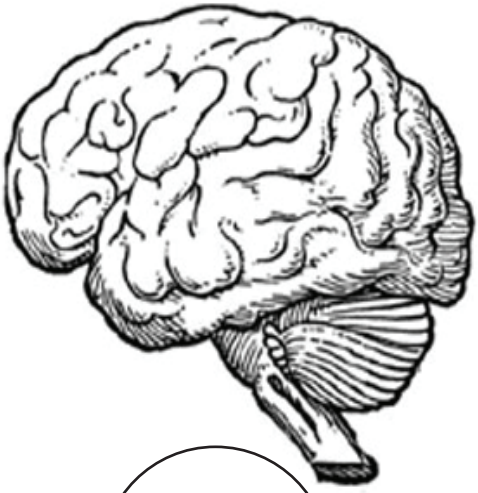
Capitolul 4

Continutul Digital

Taxonomia si Clasificarea Informatiei

Organizarea logică a informației



punctual

- Logica de clasificare se stabilește la început
- În funcție de volum aplicăm metode adecvate
- Importantă este interdependența informațiilor
- Costul clasificării este de luat în calcul

- Se lucrează cu un sistem simplu inițial, și cu un al doilea paralel care se construiește în timp

Taxis vine din limbă greaca și înseamnă ordine, aranjare. **Taxonomia se referă la clasificarea lucrurilor, conceptelor, principiilor într-o organizare logică.**

Prin 2010 lucram din când în când la un proiect « de sertar » care a prins datorită conjuncturii substanță și a intrat în analiză. Numele de cod al proiectului era Businesoo și își propunea să schimbe celebra axiomă "6 degrees of separation" în "one degree of separation" în lumea business-ului european. Pentru asta trebuia să dărâme cea mai importantă barieră a Uniunii Europene, un spațiu economic complex care se desfășoară cultural în 24 de limbi oficiale și peste 60 de limbi regionale. Sigur engleza, franceza și germana acoperă o mare parte din comunicare, însă cei care vorbesc toate cele trei limbi nu sunt foarte mulți. Aici intervenea Businesoo, care permitea fiecărui utilizator să își desfășoare activitatea informațională în limbă să maternă, urmând că interacțiunea să ia formă pentru ceilalți în limbile lor materne. Și de aici întrebarea strart-up-ului : **Cum să clasifici mii de produse și servicii într-o formă care să fie optimă pentru bazele de date și pentru operarea în multiple limbi?** 5 ani mai târziu întrebarea este încă deschisă pe

Familiile de limbi EU

Una din clasificările simple (generale) ale unui website este limba în care el se prezintă. Uniunea Europeană are 24 de limbi oficiale, toate fiind și limbi de lucru. Documentele pe care un stat membru sau o persoană aflată sub jurisdicția unui stat membru le trimite instituțiilor UE pot fi adresate în oricare din limbile oficiale. Acestea sunt: bulgară, cehă, croată, daneză, olandeză, engleză, estonă, finlandeză, franceză, germană, greacă, maghiară, irlandeză, italiană, letonă, lituaniană, malteză, poloneză, portugheză, romană, slovacă, slovenă, spaniolă, suedeză.

masă cu un răspuns aproximativ ...

Întrebarea care și-o pune real proprietarul unui website este : Chiar am nevoie de clasificarea informației ? Clientul nostru, stăpânul nostru. Utilizatorul nostru, stăpânul nostru. **Studiile au arătat că din 10 oameni puși să caute o informație online rezultă 10 moduri diferite de gândire (conceptualizare) a căutării și ajungere la informație.** E normal, creierele noastre funcționează diferit. În toate cazurile căutarea s-a orientat după REPERE prestabilite. Iar **Clasificările sunt un mare reper pentru minte, sunt Ghiduri care ne permit să nu înmagazinam enorm de multă informație**. Acesta este motivul principal pentru care trebuie să ne clasificăm informația din site, indiferent de mărimea lui.

Ce metode de clasificare folosim?

Cel mai simplu sistem de clasificare este Categoria. Albe, Negre, Roșii și fiecare după culoare în categoria să. Exprimarea vizuală a unui sistem simplu se poate face în Harta Websitului (Sitemap), o reprezentare schematică a unei informații simple. Complicarea acestui sistem este Mulți-Categoria, sau adăugarea unui al doilea, al treilea nivel de adâncime. Albe , Albe-Pure și Albe-Impure. Exprimarea vizuală a unui sistem multi-categorie se face prin intermediul **Breadcrumb** , adică a unui înșiruiri logice a categoriilor. Un exemplu de Breadcrumb : “Albe > Albe Impure > Albe Impure Partial”.

Dacă Categoriile nu sunt deja bine delimitate, apar primele semne de întrebare. Un articol poate fi clasificat **ȘI – ȘI** , iar analiza trebuie să țină cont de aceasta. Dacă vinzi articole sportive în magazinul tău online o crosă de golf semnată de Tiger Woods s-ar putea să fie clasificată atât la Produse > Golf > Crose cât și la Produse > Produse Rare > Autografe, iar **statisticile de vânzare trebuie să țină cont că produsul se află în ambele clasificări.**

Pentru a facilita clasificarea multiplă folosim conceptul de TAG sau, că să fim în trend cu social-media, de #hashtag. Practic această înseamnă o etichetare suplimentară față de orice clasificare, a oricărui produs,

articol sau item din cadrul sitului, căruia putem să îi legăm una sau mai multe etichete care ne permit clasificări suplimentare paralele. Crosa de golf poate primi etichete #tigerwoods, #autograf, #2016, #turneu, #pauortez. În măsură în care statisticile sunt flexibile, clasificarea este mai bine făcută. Iar pentru utilizator, reprezintă adăugarea unor elemente logice în plus pentru regăsirea informației.

Ataturi de etichetare o altă facilitare se poate obține și prin **localizare**, a cărei exprimare vizuală se face prin mapping, livrarea unor hărți de proximitate, populate doar cu informația dorită (harta pensiunilor disponibile dintr-o regiune specifică, harta membrilor, harta proiectelor, etc). Geolocalizarea contează mult pentru utilizator când căutarea sa vizează locația și este o clasificare primară pentru serviciile de turism și aplicațiile care fac interacțiune între utiizatori.

Ultimul tip de clasificare "simplă" se găsește în **NOR, adică nu se pre-clasifica nimic inițial, însă din utilizare, informația grupează interesele și le coagulează.** Informația o oferă utilizatorii și datele introduse în sistem. Numărul de clasificări este infinit și rezultatul este o structură DINAMICĂ de N categorii cu N subcategorii cu N taguri și N locații. Practic orice element al structurii poate reclasificat, regrupat, în același timp în altă structură paralelă, singura lege care menține sistemul fiind UTILIZAREA. **Structurile rezistă atâta timp cât sunt folosite.**

Un exemplu: să zicem că vînd cărți, am circa 300.000 titluri la vânzare și între ele sunt și cărți pentru copii, care din utilizare le pot clasifica ca și cărți de colorat, povești, benzi desenate și cărți educative. Putem face clasificarea în acest mod, dar în doi ani de la lansare constatăm că utilizatorii nu mai sunt interesați de această structură, căutând mai de-

grabă cărți de 2 ani, 3 ani, 4 ani. Decidem dacă schimbăm sau montăm paralal o a două structură dinamică (depinde de resursele aflate la dispoziție, pentru că muncă de clasificare a produselor înglobează foarte mult timp de muncă efectiv – o reclasificare 500 de titluri presupune cunoștințe bune și un volum de 10 ore de muncă).

cutiuța cu idei
reține "clasificare inițială sau ulterioară"

Nici o clasificare nu este atotcuprinzătoare. Modul în care facem clasificarea unui produs este complex : din ce este făcut, utilizarea să, destinația să, modul de funcționare, gradul de incorporare, consumabilitate sau repetivitatea folosirii, domeniul și încadrarea economică. Utilizatorii gândesc diferit și statisticile ne arată în timp calea cea mai bătută pe care aceștia o urmează. Ne putem inspiră din standardizare însă cunoașterea adâncă o propriilor produse este bază vânzării lor. Clasificarea informației se face « open », orice nouă idee fiind ușor implementată.

Ce sunt MetaDatele ?

Definiția Metadatelor ar fi "**un set de date care descriu bine conțin informații despre datele inițiale**". Adică un fel de **explicație pentru o mai bună înțelegere**. In Internet metadatele sunt foarte importante. Inițial au fost prezentate elemente de pentru motoarele de căutare, dar importanța, sau mai bine zis, rolul lor s-a schimbat timp. Ele presupun marcarea informației de baza cu : scopul creației informației, scopul datelor, timpul creației, autorul drepturile datelor, locul rețeaua de creație, standarde utilizate, mărimea fișierelor. Pentru fotografii ar putea fi extinse la mărimea pozei, culori, rezoluție, etc.

Raţiunea existenţei Metadatelor este Standardizarea. Eu fac o clasificare, tu altă clasificare, motorul de căutare trebuie cumva să ordoneze mai repede munca unui miliard de creatori. Informaţia este clasificată mult mai uşor dacă fiecare din ei foloseşte anumite standarde recomandate. O crosă de golf într-un magazin online este din punctul de vedere META un produs al cărui scop este comercial, de vânzare în Europa, pus în vânzare acum un an, 100% un produs original, şi conţine descriere foto şi video. Să nu uităm că Meta tagul <TITLE> (numele produsului, titlul articolului) este în continuare cel mai important element pentru motoarele de căutare.

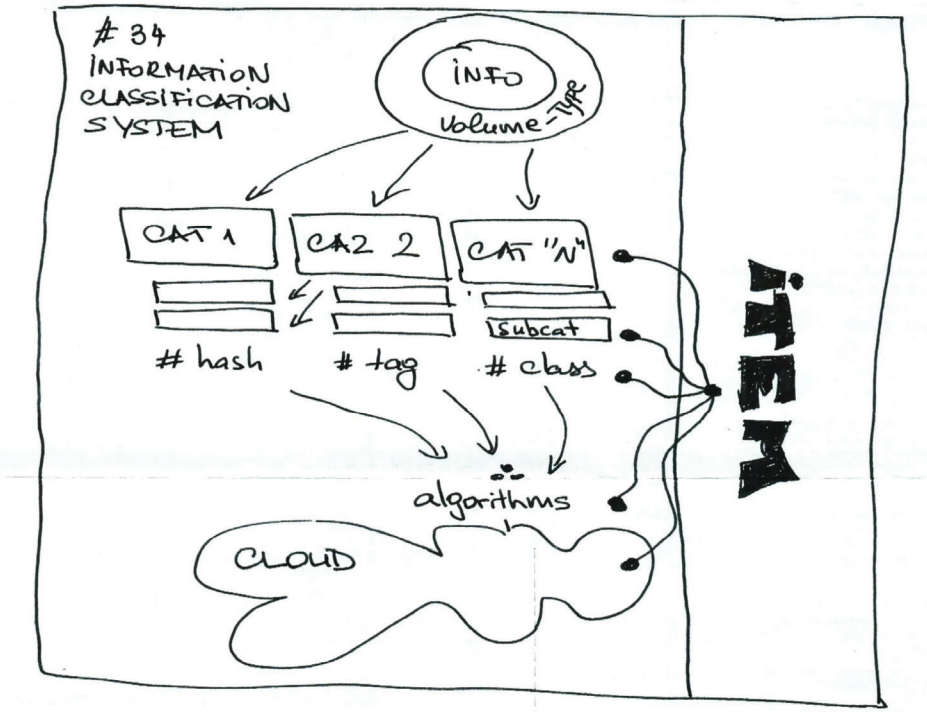
Sisteme şi algoritmi

Motorul de căutare ordonează uşor toate elementele pre-clasificate, însă deseori această muncă de clasificare nu o fac toţi, sau nu o fac standardizat. Eşti obligat să foloseşti algoritmi, mai ales când volumele de date se contorizează în milioane de Terrabytes. Aici ajungem aproape de un subiect care îmi este foarte drag şi care se numeşte “Machine Learning”, care vizează găsirea de modalităţi matemetice şi probabilistice pentru a clasifica informaţia. **Dacă sistemul ar putea învaţă singur şi să atingă Inteligenţă Artificială, atunci toate eforturile noastre de clasificare ar trebui să scadă la zero,** sistemul ştiind exact ce articole şi informaţii avem şi la ce folosesc. Avem **algoritmi de Machine Learning, algoritmi de Data Mining şi algoritmi de Clasificare** a datelor. În total **numărul lor depaşeste 200 de metode ştiinţifice diferite**. Forţa de a pune la muncă împreună multiple metode produce în timp performanţă (se vede în brevetele depuse continuu de specialişti).

Că urmare a perfecţionării motoarelor de căutare, atât interne cât şi externe, exprimarea vizuală cea mai simplă a acestor algoritmi pe care noi o cunoaştem cel mai bine este bara de căutare (google it!). Scriem şi obţinem rezultatul dorit, în majoritatea cazurilor, EXACT ce am căutat.

Pentru proiectul Businessoo am încercat adaptarea a 5 clasificări internaţionale de produse şi servicii, am încercat să definim noi propriile

clasificări, am încercat să construim după modele existente, nimic nu a ieșit perfect. Într-un eveniment al Startup West desfășurat la St Malo (France), unul din proiectele prezentate era condus de 4 doctoranzi în limbi (incidența limbilor în digital) de la IDL Atlantique din Nantes. Le-am prezentat proiectul și am cerut sfatul. După 5 ani de căutări răspunsul primit a fost dur, însă a ridicat "vălul" care îl aveam peste ochi. Mi-au spus "nu trebuie să faci nimic!". Cuuum ??? "**nu trebuie să faci nimic acum ... întâi strângi datele și abia apoi începi să le clasifici- nu ti se va potrivi nici o clasificare care o încerci**". astfel proiectul repornit.



O buna clasificare a informatiei este dupa Categorii, Taguri, Meta-Date completate de un algoritm intern.

pontul #34

verificare

- [ ] **Am determinat tipologia informatiei din site**
- [ ] **I-am determinat volumul ca sa am complexitatea**
- [ ] **Am determinat gradul de interdependenta al info**
- [ ] **Am stabilit un sistem logic initial**
- [ ] **Am calculat costul resursei umane clasificare**

Checklist: Clasificare Date

5 etape pentru o clasificare cât mai bună a datelor în produsele online

1. Să determinăm ce tipologie de informație avem pe site și care trebuie clasificată: articole scrise, produse în magazin, poze, filme video, secțiuni de discuții (forum), suport cu clienții deschis sau secțiune de întrebări și răspunsuri, etc …

Înainte de a propune clasificări utilizatorilor suntem datori să cunoaștem informația care o tranzităm în produsul nostru online pentru a ști ce avem de « aranjat pe rafturi ». Clasificările pot fi făcute după utilizarea produsului sau a informației, după scopul sau logic, după funcționalitate, după modul de funcționare, după percepție, după locul

de utilizare sau de ce nu, după algoritmi inovatori interni.

2. Să determinăm volumul de informație, că să știm cât de complex va fi sistemul de clasificare (zeci de intrări, sute, mii, milioane)
Atitudinea dă altitudinea zice o vorba circulată des în ONG-uri. Volumul dă complexitatea în materie de online. Cu cât mai multă informație, cu atât mai multă bătaie de cap în urmărirea ei că importanță. Strategia cea mai bună când este multă informație este să o lași să se clasifice singură treptat, iar capcana este că îți vine să o lași baltă datorită volumului. Trasarea principiilor inițiale te scăpa de problematica volumului.

3. Să determinăm gradul de interdependență al informației pentru a opta pentru sisteme multi-clasificare
Recomandarea este să știi de la început cum se poate lega informația. Articolele din website care sunt legate de video, de audio, de comentariile din social media și din forum, de alte websites, de motoarele de căutare, de categorii și taguri. Un desen simplu inițial ne da perspectiva de fiecare data când informația apare și se manifestă.

4. Să decidem dacă putem monta inițial un sistem logic sau dacă lăsăm operțiunea de clasificare pentru o etapă ulterioară
Chiar dacă montăm sisteme de clasificare ulterioare în paralel este obligatorie o clasificare inițială, chiar dacă este binară. Important și neimportant este un model. Vandabil sau puțin vandabil pentru magazine online (avantaj magazinele online pentru că aici produsele vin pre-clasificate prin însăși natură lor).

5. Să calculăm corect costul clasificării, în ore-resursa umană, pentru a justifica economic operațiunea. Este una din cele mai frecvente aprecieri greșite de cost pentru prodsele online. Cântărește greu costul ignoranței, pentru că de fiecare aplicarea unei noi clasificări de date se dovedește costisitoare. Volumul de muncă manuală este mare si deseori se renunță , strategia care se aplică fiind « de acum înainte ». Prevăzute corect încă de la început, sistemele flexibile, permisive, se dovedesc fi de un real ajutor antreprenorului online.