Cuprins

1. Introducere

Digitalizarea în Științele Umaniste (Digital Humanities – DH de aici înainte) este o disciplină relativ nouă, comparativ cu discipline din domenii cu o tradiție de mii de ani ca matematica sau istoria. Nu mai veche de șapte decenii, DH prezintă însă o dinamică spectaculoasă, mai ales în ultimul deceniu. Dar ce înseamnă DH?

În sensul cel mai larg, DH studiază obiectele culturale într-un mod computațional. Nu există însă o definiție unanim acceptată, ci mai degrabă o multitudine de definiții care variază în funcție de perspectiva autorului. Pentru a da un exemplu destul de extrem, pagina https://whatisdigitalhumanities.com/ oferă nu mai puțin de 817 fraze preluate de la participanții la *Day of Digital Humanities* (Ziua DH) din 2004 până în 2014, în care fiecare încearcă o definire proprie a domeniului.

Dificultatea definirii DH derivă din diversitatea disciplinară care include patrimoniul cultural, lingvistica computațională, arheologia digitală, istoria, artele, filosofia etc. Fiecare dintre aceste discipline are propriile metodologii, nevoi și instrumente specifice.

Cu toate acestea, DH reprezintă un domeniu științific, după cum indică folosirea majusculelor. Cuvântul central din construcția DH este „Humanities", iar „Digital" este doar un modificator adjectival. Ce semnifică „Humanities" și „Digital" în această alăturare?

Curentul „Umanism" s-a născut în Italia de Nord Medievală și a marcat trecerea de la o perioadă de decadență la o perioadă de renaștere, printr-un discurs cultural elevat. Curricula Umanismului cuprindea șapte arte liberale:

- trivio (gramatică, retorică și logică);
- quadrivio (aritmetică, geometrie, muzică și astronomie).

Contrar opiniei comune conform căreia științele umaniste nu au nimic de-a face cu matematica, toate aceste șapte discipline umaniste au în comun *structuri* și elemente *cuantificabile*, două concepte fundamentale de matematică. Așadar, oarecum surprinzător, umaniștii au fost întotdeauna interesați de disciplinele formale.

În ceea ce privește termenul „Digital", acesta indică un accent pe faptul că lumea se digitalizează și, în mod necesar, științele umaniste trebuie să facă același

lucru. Computerele excelează exact la capitolul cuantificare și reprezentare a structurilor. Prin urmare, umaniștii au profitat în mod natural de apariția calculatoarelor pentru a le pune să facă în locul lor munca repetitivă sau plictisitoare și să analizeze automat cantități mari de date.

Poate la fel de important ca definirea domeniului este și precizarea a ceea ce nu este DH. DH nu este doar înlocuirea analogului cu digitalul. Această concepție greșită foarte comună presupune că DH se ocupă doar de „digi**ti**zarea" bunurilor culturale, și nu de „digitalizarea" acestora. „Digitizarea" reprezintă conversia unui obiect fizic, cum ar fi un artefact cultural, într-un format care poate fi citit automat de către mașină, adică un obiect digital, în vreme ce digitalizarea reprezintă tratarea ulterioară a acestor obiecte digitale prin metode computaționale.

Obiectele digitale de interes în DH sunt de tipuri variate: text, imagine, sunet, reprezentări tridimensionale ale artefactelor, etc. În cele ce urmează vom descrie pe scurt reprezentarea acestor tipuri de obiecte digitale în cadrul DH.

Cel mai frecvent studiat tip de obiect digital în DH este textul și el va fi obiectul central și în această carte. Reprezentarea digitală a textului nu este nici pe departe atât de simplă pe cât ne-am imagina. Sistemele de scriere sunt extrem de complexe, iar reprezentarea lor digitală trebuie să țină seama de acest lucru. Există trei strategii de scriere (Sproat și Gutkin, 2021) în funcție de ce reprezintă semnele:

- scriere alfabetică – semnele reprezintă sunetele de bază ale limbii (foneme),
- silabică – semnele reprezintă silabe și
- logogramică – semnele reprezintă cuvinte.

Niciun sistem de scriere nu este pur. Fiecare sistem de scriere folosește o strategie predominanta, dar și una sau chiar amândouă dintre celelalte strategii de scriere.

Sistemele de scriere care reprezintă sunete au în general un alfabet format din 20-30 de litere cu care se reprezintă fonemele. Se pare că toate alfabetele au ca sursă alfabete care reprezentau consoanele din limbile semitice, dezvoltate în mileniul al doilea î.Hr. Cum nicio limbă nu are exact aceeași mulțime de sunete cu o altă limbă, fiecare limbă are propriul alfabet, adaptat nevoilor de reprezentare a sunetelor specifice, prin adăugarea, modificarea (combinare, alterare, adăugare de diacritice sau alte semne, ca sedila, tilda, umlautul, etc.) sau

excluderea unor litere din alfabete existente. Printre cele mai cunoscute alfabete din spațiul indo-european se numără alfabetul grecesc, cel roman, cel chirilic si cel gotic. De exemplu, engleza și româna folosesc ambele alfabetul roman, dar limba engleza are aproximativ 40 de sunete, reprezentate prin 26 de litere, în timp ce limba romana are 32 de sunete, reprezentate prin 28 de litere (+3 importate pentru reprezentarea neologismelor – q, w, y).

Sistemele de scriere în care semnele reprezintă silabe se numesc silabare. De fapt, semnele reprezintă doar silabe formate din 2 sunete, o consoana urmata de o vocala, celelalte silabe fiind reprezentate cu ajutorul acestor semne prin diverse tactici. În aceasta categorie de sisteme de scriere intră sistemul *micenian*, limba amerindiana nativa *cherokee*, limba africana *vai* sau silabarul *kana* din japoneza.

Sistemele de scriere în care semnele reprezintă cuvinte (numite și logograme) sunt cele mai vechi, cuprinzând hieroglifele egiptene, cuneiformele sumeriene sau glifele mayașe. Acestea aveau însă și o componentă puternică silabară. Astăzi printre aceste sisteme se numără scrierea tradițională chineză și cea japoneză, *kangi*.

Reprezentarea digitală a marii majorități a caracterelor din toate sistemele de scriere a fost standardizată în 1991 prin *Unicode*, o codificare care acoperă toate scripturile, alfabetele, simbolurile, emoji-urile și caracterele neimprimabile și care cuprinde aproape 150,000 de caractere.

Dar complexitatea sistemelor de scriere nu se oprește aici. Odată reprezentată o secvență scrisă, cum se pronunță aceasta? Fiecare limbă are propriile reguli de scriere și citire, iar corespondența dintre reprezentarea ortografică și cea fonetică este mai mult sau mai puțin complexă, în funcție de limbă. Pentru a putea reprezenta fonetic toate sunetele din toate limbile lumii, a fost creat sistemul de transcriere IPA, care este o reprezentare fonetică a sunetelor limbilor creată de Organizația Internațională de Fonetică (*International Phonetic Association* în engleză, în original)[1].

Cu ajutorul sistemului IPA se poate reprezenta pronunția oricărui cuvânt din orice limbă, indiferent de modul în care acesta este scris în limba respective. De exemplu, engleza nu este o limba fonetică, în sensul că nu se poate prezice pronunția unui sunet pe baza modului în care acesta este scris. Scrierea sunetului *ch*, de pildă, dă pronunții diferite în cuvinte diferite și deci se reprezintă diferit în sistemul IPA: în *chair* se pronunță [tʃ] în *chaos* [k], în *machine* [ʃ] și în

[1] https://www.internationalphoneticassociation.org/

yacht se pronunță [Ø)]. Din acest punct de vedere, limba română este „mai fonetică" decât engleza, scrierea fiind mai fidelă pronunției.

Din punct de vedere al scrierii digitale a sunetelor în IPA, există trei posibilități practice: folosirea de tastaturi fonetice fizice sau folosirea de tastaturi virtuale, fie instalate local pe calculator[2], fie tastaturi online[3].

Un alt obiect digital de interes în DH este imaginea. Ca și în cazul textului, suntem obișnuiți cu imaginile create direct în format digital, dar în cazul obiectelor culturale, imaginile pot necesita trecerea din format fizic (fotografii vechi, hărți, etc.) în format electronic, editabil. Cel mai întâlnit mod de a face acest lucru este, din nou, ca și în cazul textului vechi, scanarea. Spre deosebire de text, însă, nu există un standard unic de reprezentare digitală a imaginilor, acestea putând fi reprezentate în două moduri principale: ca vectori (*vector images*, în original, în engleză) în formate ca *svg* sau *ai* și ca pixeli (*raster images* în original, în engleză), în formate ca *gif*, *jpg* sau *png*.

Similar, sunetele ca obiecte culturale pot fi digitizate prin transformare din obiect analog (bandă magnetică, vinil, casetă) în obiect digital editabil, în formate comprimate (MP3, MP4) sau necomprimate (AIFF – standard Apple, WAWE – standard Windows).

În sfârșit, obiectele culturale tridimensionale ca artefactele (vase, sculpturi, monede, coloane, clădiri sau chiar machete de cetăți sau orașe întregi) sunt de mare interes pentru DH. Acestea se pot digitiza prin scanare tridimensională și reprezenta în două tipuri de formate editabile și imprimabile 3D: date geometrice poligonale în formate ca *obj*, *fbx*, *glTF*, *usd*, și date redate prin reprezentare a limitelor de tip CAD (acronimul de la *Computer-Aided Design* în original, în engleză) în formate ca *DXF*, *DWG* sau *SVG*.

Digitalizarea pe scară largă a obiectelor culturale a dus la o democratizare a cercetării în științele umaniste. Nu cu mult timp în urmă, cercetătorii aveau la dispoziție sute de cărți, în majoritate în format fizic, în care căutau informația manual. Astăzi, numărul acestora este mai aproape de milioane, și totuși găsirea de informații în acestea este aproape instantanee, datorită automatizării procesului. La fel se întâmplă și cu imaginile, sunetele și reprezentările tridimensionale ale artefactelor. Până nu demult, sursele documentare analoage erau greu accesibile, sensibile și împrăștiate fizic în locații diferite. Astăzi există

[2] http://inkey.freehostia.com/

[3] https://ipa.typeit.org/

nenumărate colecții digitale accesibile, nevulnerabile fizic și într-un singur loc, la dispoziția cercetătorilor.

Așadar, digitizarea obiectelor culturale a creat oportunități de cercetare fără precedent, deschizând drumul spre aplicarea metodelor computaționale pentru studiul acestora. Pe cât de importantă și laborioasă este însă această digitizare, patrimoniul cultural universal fiind imens, aceasta este doar primul pas: obținerea obiectului de studiu. Noua viață a obiectului digital ar trebui să fie scopul ultim al umanistului digital. Această etapă include, de exemplu, operațiuni precum: editare, adnotare, interogare, găsire, modificare, vizualizare, extragerea datelor, clasificarea automată, recunoașterea tiparelor, extragerea informațiilor etc. De remarcat că tratarea computațională a obiectelor digitale este specifică tipului de obiect. De pildă, pentru text, *lingvistica computați*onală și *procesarea limbajului natural* (*Natural Language Processing* – NLP) sunt sintagme mult mai vechi decât DH. Similar, *Vision* este deja un domeniu de sine stătător care se ocupă de procesarea imaginilor, cu nenumărate instrumente (ca GIS – *geographic information system*, un sistem utilizat pentru a crea, stoca, a analiza și a prelucra automat informații distribuite spațial). Generarea și recunoașterea de vorbire (s*peech generationa and recognition* în original, în engleză) este de asemenea un domeniu în sine care se ocupă de generarea și recunoașterea automată a sunetelor limbajelor naturale și de interpretarea acestora pentru numeroase aplicații ca asistenții virtuali sau chat-boții care comunică verbal în limbaj natural. Fără a le substitui, DH aduce un suflu nou, cu mult mai multă aplecare pe domeniul umanist, și cu mult mai multe analize și interpretări.

Procesarea computațională a acestor date digitale suportă trei abordări: cea programatică, tradițională, care implică scriere directă de cod în limbaje de programare, și necesită pregătire în domeniul informaticii sau al ingineriei; una alternativă, mai nouă, în care tendința este de a crea instrumente și aplicații din ce în ce mai ușor de folosit, cu interfețe grafice și funcții gata implementate, care pot fi utilizate de un segment mult mai larg de persoane interesate de domeniu, fără o pregătire formală în programare, cum sunt umaniștii; în sfârșit, în ultima jumătate de deceniu, apariția modelelor de limbaj mari (***L**arge **L**anguage **M**odels* – LLMs) (Naveed et al., 2023) a democratizat și mai mult accesul la metodele computaționale, la viteza și puterea de calcul enorme ale mașinilor, prin posibilitatea de comunicare cu acestea direct în limbaj natural, din prompt, prin ingineriea de prompt (*prompt engineering* în original, în engleză), fără intermediul scrierii de cod sau a interfețelor grafice. LLM-urile sunt astăzi

capabile să asiste pe oricine într-o gamă nelimitată de sarcini, pot scrie rapoarte, pot scrie cod, pot oferi sfaturi tehnice, pot genera ficțiune, versuri, muzică, imagini, etc.

Această stare de lucruri favorizează competențele umaniste în defavoarea celor tehnice, constituind un moment de cotitură în cercetare. De pildă, nu mai departe de acum două decenii, doar o persoană cu background tehnic putea crea și întreține o pagină web sau un site, dar astăzi oricine poate învăța *web design* într-un interval de timp rezonabil, cu ajutorul instrumentelor ca WordPress sau Wix, care elimină nevoia de a scrie direct cod html, având interfețe grafice prietenoase, cu meniuri de unde se pot selecta toate opțiunile printr-un click sau prin drag-and-drop. Mai mult, odată cu apariția modelelor de limbaj mari, este posibilă crearea unei pagini web personalizate doar din prompt-uri adresate unui astfel de model. Un alt exemplu este evoluția cartografiei, domeniu nu cu mult timp în urmă rezervat doar unui număr mic de specialiști, dar care acum a devenit mult mai inclusiv: oricine poate crea hărți personalizate sincronic sau diacronic, oricine poate căuta coordonate sau adrese în aplicații extrem de ușor de folosit ca Google Maps. Și exemplele pot continua.

Această schimbare a accentului de la capacitatea tehnică de a rezolva o problemă prin scriere de cod la capacitatea de a o rezolva prin utilizarea unui instrument, a unei biblioteci de funcții, a unei aplicații ușor de folosit sau a unui model de limbaj mare poate fi descrisă prin sintagma din ce în ce mai folosită de „procesare asistată de calculator”, unde accentul se pune nu pe implementarea tehnică, lăsată în sarcina mașinii, ci pe capacitatea umană de cunoaște bine problema, de a ști ce să ceară mașinii, de a înțelege, a îmbunătăți și a interpreta rezultatele. Așadar, umaniștii nu mai pot ignora memoria și viteza mașinilor, algoritmii și modelele puternice de inteligență artificială (cum ar fi învățarea profundă, rețelele neuronale de tip transformer, LLM-urile, etc.) pentru a procesa date culturale. Însă ar trebui să o facă profesionist, având grijă să colecteze corect datele (scalate, echilibrate), să înțeleagă principiile, mecanismele și metodologia din domeniu, pentru a lua decizii informate cu privire la ce metodă sau algoritm să aplice, ce parametrii să aleagă sau ce unealtă sau aplicație se potrivește cel mai bine cu problema cu care se confruntă și să interpreteze cu atenție rezultatele, deoarece adesea corelațiile sunt înșelătoare și, în realitate, există alte fapte care le explică.

În acest context dinamic, rolul umaniștilor digitali a crescut. Tehnicienii (informaticieni, ingineri, matematicieni) care înțeleg cu adevărat cum

funcționează sistemele de inteligență artificială sunt foarte puțini și importanți. Dar rolul persoanelor care scriu cod de bază se diminuează, mai ales acum, când sistemele generative pot prelua aceste sarcini. Mai degrabă, este nevoie de umaniști critici, creativi, adaptabili și poate chiar vizionari pentru a utiliza în siguranță sistemele inteligente la potențialul lor maxim, în beneficiul oamenilor.

Pentru a răspunde însă acestei provocări, umaniștii trebuie să facă efortul de a dobândi și stăpâni tehnici și cunoștințe digitale complet nespecifice domeniului lor de studiu, care apar în sintagme populare în limba engleză, ca *Artificial Intelligence* (Inteligență Artificială), *Large Language Models* (modele mari de limbaj), *neural nets* (rețele neuronale), *Big Data* (date masive), *Machine Learning* (învățare automată), *Deep Learning* (învățare profundă), *Data Analysis* (analiza datelor), *Data Science* (știința datelor), *Text Mining* (extragerea de informații), *Network Theory* (teoria rețelelor), or *Geoinformatics* (geoinformatică). Mai mult, au nevoie să se adapteze permanent la un mediu în care instrumentele și metodele se schimbă într-un ritm amețitor. Astfel, învățarea continuă nu este doar o mantră goală de conținut a dezvoltării personale, ci o realitate de care umaniștii trebuie să țină cont în munca lor de zi cu zi. Aceasta este o schimbare profundă la nivel social și psihologic, care afectează generațiile contemporane. Dacă acum trei-patru decenii norma era ca o persoană să iasă la pensie de la locul de muncă unde s-a angajat pentru prima dată și să folosească același soft, instrument sau aplicație la serviciu timp de ani de zile, fără modificări substanțiale și fără a avea nevoie de recalificări majore, acum acest lucru este practic imposibil. Majoritatea tinerilor de astăzi se vor confrunta de-a lungul carierei de nenumărate ori cu schimbarea serviciului, a angajatorului, vor participa la diverse tipuri de proiecte, și vor trebui să învețe softuri, instrumente și aplicații noi aproape anual, de cele mai multe ori autodidact.

Din aceste motive, ne propunem ca această carte să vină în ajutorul umaniștilor (și nu numai), constituind un punct de plecare pentru învățarea continuă. Alte cărți de introducere în DH abordează această problemă prin listarea instrumentelor și aplicațiilor relevante la data publicării. Cum acestea sunt extrem de volatile, am ales o altă perspectivă în cartea de față: vom porni de la a explica mai întâi metodologia generală a cercetării din domeniul DH, cu accent pe text ca obiect de studiu, apoi vom exemplifica această metodologie prin abordarea unor probleme de analiză concrete, cu care se poate confrunta un umanist, alegând instrumentele potrivite pentru aceste sarcini, așa cum se întâmplă de fapt în procesul de cercetare.

2. Scurtă istorie a DH

Umaniștii au fost dintotdeauna interesați de structuri, formalizări și analize cantitative ale obiectelor lor de studiu, pe care le obțineau bineînțeles manual, cu mult înainte de apariția calculatoarelor. Un astfel de exemplu este studiul lui T. C. Mendenhall (1901), la sfârșitul secolului al XIX-lea, care a angajat două persoane pentru a număra manual cuvintele de două litere, de trei litere, și așa mai departe, în textele cu autor posibil fie Shakespeare, fie Marlowe, fie Bacon, pentru a încerca determinarea autorului pe bază de frecvențe. Odată cu începutul erai digitale, aceste interese și-au găsit în calculator un aliat rapid și eficient.

Inițial, domeniul este denumit Calcul Umanist (*Humanities Computing*, în original, în engleză) până în anii 2006 această sintagmă dominând mai recentul *Digital Humanities*, după cum se poate vedea în figurile 1 (în perioada 1980-2022) și 2 (zoom pentru perioada 1970-2010), obținute cu Google Books Ngram Viewer[4]. Acesta este un motor de căutare care afișează frecvențele cuvintelor sau șirurilor de cuvinte căutate, așa cum apar acestea într-un corpus enorm de cărți care conține peste 360 de miliarde de cuvinte, de-a lungul axei timpului, anual, începând cu anul 1800 până în 2022.

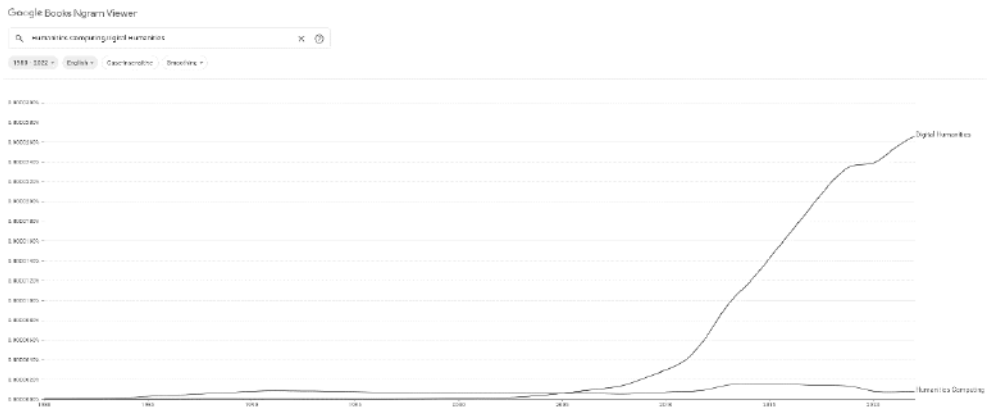

Figura 1. Evoluția în timp a denumirilor alternative *Humanities Computing* și *Digital Humanities*, conform Google NgramViewer, 1980-2022

[4] https://books.google.com/ngrams/