e-Infrastruktura: preludium cyfrowej humanistyki

MICHAŁ STARCZEWSKI

Badania potrzebują odpowiedniego zaplecza. To stwierdzenie aż razi swoją oczywistością. My, badacze z obszaru nauk humanistycznych, korzystamy z bibliotek, archiwów i muzeów, tak jak naukowcy z obszarów nauk ścisłych, przyrodniczych i medycznych korzystają z laboratoriów. Planując kolejne projekty badawcze, zakładamy, że to zaplecze istnieje. Formułujemy wobec niego oczekiwania. Im lepiej jest ono wyposażone, tym badania mogą być bardziej zaawansowane, a studia wnikliwsze. Bogato zaopatrzone biblioteki umożliwiają śledzenie dyskusji naukowych, a zbiory specjalne dostarczają źródeł. Archiwa gromadzą dokumenty. O wartości archiwów i bibliotek

świadczą jednak nie tylko zbiory, ale i jakość ich opisania. Katalogi biblioteczne, pomoce archiwalne, systemy opisywania obiektów to zbyt często niedoceniane, wypracowane przez stulecia narzędzia pozwalające sprawnie poruszać się badaczom w ogromie materiału. Cyfrowa humanistyka wymaga zaplecza odpowiadającego narzędziom wykorzystywanym w projektach z tego obszaru. Tak jak te projekty są zróżnicowane (cyfrowa humanistyka jest pojęciem szerokim i niesprecyzowanym[1]), tak i zaplecze, którego potrzebują, nie może być jednolite.

[1] Zob. M.K. Gold (red.), **Debates in the Digital Humanities**, http://dhdebates.gc.cuny.edu (15.02.2015).

e-Infrastruktura

Projekty cyfrowe potrzebują twardego sprzętu, choćby w postaci serwerów, za pomocą których przetwarza się dane oraz utrzymuje i udostępnia efekty prac badawczych. Tę infrastrukturę należy odróżnić od e-infrastruktury. Afiks „e-" w pojęciu „e-infrastruktura" ma na celu odróżnienie narzędzi, o których będzie mowa, od sprzętu, z którym częściej kojarzy się pojęcie infrastruktury. Pojęcie e-infrastruktury stosuje się do wielu dyscyplin naukowych, a cyfrowa humanistyka jest tylko jednym z wielu przypadków.

„e-Infrastruktura" jest dość często stosowanym terminem, choć unika się podawania jego precyzyjnej definicji. Kiedy już autorzy próbują dookreślić, co mają na myśli, trudno zarzucić im nadmierną szczegółowość; na przykład „e-infrastruktura – lub cyberinfrastruktura w USA – jest nowym typem infrastruktury składającej się z rozproszonych zasobów teleinformatycznych wspierających zespołową pracę badawczą"[2]. Również programy Unii Europejskiej, finansujące e-infrastrukturę badawczą, nie są w tym zakresie precyzyjne, a autorzy ich dokumentów chętnie posługują się przykładowymi wyliczeniami: „e-infrastruktury, takie jak systemy danych, systemy obliczeniowe i sieci komunikacyjne"[3].

Ta niedookreśloność nie jest przypadkowa. Do e-infrastruktury wspierającej naukę zalicza się serwisy i narzędzia adresowane do wielu różnych grup odbiorców, mające niejednorodne cele i funkcje. Znalezienie wspólnego mianownika byłoby sztuką karkołomną i raczej niepotrzebną. Nie ma konieczności dążenia do większej precyzji niż potrzebna do odróżnienia tego typu narzędzi w praktyce.

Czy wszystkie zasoby teleinformatyczne wspierające naukę to e-infrastruktura? Cytowani wcześniej autorzy artykułu *The Emerging Governance of E-Infrastructure* stawiają poprzeczkę wyżej. Interesuje ich przemiana efemerycznych projektów w stabilną e-infrastrukturę, czyli w zaplecze badawcze mające zapewnione finansowanie i utrwaloną, sformalizowaną strukturę. Po zbadaniu 16 e-infrastruktur/projektów wspierających badania naukowe z różnych obszarów wiedzy badacze ci doszli do wniosku, że tylko część z nich uzyskała tę stabilność. Zauważyli jednak, że potrzeba zbudowania e-infrastruktury w ich rozumieniu uwidacznia się choćby w skonstruowaniu w państwach Europejskiego Obszaru Badawczego formuły ERIC – European Research Infrastructure Consortium[4]. Konsorcjum DARIAH-EU, mające na celu e-infrastrukturalne wsparcie cyfrowej humanistyki, przybrało w listopadzie 2014 roku właśnie tę formę prawną, idąc w ślad za CLARIN. Choć stabilność jest bardzo ważną i pożądaną cechą zaplecza badawczego, to – inaczej niż wspomniani autorzy – ważniejsza wydaje mi się funkcja pełniona przez dany serwis.

Typy e-infrastruktury

Ze względu na funkcję e-infrastrukturę można podzielić na: 1) wspierającą bieżącą komunikację naukową, 2) gromadzącą i udostępniającą dane badawcze (*data repositories*) i programy komputerowe powstałe w ramach projektów badawczych (*software repositories*), 3) zapewniającą dostęp do cyfrowych źródeł.

Często się podkreśla, że efekty projektów z zakresu cyfrowej humanistyki nieraz przyjmują postać inną niż tradycyjne publikacje. Mogą to być bazy danych lub wizualizacje. Nie należy jednak lekceważyć artykułów, monografii i materiałów konferencyjnych. Stanowią one podstawę komunikacji naukowej we wszystkich obszarach wiedzy, nawet tych, w których intensywnie wykorzystuje się postaci alternatywne. Publikacje powinny mieć, rzecz jasna, formę elektroniczną, by wykorzystać potencjał nowoczesnych narzędzi do dystrybucji treści naukowych. Coraz częś-

[2] „E-infrastructure – or cyberinfrastructure in the US – is a new type of infrastructure consisting of distributed ICT-based resources which support collaborative research", F. Barjak, K. Eccles, E.T. Meyer, S. Robinson, R. Schroeder, **The Emerging Governance of E-Infrastructure**, „Journal of Computer-Mediated Communication" 18/2003, s. 113.

[3] „e-infrastructures, such as data and computing systems and communication networks", **Horizon 2020. Work Programme 2014–2015. 4. European research infrastructures (including e-Infrastructures). Revised**, s. 4, http://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/main/h2020-wp1415-infrastructures_en.pdf#25 (10.02.2015); zob. też np. http://ec.europa.eu/programmes/horizon2020/en/h2020-section/e-infrastructures (10.02.2015).

[4] Zob. http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=eric (10.02.2015).

ciej wydawcy sięgają po *enhanced publications* (nie stosuje się polskiego tłumaczenia), czyli publikacje, w których oprócz warstwy narracyjnej, opisującej badania i ich wyniki, dodawane są dodatkowe materiały, na przykład dane badawcze, nagrania audio i wideo, symulacje 3D[5]. Pozwalają przedstawić wyniki badań w szerszym kontekście, niż pozwala sam tekst.

e-Infrastruktura wspierająca bieżącą komunikację naukową to serwisy udostępniające czasopisma i książki. Narzędziem, które nie jest tylko cyfrowym odwzorowaniem praktyk analogowych w dystrybucji wiedzy, są repozytoria. Repozytoria naukowe to serwisy służące do deponowania prac przez samych autorów (a nie przez biblioteki lub wydawnictwa). Ważnym europejskim projektem e-infrastrukturalnym w tym zakresie jest OpenAIRE 2020. Jest to kolejna już edycja projektu polegającego na budowie sieci europejskich otwartych repozytoriów, ich agregatora oraz wypracowania wspólnych standardów wymiany metadanych[6].

Stan polskiej e-infrastruktury wspierającej bieżącą komunikację naukową został opisany w raporcie *Otwarta nauka w Polsce 2014. Diagnoza*[7]. Wciąż słabo rozwinięta jest sieć otwartych repozytoriów. Należy zauważyć, że od momentu opublikowania raportu uruchomione zostało repozytorium nauk historycznych Lectorium[8]. Jest to repozytorium dziedzinowe, które prawdopodobnie w przyszłości będzie ważnym kanałem dystrybucji wiedzy w zakresie polskiej humanistyki, także cyfrowej. Bardzo prężnie rozwija się baza czasopism humanistycznych i społecznych CEJSH[9], udostępniająca w wersji pełnotekstowej już ponad 170 czasopism. Cyfrowe kolekcje książek z otwartym dostępem ograniczają się de facto do serwisu Otwórz Książkę[10].

Udostępnianie danych badawczych jest zagadnieniem, którego znaczenie jest coraz bardziej dostrzegane. Jim Gray mówi o „czwartym paradygmacie": badania oparte na dużych danych to czwarta epoka po badaniach empirycznych, teoretycznych i komputerowych[11]. Duże dane w przypadku astronomii lub fizyki cząstek elementarnych to zbiory mierzone w terabajtach danych, dostarczanych nieustannie przez teleskopy czy zderzacz hadronów w CERN. Humaniści cyfrowi nie posługują się tak ogromnymi zbiorami danych. Jednak ich dane też wymagają gromadzenia i udostępniania. Jednym z projektów poświęconych temu zagadnieniu jest LinkedHumanities, projekt będący częścią Linked Data[12]. Dane badawcze udostępniane są na świecie poprzez otwarte repozytoria danych. W Polsce obecnie nie ma takich prawie w ogóle[13].

Do tej kategorii e-infrastruktury można zaliczyć również CLARIN (Common Language Resources and Technology Infrastructure)[14]. Celem CLARIN jest zbudowanie repozytoriów danych i dostarczanie łatwych w obsłudze narzędzi językowych. Ponieważ badanie większości źródeł odbywa się poprzez język, narzędzia takie są bardzo potrzebne. Do tej kategorii będzie prawdopodobnie można zaliczyć też efekty projektu DARIAH[15].

Dostęp do cyfrowych źródeł

Projekty digitalizacyjne rozwijają się bardzo dynamicznie na całym świecie, również w Polsce. Obejmują one manuskrypty, druki, obrazy, obiekty muzealne, fotografie, nagrania audio i filmy. W Polsce digitalizację finansuje Ministerstwo Kultury i Dziedzictwa Narodowego. Program Kultura+ jest koordynowany przez Narodowy Instytut Audiowizu-

[5] Zob. **What is an Enhanced Publication?**, OpenAIRE, https://www.openaire.eu/en/component/content/article/76-highlights/344-a-short-introduction-to-enhanced-publications (10.02.2015) oraz A. Bardi, P. Manghi, **Enhanced Publications: Data Models and Information Systems**, „Liber Quarterly. The Journal of the Association of European Research Libraries" 4/2014, s. 240–273.

[6] https://www.openaire.eu/ (10.02.2015).

[7] J. Szprot (red.), **Otwarta nauka w Polsce 2014. Diagnoza, Warszawa 2014**, rozdział III: **e-Infrastruktura otwartego dostępu w Polsce**, s. 30–53, http://pon.edu.pl/index.php/nasze-publikacje?pubid=13 (12.02.2015).

[8] http://lectorium.edu.pl/pl/ (15.02.2015).

[9] http://cejsh.icm.edu.pl (15.02.2015).

[10] http://otworzksiazke.pl/ (15.02.2015).

[11] **Jim Gray on eScience: Transformed Scientific Method**, [w:] T. Hey, S. Tansley, K. Tolle (red.), **The Fourth Paradigm. Data-Intensive Scientific Discovery**, Redmond 2009, s. XVII–XIX, http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf (10.02.2015).

[12] J. Huber i in., **LODE: Linking Digital Humanities Content to the Web of Data**, http://arxiv.org/pdf/1406.0216v1.pdf (10.02.2015).

[13] Funkcjonują przynajmniej dwa repozytoria danych pochodzących z badań społecznych.

[14] http://clarin.eu/guest-portal (14.02.2015).

[15] http://dariah.eu/ (17.02.201 5).

alny[16]. Konsekwencją takiej organizacji (skądinąd słusznej – alternatywne rozwiązania wydają się mniej uzasadnione) jest koncentrowanie się na aspekcie zabezpieczenia dziedzictwa, a nie na dostosowaniu cyfrowych postaci obiektów do badań.

Polskie biblioteki cyfrowe współpracują ze sobą. Agregator ich treści, Federacja Bibliotek Cyfrowych[17], pozwala użytkownikom przeszukiwać równocześnie zasoby ich wszystkich, a bibliotekarzom koordynować dalsze prace i unikać skanowania ponownie tych samych obiektów. Dane te są agregowane również na poziomie europejskim. Europeana to portal agregujący informacje o zdigitalizowanych obiektach dziedzictwa kultury z państw europejskich[18]. Zarówno FBC, jak i Europeana są świetnymi przykładami, jak powinno działać wymienianie się metadanymi.

Otwartość e-infrastruktury

Komunikacja naukowa podlega dynamicznym przemianom. Cyfrowe technologie sprawiają, że koszty związane są prawie wyłącznie z wytworzeniem pierwszego egzemplarza, a każda kopia wytwarzana jest w zasadzie bezkosztowo. Jednocześnie mechanizmy finansowania badań naukowych oraz bodźce motywujące do pracy badawczej sprawiają, że instytucje naukowe mogą oczekiwać udostępniania efektów finansowanych przez siebie badań bez dodatkowych opłat ze strony odbiorcy. Efektywne udostępnianie prac sprzyja rozwojowi karier akademickich. Dystrybucja prac naukowych bez barier finansowych, prawnych i technicznych nazywana jest otwartym dostępem[19].

Otwartość (rozumiana m.in. jako stosowanie otwartych modeli komunikacji naukowej) to wartość, z którą identyfikuje się wielu humanistów cyfrowych. Pojawia się ona w wypowiedziach przedstawicieli tej dyscypliny, choćby na marginesie rozważań na inny temat. Bezpośrednio tę kwestię podjęła Lisa Spiro: „wspólnota humanistów cyfrowych uznaje otwartość za swoją, zarówno ze względu na własny interes, jak i z powodów etycznych”[20]. Uderzanie w ton etyczny może wydawać się nieco górnolotne. Świadczy jednak o silnym zintegrowaniu otwartych modeli komunikacji naukowej z cyfrową humanistyką[21].

e-Infrastruktura wspierająca cyfrową humanistykę powinna umożliwiać jak najszersze stosowanie modeli otwartych. Korzyści z otwartości odnoszą sami badacze. Otwartość zwiększa szanse zarówno na cytowalność (temu zagadnieniu poświęcona jest obszerna literatura[22]), jak i na dobrą indeksację w wyszukiwarkach internetowych, zwłaszcza z popularnej w środowisku akademickim Google Scholar. Zagadnienie indeksacji przez wyszukiwarki jest jednak dość złożone[23]. Odpowiednich działań wymaga nie tylko od serwisów udostępniających treści, ale także od wydawców i autorów.

Otwartość przynosi przede wszystkim duże korzyści całej społeczności naukowców, tak jest również w przypadku cyfrowej humanistyki. Obieg treści staje się szybszy i skuteczniejszy. Łatwiej można weryfikować hipotezy, sprawdzać, czy i kto wykonał już podobną pracę. Rzadziej wyważa się otwarte drzwi.

Otwartość dotyczy nie tylko publikacji, ale również danych i oprogramowania. Otwartość poszerzona o dodatkowe uprawnienia przyznane odbiorcom prowadzi do tak zwanych wolnych treści. W tym przypadku użytkownicy są uprawnieni do ponownego wykorzystania treści. Jest to możliwe dzięki zastosowaniu wolnych licencji, najczęściej są to licencje Creative Commons CC BY lub CC BY-SA[24]. Zastosowanie wolnych licencji jest równoznaczne z tak zwanym otwartym dostępem *libre*.

[16] http://www.nina.gov.pl/instytut/programy/artyku%C5%82/2011/06/26/wieloletni-program-rzadowy-kultura- (14.02.2015).
[17] http://fbc.pionier.net.pl (14.02.2015).
[18] http://europeana.eu/ (14.02.2015).
[19] P. Suber, **Otwarty dostęp**, Warszawa 2014, http://pon.edu.pl/index.php/nasze-publikacje?pubid=14 (12.02.2015).
[20] „The digital humanities community embraces openness because of both self-interest and ethical aspirations”, L. Spiro, **“This Is Why We Fight”: Defining the Values of the Digital Humanities**, [w:] M.K. Gold (red.), **Debates in the Digital…**, op. cit.
[21] Otwartości w humanistyce w ogóle, w tym jej ekonomicznym aspektom, poświęcona jest książka Martina Paula Eve'a, **Open Access in the Humanities**, Cambridge 2014, http://ebooks.cambridge.org/ebook.jsf?bid=CBO9781316161012 (17.02.2015).
[22] Bibliografia literatury dotyczącej wpływu otwartości na cytowalność: http://opcit.eprints.org/oacitation-biblio.html (12.02.2015).
[23] Zob. na przykład T. Lewandowski, **Google Scholar a repozytoria i biblioteki cyfrowe w Polsce**, http://otwartanauka.pl/analysis/case-studies?id=945 (12.02.2015).
[24] K. Siewicz, **Otwarty dostęp do publikacji naukowych. Kwestie prawne**, Warszawa 2012, http://pon.edu.pl/index.php/nasze-publikacje?pubid=12 (12.02.2015).

Szeroki zakres uprawnień przyznawanych użytkownikom jest ważny ze względu na innowacyjne metody badawcze, wykraczające poza tradycyjne użytkowanie. Możliwość maszynowej analizy tekstów (*data mining*) w ramach dozwolonego użytku jest wciąż kwestią dyskusyjną. Prawdopodobnie będą pojawiały się kolejne metody badań. W tej chwili należy dołożyć starań, by materiały dostępne za pomocą e-infrastruktury wspierającej badania humanistyczne, ich metadane i same narzędzia nie stawiały przeszkód prawnych blokujących kolejne badania.

Interoperacyjność

Zagadnieniem częściowo pokrewnym otwartości jest interoperacyjność. Poważnym ryzykiem jest powstanie wielu serwisów wspierających badania, które nie będą ze sobą komplementarne. Paul Ell i Lorna Hughes wskazują na kilka problemów utrudniających wykorzystanie zasobów cyfrowych w humanistyce, między innymi na rozdrobnienie cyfrowych kolekcji, brak bogatych metadanych i digitalizację tylko części kolekcji zamiast całości. Wskutek tego to, co wydaje się humanistom „informacyjnym przeciążeniem", w rzeczywistości okazuje się brakiem spójnego podejścia do pracy z większymi korpusami cyfrowych zasobów w sposób zintegrowany. To oznacza, że tak bardzo oczekiwana „przemiana humanistyki" przebiega powoli[25]. Humaniści, którzy zostali wyszkoleni do pracy z materiałami drukowanymi, często korzystają ze źródeł cyfrowych tak, jakby to były tylko repliki dokumentów analogowych. Takie podejście uniemożliwia stawianie nowych pytań badawczych. Cyfrowa humanistyka posługuje się z reguły metodami ilościowymi, co wyraźnie odróżnia ją od obecnie najczęściej stosowanych metod w humanistyce „analogowej". Przygotowanie narzędzi cyfrowych, pozwalających na zbieranie i analizowanie danych zgodnie z najwyższymi standardami i szczegółowymi metadanymi, jest warunkiem koniecznym dla pojawienia się znaczących rezultatów, także takich, których nie jesteśmy w stanie w tej chwili przewidzieć. To usprawni dotychczasowe (tzn. analogowe) procesy badawcze, ale, co znacznie istotniejsze, umożliwi badaczom stawianie zupełnie nowych pytań i odpowiadanie na nie[26].

Interoperacyjność jest dużym wyzwaniem, ponieważ w grę wchodzą bardzo zróżnicowane zasoby. Zdigitalizowane manuskrypty, czasopisma, książki drukowane, zdjęcia, filmy, nagrania audio wymagają różnych rozwiązań technicznych i różnych metadanych. Zapewnienie możliwości równoczesnej pracy z wszystkimi zasobami jest trudne. Dodanie do tego choćby bieżącej komunikacji naukowej jeszcze bardziej komplikuje problem[27].

Najpierw inwestycja, potem zyski

Cyfrowość w badaniach humanistycznych niesie ze sobą obietnicę daleko idących zmian i przekształcenia całego obszaru wiedzy. Choć projekty badawcze zakrojone na niewielką skalę mogą być i są realizowane cały czas, to nie należy oczekiwać spektakularnych efektów, dopóki nie zostanie w pełni przygotowane zaplecze. e-Infrastruktura jest potrzebna cyfrowej humanistyce, tak jak biblioteki i archiwa konieczne są do badań tradycyjnych.

Należy być przygotowanym na poniesienie nakładów na potrzebne narzędzia i trzeba uzbroić się w cierpliwość. e-Infrastruktura musi być budowana z myślą o jej różnorodnym wy-korzystaniu. Dlatego istotnym zagadnieniem jest jej interoperacyjność, czyli możliwość współpracy i wymiana danych między różnymi narzędziami. Dopóki e-infrastruktura będzie zbudowana tylko częściowo, ocena dorobku cyfrowej humanistyki będzie przedwczesna. Krytycy będą mogli odnosić się zaledwie do zalążka tego, czym badania cyfrowe w humanistyce mogą się stać w przyszłości, a entuzjaści będą mogli mówić raczej o swoich wyobrażeniach niż o realnych, sprawdzonych możliwościach. |

[25] „These can result in what feels to humanists to be an 'information overload', when in fact the issue is a lack of a cohesive approach to working with a larger body of digital content in an integrated fashion. This has meant that the much-anticipated 'transformation of the humanities' is slow to emerge", P. Ell, L. Hughes, **E-infrastructure in the Humanities**, „International Journal of Humanities and Arts Computing" 1–2(7)/2013, s. 31.

[26] „It enables existing (i.e. analogue) research processes to be conducted better and/or faster, but most significantly of all, it enables researchers to ask, and answer, completely new research questions", ibidem, s. 32.

[27] Zagadnieniu interoperacyjności tylko między repozytoriami naukowymi poświęcony został dokument Confederation for Open Access Repositories (COAR) **COAR Roadmap. Future Directions for Repository Interoperability**, https://www.coar-repositories.org/files/Roadmap_final_formatted_20150203.pdf (13.02.2015).