

Odpowiednio zabezpieczajmy dla przyszłych pokoleń także te obiekty, które istnieją wyłącznie w postaci cyfrowej, zauważmy pracę archiwów społecznych i oddolne środowiska fanów gier i telewizji.

Kiedy zniknie internet

MARCIN WILKOWSKI

Mówi się, że to AI będzie podstawą radykalnej przemiany współczesnego świata. Prawdziwa zmiana nastąpi jednak z innego powodu: przez utratę cyfrowego dziedzictwa czeka nas radykalne przeformułowanie tego, w jaki sposób spoglądamy w przeszłość. Kryzys w AI nadejdzie przy okazji, kiedy z powodu braku zasobów archiwalnych modele zaczną się uczyć na treściach już przez siebie tworzonych.

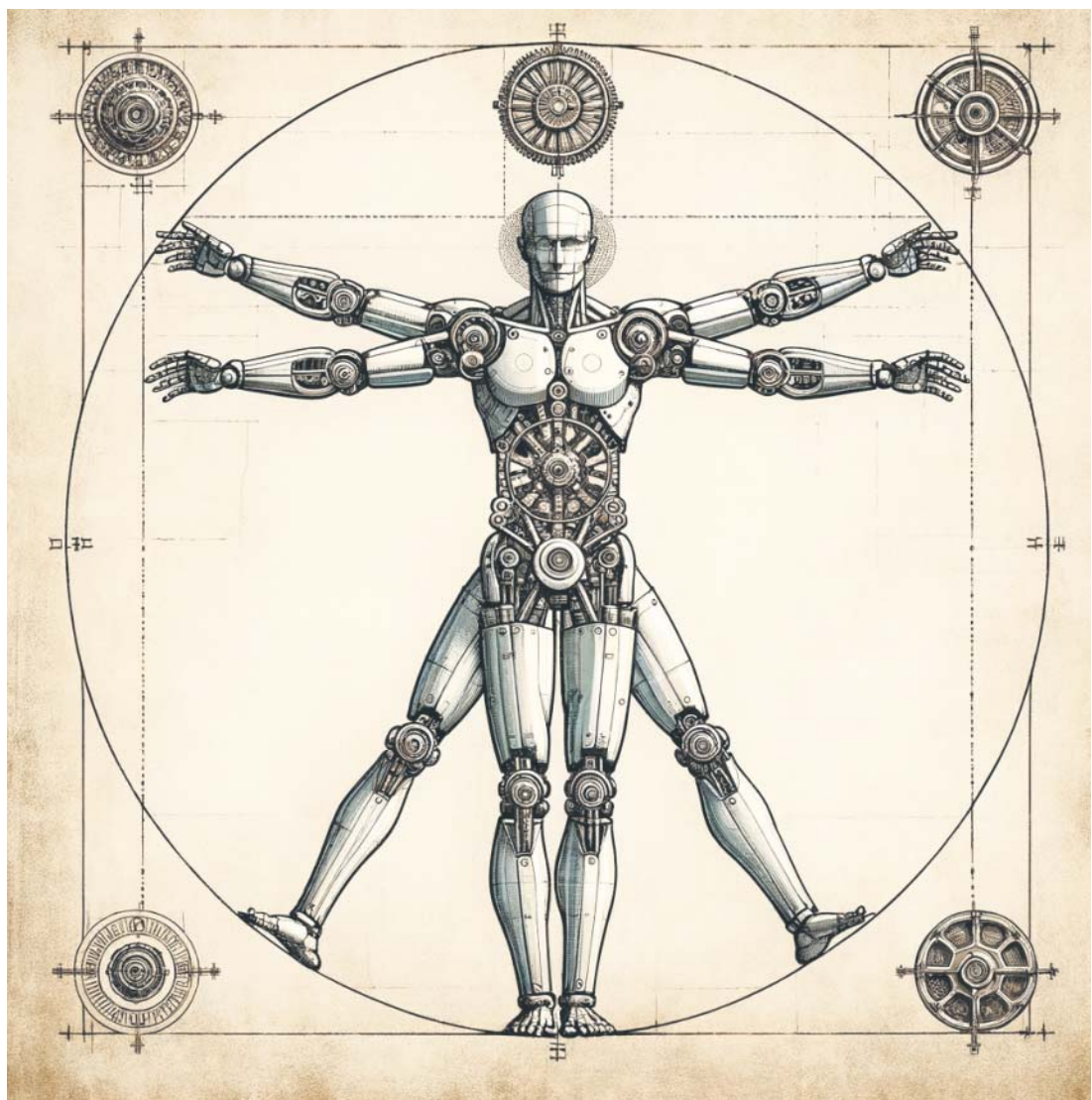
Archiwalny outsourcing

Wszystkie zachowane do dziś zasoby polskiej domeny krajowej PL z lat 1996–2001, dokumentujące przełom, jaki miał miejsce w związku z pojawieniem się World Wide Web w naszym kraju, mieszczą się w kieszeni płaszcza. To nieco ponad 200 GB, które bez problemu da się zapisać na małym dysku zewnętrznym. Wcześniejsze dane są bezpowrotnie stracone, ponieważ fundacja Internet Archive prowadzi swoje archiwum Wayback Machine dopiero od połowy lat 90. Podobnie jak 30 lat temu, wciąż wyręcza w tym polskie państwo, chociaż archiwizacja fragmentów krajowego Webu to dziś zadanie realizowane przez wiele bibliotek i archiwów narodowych na całym świecie.

Webarchiv archiwizuje czeskie zasoby internetowe od 2001 r. – Obecnie przechowujemy ponad 600 TB danych – mówi Marie Haškovcová, kierowniczka działu archiwizacji Webu w Bibliotece Narodowej Republiki Czeskiej. Kierowana przez nią jednostka stara się zabezpieczać zasoby czeskiego internetu jako części narodowego niematerialnego dziedzictwa kulturowego. Haškovcová dodaje, że na szczeblu rządowym trwają obecnie w Czechach prace nad legislacją, która ma mocniej wpisać proces archiwizacji internetu w działalność Biblioteki Narodowej. Domenę krajową archiwizują też choćby Niemcy, Portugalczycy czy Duńczycy, co pokazuje, że można to z powodzeniem robić w systemie prawnym UE.

Państwa mogą archiwizować swój internet m.in. dzięki przepisom rozszerzającym prawo egzemplarza obowiązkowego czy uprawnieniom dla archiwów, które zabezpieczają treści dokumentujące funkcjonowanie instytucji publicznych. Być może to ostatnie rozwiązanie zostanie przyjęte w Polsce, ponieważ Naczelna Dyrekcja Archiwów Państwowych pracuje nad systemem archiwizacji domeny rządowej gov.pl. W latach 2009–2011 Narodowe Archiwum Cyfrowe eksperymentowało już z zabezpieczaniem tych zasobów. Wobec bierności Biblioteki Narodowej, która podjęła tylko jedną próbę archiwizacyjną podczas wyborów do Parlamentu Europejskiego w 2009 r., zasoby polskiego internetu są archiwizowane wyłącznie przez zewnętrzne podmioty, takie jak fundacja Internet Archive. Internetowe ślady ważnych wydarzeń społecznych, takich jak katastrofa w Smoleńsku, pandemia, czarne marsze czy wojna w Ukrainie i pomoc uchodźcom, zabezpieczane są szczerkowo i chaotycznie – nie wiemy nawet, jakie strony zostały już utracone.

Tymczasem tempo utraty zasobów Webu jest porażające. Z najnowszych badań Pew Research wynika, że w 2023 r. przestało istnieć prawie 40 proc. stron internetowych, dostępnych jeszcze dziesięć lat temu. Ten pro-



Zamiast martwić się „sztuczną inteligencją niszczącą kulturę”, lepiej zastanowić się, co brak dostępu do źródeł cyfrowych i zdigitalizowanego dziedzictwa znaczy dla nas, ludzi

blem dotyczy też treści na platformach społecznościowych, które nie są przecież żadnym archiwum – 20 proc. tweetów znika w ciągu kilku miesięcy z powodu umyślnego usunięcia lub zawieszenia konta użytkownika.

Ten brak to nie tylko wyzwanie dla komunikacji naukowej (linki często umieszcza się w przypisach), pracy dziennikarskiej, białego wywiadu czy działalności strażniczej organizacji pozarządowych. To przede wszystkim realna utrata cyfrowych źródeł historycznych, które już dziś mogłyby być wykorzystywane przez historyków, oraz zagrożenie dla pamięci osobistej i rodzinnej, bo ta w dużym stopniu opiera się dziś na cyfrowych opisach, fotografiach i filmach. Internet, chociaż wydaje się to niemożliwe, jest historyczny i w żadnym wypadku nie jest tak, że „nigdy nie zapomina”.

Udręka masowej digitalizacji

Na razie zarządcy modeli AI poradzą sobie bez trudu z brakiem historycznych zasobów Webu i dziurami w głębokich silosach mediów społecznościowych – internet na bieżąco wykararmi je niemal nieskończonymi nowymi treściami. Jednak jeśli treści te w coraz większym stopniu pochodzą z maszynowych wytwórni, takich jak ChatGPT czy DALL-E, modele zostaną zainfekowane własną wytwórczością i jakość ich działania gwałtownie się obniży. Dziś generuje się już co prawda syntetyczne dane treningowe (choćby recenzje filmowe o różnym nacechowaniu emocjonalnym), którymi następnie uczy się modele, ale ich zróżnicowanie, a co za tym idzie, i potencjał są ograniczone. Dlatego tak ważne są oryginalne, wiarygodne i dobrej jakości zbiory kultury, dziedzictwa i nauki, które udostępniają muzea, biblioteki i archiwa oraz które tworzą ludzie.

Być może dla niedofinansowanych i medialnie marginalizowanych instytucji polskiej kultury i nauki to jedyna pozytywna wiadomość wobec medialnego za-

chwytu nad „sztuczną inteligencją”, z której mogą skorzystać one jedynie w ograniczony sposób ze względu na brak środków i kompetencji. Nie wymagamy zresztą innowacji od osób, których wynagrodzenie obniża się powoli do poziomu płacy minimalnej. Pocieszeniem dla wyrzuconych już na starcie z wyścigu o nowoczesność może być jedynie to, że wytwarzane i gromadzone przez nich zasoby mogą się stać wymiennie cenne – nawet jeśli nie w sensie rynkowym, to przynajmniej godnościowym. Chociaż wartość pracy naukowej czy kolekcji muzealnej w żaden sposób nie opiera się na jej popularności, można będzie czuć się docenionym, że jakaś maszyna w końcu przeczyta nasz artykuł albo odwiedzi niewielki portal ze zdigitalizowanymi zbiorami, żeby przetworzyć go na zestaw miar w przeliczeniach statystycznych. Być może w epoce clickbajtu i śmieciowej wyszukiwarki, jaką stał się Google, to jedyna uwaga, na jaką możemy liczyć.

Władcy AI dysponują dziś niemal nieograniczonymi funduszami inwestorów, próbujących odnaleźć nową nadzieję w rynku cyfrowym po kompromitacji, jaką była gwałtowna fascynacja blockchainem i NFT. Nie wiadomo jednak, czy biorą pod uwagę stopniowe wyczerpywanie się dobrej jakości danych, których można legalnie używać do trenowania modeli. Dziś nie wystarczy już treści z Wikipedii czy Common Crawl. Jak informują badacze z grupy naukowej Data Provenance Initiative, w ciągu ostatniego roku około jednej czwartej danych pochodzących z wysokiej jakości źródeł przestało być dostępne do trenowania modeli z powodu restrykcji prawnych i zabezpieczeń technologicznych. Jeśli AI ma być projektem na dekady, a nie technologiczną modą, jej modele muszą być nieustannie aktualizowane i poprawiane. Tymczasem już dziś OpenAI pozywane jest przez wydawców za bezprawne wykorzystanie treści.

Wygląda więc na to, że muszą one sięgać głębiej do zbiorów kultury i dziedzictwa. Te są przebogate, jednak tylko niewielka ich część dostępna jest w postaci cyfrowej. Podczas niedawnego Zjazdu Historyków Polskich w Białymstoku Piotr Zawilski z Narodowego Archiwum Cyfrowego przekonywał, że do tej pory zdigitalizowano

1,5 proc. zbiorów krajowych archiwów, a wynik ten jest i tak lepszy choćby w porównaniu z archiwami francuskimi. Według niego „przy obecnym tempie prac digitalizacyjnych odwzorowanie całości zasobów archiwów państwowych według stanu na rok 2024 zajęłoby ok. 600 lat”. Muzeum Narodowe w Warszawie zdigitalizowało jak dotąd nieco ponad połowę swojej kolekcji. Tyle mówimy o potrzebie tworzenia polskich, lokalnych modeli, a okazuje się, że może brakować danych, na których można je budować.

Zatrzymało się tempo digitalizacji zbiorów materialnych, o digitalizacji prawdziwie masowej nie może być już mowy ze względu na skalę i koszty. Oznacza to też, że AI, która dokonuje rewolucji w kulturze, jest jedynie projekcją zapatrzonych w technologie fantazmów. Modele generowania grafiki i tekstu wciąż zdolne będą jedynie do przeslizgiwania się po bogactwie ludzkiej kultury i wytwarzania generycznych słów – sztucznie tworzonego spamu, którego nikt nie będzie chciał oglądać i czytać. Uczenie maszynowe da się użyć w usprawnianiu wyszukiwania w katalogach czy nawet rekonstrukcji zniszczonych obiektów, ale w żaden sposób nie podważa ono sensu istnienia muzeów, archiwów i bibliotek oraz materialnej postaci zbiorów.

Gry komputerowe i doświadczenie z epoki

Zamiast martwić się „sztuczną inteligencją niszczącą kulturę”, lepiej zastanowić się, co brak dostępu do źródeł cyfrowych i zdigitalizowanego dziedzictwa znaczy dla nas, ludzi. – Choć gry wideo są bardzo istotną częścią kulturowego dziedzictwa ludzkości, nie mamy wypracowanych standardów opieki nad ich historią – mówi Bartłomiej Kluska, historyk gier wideo, autor książki „Bajty polskie”. – Brak choćby regulacji w zakresie przechowywania dawnych gier dla przyszłych pokoleń (na wzór bibliotecznego egzemplarza obowiązkowego), nie zorganizowano też jednego, ogólnie nadzorowanego i prowadzonego zgodnie z dobrą archiwalną praktyką inwentarza tytułów – dodaje.

Kluska przekonuje, że tylko oddolna aktywność fanów gier pozwala ratować tę część historii polskiej kultury. Od lat zajmują się oni wykonywaniem kopii z nietrwałych nośników, takich jak kasety magnetofonowe czy dyskiety. – Osiągnięciem oddolnych grup fanowskich jest także praktycznie kompletna digitalizacja dawnych czasopism komputerowych i growych, będących podstawowym źródłem historycznym dla tej tematyki – konkluduje. Polskie instytucje potrafią jednak dostrzec wartość w dziedzictwie cyfrowym tego typu, czego dowodem jest decyzja Ministerstwa Kultury z grudnia 2021 r., zgodnie z którą polska demoscena (społeczność użytkowników komputerów tworzących tzw. dema) została wpisana na krajową listę niematerialnego dziedzictwa kulturowego UNESCO. Działania Społecznego Komitetu Kroniki Polskiej Demosceny, którego celem jest zabezpieczanie historii i dzieł polskich twórców komputerowych przełomu lat 80. i 90., wsparł m.in. Narodowy Instytut Dziedzictwa. W ślad za tym nie idzie jednak legislacja, która pozwoliłaby bibliotekom i archiwom gromadzić i legalnie udostępniać stare gry i oprogramowanie.

W przypadku archiwizacji gier i produkcji demosceny problemem jest nie tylko dostęp do nośników z artefaktami cyfrowymi, ale też do oryginalnego sprzętu, na którym były one tworzone i odtwarzane. Francuska Biblioteka Narodowa (BnF) przechowuje ponad 20 tys. tytułów gier komputerowych, które można uruchamiać i badać nie tylko na emulatorach, specjalnym oprogramowaniu symulującym stare platformy, ale też na historycznych platformach. W Polsce podobną kolekcję rozwija działający na Uniwersytecie Jagiellońskim od 2016 r. UBU lab. Jego założyciel, prof. Piotr Marecki, podkreśla, że laboratorium działa w obszarze nauk o sztuce. – Interesuje nas, jak sprzęt komputerowy był wykorzystywany do działań kreatywnych. W laboratorium udało się zgromadzić ważne historycznie platformy 8-, 16- i 32-bitowe, takie jak ZX Spectrum, Atari, Amstrad, Amigi, monitory, telewizory i peryferia z epoki, ale także sprzęt współczesny. Nasz główny cel to możliwość uruchomienia utworów na platformach natywnych, czyli takich, na jaki dany utwór kreatywny został napisany – mówi.

Jakie znaczenie ma dostęp do sprzętu z epoki? – Jeśli założymy, że utwory programowane, jak gry czy dema, są liczone w czasie rzeczywistym, wówczas można porównać to do sztuk performatywnych, więc proces dzieje się w czasie rzeczywistym, jak teatr czy performance. W tym sensie emulator czy prezentacja utworu

jako filmiku na YouTube jest tylko rodzajem dokumentacji dzieła. Interesuje nas też to, co istotne w doświadczeniu, czyli np. odgłos uruchamianego programu, kolory, jakie pojawiają się na monitorach z epoki, gest wkładania dyskietki do stacji dysków czy użycie dżojstika – wyjaśnia dalej. Retro sprzęt wykorzystuje się tu nie tylko do dydaktyki i badań naukowych, ale też jako narzędzia współczesnej twórczości.

YouTube jako archiwum

Decydowanie gorzej jest z ochroną i dostępem do polskiego dziedzictwa audiowizualnego. Mateusz Witkowski, współtwórca popularnego podcastu Podcastex, poświęconego latom 90. i 2000., na co dzień pracuje z publikowanymi na YouTube nagraniami archiwalnych programów telewizyjnych. Ceny za dostęp do zbiorów TVP są zaporowe, a stacje komercyjne w ogóle nie chcą się dzielić swoimi archiwami. – Postępująca cyfryzacja i rozwój prywatnych mediów wyrobiły w nas przeciętne przekonanie, że w telewizji, podobnie jak w internecie (co zresztą również nie jest prawdą), nic nie ginie – stwierdza Witkowski. – Tymczasem w przypadku mnóstwa produkcji telewizyjnych opieramy się wyłącznie na wspomnieniach i często dość przypadkowych „zgrzykach” z kaset VHS, udostępnianych nie tylko na YouTube przez domorosłych archiwistów – dodaje.

Witkowski obawia się ponadto, że brak dostępu do archiwów telewizyjnych może świadczyć o tym, że określonych zbiorów – szczególnie w telewizjach prywatnych – po prostu już nie ma. Wobec takiej polityki wydawców telewizyjnych YouTube pełni rolę nieoficjalnego archiwum, w którym nie tylko brakuje standardów i metadanych, ale przede wszystkim pewności, że dostępne dziś nagranie da się obejrzeć również jutro. TVP, powołując się na prawa autorskie, latem tego roku doprowadziła do usunięcia treści kanału Dawno Temu w TV, na którym umieszczono ponad 5 tys. fragmentów archiwalnych audycji telewizyjnych.

YouTube i inne media społecznościowe to także przestrzeń aktywności wielu nieformalnych i sformalizowanych polskich archiwów społecznych. Bada je dr Magdalena Wiśniewska-Drewniak, która zwraca uwagę, że często są one mieszkanką materiałów fizycznych i cyfrowych, oryginalnych i kopii i w dowolnych kombinacjach. – Wiele archiwów społecznych decyduje się na gromadzenie obiektów w duchu post-custodial – a więc wchodząc w posiadanie kopii materiałów fizycznych oraz prawa ich używania, ale ich oryginały pozostawiając w rękach darczyńców – wyjaśnia Wiśniewska-Drewniak.

Zbiory te przechowywane są na dyskach i udostępniane w internecie, niekiedy wyłącznie na platformach społecznościowych. Zdaniem badaczki to sposób na „wyjęcie” źródeł historycznych z odmetów szaf, strychów i szuflad – danie im drugiego, „publicznego” życia. Żadna instytucja nie archiwizuje jednak zbiorów publikowanych na Facebooku, chociaż powstałe w 2020 r. Centrum Archiwistyki Społecznej wspiera te archiwa w odpowiednim opisywaniu i przechowywaniu oddolnie gromadzonych materiałów, planuje też zbudowanie narzędzi do zabezpieczania ich także w postaci cyfrowej.

Potencjał takiej oddolnej archiwizacji widać było szczególnie w czasie pandemii, kiedy nawet muzea i biblioteki, wobec braku możliwości działania we własnych przestrzeniach, podejmowały prace dokumentujące rzeczywistość stanu wyjątkowego. Szybkie gromadzenie zbiorów (Rapid Response Collecting), bo tak określa się w literaturze tego typu inicjatywy, pokazało, jak bardzo aktualną instytucją są muzea, biblioteki i archiwa – nie tylko przechowują i udostępniają historyczną wiedzę i zasoby, ale są też w stanie reagować na kryzysy i aktywnie szukać nowych zbiorów w przestrzeni cyfrowej. Granicą takiego gromadzenia jak zawsze są środki i kompetencje, jednak specyfika współczesnego internetu będącego oligopolem kilku dominujących firm także nie ułatwia tej pracy. Jeśli założyć, że społeczne reakcje Polaków na wojnę w Ukrainie i samoorganizacja pomocy uchodźcom są w znacznej mierze udokumentowane w mediach społecznościowych, np. na Twitterze, to dostęp do tych źródeł jest dziś skrajnie utrudniony. Dość powiedzieć, że dane to kapitał, a koszty miesięcznej subskrypcji pozwalającej wyeksportować z Twittera 1 mln wpisów to 5 tys. dol. Przy czym szacuje się, że na platformie, którą dziś nazywamy musimy X, publikuje się dziennie około 500 mln tweetów. Żadna uczelnia ani instytucja nie uniesie tego typu kosztów, jeśli miałaby zgodnie ze sztuką zabezpieczać takie zbiory.

Perspektywa straty

Ponieważ już dawno załamał się podział na świat online i offline, kultura i dziedzictwo ma dziś postać hybrydową. Z jakiegoś jednak powodu wymiar cyfrowy uznawany jest za mniej wartościowy lub przynajmniej na tyle bezpieczny, że nie warto martwić się jego wieczystym zabezpieczeniem. Niestety, to skrajnie niebezpieczne podejście – już dziś straciliśmy wiele potencjalnych źródeł historycznych dokumentujących pandemię i wojnę w Ukrainie, a zrozumienie współczesnych podziałów w Polsce bez dostępu badaczy do historycznych zasobów mediów społecznościowych nie będzie nigdy pełne. Co więcej, jak pokazują działania UBU labu, cyfrowe dziedzictwo może być podstawą współczesnej twórczości. Jeśli nie zostanie zachowane, nie będzie można się do niego odnieść w przyszłości w projektowaniu, grafice, literaturze czy filmie.

„Zanikająca kultura: Raport o naszym kruchym dziedzictwie kulturowym” to opublikowane niedawno przez fundację Internet Archive opracowanie zwracające uwagę na realny kryzys dziedzictwa cyfrowego. Przejawia się on nie tylko w tym, jak kruche i podatne na zniszczenie są obiekty niemające pierwotnej postaci materialnej, ale też w samych regulacjach prawnych tego, na czym w ogóle polega posiadanie takich obiektów. Podcastu z platformy streamingowej czy filmiku z YouTube’a nie możemy mieć na własność, ponieważ jedynie korzystamy z tymczasowych licencji, jakie dają nam wydawcy. To trudne wyzwanie dla bibliotek i archiwów. Jak czytamy w raporcie, zamiast być zabezpieczone poprzez trwałe nabywanie, obiekty cyfrowe kultury, takie jak nagrania dźwiękowe, książki, programy telewizyjne i filmy, są stale narażone na usunięcie z platform bez jakiegokolwiek archiwizacji.

Jak wobec takiego zagrożenia myśleć o naszych relacjach z przeszłością, doświadczaną i opisywaną w hybrydowy, materialno-cyfrowy sposób? To nie jest abstrakcyjny problem i nie dotyczy on wyłącznie instytucji i prawodawców. Spróbujmy wyobrazić sobie, co za 50 lat stanie się z naszymi fotografiami, które gromadzimy „w chmurze”, i z naszymi wpisami w mediach społecznościowych.

Dziś w tle rozmów o kulturze i dziedzictwie umieszcza się „sztuczną inteligencję” jako wyzwanie najważniejsze i najbardziej niebezpieczne. Wydaje się, że wbrew modnym prorokom AI możemy spokojnie czekać na załamanie się rynkowej mody wywołane brakiem jakościowych źródeł do trenowania dużych komercyjnych modeli. Kultura, sztuka, literatura i historia nie mogą być zagrożone wytworami modeli generatywnych, chociaż te mogą zabierać pracę niektórym twórcom. Daleki jestem przy tym od postawy ludystycznej – kulturę i dziedzictwo można przetworzyć podmiotowo i z poszanowaniem jej lokalności. Niezwykle pozytywną inicjatywą w tym względzie jest projekt budowy dużego polskiego modelu językowego PLLuM, którego twórcy (instytucje publiczne!) dbają o partnerskie relacje z krajowymi wydawcami i instytucjami kultury i nauki. Innym projektem o podobnych wartościach są WikiDane.

Jasne jest przy tym, że nie będzie tu żadnej rewolucji. Już wkrótce AI będzie się trenowała na własnych wytworach. Zamiast bajdurzyć o kolejnych rewolucjach technologicznych, wypracujmy odpowiednie finansowanie sprawdzonych już publicznych instytucji i sprawiedliwe wynagrodzenia dla ich pracowników. Odpowiednio zabezpieczajmy dla przyszłych pokoleń także te obiekty, które istnieją wyłącznie w postaci cyfrowej, zauważmy pracę archiwów społecznych i oddolne środowiska fanów gier i telewizji. Niech pamięć, historia i dziedzictwo będą dla nas dostępne przez silne i otwarte instytucje. Jeśli o to nie zadamy, będziemy korzystać z kultury zarchiwizowanej w koślawych i generycznych wytworach ChatGPT, który na prośbę o tekst „Campo di Fiori” Miłosza jest w stanie bez halucynowania wydukać jedynie dwie linijki, podczas gdy jego producent, firma OpenAI, wyceniany jest już na 157 mld dol. W tym samym czasie istniejący od 2007 r. portal Wolne Lektury, który mógłby opublikować ten wiersz w swojej dostępnej dla wszystkich za darmo bibliotece, wciąż nie doczekał się stabilnego państwowego finansowania. /©©

Marcin Wilkowski – programista w Centrum Kompetencji Cyfrowych UW, w lutym 2025 r. obronił doktorat na Wydziale Historii UW. Naukowo zajmuje się archiwistyką Webu i badaniami wczesnej polskiej WWW. Rozwija projekt edukacyjny humanistyka.dev, którego celem jest upowszechnianie wykorzystania metod i narzędzi cyfrowych w badaniach humanistycznych i pracy ze zdigitalizowanymi zbiorami kultury i dziedzictwa