczyć na miesiąc, być może przeczyta ją nie tylko ten czytelnik, ale także ktoś z jego rodziny lub znajomych, którzy znajdą się poza statystyką. Jeśli jest to ważna dla danej dyscypliny pozycja, którą można wypożyczyć najwyżej na tydzień, a potem znów trzeba się do niej ustawić w kolejce, to liczba wypożyczeń znacząco się zwiększy… albo spadnie do zera, bo szybko ktoś ją sfotografuje i udostępni w formie cyfrowej za niewygórowaną opłatą. To oczywiście nielegalne, ale żadna biblioteka nie jest w stanie temu zapobiec.

Nie postuluję, by na formularzu sprawozdania biblioteki dodać 100 nowych pól, które pozwolą na głębszą interpretację opisanych przykładowo danych statystycznych. Zwracam tylko uwagę na fakt, że liczby nie kłamią jedynie co do swojej wartości. Ich znaczenie jest

natomiast zupełnie inną sprawą, która z obiektywizmem nie ma nic wspólnego. GUS na podstawie tych liczb niczego konkretnego się nie dowiaduje, a wyciąganie wniosków opartych o tę niewiedzę jest tylko większą niewiedzą.

Jeszcze gorzej ma się rzecz z liczbami podlegającymi interpretacji. Najlepszym przykładem jest liczba odwiedzin[2]. W prostym przypadku sprawa jest oczywista. Przychodzi czytelnik, wypożycza coś lub nie, ale nas odwiedza. A jak przyjdzie dwa razy w ciągu dnia? Niektórzy przyjmują żelazną zasadę, że w jednym dniu jeden czytelnik jest liczony tylko raz. Ale właściwie dlaczego? Jak przyjdzie rano wypożyczyć książkę, a po południu poczytać czasopismo, to czemu nie liczyć dwóch odwiedzin? A jeśli zmieniła się w tym czasie obsada bibliotekarska, to skąd będzie wiedziała, że już był? Nie wszyscy mają skomputeryzowane systemy, które wykrywają takie powtórne odwiedziny. A jak najpierw wpadnie poczytać, to musi się wylegitymować kartą biblioteczną, żeby potem wyeliminować ewentualne powtórzenie? A jeśli nie ma karty?

Kiedy do akcji wkraczają „odwiedziny zdalne”[3], robi się jeszcze zabawniej. Czy rezerwacja dokonana przez OPAC to odwiedziny? A ustawienie się w kolejce? Niby nie, ale dlaczego nie? Kiedyś po to przychodziło wielu czytelników. Niektórzy wypisywali sobie coś z katalogów, inni rezerwowali książki… teraz robią to samo przez OPAC-a. A więc znów – ile razy w ciągu dnia? A jak czytelnik siedzi w nocy i jedną pozycję zarezerwuje przed północą, a drugą po północy, to odwiedził nas dwa razy? Jak widać, precyzyjne ustalenie liczby odwiedzin nie jest banalne.

Cóż zatem robi się w praktyce? Dodaje się po prostu różne liczby, które kojarzymy z odwiedzinami i jeśli wyjdzie za mało, to szukamy kolejnych liczb, które mogłyby poprawić wynik. A skąd wiemy, że jest „za mało”? Po prostu porównujemy to ze sprawozdaniem zeszłorocznym i dążymy albo do minimalnego wzrostu, albo tylko nieznacznego spadku tej liczby. Jeśli komuś się wydaje, że to niepoważne, to będę bronił tej praktyki, bo niewielu decyduje się na podawanie liczb zupełnie z sufitu. Bibliotekarze w większości starają się zmieścić w interpretacji, jaką wskazuje im formularz, a jeśli jakaś interpretacja jest dopuszczalna, to nie ma powodu jej negować.

Moglibyśmy tak punkt po punkcie omawiać popularny formularz GUS K-03, ale nie chodzi mi tu ani o jego ocenę, ani o udoskonalanie. Mam silne przekonanie, że cała ta monumentalna praca rzeszy bibliotekarzy nie służy niczemu. To byłby wbrew pozorom wniosek dość optymistyczny. Produkcja niczego to też są miejsca pracy i jakiś pożytek z tego w ogólnym rozrachunku jest. Bardziej niepokoi mnie fakt, że te dane są następnie traktowane jak wyrocznia i podstawa podejmowania różnych decyzji. Co prawda w ostatnim czasie poziom wolicjonalności decydentów osiągnął stan “boski, ale bez cudów”, więc żadne opracowania GUS-u nie mają tu znaczenia. Jednak pomniejsi urzędnicy nadal lubią mieć jakieś uzasadnienie, popularnie nazywane od chronienia miękkich, acz wstydliwych części ciała, więc chętnie sięgają do „obiektywnych wskaźników”, a te zdaje się dostarczać GUS. Moim zdaniem, z akcentem na „zdaje się”.

Po tej krótkiej charakterystyce słabości formularzowej statystyki pora byłaby na zadanie pytania, czy mamy jakąś alternatywę. Na razie nie, ale dość szybko możemy ją stworzyć.

[2] W formularzu GUS należy podać liczbę wejść do biblioteki, a nie osób [przyp. red.].
[3] W formularzu GUS nie wymaga się podawania odwiedzin wirtualnych [przyp. red.].

To jednak wymagałoby porzucenia starodawnej sztuki wypełniania formularzy na rzecz zrozumienia, co w tej sprawie może zrobić technologia. Bo może dużo więcej, niż wynika to z formularza, na którym pojawiły się rubryki związane z komputeryzacją i cyfrowymi zasobami. Pod warunkiem, że po pierwsze dostrzeżemy te możliwości, a po drugie zmienimy przepisy tak, by te możliwości wykorzystać. Osobiście to ostatnie uważam za najtrudniejsze.

Większość dużych i średnich bibliotek posiada komputerowe systemy biblioteczne. I coraz więcej małych bibliotek również. Myślę, że śmiało można by obronić tezę, że zebranie statystyk tylko z bibliotek skomputeryzowanych byłoby zupełnie wystarczające dla aproksymacji statystyk wszystkich bibliotek. Ewentualnie mógłby powstać narodowy plan skomputeryzowania każdej biblioteki, bo tego wymaga właśnie zdolność do generowania statystyk. Nie mam systemu – nic nie wypełniam – nie dostaję dotacji. I szybko znalazłyby się środki na komputeryzację, tam, gdzie jeszcze tego nie zrobiono. Niemożliwe? Kiedyś wprowadzono jednym pociągnięciem obowiązek elektronicznego składania zeznań podatkowych i posiadania kas fiskalnych dla przedsiębiorców, których było dużo więcej niż wszystkich bibliotek razem wziętych. Można? Można.

Posiadanie systemu to warunek konieczny, ale niewystarczający. System musi jeszcze generować odpowiednie statystyki i teoretycznie mógłby produkować zestaw gotowy do przeniesienia wprost na formularz K-03, gdyby nie wspomniane wcześniej problemy interpretacyjne. Powodują one, że to, co pracowicie obliczają systemy, jest potem przepuszczane przez uświęconą tradycją, jedynie słuszną interpretację każdej biblioteki. Jedne biblioteki coś dodają, inne coś ujmują, pomijają lub uwzględniają. W ten sposób, zamiast korzystać z dobrodziejstw komputeryzacji, biblioteki dokładają kolejne ogniwo do interpretacyjnego wysiłku intelektualnego tysięcy bibliotekarzy. Kiedyś interpretowali to, co wpisywali na formularze. Teraz interpretują dodatkowo jeszcze to, co uzyskują z systemu, zanim zinterpretują to, co należy wpisać na formularz. Czy ktoś ma jeszcze nadzieję, że te statystyki coś mówią o rzeczywistości?

GUS mógłby dostawać jednolite dane statystyczne, pod warunkiem, że w tej szacownej instytucji znalazłoby się grono na tyle kompetentne informatycznie, by zrozumieć, że z systemu można wyjąć tylko to, co się tam włoży. Jeśli zatem GUS interesuje liczba odwiedzin, to musi opracować algorytm, jak te odwiedziny mają być liczone. Mam jednak wrażenie, że same kategorie powinny być zrewidowane i unowocześnione, bo co to są odwiedziny w sytuacji rozwiniętych usług sieciowych, albo podział na pracujących i niepracujących w sytuacji, kiedy mamy cały wachlarz form zatrudnienia, nieciągłych w czasie i przestrzeni. Konieczność generowania tego typu statystyk czyni z bibliotek instytucje para-inwigilacyjne, bo aby posegregować te informacje, należy je najpierw zebrać. No więc bibliotekarze przepytują cierpliwie czytelników prawie jakby pracowali w urzędzie emigracyjnym. Ale co ma powiedzieć student pracujący na ⅖ etatu przez 4 miesiące każdego roku? Jest uczący się czy pracujący? A może powinien aktualizować swój status na bieżąco jak na Facebooku?

Należałoby zatem zacząć od zorientowania się, jakie informacje systemy zbierają i na tym oprzeć ewentualne statystyki zamiast wymyślać rubryki, których zawartość można obliczać na wiele sposobów. Każdy system biblioteczny radzi sobie z ewidencją zasobów, a więc jest w stanie wykazać dokładnie, ile egzemplarzy biblioteka posiada, a nawet ile wśród nich to powtórzenia tego samego tytułu, ile jest egzemplarzy poszczególnych rodzajów

materiałów oraz jak wygląda struktura księgozbioru pod kątem ich przynależności do określonego typu literatury. Aby te dane były spójne, statystyka nie może się opierać na wymyślonych rodzajach dokumentów, tylko posiłkować się rodzajami, na które wskazuje odpowiednie pole formatu MARC. To samo z typami literatury. Muszą to być ściśle określone kategorie wraz ze wskazaniem, jak traktować przypadki graniczne – fabularyzowana biografia filmowa, która doczekała się formy książki, to beletrystyka, literatura faktu czy adaptacja? I jeśli system nie odnotowuje, czy dane czasopismo zostało oprawione, to nie można takiej statystyki wymagać albo należy wskazać, gdzie i jak w systemie taka informacja ma zostać zapisana. Systemy mogą też dostarczać innych statystyk, dziś nieosiągalnych, np. który z autorów lub wydawców jest najliczniej reprezentowany w zbiorach bibliotek. Taka statystyka wymagałaby jednak, by GUS zrezygnował z mentalności poborcy podatkowego na rzecz postawy bardziej aktywnej – samodzielnego sięgania wprost do systemów bibliotecznych po potrzebne dane. Dane bibliograficzne i inwentarzowe mogą być szeroko udostępnione, bo poprzez OPAC i tak są prezentowane. GUS musiałby tylko określić, jaki zestaw danych w zakresie ewidencji go interesuje.

Systemy biblioteczne radzą sobie również z ewidencją wypożyczeń. Tutaj oczywiście otwarcie systemu musiałoby mieć ograniczenia związane z RODO, ale wciąż można by pozyskiwać statystyki dużo bardziej interesujące i trafniej opisujące rzeczywistość niż da się to osiągnąć, przepisując dane z formularzy. System może nam odpowiedzieć na pytanie o dowolny aspekt dynamiki wypożyczeń: ile z posiadanych egzemplarzy jest w ruchu, jakiego rodzaju są to materiały, jakiego typu literatura, kto ją wypożycza (w sensie płci, lokalizacji, wieku), w jakie dni tygodnia lub w jakie miesiące najwięcej, na jak długo itd. itp. To wszystko może nam wygenerować system bez podawania danych osobowych – wystarczy rok urodzenia, płeć, kod pocztowy. Jestem zdecydowanie sceptyczny co do innych aspektów, bo to zmuszałoby biblioteki do przepytywania czytelnika na różne okoliczności, których nie musi chcieć nam ujawniać. Jednak podanie kodu pocztowego, roku urodzenia i ustalenie płci raczej nie powinno sprawiać trudności i jest chyba zestawem najmniej kontrowersyjnym, a dającym duże możliwości, jeśli skorelujemy to z wypożyczeniami.

Ewidencja zbiorów i wypożyczeń to tylko dwa najbardziej oczywiste przykłady obrazujące inne podejście do statystyk. Początkiem musi być ustalenie, jakie informacje systemy zbierają, jakie jeszcze powinny zbierać (np. jak liczyć odwiedziny inne niż związane z wypożyczeniem lub zwrotem książki), jak je przetwarzać, by uzyskać żądaną statystykę (dokładny algorytm), jakie informacje powinny zostać udostępnione w formie surowych danych, by GUS mógł je zaciągnąć, zagregować i przetworzyć samodzielnie. Potrzebne są zatem specyfikacje i protokoły zamiast rubryk w formularzu. Od strony systemów bibliotecznych nie widzę bariery – każdy z producentów samodzielnie zmaga się ze statystyką biblioteczną i z ulgą powita normalizację w tym zakresie.

Docelowo statystyka powinna zniknąć jako zagadnienie i ulec całkowitej automatyzacji. Jej rzetelność przestanie być pochodną wiedzy i chęci bibliotekarzy, dzięki czemu uzyskany obraz ma szansę lepiej odzwierciedlać rzeczywistość. Działające nocami boty GUS-u wyciągałyby samodzielnie z systemów bibliotecznych odpowiednio przygotowane dane. Bibliotekom pozostałby jedynie obowiązek podania numeru IP i niewyłączania na noc serwerów bibliotecznych. W systemach chmurowych ten obowiązek przejąłby dostawca systemu i bibliotekarz nawet o to nie musiałby się martwić. Opracowanie specyfikacji

i protokołów, stworzenie odpowiedniego interfejsu po stronie systemu bibliotecznego oraz stworzenie oprogramowania dla GUS-u, które korzystałoby z tego interfejsu, można przeprowadzić relatywnie szybko. Jest wiedza na ten temat, są specjaliści, istnieją technologie. Wystarczy tę wiedzę zebrać, pozyskać specjalistów i wdrożyć technologie. Szacowałbym wykonalność takiego projektu na okres 3 lat, pod warunkiem jego sensownego (realistycznego) finansowania i nadania mu odpowiedniego priorytetu. Oszczędności czasu są ogromne, ale bardzo rozproszone. Poprawienie jakości statystyk – bezsporne. Kto zatem powinien być tym najbardziej zainteresowany?

RADWAŃSKI, A. Statystyka dłubana. *Biuletyn EBIB* [online]. 2019, nr 5 (188), Statystyki i analizy biblioteczne – nowe metody i narzędzia. [Dostęp 6.11.2019]. ISSN 1507-7187. Dostępny w: http://open.ebib.pl/ojs/index.php/ebib/article/view/730.