Fragment jedenastej tabliczki Eposu o Gilgameszu. Based on derivative work: Frédéric GilgameshTablet.jpg: Babylonian [Public domai], via Wikimedia Commons

Dlaczego ważne są technologie cyfrowe, skanowanie i digitalizacja dokumentów i książek i innych obiektów? Jakie jest uzasadnienie ogromnego wysiłku przekształcania spuścizny kulturowej w postać cyfrową? Często słyszę takie pytania - od historyków, którzy preferują zapach i dotyk oryginalnych dokumentów lub archiwistów, którzy twierdzą, że mikrofilmy są wystarczająco dobre. Czy cyfryzacja to tylko moda, która wkrótce przejdzie, czy też ma to głębsze uzasadnienie?

“Cyfrowe” jest ważne - dla archiwów, bibliotek, muzeów (GLAM) oraz dla wszystkich producentów i konsumentów dóbr kultury. Omówimy tu trzy powody przechodzenia do cyfrowego przetwarzania informacji:
Zabezpieczanie
,
Znajdywalność
(discoverablity) i
Dostęp
.

Zabezpieczanie

Układ cyfrowy jest tylko jedną z wielu implementacji dyskretnych systemów przechowywania i obróbki informacji. Większość sygnałów, które docierają do naszych zmysłów, np. widok tęczy,

symfonia lub zapach róży, można uznać za analogowe. Sygnał analogowy może przyjąć dowolną wartość, na przykład dźwięku lub koloru. Zakres jest zazwyczaj ograniczony jedynie możliwościami naszych zmysłów - nie widzimy podczerwieni, ani słyszymy ultradźwięków itp. Ale sygnał optyczny, po tym jak wpadnie do naszego oka lub kamery cyfrowej, nie jest dalej przetwarzany jako sygnał ciągły. Czujniki światła w siatkówce (czopki i pręciki) działają na zasadzie “wszystko albo nic”, podobnie dzieje się w kamerze gdzie każdy element czujnika rozkłada światło na ograniczoną liczbę poziomów. Sygnał zostaje zmieniony w informację - wkraczamy tu w sferę dyskretności. W układzie dyskretnym tylko ograniczona, przeliczalna liczba stanów jest dozwolona, nie ma nic pomiędzy. W nowoczesnych komputerach cyfrowych podstawową jednostką informacyjną jest bit, który może posiąść tylko dwa stany (zwyczajowo zwane 0 i 1). Matematyczna teoria informacji, po raz pierwszy zaproponowana przez Claude E. Shannona, również używa jako jednostki binarnego bitu, z implikacją, że informacja w naturze swojej jest dyskretna. W komputerach, pojedyncze bity są zazwyczaj ułożone w grupy: 8 bitów w określonej kolejności nazywa się bajtem. W celu utrzymania ogólnego charakteru dyskusji, najmniejsza jednostkę systemu dyskretnego będziemy dalej nazywać znakiem, a ciąg znaków słowem.

W dalszym ciągu przyjrzymy się kilku systemom dyskretnym i na ich przykładzie tym ich cechom, które są są ważne w zabezpieczeniu i zachowaniu zasobów: bezstratnemu kopiowaniu, czytelności maszynowej i korekcji błędów.

Dyskretne systemy informacyjne

Pierwsze komputery zostały zbudowane w czasie II wojny światowej, dziś informatyka obchodzi swoje 75 - lecie. Nie jest to jednak pierwszy przykład dyskretnego schematu przekazywania informacji. Ludzie opracowali wiele takich systemów, od sygnałów dymnych (znak dwu-stanowy, dym albo brak dymu) do kodu Morse'a (znak sześcio-stanowy). Największym ludzkim wynalazkiem dyskretnego przechowywania i przekazywania informacji jest jednak alfabet. Powszechnie uznaje się, że najwcześniejszy alfabet został wynaleziony przez Sumerów ok. 5200 lat temu. Sam język jest oczywiście o wiele starszy, między 2 miliony a 200,000 lat.

Liczba różnych znaków zależy od języka, od około 26 w alfabecie łacińskim do tysięcy w chińskim. Na przykładzie alfabetu łacińskiego spróbujmy oszacować liczbę stanów, jakie może przyjąć jeden znak. Jest 26 małych liter, 26 dużych, 10 cyfr arabskich, różne symbole, takie jak $, &, §, odstęp, znaki interpunkcyjne, itp. W sumie, znak podstawowego alfabetu łacińskiego może przyjąć jeden z około 120 do 150 stanów. Długość słowa jest zmienna, ale w języku

potocznym rzadko przekracza 20 znaków. Nie ma górnej granicy długości słowa - nowe konwencje nazewnictwa technicznego, np. w chemii, pozwalają na budowanie dowolnie długich ciągów znaków, tyle, ile potrzeba aby utworzyć nazwę.

Bezstratne kopiowanie

Portret Jean Miélot, (zmarł in 1472), autora, tłumacza, ilustratora and skryby. Photograph by Mike Peel (www.mikepeel.net). [CC-BY-SA-4.0 (http://creativecommons.org/licenses/by-sa/4.0)], via Wikimedia Commons

Wynalezienie alfabetu, dyskretnej reprezentacji języka, umożliwiło przechowywanie, kopiowanie i przekazywanie informacji na skalę uprzednio nie do pomyślenia przy komunikacji ustnej. Możliwość bezstratnego kopiowania jest jedną z najważniejszych konsekwencji tego wynalazku. Sygnały analogowe szybko ulegają degradacji przy kopiowaniu, tak jak w grze w głuchy telefon. Rozprzestrzenianie plotek we wsi, lub wielokrotne kopiowanie z taśmy magnetycznej wykazują ten sam efekt. Dzięki ograniczonej liczbie stanów, ciągi znaków mogą być kopiowane dokładnie (za wyjątkiem błędów ludzkich).

Nośnik, na którym teksty zostały zapisane może czasami przetrwać setki lub tysiące lat. Fragmenty glinianych tabliczek sumeryjskich i papirusów egipskich dotrwały do dziś. Na ogół jednak opieranie się na trwałości nośnika nie prowadzi do trwałości informacji. Ogień może zniszczyć bibliotekę lub archiwum, co widać na przykładzie Biblioteki Aleksandryjskiej dwa tysiące lat temu czy archiwów Imperium Osmańskiego w Sarajewie 7 lutego 2014 roku. “Epos o Gilgameszu”, jedno z najwcześniejszych zachowanych dzieł literatury, można przeczytać dziś tylko dlatego, że został skopiowany wiele razy. Praktyka kopiowania tekstu przez mnichów, a później przez zawodowych skrybów, rozkwitła w 13 i 14 wieku i przyczyniła się do przetrwania starożytnych tekstów. Kopiowanie było dokładne, z systemem recenzentów i kontroli jakości, ale w praktyce pojawiały sie nieuniknione błędy. Liczba omyłek jest jednak o wiele rzędów

wielkości mniejsza niż w kopiowaniu analogowym.

Czytelność maszynowa

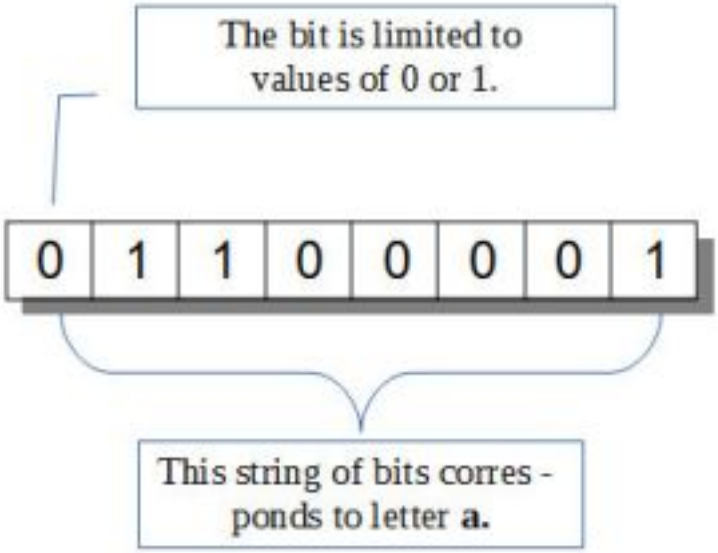

Diagram ilustruje znak binarny, bit i ciąg bitów - bajt. Przy zastosowaniu konkretnego kodowania (ASCII), ciąg ten odpowiada literze a.

Innym aspektem zapisanych tekstów jest zdolność odczytania informacji przez maszynę, w tym wypadku komputer (machine readability). Pisanie jest w swej istocie dyskretne, polegając na zapisywaniu liter tworzących dalej słowa, zdania, akapity itp. Jednak do niedawna teksty były przechowywane jako znaki na nośniku (takim jak kamień, glina, papirus lub papier) - czytelne dla ludzi, którzy znają język, lecz nie dla komputerów. Istnieją techniki, takie jak OCR (Optical Character Recognition), które mogą, choć nadal z mieszanym powodzeniem, dokonać automatycznej konwersji tekstu drukowanego w postać czytelną dla maszyny. W ogólności proces ten wymaga jeszcze wiele pracy ludzkiej. Po przekształceniu w formę czytelną dla maszyny, komputery mogą robić te wszystkie “magiczne” operacje które robią z informacją w ogólności, tworzyć indeksy, kategoryzować, tłumaczyć, przekształcać w inne formy i wiele innych.

Inne dzieła kultury ludzkiej, zawierające na przykład obrazy lub dźwięki, mają charakter analogowy i nie dają się tak łatwo zamienić w czytelne maszynowo. Jest to częściowo możliwe w przypadku dźwięku, poprzez zakodowanie każdej składowej złożonego dźwięku. Metadane , które zawierają informacje o obiekcie również posiadają zwykle czytelność maszynową.

Korekcja błędów

Alfabet jest starym wynalazkiem, ale nie jest to najstarszy znany człowiekowi dyskretny system informacji. Aby zlokalizować powstanie najstarszego, musimy cofnąć się o około 3,5 miliardów lat wstecz. Żeby umieścić tę liczbę w perspektywie, wiek Ziemi szacuje się na 4,5 miliardów lat a wiek Wszechświata w którym żyjemy na 13,8 miliardów lat. Najstarszy system informacji ma znak o czterech stanach, słowo o dokładnie 3 znakach i pozostaje bez zmian około 3,500,000,000 lat dzięki bezstratnemu kopiowaniu i korekcji błędów. Systemem tym jest kod genetyczny. Cztery stany maja postać 4 różnych związków chemicznych, skrótowo oznaczanych literami A, C, G, T. Słowo, ciąg 3 takich znaków koduje jeden z około 20 cząsteczek zwanych aminokwasami. Aminokwasy nawleczone po kolei razem tworzą białka, których jest niezliczona różnorodność - 10 mln. lub więcej.

Nośnik chemiczny kodu genetycznego - DNA - nie jest zbyt trwały. Przeżywa w komórkach organizmu, kopiowany wielokrotnie, ale niezbyt długo po śmierci osobnika. Informacje zawarte w DNA są jednak ogromnie trwałe. Losowe błędy, które naturalnie występują w organizmach żywych są naprawiane przez zestaw złożonych mechanizmów biologicznych. Wyniki działania komórkowych systemów naprawy DNA daje poziom błędów rzędu jeden na miliard lub dziesięć miliardów kopiowań. Dobór naturalny działa jako kolejna warstwa sprawdzania błędów. Rzadkie błędy wymykają się tym mechanizmom, inaczej nie byłoby nas tutaj żeby przeczytać ten blog, ale są one o wiele rzadsze od błędów w systemach komputerowych. Informacja trwa, ponieważ jest kopiowana, bardzo wiernie, z pokolenia na pokolenie.

Korekcja błędów oraz weryfikacja jest również cechą nowoczesnych komputerów cyfrowych, aktywnie wykorzystywana w kopiowaniu informacji.

Zabezpieczanie

Wracając do zabezpieczenia dziedzictwa kulturowego człowieka: nie możemy liczyć na trwałość nośnika dla długoterminowego zachowania informacji. Lekcje z naszej własnej kultury i z biologii są jednoznaczne - tylko kopiowanie informacji, z tworzeniem wielu kopii i z najlepszą możliwą korekcją błędów może zachować dziedzictwo kultury dla przyszłych pokoleń.

Znajdywalność

W pracy w archiwum, bardzo często spotykamy się z pytaniami: "Mój dziadek brał udział w bitwie pod (...), co się z nim dalej działo?" Za każdym razem staramy się wyjaśnić, że informacja ta mogą być gdzieś wśród 1,5 mln stron dokumentów w naszym archiwum. W krótkim czasie, w jakim istnieje Internet, ludzie nauczyli się polegać na Google czy Wikipedii,

aby znaleźć cokolwiek. W rzeczywistości Internet już teraz spełnia funkcje naszej " pamięci zewnętrznej
", przeszukiwalnej szybciej niż informacja w naszych głowach. Znajdowanie informacji w Internecie jest wspomagane przez zwiększenie dopływu metadanych, jak również przez wyszukiwarki ogólnego przeznaczenia.

Co w epoce książek papierowych nazwało się tworzeniem indeksów uzyskała teraz nową nazwę, "metadane". Organizacje przechowujące zasoby kulturowe coraz szerzej udostępniają metadane w Internecie, niezależnie od tego, czy sam zasób jest swobodnie dostępny czy nie. Być może trzeba jeszcze przyjść do biblioteki, aby wypożyczyć książkę, ale przynajmniej można szybko znaleźć, w której książce znajduje się poszukiwany cytat.

Istnieją dwie tendencje, obie obiecujące o wiele lepsze wyniki w znajdywaniu informacji w przyszłości. Jedną jest udoskonalanie przetwarzania języka naturalnego, co pomaga Google i innym wyszukiwarkom lepiej zrozumieć zarówno nasze pytania jak odpowiedzi ukryte w złożonych zdaniach. Wyszukiwanie w Internecie już teraz wykracza poza proste hasła, daje lepsze odpowiedzi na proste pytania w języku angielskim (inne języki, jak polski, nadal pozostają w tyle) i można się spodziewać rosnącego wyrafinowania takich narzędzi. Drugą tendencją jest rosnąca podaż informacji ustrukturyzowanej, czyli metadanych. Pisałem wcześniej o Linked Data , idei która zakłada, że jeśli oznakujemy dane i relacje między przedmiotami, chmura Internetowa będzie w stanie udzielić odpowiedzi na znacznie bardziej skomplikowane pytania. Jeśli tylko informacja istnieje w postaci cyfrowej, będziemy w stanie do niej dotrzeć.

Dostęp

Dwudziestowieczny model dostępu do zasobów kulturowych zmienił się drastycznie w ciągu ostatnich dziesięcioleci. Przed Internetem były książki, księgarnie i biblioteki. Jeśli byłeś kolekcjonerem, mogłeś budować własną bibliotekę, a jeśli nie mogłeś sobie na to pozwolić lub brakło ci miejsca, mogłeś pójść do biblioteki - zwykle otwartej dla publiczności. Sztuka, obiekty historyczne lub archeologiczne były do obejrzenia w muzeach (lub przechowywane w ich podziemiach). Tylko najbardziej popularne książki były powszechnie czytane, starsze, o wyczerpanym nakładzie, znikały z obiegu. W 21 wieku, na skutek kolizji nowych technologii cyfrowych i starego prawa autorskiego, nadal można kupić książkę papierową, ale opcje dla wersji cyfrowej są bardzo ograniczone. Nie można kupić e-książki, można tylko uzyskać ograniczoną licencje na jej używanie. Biblioteki zaczynają wypożyczać e-książki, ale istnieją duże ograniczenia. Powstanie Internetu zmieniło krajobraz w sposób dramatyczny. Stare książki, stare filmy i sztuka są dziś o wiele bardziej dostępne i przechodzą renesans. Jest segment ludzkiego działalności kulturalnej, który jest swobodnie dostępny, i segment który na skutek ograniczeń praw autorskich pozostaje w tyle. Nowa generacja opiera się prawie wyłącznie na Internecie w dostępie do informacji i dóbr kultury.

Gdy jakiś zasób jest "umieszczony w Internecie", można do niego dotrzeć z dowolnego miejsca na świecie. Wzrost dostępności jest oszałamiający. Google ma rzędu 500 milionów wyszukiwań dziennie, artykuły w Wikipedii są czytane miliony razy w ciągu dnia. Niewielka część zasobu archiwum Instytutu Piłsudskiego, która została do tej pory zdigitalizowana i jest dostępna przez Internet cieszy się około 200 - krotnym wzrostem liczby odwiedzających (ok 40 tysięcy rocznie). Wzrost ten jest nie tylko w wartości bezwzględnej, ale także w zasięgu geograficznym. Bez konieczności podróżowania, trafienia w stronę archiwów internetowych Instytutu w poszukiwaniu informacji pochodzą z około 2,500 różnych miejsc na całym świecie. Jest to możliwe, ponieważ zasoby są zdigitalizowane, zindeksowane i dostępne w sposób otwarty. Przyszłością dostępu do tekstu, zdjęć, muzyki, obrazów ruchomych i innych wytworów ludzkiej kultury jest niezaprzeczalnie technologia cyfrowa.

Czytaj więcej

- Manuscript Culture - artykuł w Wikipedii o kulturze manuskryptowej w średniowieczu.
- Naprawa DNA - artykuł o mechanizmach komórkowych naprawy DNA.
- Search Engines Change How Memory Works - artykuł w Wired.
- Google Search scratches its brain 500 million times a day - artykuł w CNET.
- Paper Rules: Why Borrowing an e-book from your library is so difficult - artykuł w Digital Trends.
- Semantic Search - artykuł w Wikipedii o przeszukiwaniu z użyciem zdań języka naturalnego.

Marek Zieliński

Artykuł ukazał się 4 kwietnia 2014 *Blogu archiwistów i bibliotekarzy Instytutu Piłsudskiego*

Może Cię też zainteresować

- Czy jesteś GLAM?
- Projekty digitalizacji
- Wstęp do Linked Data
- Digitalizacja a cyfryzacja

- Wstęp do standardów metadanych

{plusone}