Biblioteci digitale pe structuri GRID

Prezentator: Gheorghe Sebestyen

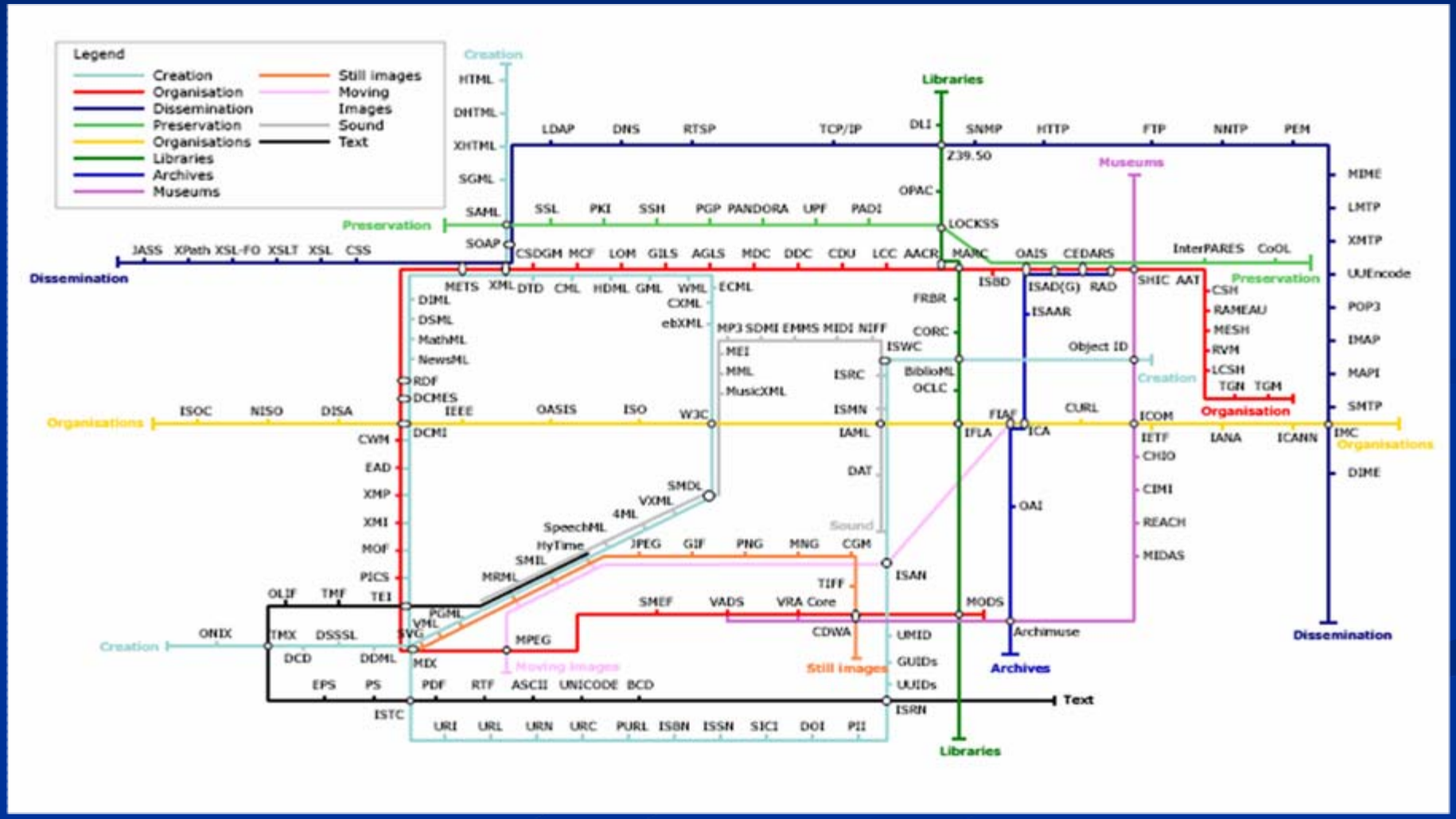
Continut

- Biblioteci clasice vs. biblioteci digitale
- Cercetari recente in domeniul bibliotecilor digitale (Digital Libraries - DLs)
- Obiective si cerinte de proiectare pentru bibliotecile digitale
- Biblotecile digitale raportate la Sistemele de management a continutului digital
- Biblioteci digitale bazate pe ontologie – biblioteci semantice
- "Grid-ificarea" bibliotecilor digitale
- Modelul unei Biblioteci digitale bazata pe o infrastructura GRID
- Rezultate experimentale –
  - Cautare pe baza de chei
  - Tehnici de cautare si clasificare semantica
- Concluzii

Biblioteci clasice si digitale

- Biblioteca clasica
  - o arhiva de cunostinte/informatii pe suport de hartie
  - Masura a gradului de civilizatie a unei societati
- Biblioteca digitala
  - Nu numai o versiune digitizata a unei biblioteci
  - Un set nou de functionalitati si servicii (controlul accesului, alocarea si managementul resurselor, servicii complexe de cautare si regasire)
  - Un mediu pentru schimb de informatii si cooperare
  - Contine o mare varietate de date in diverse formate (text, audio, video, documente compuse, obiecte digitale si colectii)
  - Bibliotecile digitale sunt sisteme informatice complexe care acopera toate aspectele legate de crearea, stocarea, procesarea, distributia si accesul la la date

Tehnologii IT si de comunicatie implicate in implementarea bibliotecilor digitale



http://mapageweb.umontreal.ca/turner/meta/english/metamap.html

Obiective pentru o biblioteca digitala moderna

- Viziunea proiectului DELOS –
  - "sa permita oricarei persoane accesul la orice informatie (cunostinte) oriunde si oricand, intr-un mod prietenos, eficient efectiv si multi-modal prin eliminarea barierelor de distanta, limba, si cultura si prin utilizarea de dispozitive interconectate pe Internet"
  - Biblioteca digitala = o ***arhiva de cunostinte*** si o ***infrastructura pentru schimbul de informatii*** care permite generarea, stocarea si accesul usor la date independent de distributia resurselor fizice, a bazelor de date si a persoanelor.
- Implementarea unei biblioteci digitale necesita infrastructura si servicii de calcul si de comunicatie de inalta performanta

Cercetari in domeniul Bibliotecilor digitale

- Delos Network of Excellence –
  - Obiectivul: definirea si implementarea de biblioteci digitale pe tehnologii noi de calcul si de comunicatie
  - Realizari: definirea ***cerintelor functionale si arhitecturale*** pentru o biblioteca digitala
- Proiectul BRICKS
  - Obiectiv: proiectarea unui spatiu orientat pe utilizator si pe servicii pentru ***utilizarea in comun a cunostintelor si a resurselor*** intr-un context multi-cultural
  - Realizari:
    - Definirea unei arhitecturi de biblioteca pentru o comunitate forte mare si eterogena de utilizatori,
    - functii automate de adnotare si indexare a continutului
- Proiectul OpenDlib
  - Obiectiv: dezvoltarea unui instrument software (toolkit) pentru gererarea de biblioteci digitale dedicate
  - Realizari: instrumente pentru **culegerea de continut digital (content harvesting)** din resurse existente
- Fedora, DSpace – software de tip “open source” pentru biblioteci digitale

Cercetari in domeniul Bibliotecilor digitale

- Proiectul Diligent (parte a proiectului EGEE)
  - Obiectiv: utilizarea infrastructurilor Grid pentru implementarea bibliotecilor digitale
  - Realizari: o noua viziune privind conceptul de biblioteca digitala:
    - Biblioteca digitala = un sistem dinamic de de stocare si management a continutului digital destinat unui scop bine definit (ex: proiect, curs, colectie de arta, etc.)
    - Definirea de servicii generice de biblioteca mapate pe servicii Grid
    - Experiment de catalogare automata a tuturor imaginilor existente pe un portal de imagini
- Proiectul Sinred – un proiect national in cadrul Programului de excelenta
  - Obiectiv: dezvoltarea unui cadru/model national pentru biblioteci digitale destinate domeniilor stiintifice si tehnice
  - Realizari:
    - evaluarea cerintelor, evaluarea produselor software existente
    - dezvoltarea unei infrastructuri Grid,
    - definirea unui model generic de biblioteca digitala,
    - implementare si experimente de cautare si regasire in biblioteci digitale pe Grid

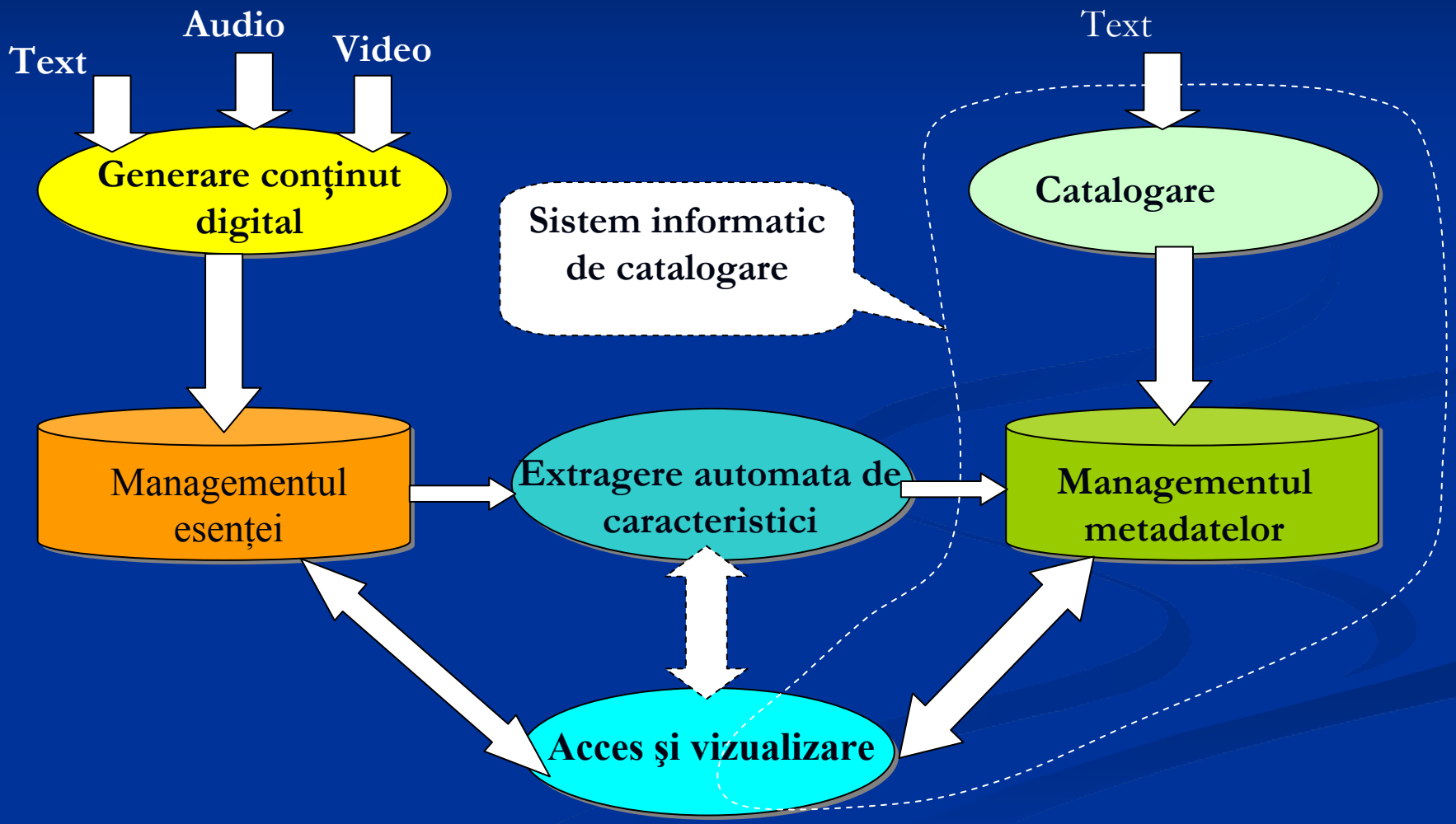
Cerinte pentru un sistem de biblioteca digitala

- **Cerinte arhitecturale**:
  - Natura distribuita a resurselor de stocare, procesare si de acces
  - Scalabilitate, interoperabilitate si flexibilitate
- **Cerinte functionale**:
  - **Functii de baza**: stocare, indexare si adnotare, cautare, regasire de continut, managementul utilizatorilor si a resurselor
  - Organizarea continutului trebuie sa reflecte conexiunile semantice existente
- **Facilitati de procesare**
  - Servicii de procesare a datelor – specializate pentru diferite domenii
  - Identificarea modelelor (pattern-urilor) de cautare si regasirea informatiilor pe baza acestora (de la chei de cautare la cautare semantica)
- **Cerinte de calitate a serviciilor (QoS)**
  - Siguranta datelor si a accesului
  - Timp rezonabil de regasire a informatiilor relevante
- **Manamenentul utilizatorilor si controlul accesului**
  - Promovarea ideii de Organizatie virtuala

Biblioteci digitale si/sau Sisteme de gestiune a continutului digital

- Sistem de management al continutului:
  - Sistem informatic destinat pentru stocarea, indexarea si clasificarea, vizualizarea si transmiterea datelor relevante pentru un anumit domeniu sau sfera de activitate
  - Gestionarea de formate foarte variate (continut web, multimedia, documente tehnice, rapoarte economice, etc.)
  - Exemple:
    - eGovernment and eAdministration,
    - Furnizare de continut Multi-media (muzica, film)
    - Date de administrare a companiilor
    - Continut stiintific si tehnic: standarde, conferinte, cursuri (eLearning)
- Biblioteci digitale:
  - Arhiva de continut digital
  - Un tip de Sistem de management a continutului
  - Asigura un acces mai larg si deserveste obiective mai generale (ex: cel de informare)
- Cele doua concepte sunt dificil de delimitat
  - In viitor, mai multe biblioteci digitale cu un scop bine definit

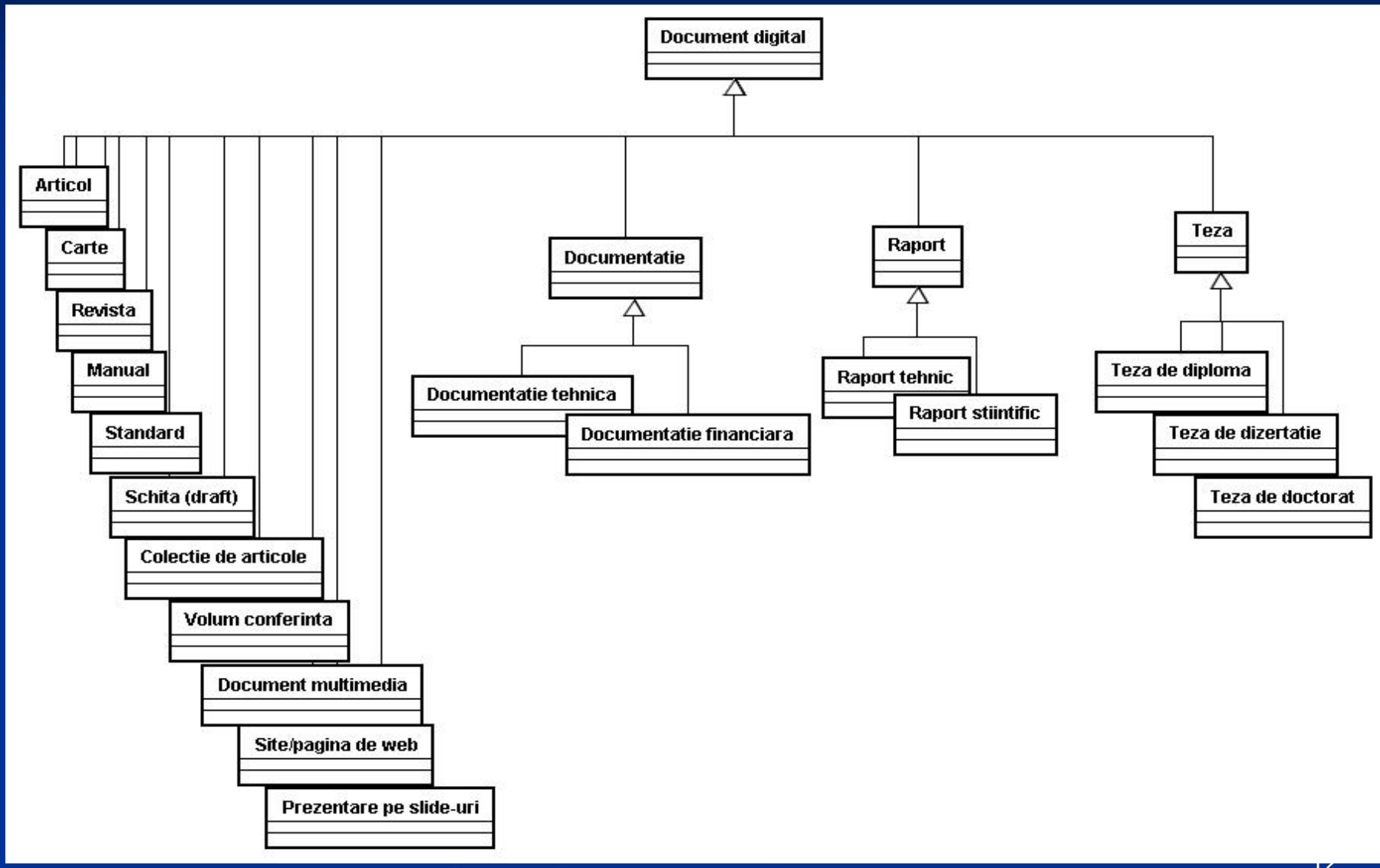
Schema de principiu a unui SMCD



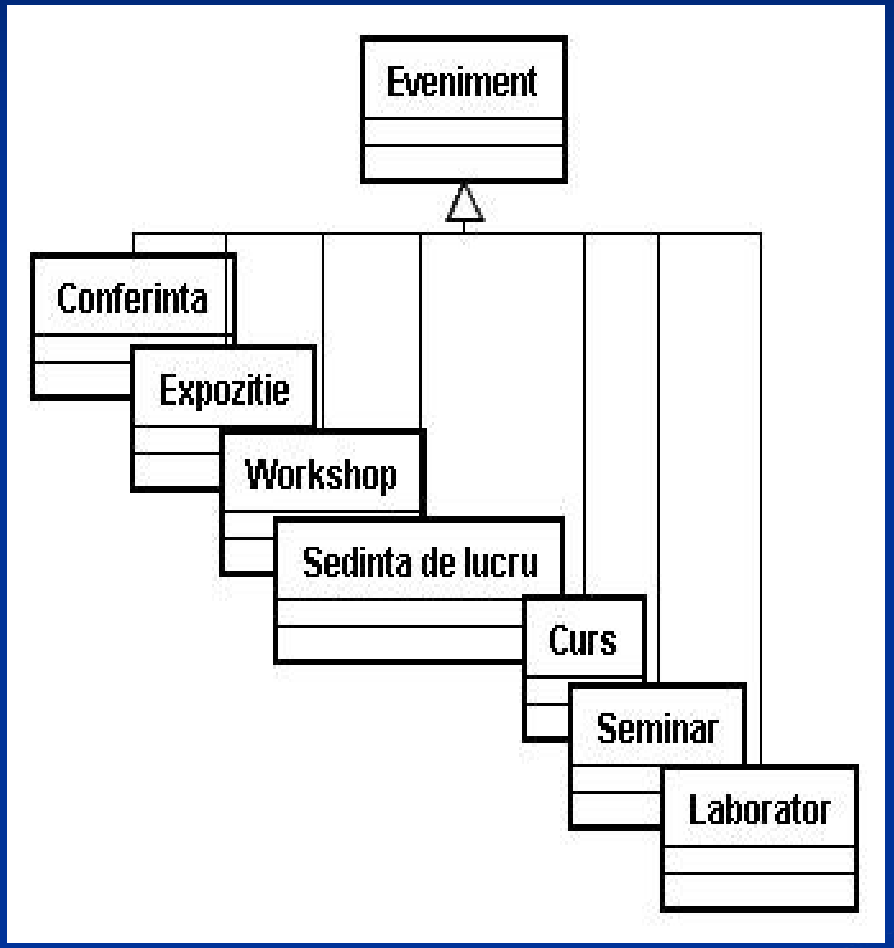
Abordarea pe baza de ontologii a Bibliotecilor digitale

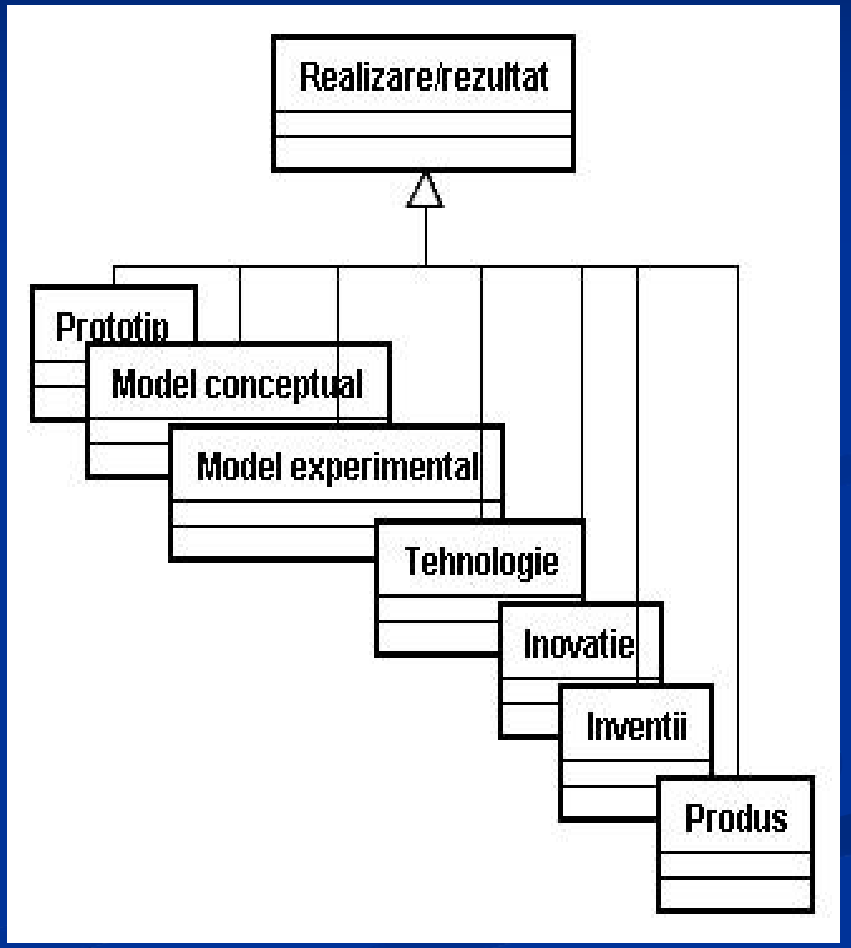
- Ontologie: concepte si relatii intre ele la un nivel mai abstract
- Ontologie pentru domeniul stiintific si tehnic
  - Concepte de baza:
    - Obiecte digitale:
      - Asociere de continut, metadate si proceduri de prelucrare si de acces a procedurilor
    - Colectii digitale:
      - Asocierea pe baza unui anumit criteriu a mai multor obiecte digitale
    - Evenimente:
      - continut asociat unei anumite manifestari (de scurta durata)
      - Exemple: Conferinte, workshop-uri, seminarii
    - Procese:
      - continut asociat unei activitati de durata
      - Exemple: Proiecte, Cursuri
    - Organizatii virtuale
      - Roluri
      - Utilizatori

Taxomonia documentelor digitale in stiinta si tehnica

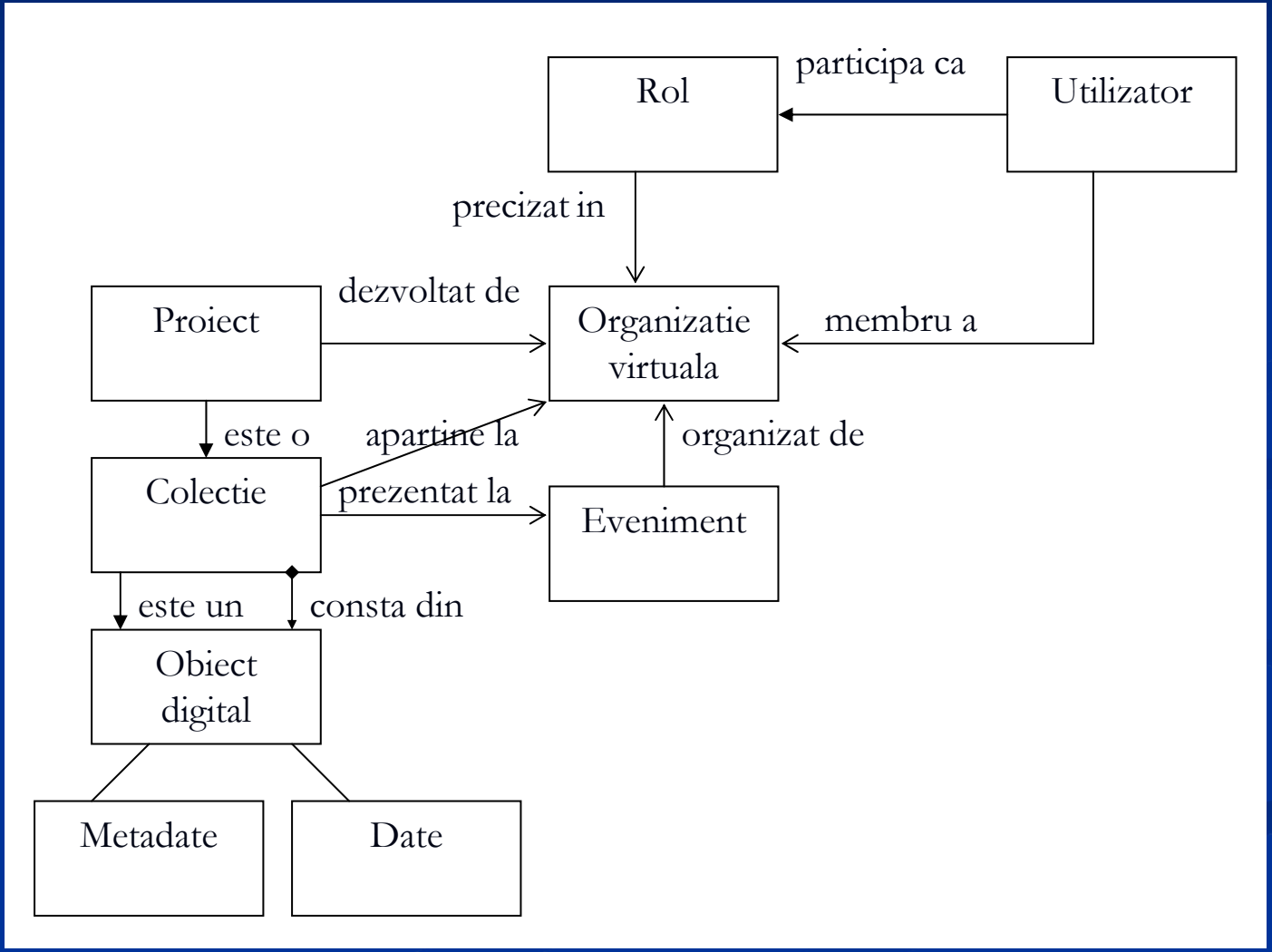


Alte taxonomii

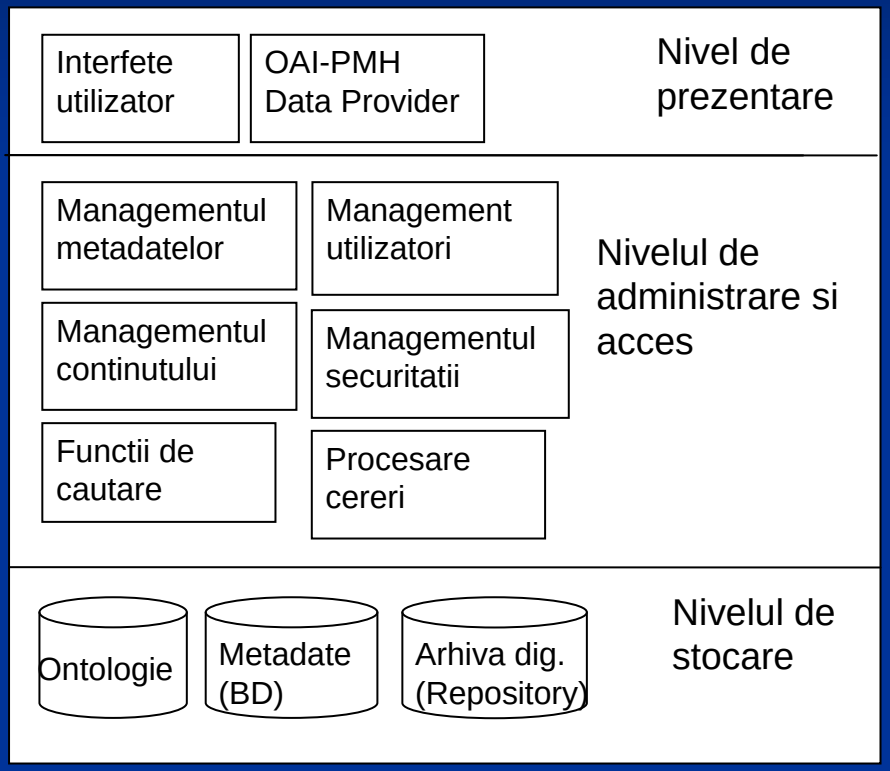




Concepte si relatii



Modelul de Biblioteca digitala



- Nivelul de prezentare - componente care comunica cu lumea in afara sistemului
- Nivelul de administrare si acces – manipuleaza continutul, utilizatorii si organizatiile vituale
- Nivelul de stocare – stocarea metadatelor si a continutului

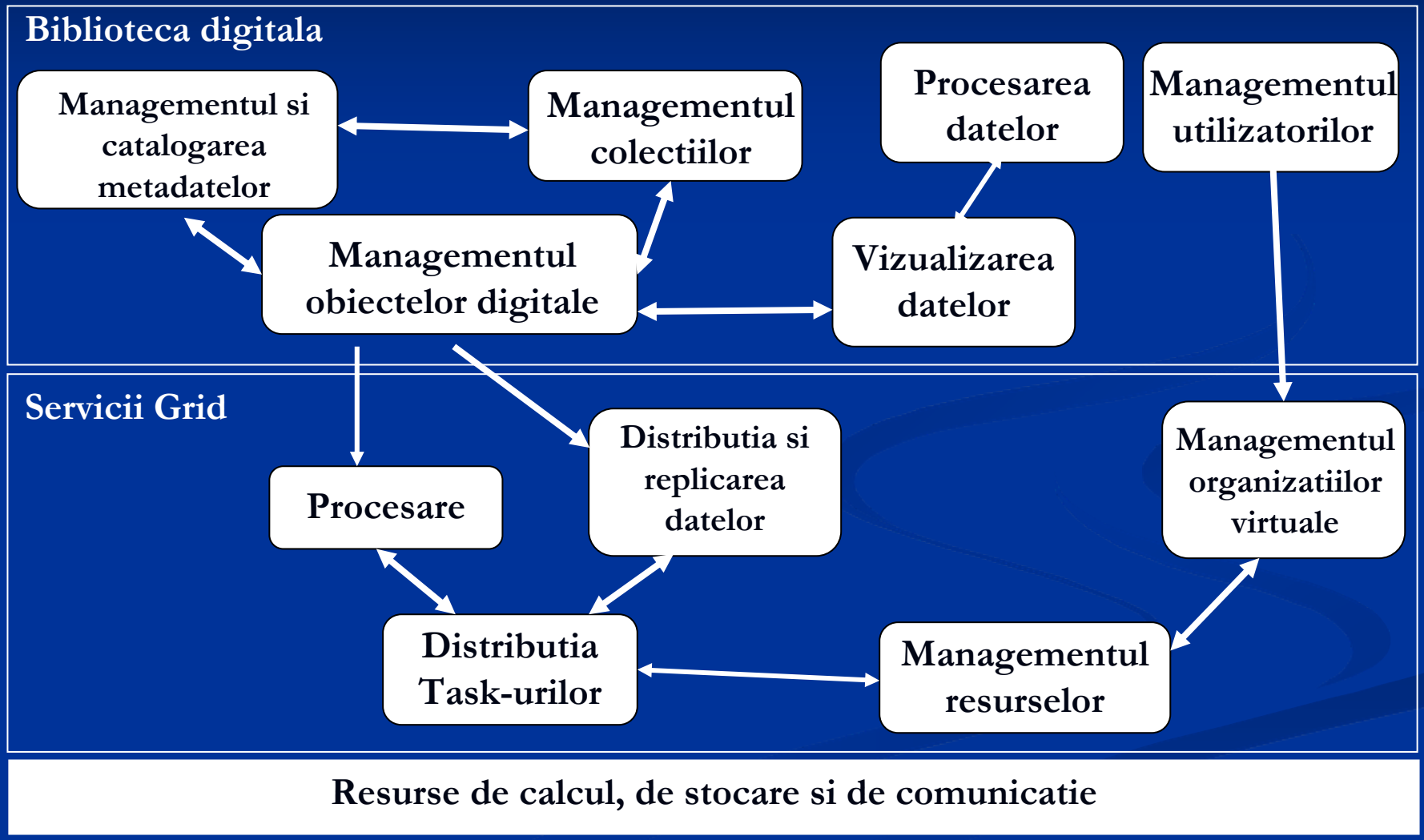
Servicii de biblioteca digitala pe GRID

- De ce Biblioteci digitale pe GRID?
  - Un volul imens de documente digitale
  - Acces concurent si motoare multiple de cautare (vezi Google)
  - Furnizare de continut multimedia on-line (Multimedia streaming)
  - Indexare, catalogare si adnotare automata
  - Procesari complexe de date (ex: recunoasterea si catalogarea automata a continutului multi-media) necesita timp de executie prohibitiv de mare
  - Managementul utilizatorilor si alocarea resurselor prin Organizatii virtuale
  - Facilitati de distribuire a sarcinilor oferite de serviciile Grid

“Grid-ificarea” modelului de biblioteca digitala

- Distribuirea continutului si replicare
- Controlul accesului la date prin:
  - Organizatii virtuale,
  - Certificarea si autentificarea utilizatorilor
  - Atribuirea de roluri
- Executia paralela a procedurilor de cautare si clasificare
  - Aceeasi procedura de cautare aplicata in paralel pe mai multe documente, pe mai multe noduri Grid
  - Distribuirea fazelor de executie ale unei proceduri de cautare (parsare, calculul vectorilor de caracteristici, identificare si selectie, clasificare) ????

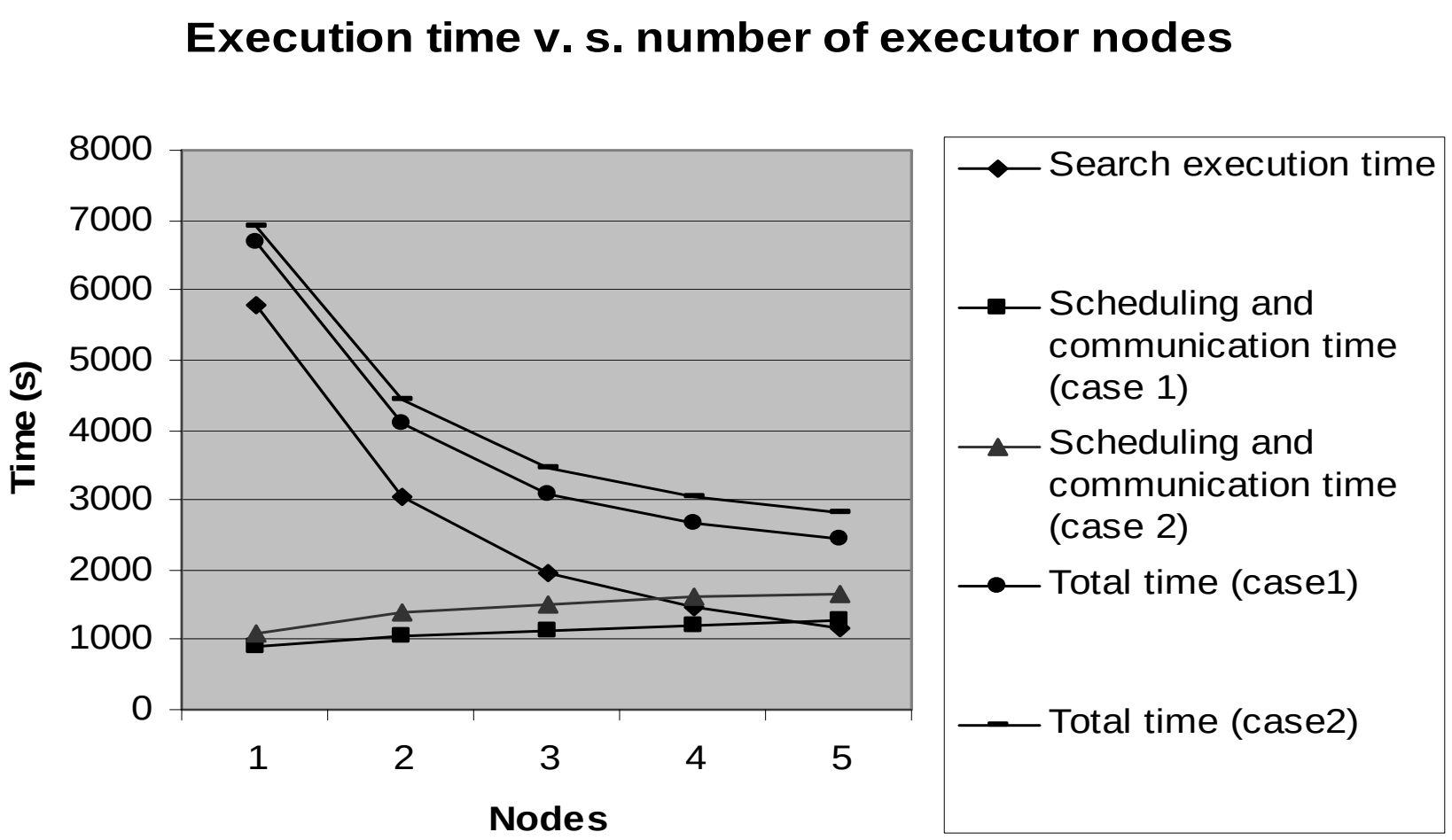
Modelul de Biblioteca digitala pe o infrastructura Grid



Experimente

- Implementarea unei biblioteci digitale pe platforma Alchemi Grid (Microsoft)
  - Distributia sarcinilor la nivel de fire de executie (thread-uri)
  - Programare Grid explicita
  - Experimente de furnizare in paralel de continut multimedia (multimedia content streaming)
- Implementarea unei biblioteci digitale pe platforma Condor Grid (Open source)
  - Distributia sarcinilor la nivel de task-uri
  - Distributia sarcinilor si a datelor este transparenta pentru aplicatia de biblioteca (distributia se face prin script-uri)
  - Experimente de cautare de documente pe baza de cuvinte cheie (cautare in continut si nu in catalogul de metadate)
    - Timpul de executie scade cu numarul de noduri executoare utilizate
    - Pentru mai mult de 5 executoare timpul de planificare si comunicatie devine comparabil cu cel de procesare
- Cautare statistica si semantica

Experimente



Cautare statistica si semantica

- Ideea:
  - regasirea sau catalogarea unor documente pe baza unor documente date ca exemplu
  - Regasire prin similaritate
- Algoritmi:
  - Algoritm de cautare de tip “Naive Bayesian”
  - Algoritm Topic-Based Vector Space Model (TVSM)
- Beneficiile implementarii acestor algoritmi folosind sisteme Grid:
  - Performante mai bune la timpul de procesare
  - Distributia documentelor

Algoritmul Naive Bayes – 1

- Scop:
  - Clasificarea datelor neetichetate cu ajutorul unor estimari folosind date de antrenare etichetate
- Conform cu teorema Bayes se poate obtine probabilitatea posterioara cunoscand
  - probabilitatea anterioara
    - probabilitatea ca un document sa apartina la un subiect
  - Probabilitatile pentru noile date de antrenare ale unui clasificator (evidence)

    $P(D|T)/(P(D|\overline{T})$ unde: D– document, T – topic

Algoritmul Naive Bayes – 2

- Estimarea acestor probabilitati se face prin masurarea frecventei de aparitie a cuvintelor intr-un set de documente de antrenare.

$$\frac{P(D|T)}{P(D|\overline{T})} = \frac{P(w_1, w_2 \ldots\ldots w_n|T)}{P(w_1, w_2 \ldots\ldots w_n|\overline{T})} \approx \frac{P(w_1|T)}{P(w_1|\overline{T})} * \frac{P(w_2|T)}{P(w_2|\overline{T})} * \ldots\ldots * \frac{P(w_n|T)}{P(w_n|\overline{T})}$$

w_k este cuvantul k din cele n cuvinte ale documententului D

- Documentele neetichetate se folosesc pentru a imbunatati setul de documente de antrenare
- Cuvintele din document sunt independente de context

Topic-Based Vector Space Model (TVSM) - 1

- Scop
  - Clasificarea documentelor folosind o abordare bazat pe spatii vectoriale
- Pasi de procesare:
  - *Eliminarea cuvintelor nerelevante (Stopwords)* – acesti termeni nu influenteaza sensul documentului
    - Exemple: si, in, ca, pana, cand,…
  - *Stemming* – reducerea formei cuvintelor la radacina
    - Exemplu: “software” -> “soft”
  - *Substitutia tezaurului de cuvinte* – inlocuirea sinonimelor cu un cuvant cheie

Topic-Based Vector Space Model (TVSM) - 2

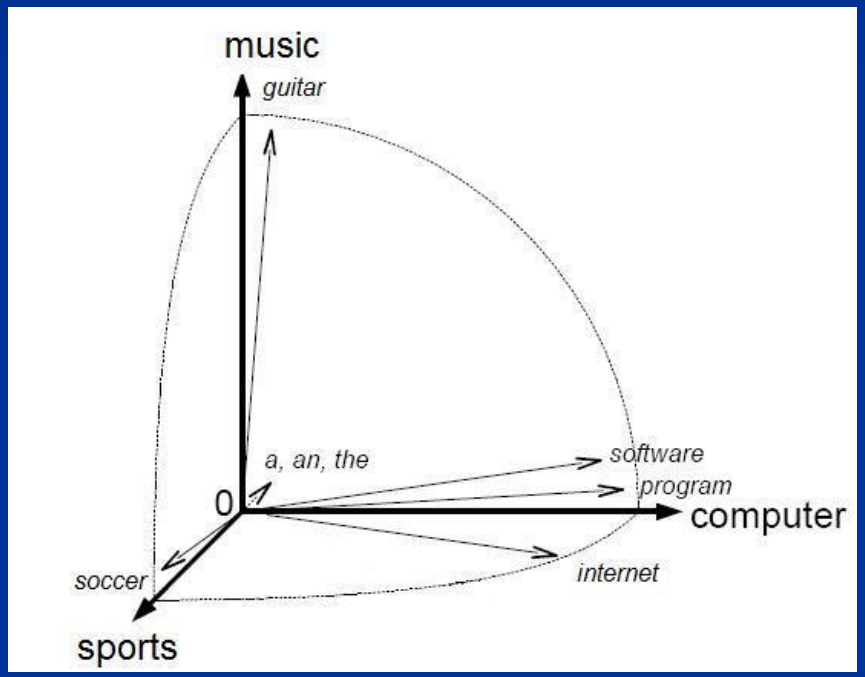
- Descriere algoritm:
  - Utilizatorul defineste un profil prin asocieaza unor documentele la clase predefinite
  - Restul documentelor se clasifica in concordanta cu documentele similare
  - Documentele noi clasificate imbunatatesc profilul
- Se presupune ca termenii (cuvintele) sunt elementele atomice ale unui document
- Similaritatea dintre doi termeni:

$$Sim(i,j) = \cos \omega_{i,j} \in [0,1].$$

$\omega_{i,j}$ – unghiul dintre vectorii termenilor *i* si *j*

- Cuvintele care apartin unui subiect anume au lungimea de vector aproape de 1

Topic-Based Vector Space Model (TVSM) - 2



- Axele
  - reprezinta subiecte elementare
  - pot avea doar valori pozitive
- Fiecarui document *k* i se asocieaza un vector d_k
- Asemanarea bazata pe subiect sim(*k*,*l*) dintre doua documente *k* si *l* este data de produsul scalar dintre vectorii documentelor respective

Concluzii

- Bibliotecile digitale sunt sisteme informatice complexe de management a continutului care extind functionalitatile bibliotecilor clasice:
  - Mediu pentru schimb de informatii si cooperare
  - Organizarea semantica a unor informatii diverse ca format
  - Acces controlat la date distribuite
- Infrastructurile Grid pot sa ofere un suport de implementare fezabil pentru bibliotecile digitale
  - Pentru distribuirea automata a datelor si a sarcinilor de procesare
  - Pentru transfer eficient de date si sincronizare
  - Pentru managementul utilizatorilor si controlul accesului
- Probleme:
  - Multe platforme GRID adopta un stil de procesare de tip “prelucrare pe loturi (batch)” in care lipseste interactivitatea
  - Programatorul aplicatiei de biblioteca este implicat in mica masura in procesul de grid-ificare (executie pe Grid).