

Índex de continguts

1. Introducció, motivacions, propòsit i objectius del projecte.....	3
1.1 Introducció.....	3
1.2 El portal de cerca conjunta: DUGi 1.0.....	4
1.3. Motivacions.....	6
1.4 Objectius del projecte.....	7
2. Estudi de viabilitat.....	8
3. Metodologia.....	9
4. Planificació.....	11
5. Marc de treball i conceptes previs.....	13
5.1 Què és un repositori.....	13
5.2 Repositoris de la Universitat de Girona.....	13
5.3 Problemàtiques del DUGi 1.0.....	15
5.4 Metadades.....	16
5.4.1 L'estàndard Dublin Core.....	17
6. Requisits del sistema.....	19
7. Estudis i decisions.....	20
7.1 Decisions principals.....	20
7.2 El protocol OAI-PMH.....	20
7.2.1 Conceptes bàsics.....	20
7.2.2 Els diferents verbs.....	22
7.2.3 Els formats de metadades suportats per OAI-PMH.....	23
7.3 Motors de cerca o Bases de Dades?.....	31
7.3.1 Índexs de cerca.....	32
7.4 Lucene.....	38
7.5 SOLR.....	40
7.5.1 Característiques.....	40
7.6 Elasticsearch.....	42
7.7 REDIS.....	43
7.8 Sistema Vista-controlador.....	44
8. Anàlisi i disseny del sistema.....	46
8.1 Anàlisi.....	47
8.1.1 Casos d'ús.....	47
8.1.2 Fitxes de casos d'ús.....	50
8.1.3 Diagrames d'activitat.....	60
8.1.4 Model de dades.....	67
8.2 Disseny.....	70
8.2.1 Diagrames de classes de la part del sistema.....	70
8.2.2 Model de navegació.....	73
8.2.3 Model de presentació.....	76
8.2.4 Model de procés.....	80
8.2.5 Diagrames de seqüència.....	84
8.2.6 Patrons.....	85
8.2.7 Interfícies d'usuari.....	87
9. Implementació i proves.....	92
9.1 Implementació.....	92
9.1.1 Detectar les autoritats conflictives.....	92
9.1.2 Llistar autors i matèries alfabèticament.....	94
9.1.3 Dates.....	95
9.1.4 Estadístiques sense Base de Dades.....	96
9.1.5 Cache.....	97

9.1.6 Multi-llenguatge.....	98
9.1.7 Sistema de logs.....	100
9.1.7 Seguretat.....	101
9.1.8 Obtenció de la url dels documents.....	103
9.1.9 SEO (Optimització per cercadors).....	104
9.1.10 Optimització.....	104
9.2 Proves.....	105
9.2.1 Detecció d'autoritats.....	105
9.2.2 Ordenació d'autors i matèries.....	108
9.2.3 Dates.....	113
9.2.3 Recol·lecció.....	116
9.2.4 Mod_rewrite.....	119
9.2.5 SSL a l'administració.....	120
9.2.6 Compresió de la sortida.....	121
9.2.7 Mod_expires.....	122
10. Implantació i resultats.....	124
10.1. Instal·lació dels serveis.....	124
10.2 Instal·lació del DUGi.....	132
10.3 Resultats.....	135
10.3.1 Resultats de la implantació.....	135
10.3.2 Grau d'assoliment dels objectius.....	135
11. Conclusions.....	139
11.1 Conclusions sobre els requisits.....	139
11.2 Desviacions del plantejament inicial.....	144
11.3 Crítica dels resultats.....	144
11.4 Valoracions personals.....	145
12. Treball futur.....	146
13. Bibliografia.....	148
14. Annexos.....	152
14.1 Annex Verbs OAI-PHM.....	152
14.2 Annex APACHE SOLR.....	160
14.2.1. Pre-requisits.....	160
14.2.2. Descarrega i instal·lació.....	160
14.2.3. Paràmetres d'indexació del camps.....	160
14.2.4. Posada en funcionament.....	161
14.2.5. Indexar ítems.....	162
14.2.6. Actualitzar l'índex.....	165
14.2.7. Consultar l'índex.....	165
14.2.8. Facetes.....	167

1. Introducció, motivacions, propòsit i objectius del projecte

1.1 Introducció

La Biblioteca de la Universitat de Girona és un dels serveis de la universitat que gaudeix de més prestigi i popularitat entre els seus usuaris. La seva missió és donar resposta a les necessitats informacionals de tots els membres de la comunitat universitària, és a dir, professors, estudiants i personal d'administració i serveis (PAS).

El seu funcionament és com a Biblioteca única. Ara bé, per donar un millor servei a l'usuari i poder funcionar realment com una biblioteca única sota unes mateixes directrius, s'estructura en tres biblioteques de Campus que donen servei als seus corresponents usuaris finals: Barri Vell, Montilivi i Campus Centre o, més coneguda com a Emili Grahí.

Per coordinar les tres biblioteques i funcionar d'enllaç amb les altres entitats amb qui col·labora en diversos projectes (CBUC¹, REBIUN², etc.) hi ha els Serveis Tècnics de Biblioteca, on es troba la Direcció i els departaments de Projectes i Comunicació, Procés Tècnic, Documentació Digital, Gestió econòmica i Informàtica.

El departament de Projectes i Comunicació és l'encarregat de desenvolupar i coordinar els projectes que impulsa la Biblioteca o aquells en els que hi participa. Entre aquests projectes trobem els repositoris, tant els institucionals com els consorciats amb el CBUC, i el Servei de Gravació i Imatge, encarregat d'enregistrar tots els esdeveniments acadèmics relacionats amb la Universitat de Girona per a la seva posterior difusió i preservació. Així mateix, aquest departament s'encarrega de les tasques de comunicació de la Biblioteca: elaboració de la memòria, tramitació de les sol·licituds d'ajuts públics per al desenvolupament de projectes i creació de les eines de difusió de la Biblioteca (punts de llibre, ex-libris dels fons especials, imatge del web de la Biblioteca i marxandatge de la Biblioteca).

El departament de Procés Tècnic és el responsable de vetllar per l'aplicació de les pautes i normatives catalogràfiques acceptades pels membres del CCUC³, així com del manteniment de la qualitat del catàleg UdG⁴, mitjançant la resolució d'incidències, actuacions de qualitat diverses i el control d'autoritats. Aquest departament també s'ocupa de la publicació en el repositori consorciat TDX⁵ de les tesis lligades a la Universitat de Girona i de la coordinació de la gestió dels Fons i col·leccions especials⁶ que custodia la Biblioteca.

El departament de Documentació Digital gestiona les subscripcions dels recursos electrònics i l'accés a la Biblioteca digital⁷, així com l'intercanvi de publicacions amb d'altres institucions.

El departament de Gestió Econòmica és l'encarregat de la gestió dels pressupostos anuals de la Biblioteca, així com l'administració dels diferents centres de cost dels departaments, pel que fa a les adquisicions de fons bibliogràfics.

Finalment, el departament d'Informàtica que, malgrat que la Universitat de Girona ja disposa d'un Servei Informàtic que dona servei a tota la comunitat universitària, es dedica a donar resposta a les

1 Consorci de Biblioteques Universitàries de Catalunya <http://www.cbuc.cat>

2 Red de Bibliotecas Universitarias <http://www.rebiun.org/>

3 Catàleg Col·lectiu de les Universitats de Catalunya <http://ccuc.cbuc.cat/>

4 Catàleg UdG <http://cataleg.udg.edu/>

5 TDX (Tesis Doctorals en Xarxa) <http://www.tdx.cat/>

6 Fons Especials UdG <http://www.udg.edu/fonsespecialsbiblioteca>

7 Biblioteca digital UdG <http://www.udg.edu/biblioteca/Bibliotecadigital>

necessitats específiques de la Biblioteca, tant de manteniment del nombrós parc informàtic de què disposen les biblioteques, com de l'administració dels sistemes de gestió integrats de la Biblioteca, el manteniment dels servidors, així com també dels nombrosos projectes en què participa la Biblioteca.

Un dels projectes més ambiciosos i més recents en els que participa la Biblioteca és en la recollida, preservació i difusió de la producció científica realitzada a la Universitat de Girona. Això és possible a través de la publicació d'aquesta producció en repositoris d'accés obert.

La Biblioteca participa en dos tipus de repositoris, en primer lloc els institucionals, que són els creats i gestionats íntegrament per ella mateixa, i en segon lloc els repositoris consorciats en els que participa juntament amb d'altres institucions.

La necessitat de crear repositoris institucionals neix de la constatació que la recerca produïda dins el marc d'institucions públiques, com les universitats, i amb finançament públic era captiva del mercat dels grans distribuïdors privats de revistes i publicacions científiques. Es tractava de poder donar una difusió molt més àmplia a aquesta recerca, ajudar als autors a la seva publicació més enllà de les limitacions de criteris d'impacte bibliomètric, tot ajudant a generar més coneixement i recerca amb la compartició d'aquesta recerca i els seus resultats en obert. Aquesta forma de publicació, molt més directa i amb la participació activa de la universitat, contribueix també al prestigi institucional, cosa que fins ara quedava molt diluïda pels distribuïdors comercials.

El portal de cerca conjunta als documents digitals de la UdG (DUGi), és el punt d'accés als materials en obert, té el seu origen en la necessitat d'unificar tota aquesta producció universitària sota un punt d'accés comú per facilitar-ne la consulta per part dels usuaris i millorar-ne la visibilitat.

1.2 El portal de cerca conjunta: DUGi 1.0

El DUGi sorgeix de la necessitat de donar als nostres usuaris un sol punt de cerca i recuperació als repositoris de la Universitat de Girona, és a dir, un lloc web on es puguin trobar tots els nostres recursos sense haver d'estar saltant de repositori en repositori. L'usuari final, desconeix l'organització interna dels repositoris i les seves diferències. Ell només vol trobar un document independentment d'on sigui.

Per donar aquest accés únic, el portal necessita indexar les metadades descriptives de cada ítem, però no necessita el document. Aquest estarà preservat al repositori. Per tant, només en necessitem el **handle** per poder apuntar-lo.

La idea no és nova, i per això com a títol d'aquest projecte proposem “DUGi 2.0”. Això fa entreveure que en algun moment hi ha hagut un DUGi 1.0, que intentarem explicar seguidament.

Recapitulant una mica, tenim 3 repositoris independents gestionats per la Biblioteca de la Universitat de Girona, i 2 més de consorciats on dipositem les nostres Tesis i les nostres Revistes.

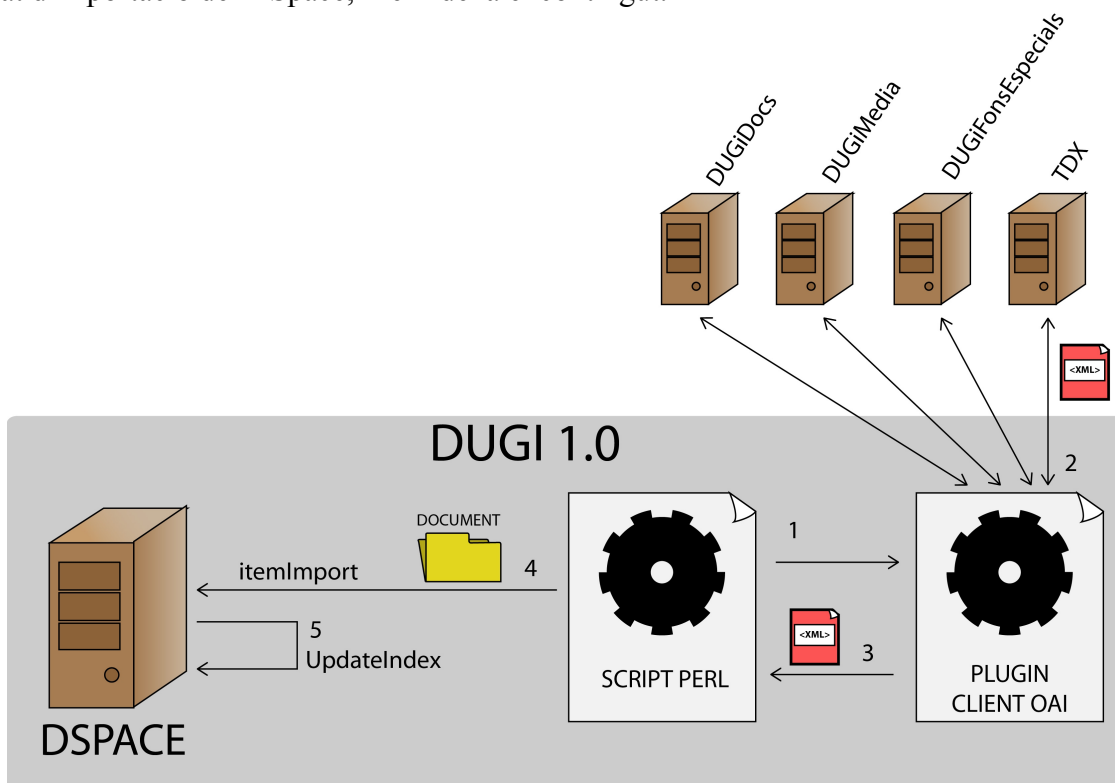
Al DUGi 1.0, la solució proposada va ser utilitzar el mateix programari **DSpace**, que utilitzem per als nostres repositoris, per indexar i servir les dades.

Però necessitàvem una manera de recollir les dades dels repositoris i mantenir-nos al dia de les possibles actualitzacions d'aquestes dades.

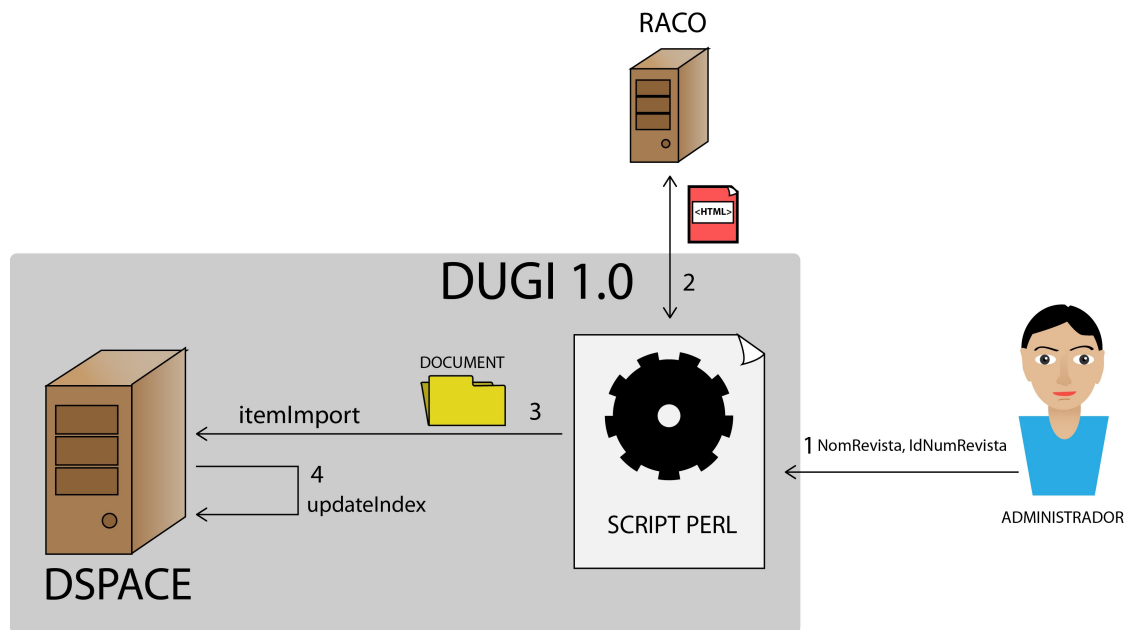
Vam trobar un connector que implementava un client del protocol de **harvesting OAI-PMH** (Open Archives Initiative-Protocol Metadata Harvesting), que permet als repositoris servir les seves dades utilitzant fitxers **XML**.

Finalment vam implementar un script que cridava aquest connector, en recollia les dades, creava els fitxers i directoris necessaris i finalment feia una importació al repositori **Dspace**. Tot plegat lligat amb el dimoni **CRON**, temporitzador del sistema, per aconseguir anar recol·lectant tots els nous ítems.

A la gràfica següent podem veure el procés. L'script de perl que controla tot el procés de recol·lecció. Primer crida el plugin que fa de client d'OAI, i recol·lecta cada repositori. Els repositoris retornen un document en format XML que el plugin processa i retorna les metadades de cada ítem. Seguidament, l'script de perl processa cada ítem i genera una carpeta amb el fitxer d'XML en el format de metadades d'importació del DSpace. Finalment importa els ítems amb l'utilitat d'importació del DSpace, i re-indexa el contingut.



En el cas de la recol·lecció del repositori RACO, on tenim dipositades les nostres revistes, el procés és una mica diferent, ja que el repositori no disposa de servidor d'OAI-PMH. En aquest cas, el procés s'ha de fer manualment. A la gràfica següent s'il·lustra aquesta recol·lecció, que comença quan l'administrador crida l'script de perl passant-li per paràmetre el nom de la revista i l'identificador del número de revista. Amb aquestes dades l'script de perl parseja l'HTML del RACO i genera una carpeta amb un fitxer XML amb el format de metadades d'importació de DSpace, que finalment s'importa i es re-indexa el contingut.



El DUGi 1.0 té algunes problemàtiques que veurem detalladament a l'apartat 5.3 d'aquesta memòria. Aquestes problemàtiques ens han portat a buscar una nova solució que és la que es presenta en aquest projecte de final de carrera. De forma resumida aquests problemes són:

- El client no entén els atributs d'alguns tags d'XML.
- El client no detecta més d'un tag d'XML en una mateixa línia.
- Sobren moltes característiques del DSpace.
- El DSpace consumeix masses recursos.
- Necessitem acoblar masses tecnologies.
- Es perd el qualificador de les metadades.
- La interfície no és clara ni usable.
- No existeix un control d'autoritats.
- Les opcions de cerca no ens convencen.

1.3. Motivacions

Per tal d'oferir el millor servei possible, tant als usuaris com als autors, hem cregut necessari implementar un portal que permeti trobar tots els materials d'un autor o una temàtica determinats agrupats en un mateix lloc web.

La feina de proporcionar a aquests materials unes metadades que ens permetin tenir-los agrupats per unes determinades característiques, i a més a més de proporcionar-ne resums, descripcions de formats, documents relacionats, ja està feta pels bibliotecaris en el moment d'entrar els materials als repositoris i pretenem aprofitar-la per augmentar la capacitat de recuperació i enriquir la consulta d'aquests documents.

També creiem que per augmentar el prestigi tant dels nostres professors com de la universitat mateixa, hem de proporcionar una eina que permeti als principals cercadors trobar la informació necessària per indexar aquests materials i que tinguin una molt alta visibilitat dins del web.

1.4 Objectius del projecte

L'objectiu d'aquest PFC és desenvolupar un portal web amb un sistema d'indexació que permeti la cerca per diferents criteris com, per exemple, autor, matèria i títol.

L'entorn ha de tenir un disseny agradable per a l'usuari final i ha de permetre obtenir les dades del repositori que se li indiquin. També ha de ser capaç de detectar els possibles duplicats en els criteris de cerca i notificar-ho a l'administrador.

Per tal d'aconseguir aquest objectiu, ens hem proposat les següents tasques:

- Fer un estudi de les eines més interessants per implementar aquest projecte.
- Trobar una metodologia que ens pugui servir per dissenyar i implementar el portal d'una manera eficient.
- Dissenyar l'aplicació
- Implementar-la
- Fer la instal·lació tant de l'entorn de treball com de l'entorn de producció
- Fer el testeig
- Fer una optimització
- Fer un estudi de les principals amenaces de seguretat que ens poden afectar, i en cas d'haver-n'hi corregir-les.

2. Estudi de viabilitat

Pel que fa a la viabilitat econòmica del projecte, la Biblioteca, com tots els serveis de la Universitat, té assignat un pressupost anual. Aquest pressupost es distribueix entre els projectes que es creu que poden ser interessants de desenvolupar, com per exemple, les implementacions de nous serveis, els projectes de digitalització, l'adquisició de fons bibliogràfics, les actualitzacions de maquinari, etc.

Quan es va pensar en la necessitat d'aquest portal, es va decidir donar-hi prioritat màxima, però sense deixar de banda les altres necessitats del departament, com els canvis de versió dels repositoris existents.

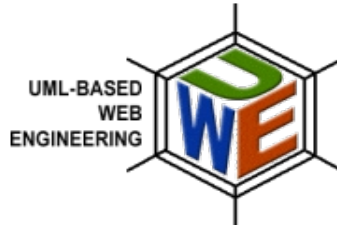
En quant a la viabilitat tecnològica, la idea era utilitzar les eines lliures que es troben disponibles al web, i pel que fa a l'entorn de producció i desenvolupament, es comptava amb un servidor de màquines virtuals que ens permetia disposar de les màquines necessàries sense que suposés gaire esforç.

Així doncs, en veure que el projecte era viable tant econòmicament com tecnològicament es va decidir de tirar-lo endavant.

3. Metodologia

Al ser una aplicació web, necessitem una metodologia que ens permeti definir bé tant la interfície com la navegació, ja que en el nostre cas aquesta és molt important.

Amb aquests requeriments, la metodologia escollida és UWE (UML-Based Web Engineering).

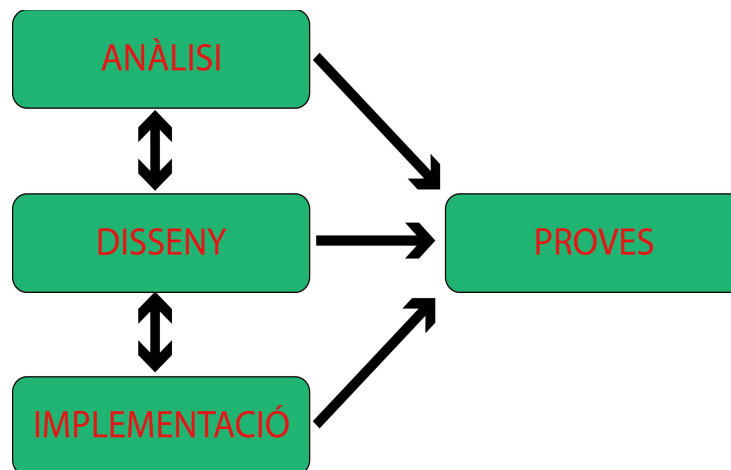


L'UWE és una proposta basada en UML per modelar aplicacions web. Les fases proposades per un projecte definit en UWE són:

- Anàlisi
- Disseny
- Implementació
- Proves

Per modelar el nostre projecte, ens anirem movent d'anàlisi a disseny i a implementació iterativament, perquè normalment ens cal redefinir la capa superior quan ens trobem amb alguna nova necessitat.

El diagrama següent mostra la relació entre les fases del projecte.



A la fase d'anàlisi l'UWE ens aporta uns estereotips nous per als diagrames de casos d'ús i per als diagrames d'activitat.

A la imatge següent tenim els estereotips que s'afegeixen als diagrames de casos d'ús.



WebUseCase

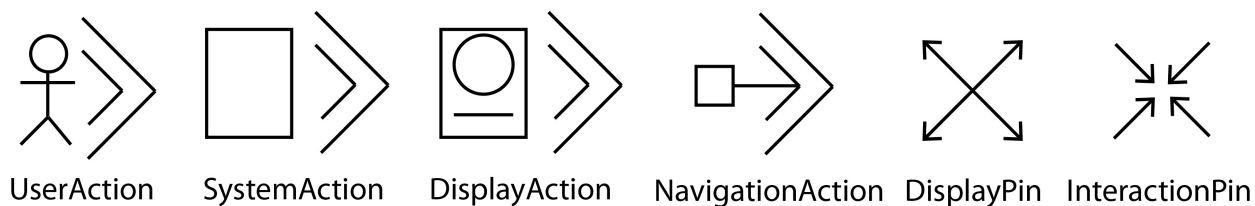


Processing



Browsing

En aquest altre diagrama hi ha els estereotips que s'utilitzen per als diagrames d'activitat de l'anàlisi.

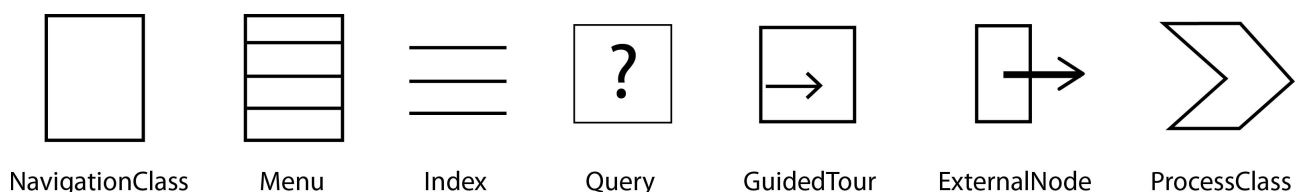


A la fase de disseny l'UWE defineix 4 lògiques:

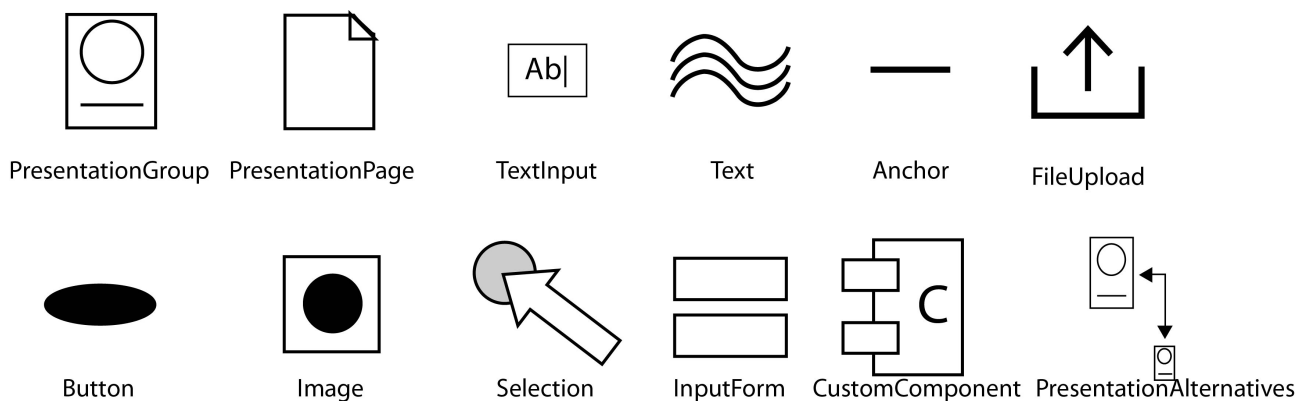
- Lògica de contingut
- Lògica de navegació
- Lògica de presentació
- Lògica de procés

Els tres primers models estan formats per diagrames de classes i en el model de procés hi tenim diagrames d'activitat, però tant en el de navegació, de presentació i procés tenim uns estereotips nous que ens permeten definir millor aquestes parts.

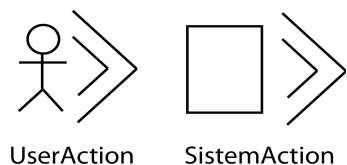
La imatge següent mostra els estereotips que s'afegeixen per a definir la navegació.



Aquesta altra mostra els que ens permeten definir la capa de presentació.



Finalment aquests són els estereotips utilitzats per definir la capa de procés.



Per al nostre projecte utilitzarem UML per a la part de recol·lecció de metadades dels repositoris i UWE per a la part de consulta.

4. Planificació

Per tal de no descuidar les meves tasques habituals al departament de Projectes de la Biblioteca, podré dedicar-hi 3 matins a la setmana aproximadament de les 8 del matí a les dues de la tarda. Els dies escollits per poder alternar aquest projecte amb les meves feines habituals seran dilluns, dimecres i dijous. Així doncs, a partir d'ara això serà una setmana dins d'aquesta planificació.

Les tasques necessàries per a dur a terme aquest projecte són les següents:

1. Anàlisi dels requisits del sistema (1 setmana)

Esperem definir totalment les funcionalitats necessàries per al nostre sistema, així com crear els casos d'ús i les fitxes de casos d'ús que ens serviran per al disseny de l'aplicació

2. Estudi de les eines necessàries (2 setmanes)

Pretenem escollir les eines que ens serviran per a la implementació del sistema. Entre elles en destaquem la tria del llenguatge de programació, així com un sistema de plantilles per separar la capa de negoci de la de presentació, i la tria d'un sistema de dades que ens permeti indexar i recuperar la informació d'una manera ràpida.

En aquest punt també entraria l'estudi de les metadades que utilitzarem per a definir els materials, però no ho comptabilitzem ja que aquest estudi ha estat fet durant els últims 3 anys al ser jo personalment l'administrador dels repositoris que gestiona la biblioteca.

Proposem aquest punt en aquesta posició i just abans de la implementació perquè l'elecció del sistema de base de dades ens repercutirà molt sobre el disseny conceptual.

3. Disseny conceptual (2 setmanes)

Intentarem dissenyar l'estructura de la informació amb la qual interactuarà la nostra aplicació. També hauran de quedar definits els mòduls que formaran l'aplicació.

El nostre objectiu en aquest punt és tenir fets els diagrames que ens permetran implementar la capa de negoci.

4. Disseny navegacional (3 setmanes)

Farem els diagrames que ens representaran la interacció de l'usuari amb la nostra aplicació.

Aquí definirem les pàgines i menús necessaris per oferir la informació d'una manera senzilla i amena.

5. Disseny de la presentació (3 setmanes)

En principi proposem 3 setmanes per a aquesta fase en la que també hi inclourem els diagrames de la capa de procés, ja que aquests seran poquets. A més a més, intentarem aplicar tècniques de disseny gràfic per a dissenyar una interfície agradable a la vista i col·locant la informació i els enllaços pensant molt en la usabilitat.

6. Instal·lació de la màquina (1 setmana)

Un cop escollides les eines, no ha de ser complicat instal·lar el servidor web, el mòdul per a lligar-lo amb el llenguatge de programació, ni la base de dades.

A no ser que ens sorgeixi algun problema inesperat en la configuració no hauríem de tenir cap problema per muntar-ho tot en una setmana.

La definició del model de dades dins la base de dades, la considerem part de la implementació.

7. Implementació (6-7 setmanes)

Crearem el model de dades necessari dins del sistema de base de dades escollit.

Implementarem el cos principal del sistema, així com tots els mòduls necessaris definits en el disseny conceptual.

Intentarem aplicar totes les normes per aconseguir un software de qualitat i sense errors.

Aquesta pot ser la part més llarga del projecte, ja que tot i que algunes parts seran molt repetitives caldrà fer diversos mòduls bastant diferenciats.

8. Testeig (1-2 setmana)

En una aplicació web és molt important aquesta fase ja que aquesta estarà oberta a tothom, i no hi poden haver errors que permetin una intrusió al sistema.

Creiem que en una entitat com una universitat cal assegurar la integritat i l'accés solament als usuaris autoritzats a la nostra aplicació ja que podria ser un punt d'entrada a altres sistemes més importants per al funcionament de la universitat, i no ens podem permetre deixar-ho a l'atzar.

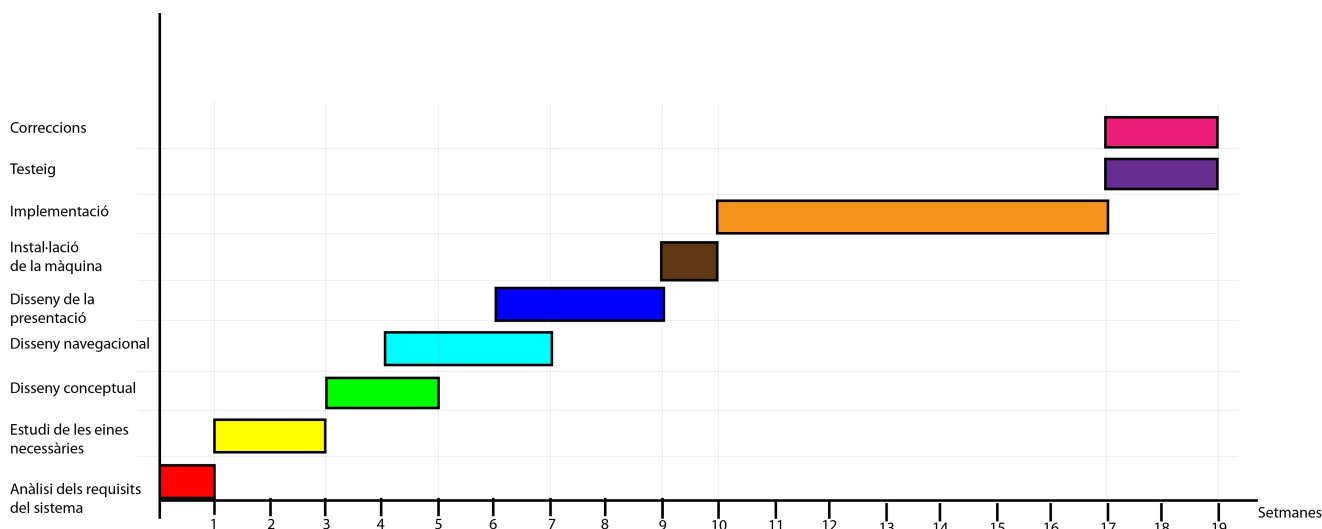
Així doncs, testejarem bé totes les pantalles de l'aplicació i comprovarem els temps de resposta per a assegurar-nos que la informació es retorna en un temps raonable.

També passarem una aplicació per a la detecció d'errors comuns a l'hora de programar amb el llenguatge escollit.

9. Correccions (0-2 setmanes)

En cas d'haver trobat coses a millorar en el punt anterior farem les correccions necessàries en aquest punt. En el millor dels casos no ens caldrà fer aquesta fase.

Els següent diagrames de Gantt ens mostra la planificació del projecte per setmanes i les tasques en paral·lel.



5. Marc de treball i conceptes previs

5.1 Què és un repositori

Un repositori és un sistema informàtic que permet emmagatzemar i difondre informació i alhora n'ha de garantir la preservació.

Aquesta informació s'ha d'organitzar d'alguna manera perquè la seva recuperació sigui fàcil i ràpida. En aquest punt sorgeix la necessitat d'utilitzar les metadades descriptives.

Al mateix temps, per garantir la preservació necessitem tenir ben definits uns protocols i sistemes de verificació, és a dir, les funcions de resum (hash⁸), les sumes de verificació (checksum), les metadades de preservació i les còpies de seguretat múltiples.

També necessitem un sistema d'indexació i cerca de les metadades i un sistema per servir-les (per exemple, un servidor web, front-end).

Anomenarem ítem a cada objecte digital format per aquestes metadades descriptives i documents en qüestió.

Per tal de poder recuperar-los, aquests ítems tenen un **identificador uniforme de recursos** (URI), únic i permanent, que a partir d'ara anomenarem **handle**.

5.2 Repositoris de la Universitat de Girona

En el marc universitari els repositoris preserven i difonen la producció científica de la universitat. La universitat de Girona té dos repositoris: DUGiDocs <<http://dugi-doc.udg.edu/>> i DUGiMedia <<http://diobma.udg.edu/>>.

El DUGiDocs conté articles de recerca, projectes de final de carrera, presentacions, materials docents, publicacions i documents institucionals.

El DUGiMedia conté els àudios i vídeos dels actes que es celebren a la UdG, així com algunes classes i materials per ajudar a la docència.

A més a més, la Biblioteca té un altre repositori on hi guarda els seus fons especials per a donar-los difusió quan estan digitalitzats. Aquest repositori s'anomena DUGiFonsEspecials <<http://dugifonsespecials.udg.edu/>>. Normalment amb aquests fons es necessiten permisos especials, sobretot en el cas de les correspondències, i es van obrir a mesura que s'obtenen els permisos.

Finalment tenim un altre repositori d'imatges, que encara està en desenvolupament. Aquest s'anomena DUGiImatges, i contindrà digitalitzacions de fons gràfics com per exemple pòsters i diapositives.

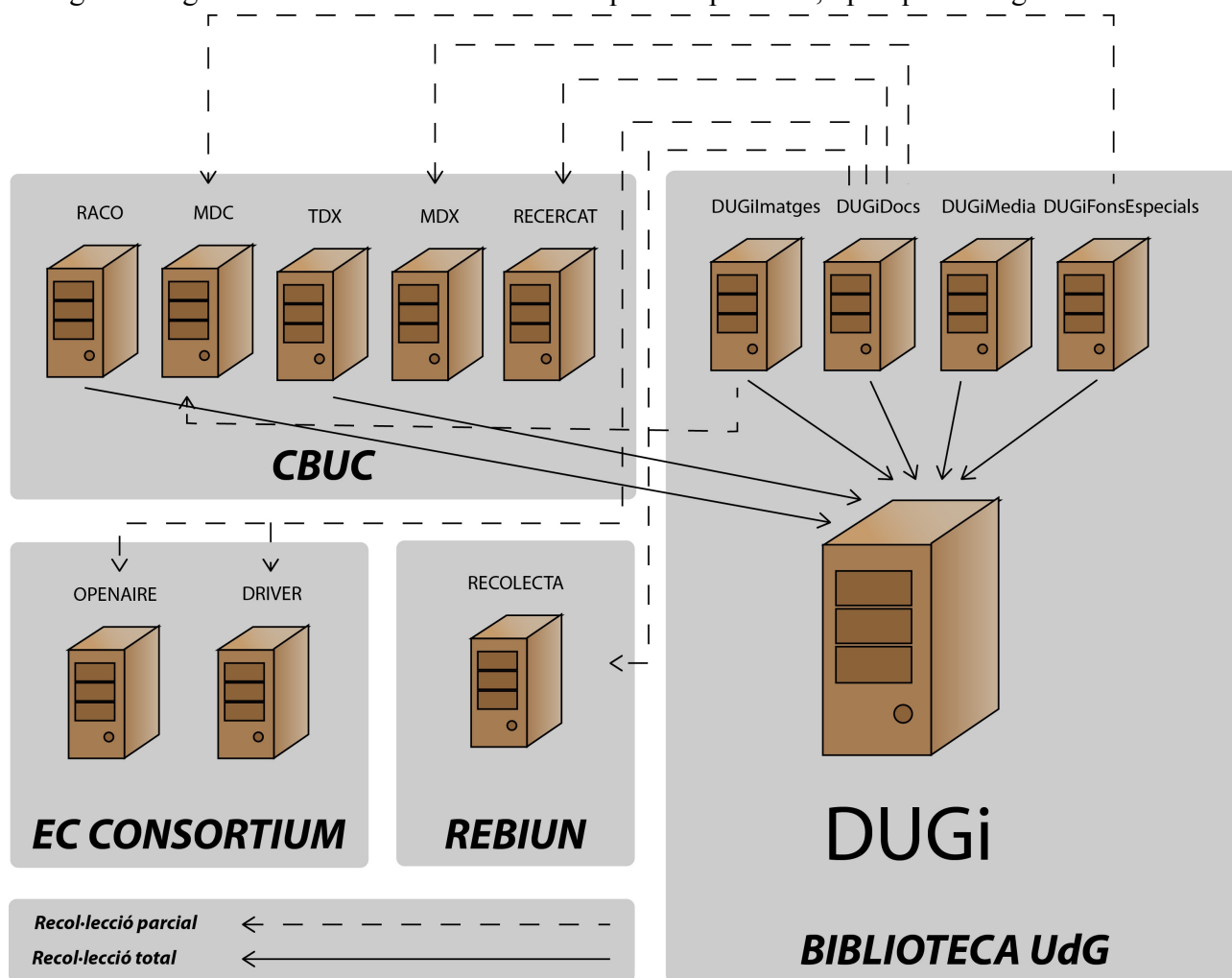
⁸ Les funcions de Hash, també anomenades funcions de resum, proporcionen un resultat numèric sobre una cadena de caràcters o un fitxer. Es solen utilitzar tant per comprovar la integritat de fitxers, com per guardar contrassenyes quan són funcions d'un sol sentit. El seu principal problema és que poden tenir col·lisions. Les més utilitzades són CRC32, MD5, SHA-1, SHA-2 i normalment es codifica la sortida en base16 o hexadecimal.

La Biblioteca com a representant de la Universitat també gestiona els continguts d'altres repositoris com són els consorciats. A nivell català tenim:

- TDX (Tesis Doctorals en Xarxa) <<http://www.tdx.cat/>> que conté les tesis llegides a la UdG.
- RACO (Revistes Catalanes en Accés Obert) <<http://www.raco.cat>> que conté les revistes de producció gironina en general.
- Recercat (Dipòsit de Recerca de Catalunya) <<http://www.recercat.cat/>> que recol·lecta els materials de recerca del repositoris de les universitats.
- MDX (Materials docents en Xarxa) <<http://www.mdx.cat/>> que recol·lecta els materials docents dels repositoris de les universitats.
- MDC (Memòria Digital de Catalunya) <<http://mdc.cbuc.cat/>> que conté col·leccions digitalitzades relacionades amb Catalunya i el seu patrimoni.

A més a més, els documents dels repositoris també són recol·lectats per altres repositoris com són Recolecta (Recolector de ciencia abierta) <<http://www.recolecta.net>> gestionat per REBIUN (consorci de biblioteques universitàries espanyoles), Driver (Digital Repository Infrastructure Vision for European Research) <<http://www.driver-repository.eu/>> i OPENAIRE (Open Access Infrastructure Research for Europe) <<http://www.openaire.eu/>>, aquests dos últims d'àmbit europeu.

El següent diagrama mostra les relacions entre aquests repositoris, i per qui estan gestionats.



5.3 Problemàtiques del DUGi 1.0

El DUGi 1.0 té algunes problemàtiques que veurem detalladament, en aquest punt, i que ens han fet acabar decidint-nos per buscar una nova solució.

Entre elles podríem destacar:

- El client d'OAI-PMH no entén els atributs dels tags d'XML, i per tant no detecta els ítems esborrats del repositori.

Exemple:

```
<record>
<header status="deleted">
<identifier>oai:diobma.udg.edu:10256.1/4</identifier>
<timestamp>2011-09-09T12:39:07Z</timestamp>
<setSpec>hdl_10256.1_3</setSpec>
</header>
</record>
```

En l'exemple anterior es pot veure com el tag **header** utilitza l'atribut **status** amb el valor **deleted**, que permet detectar els ítems que han estat eliminats. El nostre client, al no parsejar els atributs, no ens permet detectar aquests ítems.

- El client d'OAI-PMH falla per alguns repositoris, en que el seu servidor d'OAI-PMH no feia servir salts de línia entre els tags XML i els deixava buits en cas de no contenir una certa metadada, enlloc de no mostrar-los.

Exemple:

```
<dc:date>2004-01-11</dc:date><dc:type>info:eu-
repo/semantics/article</dc:type><dc:type>info:eu-
repo/semantics/publishedVersion</dc:type><dc:format>application/pdf</dc:format><dc:sub
ject/><dc:subject/><dc:subject/>
```

En l'exemple anterior podem veure com no s'utilitzen salts de línia per separar cada tag. Un client que no parsegi l'XML correctament, podria agafar els tres tags **dc:type** com si fossin un de sol. A més a més, es pot observar que el servidor ens envia tags buits (**dc:subject**), que no caldria posar-los, i el client al parsejar-los ens crea metadades buides.

- Sobren moltes característiques del repositori Dspace.

Per exemple:

- Flux de submissió
- Preservació
- Comunitats i col·leccions
- Permisos, usuaris i grups
- Identificació
- Servidor d'OAI-PMH

Estem utilitzant només un 20% del seu potencial.

- El repositori Dspace consumeix molts recursos tot i no utilitza totes les seves

característiques.

Necessita un servidor web Tomcat, una base de dades Postgres...

- La solució és com un puzle, tenim un Dspace per un cantó, un client per recol·lectar i un script en perl que fa la submissió al repositori.
- No existeix un control d'autoritats per evitar la creació d'entrades repetides.

Martínez, Pilar

Martínez García, Pilar

Martinez, Pilar

...

- Les opcions de cerca són pobres i farragoses.

La cerca simple pot no ser suficient, en canvi la cerca avançada és molt lenta i complicada.

- La interfície no és amable ni gaire usable. Molts usuaris es perden.
- La versió d'OAI-PMH perd els qualificadors de les metadades, i per tant es creen confusions.

Exemple:

```
<dc:format>200286810 bytes</dc:format>  
<dc:format>video/x-ms-wmv</dc:format>
```

En aquest exemple, al crear l'ítem, confondríem les metadades **dc.format** amb **dc.format.mimetype**.

En aquest projecte de final de carrera que es presenta es desenvoluparà un portal web que permetrà solucionar el sistema actual.

5.4 Metadades

Per tal de descriure els documents dipositats als repositoris d'utilitzen les metadades. Així doncs tenim que un ítem és el conjunt d'uns documents més unes metadades tant descriptives com de preservació.

Normalment les metadades de preservació són generades per el mateix repositori, ja que són necessàries per cada document que formi part de l'ítem. Aquestes poden ser el format de l'arxiu, el **checksum** per controlar-ne la integritat, etc.

Hi ha molt tipus de metadades de descripció segons el nivell descriptiu que necessitem. Per exemple en els catàlegs bibliogràfics es necessita una descripció molt més acurada i s'usa l'estàndard MARC⁹. En el món dels repositoris normalment s'utilitza l'estàndard Dublin Core, que conté les metadades per cercar i descriure l'ítem correctament.

⁹ Estàndard de metadades que proporciona tota la informació d'un registre bibliogràfic. És l'estàndard utilitzat en els catàlegs bibliogràfics en els seus diferents formats. Les seves sigles responen a MACHine-Readable Cataloguing.

5.4.1 L'estàndard Dublin Core

El Dublin Core va ser creat el 4 d'octubre de 2005 per Pete Johnston amb el propòsit de poder definir els objectes digitals. Consta de 15 definicions semàntiques descriptives que ens permeten, com ja hem dit, definir semànticament un ítem.

Cada una d'aquestes definicions és opcional, es pot repetir i pot aparèixer en qualsevol ordre. També pot contenir qualificadors per a descriure una mica més mesuradament aquesta característica.

Aquestes metadades estan agrupades segons el seu propòsit. Poden ser metadades de contingut, de propietat intel·lectual o de instanciació del recurs. Vegem-les:

Metadades de contingut:

- **dc.title**
Conté el nom del recurs.
Per exemple: títol, títol secundari, títol en anglès...
- **dc.subject**
Descriu les paraules clau o matèries del recurs.
Normalment estan dins d'un sistema de vocabulari controlat.
- **dc.description**
Conté una descripció textual del recurs.
Per exemple: resum, taula de continguts...
- **dc.source**
Conté el nom del recurs del qual deriva el recurs descrit.
Per exemple: identificador o URI del recurs pare.
- **dc.language**
El llenguatge del recurs.
Per exemple: cat, jap, eng o la ISO del format utilitzat per descriure la llengua.
- **dc.relation**
Els recursos relacionats.
Per exemple: URI de l'objecte relacionat.
- **dc.coverage**
Representa la cobertura espacial o temporal del contingut intel·lectual del recurs.
Per exemple: Ubicació geospacial, període cronològic, mostra de recerca (gènere, edat, etc.)

Metadades de propietat intel·lectual:

- **dc.creator**
La persona o entitats creadores del recurs.
Per exemple: autor en el cas d'un article, fotògraf en el cas d'una imatge...
- **dc.contributor**
Una persona o entitat que hagi col·laborat significativament en la creació del recurs, però

que la seva col·laboració sigui secundària en relació a la del creador.
Per exemple: Director del projecte, il·lustradors, traductors...

- **dc.publisher**
L'editor del recurs. L'entitat responsable de que el recurs sigui accessible en el seu format actual.
- **dc.rights**
Ens dona informació sobre els termes i condicions d'accés al recurs.
Per exemple: Llicència Creative Commons, Tots els drets reservats...

Metadades de instanciació:

- **dc.format**
El format de dades del recurs descrit.
Per exemple: Document PDF, Imatge JPEG, Vídeo Flash, Vídeo MP4, Àudio MP3...
- **dc.type**
El tipus de recurs descrit.
Per exemple: Imatge, vídeo, àudio, article, ressenya...
- **dc.identifier**
Engloba tots els identificadors del recurs.
Per exemple: URIs, ISBN, ISSN, DOI, cita...
- **dc.date**
La data en la qual el recurs es va posar a disposició de l'usuari en la seva forma actual.
Per exemple: Data de creació, data de publicació...

6. Requisits del sistema

Veient les problemàtiques del sistema anterior ens en donem compte de que necessitem un sistema que sigui capaç de solucionar tots aquests problemes. Per això els requisits que hem definit per aquest projecte són:

- Crear un sistema que ens permeti unificar les entrades d'autors per evitar les confusions en els resultats.
- Implementar un client d'OAI-PMH completament estàndard i que ens permeti aprofitar totes les opcions del protocol.
- Enriquir les cerques donant opcions de filtre i refinament.
- Proporcionar un protocol de cerca prou potent per anar de la cerca més senzilla a la més complexa.
- Simplificar al màxim la interfície per aconseguir una alta usabilitat.
- Donar la màxima visibilitat als autors, matèries i documents.
- Implementar un sistema de memòria cau per fer les cerques més ràpides.
- Ha de tenir una interfície multilingüe. Com a mínim ha de donar suport per als idiomes català, espanyol i anglès.
- Ha de ser segur.
- Ha de consumir pocs recursos.
- Ha d'estar format de mòduls reutilitzables.

7. Estudis i decisions

7.1 Decisions principals

Utilitzarem un Ubuntu 10.04 LTS que és molt estable i deriva del sistema Debian, amb lo qual té les mateixes facilitats d'instal·lació i configuració.

Com a servidor web utilitzarem Apache2 perquè no consumeix gaires recursos, és altament configurable i es pot instal·lar fàcilment els mòduls que necessitem.

Com a llenguatge de programació hem triat PHP 5.2 ja que és un llenguatge totalment pensat per al web, molt potent i orientat a objectes.

Per tal de dur a bon terme aquest projecte necessitem estudiar les següents eines:

- Protocol OAI-PMH. (Apartat 7.2)
- Base de Dades i Índexs. (Apartats 7.3, 7.4 i 7.5)
- Un sistema per mostrar plantilles senzill, per crear el sistema vista-controlador. (Apartat 7.8)

7.2 El protocol OAI-PMH

El protocol OAI-PMH és una part molt important del nostre projecte, més o menys la meitat del nostre projecte es basa en aquest protocol. Per això ens interessa fer-ne un estudi bastant profund.

7.2.1 Conceptes bàsics

L'Open Archives Initiative Protocol for Metadata Harvesting, a partir d'ara OAI-PMH, proporciona un llenguatge per a servir i recol·lectar metadades. Les aplicacions participants poden interactuar entre elles en dos rols: servidors o clients.

El client, o també anomenat recol·lector (harvester en anglès), fa peticions al servidor, recopila les dades, i les utilitza.

El servidor, normalment un repositori, està preparat per respondre a les 6 peticions que estableix el protocol d'OAI-PMH. L'objectiu principal del protocol és servir els ítems del repositori, així doncs l'OAI-PMH estableix que un registre d'ítem conté una capçalera o header amb un identificador únic (d'ús només per a OAI-PMH) i un cos que contindrà les respectives metadades de cada ítem.

Totes les dades es serveixen en format XML, sota el protocol HTTP i tant els noms de les peticions, que anomenarem **verbs**, com tots els altres paràmetres de les URL dels servidors d'OAI-PMH distingeixen majúscules de minúscules és a dir, són **case sensitive**.

URL de petició al servidor OAI-PMH té aquesta estructura:

http://(servidor)/(camí a l'aplicació)/(script servidor)?verb=(verb)[&altres paràmetres variables segons els verbs]

Aquests paràmetres variables són:

- identifier
- from
- until
- set
- metadataPrefix
- resumptionToken.

Els explicarem a mesura que anem explicant els verbs en els quals participen.

Totes les respostes venen encapsulades de la següent forma:

```
<OAI-PMH ...>
<responseDate>DATA</responseDate>
<request>URL</request>
COS
</OAI-PMH>
```

El cos, evidentment, canvia segons la petició. Tenim 7 possibles tags de cos, que són els mateixos que els verbs més el tag d'error.

Com hem explicat anteriorment, el protocol és case sensitive. Anem a provocar-li un error per mostrar-ne la sortida.

Fent la petició: <http://diobma.udg.edu/dspace-oai/request?verb=ListmetadataFormats>
(amb m minúscula)

Ens repon:

```
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2012-11-07T16:48:13Z</responseDate>
  <request>http://diobma.udg.edu/dspace-oai/request</request>
  <error code="badVerb">Illegal verb</error>
</OAI-PMH>
```

Els codis d'error que proporciona el servidor OAI-PMH són els següents:

- **cannotDisseminateFormat**
El retorna quan no és capaç de trobar cap format de metadades que correspongui amb el proporcionat.
- **idDoesNotExist**
Quan l'identificador que buscàvem no existeix.

- **badArgument**
Quan falten un o més atributs o el seu format no és correcte.
- **badVerb**
Quan utilitzem un verb que no està definit, o hem fat algun error tipogràfic.
- **noMetadataFormats**
Si el servidor no té cap format de metadades configurat.
- **noRecordsMatch**
Aquest error ens el retorna quan no hi ha cap ítem que satisfaci la nostra selecció.
- **badResumptionToken**
Quan el nostre resumptionToken no té l'estructura adequada, o les seves parts no respecten el format dels atributs als quals representa.
- **noSetHierarchy**
Els sets poden agrupar-se en una estructura jeràrquica, si intentem accedir la jerarquia d'un set que no en té, ens retorna l'error anterior.

7.2.2 Els diferents verbs

L'OAI-PMH ens defineix els següents 7 verbs:

- Identify
- ListMetadataFormats
- ListIdentifiers
- ListSets
- ListRecords
- GetRecord

En l'ANNEX1 es fa una descripció detallada de cadascun d'ells.

En el nostre projecte només utilitzarem ListRecords i GetRecord, ja que, tant la *identificació* com els *formats de metadades* suportats i els *sets* continguts, ja ens seran coneguts en el moment de la configuració del servidor per a la seva recol·lecció. Tampoc tenim necessitat de llistar els identificadors ja que els recol·lectem directament.

7.2.3 Els formats de metadades suportats per OAI-PMH

Un servidor d'OAI-PMH pot suportar tots els formats de metadades que es vulgui, sempre i quan es proporcioni un fixer d'estils (XSL) per transformar les metadades del repositori al format escollit i a vegades un connector.

Normalment els formats més utilitzats són OAI-PMH, DIDL i METS amb PREMIS.

OAI-DC

L'OAI_DC neix el 4 d'octubre del 2005 de la mà de Pete Johnston, el creador de Dublin Core, i és un format XML per a la serialització de metadades Dublin Core simples i les seves sigles representen OAI Dublin Core. Va estar dissenyat explícitament per ser usat com a format de metadades de l'OAI-PMH.

És el format més utilitzat, i el més senzill.

Les metadades s'encapsulen dins de 15 tags definits pel nom de la metadada DublinCore.

Per exemple:

```
<dc:creator>Olivé Comadira, Oriol</dc:creator>
```

Com es pot deduir, i ja hem comentat anteriorment, el principal problema d'aquest format és que es perd el qualificador de la metadada, d'aquí que es digui que serialitza Dublin Core simple.

A més a més, aquest format només llista les metadades de l'ítem, en cap cas els noms ni URLS dels fitxers continguts a l'ítem.

Exemple de registre OAI_DC:

```
<metadata>
  <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
  <dc:title>
    Anàlisi del Cicle de Vida (ACV) del Monoestearat de Glicerina a CRODA Ibérica, S.A.
  - Mevisa Site
  </dc:title>
  <dc:creator>Pacheco Bardera, Albert</dc:creator>
  <dc:creator>Oliver Serra, Arnau</dc:creator>
  <dc:creator>Padullés Cubino, Josep</dc:creator>
  <dc:creator>Ballús i Edo, Ignasi</dc:creator>
  <dc:subject>Cicle de vida del producte -- Avaluació</dc:subject>
  <dc:subject>Cicle de vida del producte -- Aspectes ambientals</dc:subject>
  <dc:subject>Glicerina -- Aspectes ambientals</dc:subject>
  <dc:subject>Product life cycle -- Evaluation</dc:subject>
  <dc:subject>Product life cycle -- Environmental aspects</dc:subject>
  <dc:subject>Glycerin -- Environmental aspects</dc:subject>
  <dc:description>
    L'estudi de l'Anàlisi del Cicle de Vida (ACV) ha estat realitzat a l'empresa CRODA
    Ibèrica S.A - Mevisa Site, en concret a la línia productiva on s'hi fabrica el
    monoestearat de glicerina utilitzant una eina informàtica comercial denominada SIMAPRO
    que analitza i compara els aspectes mediambientals d'un producte d'una manera
    sistemàtica i consistent seguint les recomanacions de les normes ISO 14040
```

```

</dc:description>
<dc:contributor>Universitat de Girona. Facultat de Ciències</dc:contributor>
<dc:date>2007-12-14</dc:date>
<dc:type>info:eu-repo/semantics/bachelorThesis</dc:type>
<dc:format>application/pdf</dc:format>
<dc:identifier>http://hdl.handle.net/10256/40</dc:identifier>
<dc:language>cat</dc:language>
<dc:rights>
Aquest document està subjecte a una llicència Creative Commons: Reconeixement - No
comercial - Sense obra derivada (by-nc-nd)
</dc:rights>
<dc:rights>
http://creativecommons.org/licenses/by-nc-nd/3.0/deed.ca
</dc:rights>
</oai_dc:dc>
</metadata>

```

Podem veure clarament que no es pot distingir pel tag quin tipus de **dc.rights** seria cadascun, ni quin tipus de **dc.date**, ni **dc.identifier**, ni **dc.format**...

DIDL

Neix el 25 d'octubre del 2006 creat per Julie Allison del UKOLN com a llenguatge pensat per a la representació d'objectes digitals (DID) en un format basat en XML, i les seves sigles signifiquen Digital Item Declaration Language. Està orientat a la codificació d'àudio i vídeo.

El format DIDL aconsegueix superar alguns problemes de l'OAI_DC encapsulant-lo, i en aquesta càpsula, afegeix metadades més acurades de dates, i a més a més proporciona la URL dels fitxers inclosos dins de l'ítem.

És una solució interessant, però no acaba de proporcionar solucions per a totes les metadades, i no el tenen implementat tots els repositoris.

Un registre en format DIDL té dues parts: DIDLInfo i Item, dins del qual trobem les metadades en OAI_DC.

Vegem-ne un exemple:

```

<metadata>
  <didl:DIDL xmlns:didl="urn:mpeg:mpeg21:2002:02-DIDL-NS"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="urn:mpeg:mpeg21:2002:02-DIDL-NS
http://standards.iso.org/ittf/PubliclyAvailableStandards/MPEG-
21_schema_files/did/didl.xsd ">
    <didl:DIDLInfo>
      <dcterms:created xmlns:dcterms="http://purl.org/dc/terms/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://purl.org/dc/terms/">2012-11-07T19:35:16Z</dcterms:created>
    </didl:DIDLInfo>
    <didl:Item id="uuid-0540394f-27b4-4d01-8687-8c6d49fe32e4">
      <didl:Descriptor>
        <didl:Statement mimeType="application/xml; charset=utf-8">
          <dii:Identifier xmlns:dii="urn:mpeg:mpeg21:2002:01-DII-NS"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="urn:mpeg:mpeg21:2002:01-DII-NS
http://standards.iso.org/ittf/PubliclyAvailableStandards/MPEG-

```

```

21_schema_files/dii/dii.xsd">urn:hdl:10256/40</dii:Identifier>
  </didl:Statement>
  </didl:Descriptor>
  <didl:Descriptor>
    <didl:Statement mimeType="application/xml; charset=utf-8">
      <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
        <dc:title>
          Anàlisi del Cicle de Vida (ACV) del Monoestearat de Glicerina a CRODA Ibérica, S.A.
- Mevisa Site
        </dc:title>
        <dc:creator>Pacheco Bardera, Albert</dc:creator>
        <dc:creator>Oliver Serra, Arnau</dc:creator>
        <dc:creator>Padullés Cubino, Josep</dc:creator>
        <dc:creator>Ballús i Edo, Ignasi</dc:creator>
        <dc:subject>Cicle de vida del producte -- Avaluació</dc:subject>
        <dc:subject>Cicle de vida del producte -- Aspectes ambientals</dc:subject>
        <dc:subject>Glicerina -- Aspectes ambientals</dc:subject>
        <dc:subject>Product life cycle -- Evaluation</dc:subject>
        <dc:subject>Product life cycle -- Environmental aspects</dc:subject>
        <dc:subject>Glycerin -- Environmental aspects</dc:subject>
        <dc:description>
          L'estudi de l'Anàlisi del Cicle de Vida (ACV) ha estat realitzat a l'empresa CRODA
Ibèrica S.A - Mevisa Site, en concret a la línia productiva on s'hi fabrica el
monoestearat de glicerina utilitzant una eina informàtica comercial denominada SIMAPRO
que analitza i compara els aspectes mediambientals d'un producte d'una manera
sistemàtica i consistent seguint les recomanacions de les normes ISO 14040
        </dc:description>
        <dc:contributor>Universitat de Girona. Facultat de Ciències</dc:contributor>
        <dc:date>2007-12-14</dc:date>
        <dc:type>info:eu-repo/semantics/bachelorThesis</dc:type>
        <dc:format>application/pdf</dc:format>
        <dc:identifier>http://hdl.handle.net/10256/40</dc:identifier>
        <dc:language>cat</dc:language>
        <dc:rights>
          Aquest document està subjecte a una llicència Creative Commons: Reconeixement - No
comercial - Sense obra derivada (by-nc-nd)
        </dc:rights>
        <dc:rights>
          http://creativecommons.org/licenses/by-nc-nd/3.0/deed.ca
        </dc:rights>
      </oai_dc:dc>
    </didl:Statement>
  </didl:Descriptor>
  <didl:Component id="uuid-d75fb4e3-6b44-4ccc-9b4b-bf462f949c53">
    <didl:Resource ref="http://dugi-doc.udg.edu/bitstream/10256/40/1/Projecte ACV-
MEG.pdf" mimeType="application/pdf"/>
  </didl:Component>
</didl:Item>
</didl:DIDL>
</metadata>

```

Seguim tenint el problema amb els qualificadors, però podem distingir la data de creació, i ens proporciona la URL de descarrega del document.

METS

Va ser creat per Bryan Kelly, també de UKOLN, el 7 d'octubre de 2005. Les seves sigles volen dir Metadata Encoding and Transmission Standard i serveix per codificar metadades administratives, descriptives i estructurals dels objectes continguts en una biblioteca digital en format XML.

Un registre METS està format per les següents seccions: una capçalera, una part de metadades descriptives, una part amb les metadades administratives, un amb els fitxers, i una amb l'estructura de fitxers del registre.

La capçalera conté dades de creació del registre, les metadades descriptives són les mateixes del repositori, i poden estar encapsulades en diferents formats com OAI_DC, DIM¹⁰ o MODS, pot contenir una part administrativa per cada fitxer que és interessant que estigui en el format PREMIS que explicarem tot seguit.

També conté una secció amb els fitxers agrupats en grups segons al paquet de l'ítem que pertanyin, com per exemple els arxius principals, les miniatures i els fitxers de llicència.

Finalment, la última part està descriu l'estructura de fitxers de l'ítem, per saber per exemple, si aquest està format per carpetes i quina jerarquia conté.

El principal problema, és que tothom que el té implementat en el seu repositori, el té implementat amb el format de metadades intern que prefereix, i això en dificulta la seva extensió.

Un exemple de format METS amb la petició:

<http://dugi-doc.udg.edu/dspace-oai/request?verb=ListRecords&metadataPrefix=mets>

```
<record>
<header>
<identifier>oai:dugi-doc.udg.edu:10256/40</identifier>
<datestamp>2012-06-28T12:30:36Z</datestamp>
<setSpec>hdl_10256_34</setSpec>
</header>
<metadata>
<mets xmlns="http://www.loc.gov/METS/" xmlns:xlink="http://www.w3.org/1999/xlink"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" ID="DSpace_ITEM_10256-40"
OBJID="hdl:10256/40" TYPE="DSpace ITEM" PROFILE="DSpace METS SIP Profile 1.0"
xsi:schemaLocation="http://www.loc.gov/METS/
http://www.loc.gov/standards/mets/mets.xsd">
<metsHdr CREATEDATE="2012-11-08T18:18:56">
<agent ROLE="CUSTODIAN" TYPE="ORGANIZATION">
<name>DUGiDocs</name>
</agent>
</metsHdr>
<dmdSec ID="dmdSec_1">
<mdWrap MDTYPE="MODS">
<xmlData xmlns:mods="http://www.loc.gov/mods/v3"
xsi:schemaLocation="http://www.loc.gov/mods/v3
http://www.loc.gov/standards/mods/v3/mods-3-1.xsd">
<mods:mods xsi:schemaLocation="http://www.loc.gov/mods/v3
http://www.loc.gov/standards/mods/v3/mods-3-1.xsd">
<mods:name>
<mods:role>
<mods:roleTerm type="text">author</mods:roleTerm>
```

¹⁰ És un format creat per Dspace, en principi per al seu ús intern. Sempre està encapsulat dins d'un registre METS. Des del meu punt de vista és un format perfecte per a serialitzar les metadades de Dublin Core, ja que les metadades són definides dins dels atributs d'XML element i qualifier. Una llàstima que no el facin un estàndard i l'utilitzin altres repositoris.

```

</mods:role>
<mods:namePart>Pacheco Bardera, Albert</mods:namePart>
</mods:name>
<mods:name>
<mods:role>
<mods:roleTerm type="text">author</mods:roleTerm>
</mods:role>
<mods:namePart>Oliver Serra, Arnau</mods:namePart>
</mods:name>
<mods:name>
<mods:role>
<mods:roleTerm type="text">author</mods:roleTerm>
</mods:role>
<mods:namePart>Padullés Cubino, Josep</mods:namePart>
</mods:name>
<mods:name>
<mods:role>
<mods:roleTerm type="text">author</mods:roleTerm>
</mods:role>
<mods:namePart>Ballús i Edo, Ignasi</mods:namePart>
</mods:name>
<mods:name>
<mods:role>
<mods:roleTerm type="text">other</mods:roleTerm>
</mods:role>
<mods:namePart>Universitat de Girona. Facultat de Ciències</mods:namePart>
</mods:name>
<mods:extension>
<mods:dateAccessioned encoding="iso8601">2007-12-14T08:05:15Z</mods:dateAccessioned>
</mods:extension>
<mods:extension>
<mods:dateAvailable encoding="iso8601">2007-12-14T08:05:15Z</mods:dateAvailable>
</mods:extension>
<mods:originInfo>
<mods:dateIssued encoding="iso8601">2007-12-14</mods:dateIssued>
</mods:originInfo>
<mods:identifier type="uri">http://hdl.handle.net/10256/40</mods:identifier>
<mods:abstract>
L'estudi de l'Anàlisi del Cicle de Vida (ACV) ha estat realitzat a l'empresa CRODA
Ibèrica S.A - Mevisa Site, en concret a la línia productiva on s'hi fabrica el
monoestearat de glicerina utilitzant una eina informàtica comercial denominada SIMAPRO
que analitza i compara els aspectes mediambientals d'un producte d'una manera
sistemàtica i consistent seguint les recomanacions de les normes ISO 14040
</mods:abstract>
<mods:physicalDescription>
<mods:internetMediaType>application/pdf</mods:internetMediaType>
</mods:physicalDescription>
<mods:language>
<mods:languageTerm authority="rfc3066">cat</mods:languageTerm>
</mods:language>
<mods:accessCondition type="useAndReproduction">
Aquest document està subjecte a una llicència Creative Commons: Reconeixement - No
comercial - Sense obra derivada (by-nc-nd)
</mods:accessCondition>
<mods:accessCondition xlink:simpleLink="http://creativecommons.org/licenses/by-nc-nd/3.0/deed.ca">
http://creativecommons.org/licenses/by-nc-nd/3.0/deed.ca
</mods:accessCondition>
<mods:subject>
<mods:topic>Cicle de vida del producte -- Avaluació</mods:topic>

```

```

</mods:subject>
<mods:subject>
<mods:topic>Cicle de vida del producte -- Aspectes ambientals</mods:topic>
</mods:subject>
<mods:subject>
<mods:topic>Glicerina -- Aspectes ambientals</mods:topic>
</mods:subject>
<mods:subject>
<mods:topic>Product life cycle -- Evaluation</mods:topic>
</mods:subject>
<mods:subject>
<mods:topic>Product life cycle -- Environmental aspects</mods:topic>
</mods:subject>
<mods:subject>
<mods:topic>Glycerin -- Environmental aspects</mods:topic>
</mods:subject>
<mods:titleInfo>
<mods:title>
Anàlisi del Cicle de Vida (ACV) del Monoestearat de Glicerina a CRODA Ibérica, S.A.
- Mevisa Site
</mods:title>
</mods:titleInfo>
<mods:genre>info:eu-repo/semantics/bachelorThesis</mods:genre>
</mods:mods>
</xmlData>
</mdWrap>
</dmdSec>
<amdSec ID="amd_2">...</amdSec>
<amdSec ID="amd_7">
<techMD ID="techMD_8">
<mdWrap MDTYPE="PREMIS">
<xmlData xmlns:premis="http://www.loc.gov/standards/premis"
xsi:schemaLocation="http://www.loc.gov/standards/premis
http://www.loc.gov/standards/premis/PREMIS-v1-0.xsd">
<premis:premis>
<premis:object>
<premis:objectIdentifier>
<premis:objectIdentifierType>URL</premis:objectIdentifierType>
<premis:objectIdentifierValue>
http://dugi-doc.udg.edu/bitstream/10256%2F40/1/Projecte+ACV-MEG.pdf
</premis:objectIdentifierValue>
</premis:objectIdentifier>
<premis:objectCategory>File</premis:objectCategory>
<premis:objectCharacteristics>
<premis:fixity>
<premis:messageDigestAlgorithm>MD5</premis:messageDigestAlgorithm>
<premis:messageDigest>5e809113a05355eb43bb0ae8a1e75007</premis:messageDigest>
</premis:fixity>
<premis:size>1482895</premis:size>
<premis:format>
<premis:formatDesignation>
<premis:formatName>application/pdf</premis:formatName>
</premis:formatDesignation>
</premis:format>
</premis:objectCharacteristics>
<premis:originalName>Projecte ACV-MEG.pdf</premis:originalName>
</premis:object>
</premis:premis>
</xmlData>
</mdWrap>

```



```

</techMD>
</amdSec>
<amdSec ID="amd_9">
<techMD ID="techMD_10">
<mdWrap MDTYPE="PREMIS">
<xmlData xmlns:premis="http://www.loc.gov/standards/premis"
xsi:schemaLocation="http://www.loc.gov/standards/premis
http://www.loc.gov/standards/premis/PREMIS-v1-0.xsd">
  <premis:premis>
    <premis:object>
      <premis:objectIdentifier>
        <premis:objectIdentifierType>URL</premis:objectIdentifierType>
        <premis:objectIdentifierValue>
          http://dugi-doc.udg.edu/bitstream/10256%2F40/13/Projecte+ACV-MEG.pdf.txt
        </premis:objectIdentifierValue>
      </premis:objectIdentifier>
      <premis:objectCategory>File</premis:objectCategory>
      <premis:objectCharacteristics>
        <premis:fixity>
          <premis:messageDigestAlgorithm>MD5</premis:messageDigestAlgorithm>
          <premis:messageDigest>f5b9878e760df96fb3c3cb72511176c5</premis:messageDigest>
        </premis:fixity>
        <premis:size>93108</premis:size>
        <premis:format>
          <premis:formatDesignation>
            <premis:formatName>text/plain</premis:formatName>
          </premis:formatDesignation>
        </premis:format>
      </premis:objectCharacteristics>
      <premis:originalName>Projecte ACV-MEG.pdf.txt</premis:originalName>
    </premis:object>
  </premis:premis>
</xmlData>
</mdWrap>
</techMD>
</amdSec>
<amdSec ID="amd_11">...</amdSec>
<fileSec>
<fileGrp USE="CONTENT">
  <file ID="bitstream_1" MIMETYPE="application/pdf" SEQ="1" SIZE="1482895"
CHECKSUM="5e809113a05355eb43bb0ae8a1e75007" CHECKSUMTYPE="MD5" ADMID="amd_7"
GROUPID="GROUP_bitstream_1">
    <FLocat LOCTYPE="URL" xlink:type="simple" xlink:href="http://dugi-
doc.udg.edu/bitstream/10256/40/1/Projecte+ACV-MEG.pdf"/>
  </file>
</fileGrp>
<fileGrp USE="TEXT">
  <file ID="bitstream_13" MIMETYPE="text/plain" SEQ="13" SIZE="93108"
CHECKSUM="f5b9878e760df96fb3c3cb72511176c5" CHECKSUMTYPE="MD5" ADMID="amd_9"
GROUPID="GROUP_bitstream_1">
    <FLocat LOCTYPE="URL" xlink:type="simple" xlink:href="http://dugi-
doc.udg.edu/bitstream/10256/40/13/Projecte+ACV-MEG.pdf.txt"/>
  </file>
</fileGrp>
<fileGrp USE="THUMBNAIL">
  <file ID="bitstream_14" MIMETYPE="image/jpeg" SEQ="14" SIZE="1141"
CHECKSUM="0000e8a539e704169a2f2c93a8d8140a" CHECKSUMTYPE="MD5" ADMID="amd_11"
GROUPID="GROUP_bitstream_1">
    <FLocat LOCTYPE="URL" xlink:type="simple" xlink:href="http://dugi-
doc.udg.edu/bitstream/10256/40/14/Projecte+ACV-MEG.pdf.jpg"/>
  </file>
</fileGrp>

```

```

</file>
</fileGrp>
</fileSec>
<structMap ID="struct_4" LABEL="DSpace Object" TYPE="LOGICAL">
<div ID="div_5" DMDID="dmdSec_1" ADMID="amd_2" TYPE="DSpace Object Contents">
<div ID="div_6" TYPE="DSpace BITSTREAM">
<fptr FILEID="bitstream_1"/>
</div>
</div>
</structMap>
</mets>
</metadata>
</record>

```

Es pot observar que el format de metadades descriptives utilitzat en aquest cas és MODS, però en cap moment ens el deixa escollir com a clients.

És un format molt més complex que OAI_DC, i pot presentar moltes variacions que en fan molt difícil la seva aplicació generalitzada.

MODS

Creat el 2 d'octubre de 2006 per Ann Chapman, el format MODS pretén ser una versió intermitja, en quan a complexitat, entre MARC i Dublin Core. Té un set de metadades més ric que Dublin Core, i més simple que el MARC. Està més orientat cap a l'intercanvi de registres entre catàlegs i és molt més complex en quant a estructura de l'XML que l'OAI_DC i no utilitza els mateixos elements ni qualificadors que el Dublin Core, lo qual el fa molt més difícil de parsejar.

PREMIS

És un estàndard de metadades de preservació en format XML. Va ser creat el 2 d'octubre de 2006 per Ann Chapman i les seves sigles signifiquen **PRE**servation **Meta**data **Im**plementation **Str**ategies. Proporciona les metadades de format del fitxer, tamany, nom, URL, un hash per a comprovar-ne la integritat, i a vegades, fins i tot, el fitxer enter codificat en base64¹¹. És un format que pot arribar a ser molt complex segons els objectes que descriu i en el cas de la recoll·lecció sempre estarà encapsulat per METS, però pot ser molt interessant ja que ens proporciona la URL dels documents.

Altres formats i solucions

Recentment Europeana <<http://www.europeana.eu>> ha definit l'estàndard de metadades per a la recoll·lecció Europeana Semantic Elements (ESE).

Aquest format pretén acabar amb les limitacions de perdre el qualificador, estipulant un format en el valor de les metadades per discernir sobre el significat semàntic d'aquestes.

Per configurar-lo, en aquest cas, ha calgut instal·lar un plugin a més a més del fitxer d'estils XSL.

Driver per la seva banda, ha definit una especificació, dins d'OAI_DC o DIDL, afegint solament un prefix en el valor de certes metadades conflictives, que en permet ampliar el seu significat.

Com que la majoria de repositoris són recoll·lectats per Driver, i hi estan molt interessants, la majoria de repositoris implementen aquesta definició dins del seu repositori, afegint les metadades amb els prefixos que requereix Driver.

¹¹ És un sistema de codificació de caràcters que té la peculiaritat que només utilitza 64 caràcters, representats en el rang A-Za-z0-9. És molt útil per codificar fitxers binaris en mode text.

La llàstima és que només està definit per unes certes metadades que els hi permeten identificar per exemple els tipus de drets de l'ítem, els tipus d'ítems, etc.

7.3 Motors de cerca o Bases de Dades?

Avui en dia, els motors de cerca i les bases de dades NOSQL estan guanyant força enfront les bases de dades relacionals tradicionals.

En aquest punt intentarem estudiar els avantatges i inconvenients del dos sistemes.

La nostra aplicació és en definitiva un cercador. Per això tenim els següents requeriments en quan al nostre sistema gestor de bases de dades:

- **Velocitat de consulta**

Ens interessa, sobretot, poder proporcionar els ítems que satisfacin els paràmetres de la cerca en un temps molt petit. No podem tenir l'usuari esperant el resultat d'una cerca.

- **Flexibilitat**

Necessitem oferir tant la possibilitat de fer una cerca molt senzilla com de fer-ne una de molt complexa, sense que els temps de resposta augmentin gaire.

- **Facetes**

Necessitem donar la possibilitat de refinar els resultats. Un cop tenim el resultat d'una cerca, s'han de poder filtrar els resultat sobre certs criteris, com per exemple autor, matèria, tipus de document, repositori al qual pertany, data, etc.

- **Cerca a text complet**

El sistema ha de permetre buscar una paraula dins de qualsevol de les metadades de l'ítem. Tenim metadades més extenses com per exemple els resums, i és una de les metadades on pot ser més interessant cercar.

Com a contra, no ens és tant important:

- **Mida**

L'espai de disc no ens representa un problema. És un tipus de memòria barat, i en el nostre cas al fer servir màquines virtuals, ens és fàcil ampliar l'espai de disc si veiem que ens en fa falta més.

- **Velocitat de transacció**

El volum d'entrada de nous ítems serà molt més petit que el volum de consulta. A més a més es farà a una hora en la que els nostres potencials usuaris no necessitin consultar-nos. Per tant no descartem un sistema que necessiti més temps per afegir ítems, si a canvi ens garanteix més velocitat de consulta.

- **Relacions/redundància de dades?**

La nostra estructura de dades serà molt senzilla. Només guardarem ítems, i cercarem ítem. El fet que tinguem un autor de varis ítems repetit en cada ítem, no representa cap problema. Al contrari, sempre es tradueix en més velocitat de resposta al estalviar-nos els JOINS. La redundància va molt lligada a la mida de la base de dades.

- **Integritat**

Si modificar el valor d'un camp requereix modificar l'ítem enter, per un grup gran d'ítems pot portar molta feina, però donat el cas la modificació serà automatitzada mitjançant una rutina de programació, per tant aquest handicap tampoc ens representa un problema.

7.3.1 Índexs de cerca

Com que les nostres necessitats s'allunyen dels punts forts de les bases de dades, com són la integritat de les dades, l'esquema de dades utilitzant taules i relacions, una mida petita en relació a la quantitat de dades que emmagatzemen i un temps molt gran de resposta en el cas de fer consultes molt complexes, escollim els índexs de cerca.

Un índex de cerca guarda dades parsejades per aconseguir una ràpida i acurada recuperació d'informació. Si no existissin els índexs de cerca, els motors de cerca haurien d'escanejar un a un tots els ítems, i es podria tardar hores per fer una cerca. Per contra, necessita un temps bastant més gran i molt més espai per emmagatzemar un registre.

Es podria dir que els índexs es basen en la mateixa idea que els hash¹² en l'àmbit d'estructura de dades.

La majoria d'índexs utilitzats per als motors de cerca estan formats de dues parts:

- **Índexs directes**

Els índexs directes emmagatzemen els registres assignant-los un identificador que farà de clau. Cada registre nou es posa a la última posició en el mateix ordre d'entrada. Com que aquests índexs actuen com una taula d'una base de dades relacional, si volguéssim saber quins registres contenen una determinada paraula hauríem de recórrer tot l'índex enter per trobar-los tots.

Vegem-ne un exemple:

CLAU	REGISTRE
1	La vaca pastura pel camp
2	La vaca dóna la llet més bona
3	Amb la llet es fa formatge
4	Formatge de cabra
5	La llet té molt de calç

S'entreveu que necessitem alguna cosa més per a recuperar les dades d'una manera eficient. Aquesta solució només ens permet emmagatzemar les dades.

- **Índexs invertits**

Per tal d'afegir la opció de cercar a text complet amb un temps raonable, es fan servir els índexs invertits, que com el seu nom indica, són el contrari dels índexs directes.

Un índex invertit conté cadenes de caràcters, actuant com a clau, que apunten als documents que les contenen, i a més a més, també poden contenir el nombre d'entrades o la posició en el document.

¹² No s'han de confondre amb les funcions de hash. Són una estructura de dades que proporciona molta velocitat per recuperar les dades al no haver de fer un recorregut fins a trobar el valor. Usen el sistema de clau -> valor.

Així recuperar els registres que contenen una cadena de caràcters determinada és molt més ràpid.

En l'exemple següent, cada cadena de caràcters és única, i està representada per una fila. Aquestes cadenes apunten als registres que les contenen i en guardem, per exemple, també la posició que ocupen en el registre, el nombre de documents en els que apareix i el nombre d'aparicions totals en tot l'índex. Els registres corresponen als de l'índex directe, del qual es recuperarien les dades.

CADENA	(REGISTRE,[POSICIONS])	APARICIONS	NUM REGS
la	(1,[1]),(2,[1,4]),(3,[2]),(5,[1])	5	4
vaca	(1,[2]),(2,[2])	2	2
llet	(2,[5]),(3,[3]),(5,[1])	3	3
formatge	(3,[6]),(4,[1])	2	2

En un exemple més acurat, l'índex directe apuntaria a un ítem enter que seria un registre, amb totes les seves metadades indexades com a camps del registre. I en el cas de l'índex invertit, aquest podria contenir també en quin camp del registre s'ha trobat la cadena.

Aquest petit detall pot ser molt interessant, ja que ens pot permetre afegir un criteri de rellevància als resultats. És evident que trobar una cadena de caràcters en la metadada del títol no ha de tenir el mateix valor que trobar-la en el resum.

Cal dir que aquest és un exemple molt simple intentant explicar la idea de l'índex invertit d'una manera molt gràfica.

- **Sistema de consulta**

Un índex sol, no serveix de res. Necessitem un sistema que ens permeti recuperar les dades. És per això que tots els motors de cerca estan formats per un índex i un sistema de consulta.

Aquest sistema de consulta ens defineix un llenguatge, més o menys complex, amb el qual podem interrogar l'índex, filtrar els resultats i ordenar-los segons els criteris que ens convinguin.

PROCESSAMENT DE TEXT

En el fons, per una màquina, un text no són més que un conjunt de caràcters. Per poder-lo utilitzar, necessitem entendre'l.

Per això necessitem trobar les paraules del text, i classificar-les segons la seva importància.

Per tal d'aconseguir-ho, podem dur a terme una sèrie de processos que veurem a continuació.

Comencem per classificar les paraules segons la seva importància.

- Les paraules

Tant per crear un bon índex com per poder-lo consultar correctament necessitem tenir en compte la importància de les paraules.

Segons la importància les podríem dividir en 3 grups:

- **Alta importància:**

- Substantius

Els noms sempre són molt importants a l'hora de cercar. Normalment representen més del 80% de les cerques. Per exemple, si busquem un bar on anar a prendre una bona cervesa, no cercarem així: “bar on poder anar a prendre una bona cervesa”, ho farem més aviat així: “bar cervesa qualitat”. El mateix hem de fer a l'hora d'indexar un document.

Cal dir, que els documents estan escrits en llenguatge natural¹³, i cada dia més gent cerca utilitzant llenguatge natural també, per això la importància de localitzar els substantius que després es convertiran en paraules clau.

- Grups de paraules

Serien els grups de paraules que tenen un significat propi o diferent al estar juntes.

Per exemple: Medi ambient, anàlisi de conducta...

És important detectar aquests grups de paraules per oferir resultats prou rellevants.

- **Baixa importància:**

Preposicions, articles, adverbis, etc. Totes aquestes són considerades paraules buides. Són les més corrents en totes les llengües, i el seu aport al significat global de la frase és nul.

- **Variable**

- verbs

La importància d'un verb en quan al significat d'una frase pot ser molt variable. En totes les llengües hi ha verbs “auxiliars”, que no aporten contingut, ja que apareixen per formar un determinat temps verbal, la veu passiva, una perífrasi verbal¹⁴, i altres estructures similars. En aquestes estructures, el verb auxiliar, es buida de contingut semàntic, que recau sobre l'altre verb. Aquests verbs en català són: Ser (per la veu passiva), haver (per als temps compostos), anar (per al perifràstic) i tots els verbs capaços de formar perífrasi, especialment els modals. Per exemple, en la frase “Vaig anar a pescar” el significat més important el dona el verb pescar.

Com a paraula clau, no ens serveixen totes les paraules que hem categoritzat com a paraules d'alta importància. Les paraules massa freqüents no ens interessen, ja que retornarien masses resultats.

Anem a veure ara, els processos que ens poden permetre transformar aquest conjunt de caràcters sense sentit en un grup de paraules aptes per a crear un índex o una consulta.

Els processos serien els següents, i la millor manera d'aplicar-los seria amb aquest ordre mateix que els presentem:

1. Tokenitzar
2. Eliminar paraules buides
3. Lematitzar
4. Dictionaris

1. Tokenitzar

Per a *tokenitzar* ens referim a l'acció de particionar o segmentar un text en unitats indivisibles, en anglès *tokens*.

¹³ Manera com parlem habitualment, sense utilitzar només paraules clau...

¹⁴ Les perífrasis verbals són estructures formades per un verb auxiliar (conjugat), una partícula d'enllaç (optativa) i un verb en forma verbal impersonal (infinitiu, gerundi o participi de passat). El conjunt té un significat unitari que va més enllà del sentit de cada component i que s'ha d'analitzar com un tot.

Aquesta és la primera acció que es du a terme per tal de processar un text. La resta de processos es portaran a terme sobre aquests segments que ha generat el tokenitzador.

Els caràcters que formaran part dels tokens són: a-zA-Z0-9

Els que serviran de separadors: [] “()?¿¡!.,;:-

Sovint, aquest procés, pot ser complicat. Tant les paraules compostes com les abreviacions poden portar problemes.

Per exemple: U.K., O.N.U, ziga-zaga, tiqui-taca, Santa Claus, Pare Noël...

A més a més, segons l'idioma hi ha diferents caràcters separadors. Per exemple en català, i en altres llengües romàniques, utilitzem l'apòstrof (') com a separador. I per contra, l'ela geminada (l·l) no ha de ser considerat un caràcter de separació en català.

2. Eliminar les paraules buides

El segon pas, és eliminar les paraules que no aporten significat per si mateixes que solen ser preposicions, articles, conjuncions, pronomes i verbs auxiliars. Aquestes varien segons el llenguatge.

Per tant, aquest procés tampoc és trivial, ja que també requereix la prèvia detecció de l'idioma del text.

Una altra possibilitat és treure totes les paraules buides de tots els idiomes, però podria provocar molts “falsos-positius”...

Per exemple en català la paraula be és un xai, i en anglès és un verb auxiliar.

3. Lematitzar

Un cop hem eliminat les paraules buides, necessitem donada una paraula trobar el seu lema. I entenem com a lema la forma que s'accepta com a representant de totes les formes flexionades d'una paraula. Aquesta és la forma que apareix als diccionaris.

Aquest procés ens servirà per incloure totes les variants d'una paraula, i així augmentar la possibilitat de recuperar informació rellevant, i a més a més reduir considerablement el tamany dels índexs.

Sovint, enlloc de lematitzar, es parla d'stemming, que és el procés de trobar l'arrel d'una paraula. Steamitzar i lematitzar no són el mateix, sobretot en les llengües romàniques. En l'anglès sovint coincideixen.

Al lematitzar/stemitzar podem cometre dos errors, coneguts en anglès com: over-steaming i under-steaming.

- **over-steaming (hiperlematitzar):** És l'error que es produeix com a resultat de treure masses caràcters del terme a l'hora d'ajuntar paraules. Aquest procés ajuntarà paraules de diferents dominis, causant llavors recuperació de documents NO RELLEVANTS.

- **under-steaming (hipolematitzar):** Aquest altre error és produït per la raó contrària. Eliminem massa pocs caràcters, i per això paraules d'un mateix domini no s'ajuntaran, causant la NO RECUPERACIÓ de documents rellevants.

Les possibles variants són: formes verbals en els verbs, plurals, singulars, masculins i femenins en substantius i prefixos i sufixos en ambdós casos. Vegem-ne un exemple: amic, amiga, amics, amigues: aquí el lema és amic.

Una altra vegada, aquesta no és una tasca senzilla. Podem trobar moltes ambigüitats. Sobretot en els casos de verbs i substantius amb la mateixa arrel. Per exemple: ama de casa, verb amar (espanyol); creu (tant verb com substantiu), creure i creuar

Els verbs irregulars també presenten molts problemes alhora de trobar aquest lema.

Per exemple:

Anar, venir, van, va, vaig, etc: el lema aquí és anar.

Sóc, ets, és, son, sou, són, etc: el lema és ser.

En aquest cas, es pot veure clar que cal avançar l'ús del diccionari per poder trobar aquest lema.

Com lematitzar?

Una opció és fer servir un diccionari, però aquesta opció només s'hauria d'utilitzar en els casos que hem vist anteriorment, ja que és una opció costosa. Es requereixen diccionaris molt grans, i han d'estar molt ben indexats utilitzant estructures de dades molt eficients. Sobretot tenint en compte que aquests processos es realitzen tant a l'hora d'indexar com a l'hora de consultar.

Una altra opció és dividir les paraules en n-grames, i anar comparant per trobar els que comparteixen molts n-grames.

Per a poder quantificar aquesta semblança, calculem el coeficient de similaritat de Dice: $D=2C/(A+B)$. On C és el nombre de n-grames que comparteixen, i A i B són el nombre de n-grames de cada paraula. El coeficient sempre serà un valor dins de l'interval [0-1].

Per exemple per trobar el coeficient de similaritat entre tomata i tomatera:

Si treballem amb bi-grames

tomata: to-om-ma-at-ta = 5

tomatera: to-om-ma-at-te-er-ra = 7

$D=2*4/(5+7)=8/12=2/3=0.666$

La mateixa comparativa amb tri-grames

tomata: tom-oma-mat-ata = 4

tomatera: tom-oma-mat-ate-ter-era = 6

$D=2*3/(4+6)=6/10=3/5=0.6$

El resultat és bastant semblant. I sembla que com més petits siguin els trossets més exacte serà el mètode.

De totes maneres necessitem un valor llinar que ens permeti descartar falsos-positius.

La principal pega d'aquest mètode és que no ens permet trobar el lema directament, si no fer un conjunt de paraules que pertanyen a una mateixa família.

El principal avantatge és la no dependència del llenguatge del text. I a més a més, agrupa paraules amb el mateix similars tot i que tinguin prefixos i sufixos, cosa que no passa amb altres mètodes.

Una tercera tècnica és la d'eliminar afixos. Per això es necessita aplicar una sèrie de regles en un ordre determinat. Si només eliminem afixos, segurament stemitzarem, en canvi si els substituïm podem arribar a obtenir el lema. Es basen en regles gramaticals aplicades al revés.

El lematitzador de Porter segurament n'és l'exemple més famós. La pega que té és que només elimina sufixos, i només funciona per l'anglès.

Els principals inconvenients són sobretot la dependència de l'idioma i que s'ha de tenir unes regles molt bones i molt estudiat l'ordre.

I la quarta tècnica és la de varietat de successors. Aquesta tècnica consisteix en agrupar paraules amb una mateixa arrel, eliminant-ne els sufixos.

Donat un corpus o conjunt de paraules **D**, una paraula **P** de longitud n i un prefix de **P** de longitud i que anomenarem **P_i**, la varietat de successors del prefix **P_i**, coneguda com **S(P_i)**, es defineix com el nombre de lletres diferents que ocupen la posició $i+1$ en les paraules del corpus **D(P_i)**.

Així doncs, les varietats de successors que tindrà una paraula de longitud n seràn n : **S(P₁)**, **S(P₂)**, ..., **S(P_n)**.

Arribats a un punt determinat, la varietat de successors començarà a augmentar. Aquest és el punt que marcarà l'arrel.

Ex:

Donat el corpus de paraules “escola”, “escoles”, “escoleta”, “escolar” i “estat”

escola: e,s,c-t,o,l,a-e,r-blanc

escoles: e,s,c-t,o,l,a-e-s

escoleta: e,s,c-t,o,l,a-e,s-t,a

escolar: e,s,c-t,o,l,a-e,r-blanc

estat: e,s,c-t,a,t

En aquest cas, triarem com a lema, d'escola, escoleta, escoles i escolar, “escol”. Podem veure que totes coincideixen fins al punt on la varietat de successors s'incrementa a 2 amb les possibilitats de “a” i “e”. Si escollíssim com a lema “es”, estariem cometent un error, i relacionant estat amb escola...

Els principals inconvenients d'aquest mètode són el problema d'elegir un bon valor de tall, i que només elimina sufixos, i l'elecció del corpus.

4. Diccionaris

Els diccionaris ens permeten una indexació més precisa. Sobretot ens aporten la possibilitat de relacionar termes.

Gràcies a ells podem trobar:

- *sinònims*

ex: cotxe – automòbil

- *generalització* (per ampliar consultes)

ex: dòlar < unitat monetària < unitat de mesura < quantitat definida < abstracció

cavall < equí < mamífer < ser viu < entitat

- *especificitat* (restringir la cerca)

ex: unitat monetària > dòlar

equí > cavall

7.4 Lucene

Tot i que quan es parla de Lucene, sempre es parla d'ella com un índex de cerca, Lucene és una API, és a dir, un conjunt de llibreries que ens permeten afegir capacitat d'indexació i de recuperació d'informació a qualsevol sistema que s'estigui desenvolupant.

Originalment està desenvolupada en JAVA, tot i que se n'ha fet versions en quasi tots els llenguatges de programació.

Actualment forma part de la família de codi obert de la fundació Jakarta, que forma part d'Apache.

Les seves principals característiques són les següents:

- indexació incremental

Permet afegir nous ítems a un índex creat anteriorment, sense necessitat d'haver de re-indexar tots els ítems entrats anteriorment com faria un sistema d'indexació per lots.

- Origen de les dades

Pot indexar tant documents com dades procedents d'una base de dades.

- Contingut etiquetat

Per tal de donar un major contingut semàntic, Lucene permet dividir el contingut en camps. Això permetrà cercar per aquests camps, i donar diferents pesos als resultats segons el camp on apareixin els termes de la cerca.

- Stop words

Lucene permet indexar els ítems tenint en compte una llista de paraules buides.

- Concurrencia

Permet que varis usuaris estiguin consultat l'índex en un mateix moment, així com modificant-lo.

- Elecció de l'idioma

Es pot definir l'idioma dels ítems. Com hem vist al punt anterior, conèixer l'idioma dels documents ens pot permetre fer una lematització més acurada.

- Multiplataforma

No només és multiplataforma per tenir versions en diferents llenguatges de programació, si no que l'índex creat serà totalment funcional en qualsevol sistema operatiu on es vulgui utilitzar l'índex. Alhora el llenguatge JAVA ja és multiplataforma.

- Permet organitzar els resultats per rellevància

Això és possible si es troba un terme un cert nombre de vegades, així com a partir dels pesos assignats a cada camp.

- Steaming

Incorpora la funció d'steaming/lematitzar per relacionar termes i donar més resultats interessants.

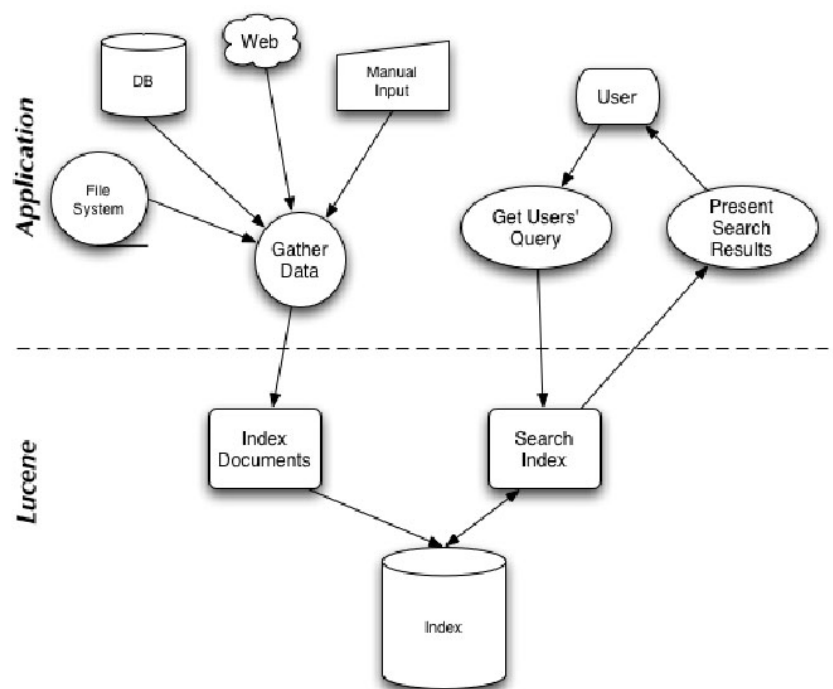
- Suport per a indexació de diferents formats de documents

Permet indexar documents amb format TXT, PDF, DOC, RTF, XML, PPT i HTML

- Ordenació

Lucene permet ordenar els resultats per qualsevol dels camps definits.

En el diagrama següent podem veure com la nostra aplicació pot indexar documents de varies fonts, com per exemple de bases de dades, de fitxers, etc, i com l'usuari pot consultar l'aplicació i recuperar-ne els resultats.



El principal inconvenient que trobo a utilitzar directament Lucene, és que l'índex queda completament integrat a la nostra aplicació, i això ens dificultaria la possibilitat de cercar a l'índex des de altres aplicacions.

Un altre inconvenient és que en el cas de programar en PHP ens obliga a utilitzar el Zend Framework, que té implementada la llibreria.

7.5 SOLR

A diferència de Lucene, Solr és un servidor de cerca a text complet que usa Lucene i l'exten. Solr permet indexar i servir documents via HTTP, i suporta diferents formats tant d'entrada com de sortida. Pot treballar tant sol com dins d'un servidor web de servlets com Tomcat.

Solr està escrit en JAVA, i forma part de Apache Software Foundation.

7.5.1 Característiques

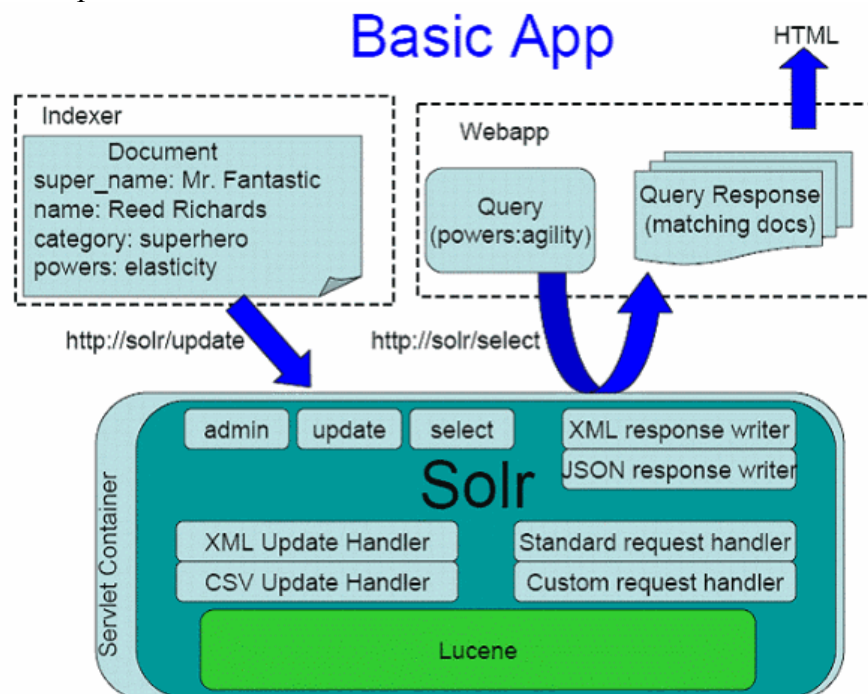
- Multiples formats d'entrada: XML, JSON, CSV, binary.
- Multiples formats de sortida: XML, JSON, CSV, binay, PHP, PHPS, PYTHON...
- Capacitats de cerca a text-complet avançades.
- Optimitzat per grans volums de trafic.
- Panell d'administració via web.
- Utilitza estàndards lliures (Standards Based Open Interfaces): XSMML, JSON, HTTP...
- Servidor d'estadístiques monitoritzant sobre JMX (Java Management Extensions)
- Escalable linealment.
- Replicació de l'index, recuperació i control d'errades automàtics.
- Indexació eficient pròxima a temps real.
- Flexible i adaptable amb configuració en format XML.
- Arquitectura de plugins extensible.
- Esquema de tipus de dades real i extensible: Nombres, camps dinàmics, claus úniques...
- Extensions potents al llenguatge de cerca de Lucene (Lucene Query Language).
- Cerca i Filtre per Facetes.
- Cerca Geoespacial amb suport per múltiples punts per document i geo-polígons.
- Anàlisi de Text avançada i configurable.
- Cache altament configurable i extensible.
- Optimització del rendiment.
- Configuració externa via XML.
- Administració basada en AJAX.
- Logging monitoritzable.
- Ràpida indexació incremental i replicació de l'índex pròxima al temps real.
- Cerca distribuïda altament escalable, amb un índex fragmentat en diversos hosts.
- Forma fàcil d'introduir dades des de bases de dades, fitxers XML locals i fonts HTTP.
- Parsejador i indexador de documents rics (PDF, Word, HTML, etc) usant Apache Tika.
- Integració d'Apache UIMA¹⁵ per configurar l'extracció de metadades.
- Múltiples índexs de cerca.

¹⁵ **Unstructured Information Management applications**, són sistemes de programari que analitzen grans volums d'informació no estructurada per tal de descobrir el coneixement que sigui rellevant per a l'usuari final. Una aplicació UIM exemple podria ingerir text i identificar les entitats, com a persones, llocs, organitzacions, etc.

Entre els principals avantatges que té respecte a utilitzar Lucene directament, s'ha de destacar la capacitat d'utilitzar-lo i configurar-lo sense haver d'escriure una sola línia de JAVA. També és important poder balancejar la càrrega i fragmentar l'índex, tot i que només és interessant a partir d'unes mides, grans volums de tràfic i nombre de clients connectats alhora molt elevats que molt probablement nosaltres no tindrem. Una altra característica a destacar és que tant el fitxer de definició dels camps com el de configuració general del SOLR utilitzen el format XML, lo qual en facilita la configuració sense haver d'utilitzar un intèrpret.

Per al nostre projecte escollirem SOLR sobretot perquè enriqueix el llenguatge de cerca de Lucene i ens ofereix la capacitat de llistar facetes, amb lo qual, ens podem estalviar l'ús d'una base de dades, cosa que en el cas d'utilitzar Lucene directament, seria pràcticament impossible. A més a més, al ser SOLR un servei ell mateix, ens permet atacar-lo des de diferents projectes, possibilitant-nos per exemple la cerca al DUGi des de la web de la Biblioteca.

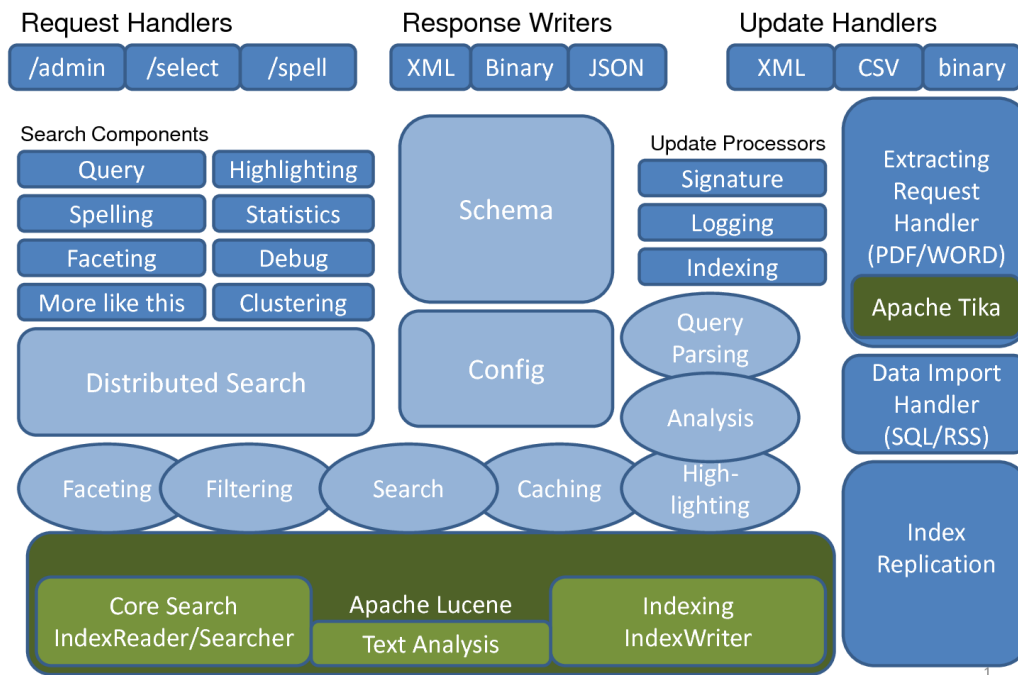
A la imatge següent podem veure un esquema d'una aplicació d'exemple que utilitza SOLR, es pot apreciar com SOLR encapsula LUCENE donant una interfície per indexar i cercar els documents. A més a més ofereix la possibilitat d'utilitzar diferents formats tant d'entrada com de sortida.



Font: <http://www.yeeach.com/wp-content/uploads/2008/07/clip-image002.gif>

La imatge següent mostra molt detalladament l'arquitectura de Lucene i de SOLR encapsulant-lo. També podem apreciar tots els mòduls que ens aporta tant per indexar com per cercar, i a part els mòduls de replicació i importació de dades des d'altres fonts.

Lucene/Solr Architecture



Font: <http://img.docstoccdn.com/thumb/orig/98318767.png>

7.5.2 Instal·lació i proves de funcionament

A l'ANNEX 2 es detalla la instal·lació del SOLR i les proves de funcionament.

7.6 Elasticsearch

Elasticsearch, al igual que SOLR, és un servidor de cerca basat en Lucene. La principal diferència i avantatge, és que ens permet molt bons temps de resposta en el cas de tenir moltes actualitzacions de l'índex simultàniament que ens fan consultes.

Vegem-ne les característiques:

- Actualització en temps real:

Elasticsearch ens permet anar indexant dades noves tota l'estona sense que aquest increment de l'índex i de les peticions ens facin augmentar el temps d'indexació.

- Consulta en temps real:

Tot i que l'índex s'estigui actualitzant i consultant constantment, el temps de resposta a les consultes es manté en temps real.

- Distribuït:

Elasticsearch ens ofereix la opció de distribuir l'índex en diferents servidors.

- Alta disponibilitat:

La xarxa de clústers d'Elasticsearch estan sempre pendents de redirigir el tràfic en cas de que falli algun dels servidors.

- Cerca a text complet:
Al utilitzar un índex Lucene, elasticsearch ens permet cercar a text complet.
- Orientat a document:
Ens permet indexar documents directament.
- Múltiples col·leccions:
Podem tenir moltes col·leccions diferents en un mateix servidor d'elasticsearch.
- Gestió de conflictes:
Al permetre actualitzacions en temps real, és important mantenir la integritat, així doncs elasticsearch ofereix la possibilitat de fer un control de versions optimista per no perdre dades.
- Sense esquema:
No cal tenir un esquema de dades definit prèviament, es pot guardar qualsevol contingut sense restriccions de camps.
- API:
Elasticsearch està construït com una API, totes les accions es poden executar fent servir aquesta API disponible en varis llenguatges usant JSON sobre HTTP.
- Persistència per operació:
Totes les transaccions queden registrades en els servidors d'elasticsearch per assegurar les dades.

Elasticsearch és una opció molt interessant per una aplicació web que tingui constants actualitzacions. Està sent utilitzada per aplicacions com Forsquare, soundcloud, github, stackoverflow entre altres.

De totes maneres, les nostres dades no necessiten ser actualitzades tant constantment, amb actualitzar l'índex a les nits en tenim prou, i per contra ens interessa molt que les consultes siguin el més ràpid possible, i en aquests casos SOLR és més ràpid.

7.7 REDIS

Redis és un altre base de dades NOSQL. En aquest cas, la principal característica és que l'emmagatzematge està basat en estructures de dades de tipus hash. Així doncs, és un sistema molt eficient tant per emmagatzemar dades com per recuperar-les.

Hi ha llibreries de redis en la majoria dels llenguatges de programació més utilitzats, i també té un protocol per a la inserció de massiva de dades a gran velocitat.

La llàstima és que no és un bon sistema de cerca, ja que només conté la possibilitat de cercar claus de hash utilitzant un comodí.

De totes maneres, pot ser una gran eina per a guardar altres tipus de dades com per exemple les estadístiques.

7.8 Sistema Vista-controlador

Per tal de separar la lògica de l'aplicació de la vista, creiem que pot ser interessant utilitzar un sistema de vista-controlador que ens permeti treballar amb plantilles a l'hora de crear les vistes.

El PHP és un llenguatge pensat per estar incrustat dins de l'HTML, per tant, ell mateix ja quasi és un sistema de plantilles.

De totes maneres, vegem les millors opcions que tenim:

- **SMARTY:**

<http://www.smarty.net/>

És probablement el sistema de plantilles de PHP més utilitzat. Les seves característiques són:

- Té una gran comunitat.
- Afegeix llegibilitat a les plantilles.
- Té un sistema de memòria cau (cache) molt robust.
- Necessita la creació de nous plugins per anar afegint funcionalitats.
- Dificil d'integrar en gestors de continguts, frameworks i editors de text.
- Redueix el rendiment de les aplicacions al haver de parsejar les plantilles per convertir-les a PHP (Quan ja les té en cache ja no passa)
- Es necessari aprendre un llenguatge nou.
- Sintaxi bastant complexa.
- Segur al implementar funcions per escapar les variables.
- Fàcil de mantenir.

- **TWIG:**

<http://twig.sensiolabs.org/>

El sistema de plantilles flexible, ràpid i segur per a PHP. Les seves característiques són:

- Afegeix llegibilitat a les plantilles.
- Relativament nou, però cada dia més utilitzat.
- Molt ben documentat.
- Més ràpid que SMARTY.
- Es necessari aprendre un llenguatge nou, tot i que més senzill que SMARTY.
- Sintaxi molt amable.
- Del creador del framework Symfony¹⁶.
- No sembla portar un sistema de cache incorporat. Probablement al estar pensat per formar part del framework Symfony que ja proporciona moltes caches.

¹⁶ Symfony Framework és un dels marcs de treball en PHP més utilitzats. Proporciona una total flexibilitat a l'hora de desenvolupar aplicacions web amb PHP. El seu punt més fluïd, al meu entendre, és la gran corba d'aprenentatge que té. <http://symfony.com/>

- Flexible.
- Extensible.
- Segur. Ja que proporciona formes d'escapar les variables.

- **HAANGA:**

<http://haanga.org/>

Plantilles estil Django¹⁷ per PHP. Les seves característiques són:

- Rapidíssim. És inclús més ràpid que Twig.
- Molt eficient. Necessita la meitat de memòria que TWIG.
- Sintaxi molt agradable. Concretament la mateixa que TWIG, ja que és la de Django.
- Necessita aprendre un llenguatge nou. Més senzill que SMARTY.
- Creat per la gent de Meneame (<http://meneame.net>) i convertit després en projecte independent.
- Extensible.
- Segur. Perquè proporciona formes d'escapar les variables.
- Sistema de cache incorporat.

- **SAVANT3:**

<http://phpsavant.com/>

El sistema de plantilles simple, elegant per PHP. Les seves característiques són:

- No és necessari aprendre cap llenguatge nou ja que les plantilles es creen utilitzant PHP.
- Molt ràpid ja que no li cal compilar les plantilles com els altres sistemes. Ja estan en PHP.
- No té sistema de cache propi.
- Les plantilles són més “brutes” (difícils de llegir) al haver d'escriure tot el tag d'incrustació del PHP.
- Segur. Perquè proporciona funcions per escapar les variables.
- I sobretot, molt simple.

El nostre sistema necessita no guardar en cache certes pàgines que poden canviar molt sovint. Això ens fa descartar els sistemes de plantilles Smarty i Haanga, que al portar incorporat un sistema de cache ens seria complicat controlar-lo nosaltres.

El sistema Twig és segurament el més amable per treballar amb dissenyadors, ja que no requereix tenir cap mena de coneixement de PHP per a fer les plantilles. El controlador és l'únic que contindrà codi PHP. Però requereix aprendre un llenguatge nou, i al no disposar de dissenyador per al projecte, ens quedem amb el sistema Savant, que ens dona la possibilitat de separar la lògica de la capa de presentació sense sortir-nos del PHP.

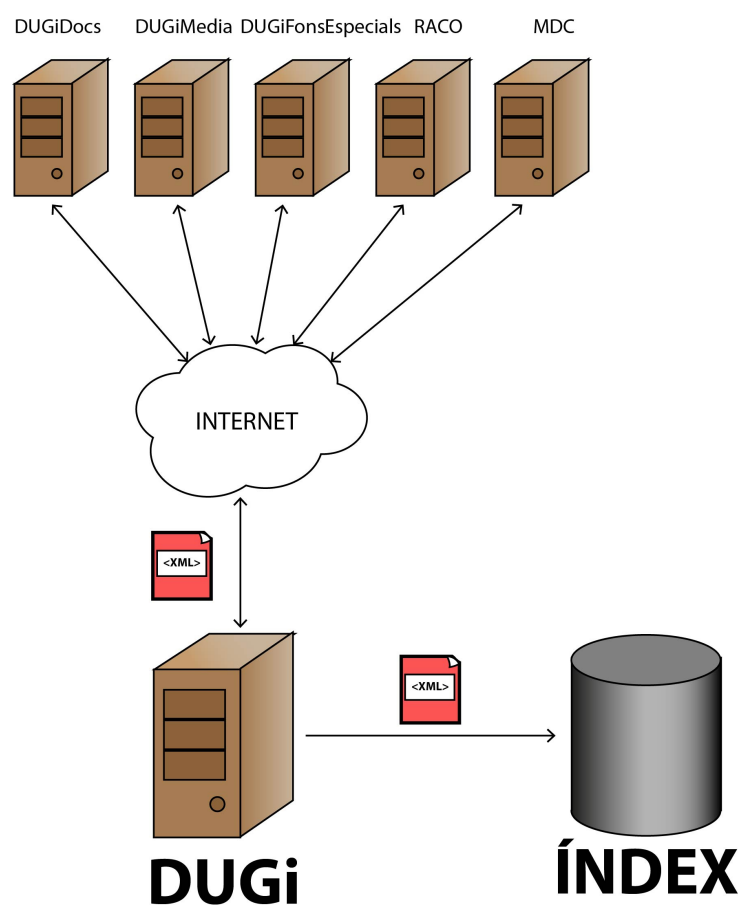
¹⁷ Marc de treball per al desenvolupament d'aplicacions en Python. <https://www.djangoproject.com/>

8. Anàlisi i disseny del sistema

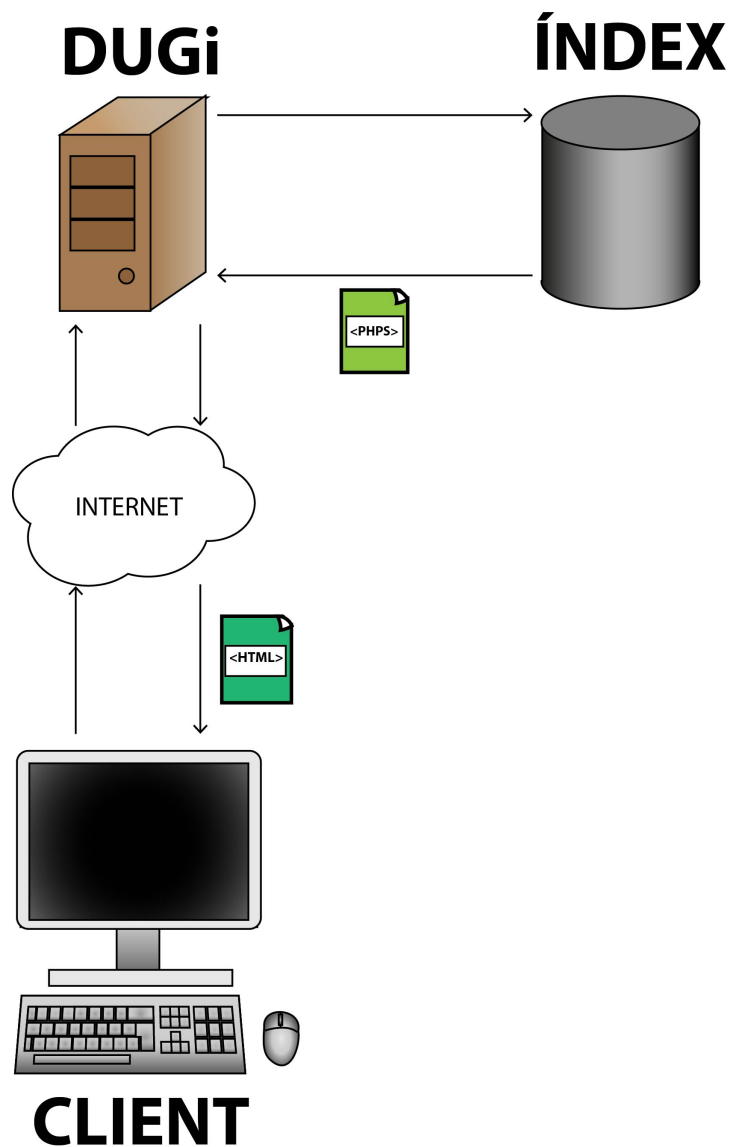
El nostre sistema es divideix en dues parts, la part de recollida de dades, i la part de consulta.

La primera es dedica a recollir els ítems que s'hagin afegit nous als repositoris, indexar-los i detectar les autoritats duplicades, i la segona és la que permet consultar i gestionar l'índex on hem afegit els ítems la primera part.

La imatge següent és una representació de la primera part del sistema, la que s'ocupa de la recollida. Podem veure com el DUGi recollirà els repositoris utilitzant el protocol OAI-PMH, que ens retornarà un fitxer en format XML, que un cop parsejat es transformarà al format XML d'inserció al repositori SOLR.



La imatge següent representa la segona part del sistema, que permet consultar l'índex. L'usuari farà una consulta al DUGi, aquest la parsejarà i la farà al seu índex SOLR, que retornarà els resultats en format d'objectes PHP serialitzats, amb els que el DUGi dibuixarà la pàgina HTML que retornarà a l'usuari.



8.1 Anàlisi

Per tal de desenvolupar l'anàlisi del nostre sistema, a partir dels requeriments que havia de satisfer la nostra aplicació, hem realitzat els següents casos d'ús i fitxes de casos d'ús.

8.1.1 Casos d'ús

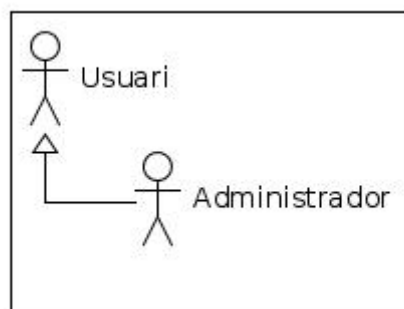
En el nostre sistema tindrem 3 actors diferents:

- Sistema
- Administrador
- Usuari

Sistema: Serà l'encarregat de fer les tasques automatitzades (recol·lectar, indexar...)

Administrador: Donarà el vist i plau a les tasques que requereixin validació, així com tasques de manteniment i adició dels registres. És un usuari que necessita autenticació. Hereda de Usuari.

Usuari: Podrà cercar i comparar registres. No és un usuari autènticat.



Sistema. Els casos d'ús de l'actor sistema són els següents:

Cas d'ús 1.1: recol·lectar

Cas d'ús 1.2: indexar

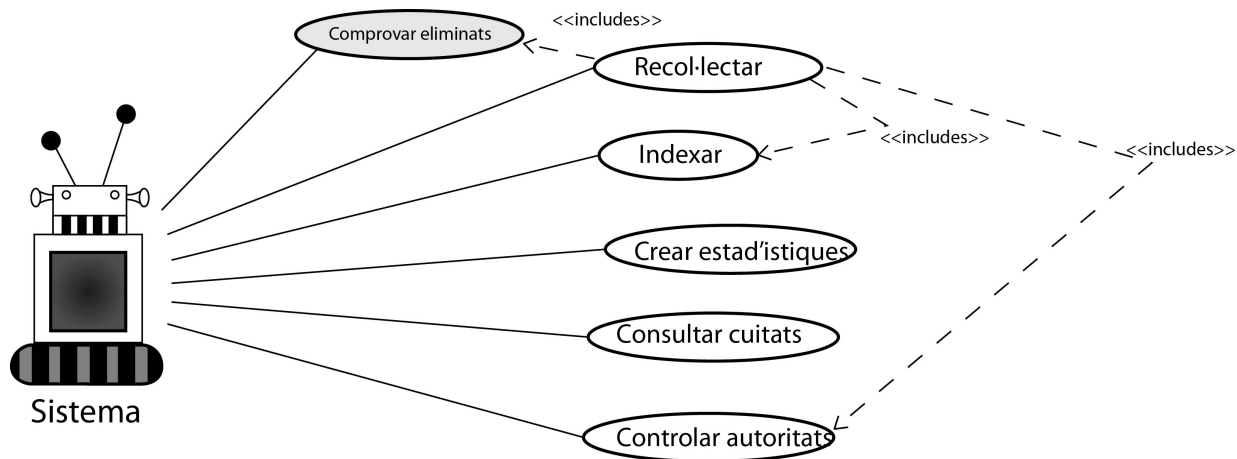
Cas d'ús 1.5: control d'autoritats

Cas d'ús 1.6: comprovar eliminats

Cas d'ús 1.3: crear estadístiques

Cas d'ús 1.4: consultar ciutats

El diagrama següent mostra els casos d'ús que farà el sistema. Tots ells es fan automàticament amb l'ajuda del dimoni cron. El cas d'ús comprovar eliminats ha sigut afegit a posteriori, durant una de les iteracions del procés. Més endavant s'explicarà el perquè d'haver-lo afegit.



Usuari. Els casos d'ús de l'actor usuari són els següents:

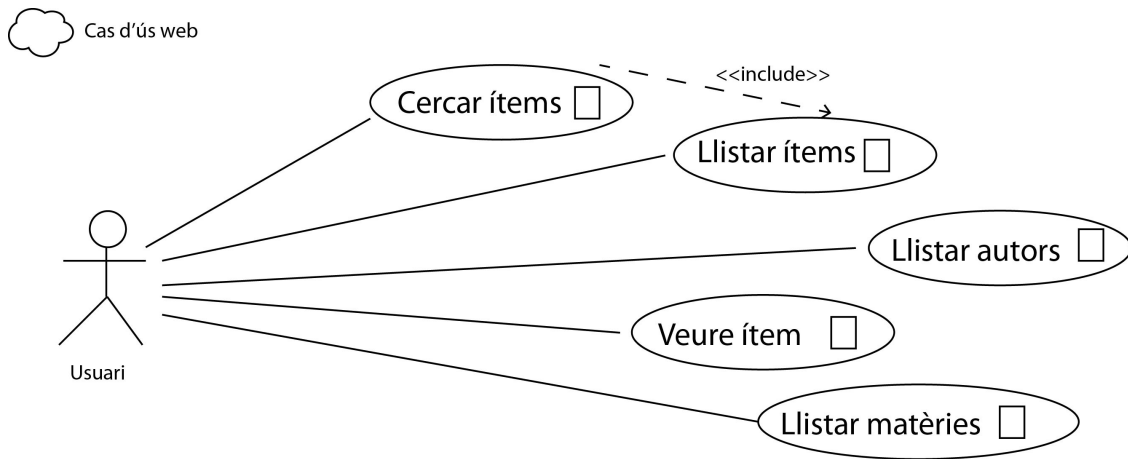
Cas d'ús 2.1: cercar ítems

Cas d'ús 2.2: llistar ítems

Cas d'ús 2.3: llistar per autor

Cas d'ús 2.4: llistar per matèria

Cas d'ús 2.5: veure ítem

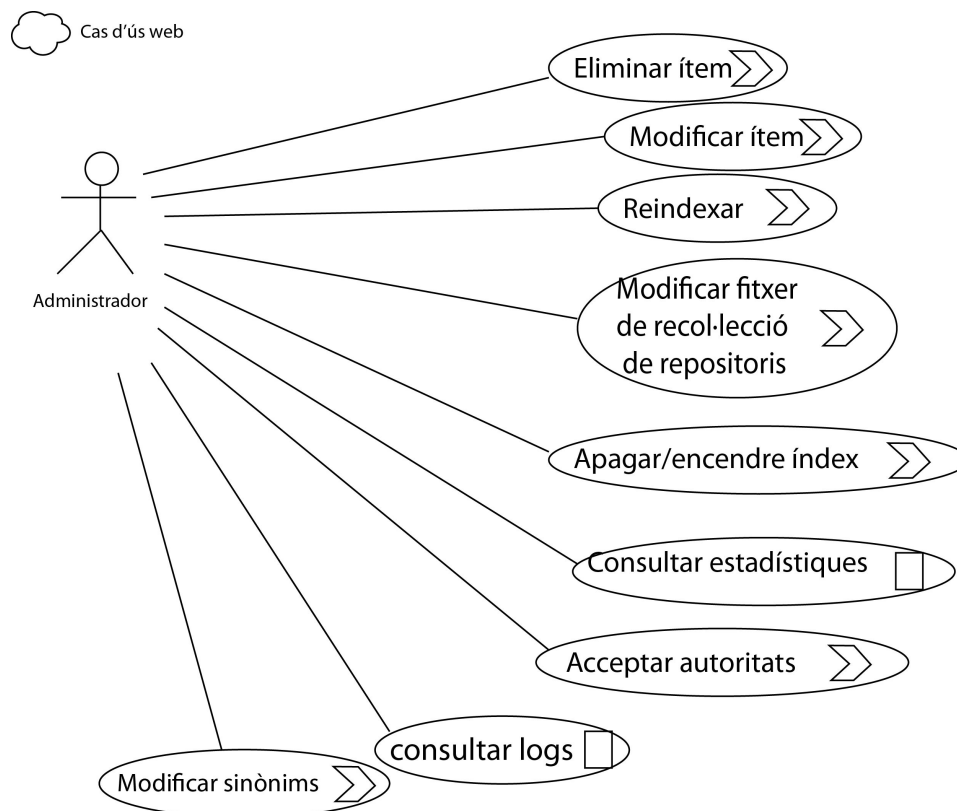


En el cas de l'actor usuari, tots els casos d'ús són de navegació. No se li permet fer cap procés.

Administrador. Els casos d'ús de l'actor administrador són els següents:

- Cas d'ús 3.1:modificar registres
- Cas d'ús 3.2:eliminar registres
- Cas d'ús 3.3:reindexar
- Cas d'ús 3.4:modificar fitxer de repositoris
- Cas d'ús 3.5:apagar/encendre index
- Cas d'ús 3.6:consultar estadístiques
- Cas d'ús 3.7:consultar logs
- Cas d'ús 3.8:acceptar autoritzacions

El diagrama següent ens mostra els casos d'ús que afecten a l'actor administrador. En el seu cas, la majoria de casos d'ús són de procés. Només dos, el de consultar estadístiques i el de consultar logs són de tipus navegació (*browsing*).



8.1.2 Fitxes de casos d'ús

Fitxes cas d'ús: sistema.

Nom:	1.1 recol·lectar
Autor:	Oriol Olivé
Data:	24/01/12
Descripció:	El sistema recol·lecta els ítems nous dels repositoris, crea els ítems dins l'índex .
Actors:	Sistema
Precondicions:	cap
Flux normal:	1. Els sistema llegeix el fitxer de recol·lecció 2. Per cada repositori 2.1 Preguntem quins ítems nous hi ha a partir de la data 2.2 parsegem els ítems 2.3 Comprovem autoritats 2.4 Comprovem els eliminats 3. Indexar
Flux alternatiu:	2.1 no hi ha ítems per indexar
Postcondicions:	Tots els ítems han estan entrats i indexats

Nom:	1.2 indexar
Autor:	Oriol Olivé
Data:	24/01/12
Descripció:	El sistema crea els ítems dins l'índex i dps optimitza l'índex.
Actors:	Sistema
Precondicions:	cap
Flux normal:	1. Per cada ítem 1.1 Transformar les metadades recol·lectades a les nostres 1.2 inserir l'ítem a l'índex 1.3 Optimitzar l'índex
Flux alternatiu:	

cap
Postcondicions: Tots els ítems han estan indexats i l'índex està optimitzat

Nom:	1.3 Crear estadístiques
Autor:	Oriol Olivé
Data:	24/01/12
Descripció:	El sistema recol·lecta les estadístiques parsejant els logs
Actors:	Sistema
Precondicions:	cap
Flux normal:	- Per cada fitxer de log - Per cada entrada - Mirar si tenim la ciutat guardada - Sumar Entradades - Afegir a fitxers mensuals,diaris i per ip
Flux alternatiu:	- Afegir ip llista per consultar ciutats
Postcondicions:	Els logs estan llegits

Es van consultant els logs i es genera un arbre de cerca amb les dates de consulta de cada ip, el tipus de consulta i la pàgina consultada. A més a més, es guarden les dades de les consultes diàries, mensuals i anuals.

Nom:	1.4 Consultar ciutats
Autor:	Oriol Olivé
Data:	24/01/12
Descripció:	El sistema consulta les ciutats per ip.
Actors:	Sistema
Precondicions:	Hi ha ips pendents de consulta
Flux normal:	- Llegim el fitxer d'ips pendents - Per cada ip - Preguntem a quina ciutat està

2. Entrem la ciutat al registre d'ips 3. Eliminem la ip de la llista
Flux alternatiu: 2. no se sap d'on és la ip 2.1 Eliminem la ip de la llista 2. No ens queden intents
Postcondicions: No queden més ciutats per consultar o no tenim intents

Per consultar les dades d'una ip, utilitzem una API remota que ens permet consultar 20 ips cada hora. Al final es va creant un arbre de cerca amb les dades de les ips.

Nom:	1.5 Controlar autoritats
Autor:	Oriol Olivé
Data:	24/01/12
Descripció: El sistema busca si hi ha autors amb un nom prou semblant com per ser la mateixa persona	
Actors: Sistema	
Precondicions: Tenim un ítem nou	
Flux normal: 1. Creem el patró de cerca amb el camp autor 2. Llistem a l'índex els possibles autors 3. Comparem els autors com el nostre patró 4. guardem els autors sospitosos al fitxer d'autoritats	
Flux alternatiu: 2. no hi ha cap possible autor 3. cap autor compleix el mínim de semblança demanat	
Postcondicions: cap	

Nom:	1.6 Comprovar eliminats
Autor:	Oriol Olivé
Data:	06/04/13
Descripció: El sistema busca si hi ha ítems per indexar que hagin estat eliminats anteriorment.	
Actors: Sistema	
Precondicions: Tenim ítems nous	

Flux normal:
1. Cerquem els identificadors dins el llistat d'ítems eliminats 2. Els ítems amb aquest identificador no s'entren al sistema
Flux alternatiu:
2. No s'ha trobat cap ítem que necessiti ser eliminat
Postcondicions:
cap

Aquest cas d'ús s'ha fet necessari al adonar-nos de que tot i que l'administrador elimini un ítem, si en algun dels repositoris que recol·lectem es modifica aquest ítem, el nostre sistema el tornaria a recol·lectar, i l'administrador quan se'n donés compte l'hauria de tornar a eliminar i així indefinidament.

Fitxes cas d'ús: Usuari

Nom:	2.1 Cercar ítems
Autor:	Oriol Olivé
Data:	24/01/12
Descripció:	L'usuari cerca al sistema
Actors:	Usuari
Precondicions:	cap
Flux normal:	1. L'usuari introdueix la seva cerca a la caixa de cerca 2. El sistema retorna un llistat d'ítems 3. L'usuari pot acotar la cerca
Flux alternatiu:	2. Cap ítem concorda amb la cerca
Postcondicions:	Cap

Nom:	2.2 Llistar ítems
Autor:	Oriol Olivé
Data:	24/01/12
Descripció:	L'usuari llista ítem del sistema
Actors:	Usuari
Precondicions:	cap

Flux normal:
1. El sistema mostra un llistat d'ítems 2. L'usuari pot veure cada ítem o avançar/retrocedir pàgina
Flux alternatiu:
cap
Postcondicions:
Cap

Nom:	2.3 Llistat d'autors
Autor:	Oriol Olivé
Data:	24/01/12
Descripció:	L'usuari llista els autors del sistema
Actors:	Usuari
Precondicions:	cap
Flux normal:	1. El sistema mostra un llistat d'autors ordenats alfabèticament 2. L'usuari pot clicar un autor o filtrar els resultats 3. El sistema retorna un llistat dels ítems d'aquest autor
Flux alternatiu:	cap
Postcondicions:	Cap

Nom:	2.4 Llistat de matèries
Autor:	Oriol Olivé
Data:	24/01/12
Descripció:	L'usuari llista les matèries del sistema
Actors:	Usuari
Precondicions:	cap
Flux normal:	4. El sistema mostra un llistat de matèries ordenades alfabèticament 5. L'usuari pot clicar una matèria o filtrar els resultats 6. El sistema retorna un llistat dels ítems d'aquesta matèria
Flux alternatiu:	cap

Postcondicions: Cap

Nom:	2.5 Veure ítem
Autor:	Oriol Olivé
Data:	24/01/12
Descripció: L'usuari veu les dades d'un ítem en particular.	
Actors: Usuari	
Precondicions: cap	
Flux normal: 1. L'usuari clica un ítem 2. El sistema mostra l'ítem 3. L'usuari pot accedir al text complet	
Flux alternatiu: 3. L'usuari tanca l'ítem 4. L'usuari clica un altre ítem	
Postcondicions: Cap	

Fitxes casos d'ús: Administrador.

Nom:	3.1 Modificar ítem
Autor:	Oriol Olivé
Data:	24/01/12
Descripció: L'administrador modifica les dades d'un ítem entrat dins del sistema.	
Actors: Administrador	
Precondicions: L'usuari té privilegis d'administrador	
Flux normal: 1. L'administrador cerca l'ítem 2. El sistema li mostra un llistat d'ítems 3. L'administrador clica l'ítem que vol modificar 4. L'administrador guarda l'ítem 5. L'índex re-indexa l'ítem 6. L'índex s'optimitza	
Flux alternatiu:	

cap
Postcondicions: Les dades de l'ítem han estat canviades, l'ítem està indexat i l'índex està optimitzar

Nom:	3.2 Eliminar ítem
Autor:	Oriol Olivé
Data:	24/01/12
Descripció:	L'administrador elimina un ítem entrat dins del sistema.
Actors:	Administrador
Precondicions:	L'usuari té privilegis d'administrador
Flux normal:	1. L'administrador cerca l'ítem 2. El sistema li mostra un llistat d'ítems 3. L'administrador clica l'ítem que vol eliminar 4. L'índex elimina l'ítem 4.1 El sistema guarda l'identificador de l'ítem eliminat 5. L'índex s'optimitza
Flux alternatiu:	cap
Postcondicions:	L'ítem ja no existeix al sistema i l'índex està optimitzat

Nom:	3.3 Re-indexar
Autor:	Oriol Olivé
Data:	24/01/12
Descripció:	L'administrador re-indexa els ítems d'un determinat repositori
Actors:	Administrador
Precondicions:	L'usuari té privilegis d'administrador
Flux normal:	1. El sistema mostra un llistat dels repositoris 2. L'administrador clica un dels repositoris 3. L'administrador pot escollir la data a partir de la qual vol recol·lectar 4. El sistema recol·lecta el repositori particular a partir del fitxer de recol·lecció
Flux alternatiu:	cap

Postcondicions: Els ítems del repositori estan reindexats de nou i l'índex està optimitzat.

Nom:	3.4 Modificar fitxer de recol·lecció de repositoris
Autor:	Oriol Olivé
Data:	24/01/12
Descripció: L'administrador modifica les dades dels repositoris i de les metadades al fitxer de recol·lecció	
Actors: Administrador	
Precondicions: L'usuari té privilegis d'administrador	
Flux normal: 1. El sistema mostra un llistat dels repositoris 2. L'administrador clica al repositori al qual li vol canviar les dades 3. L'administrador pot canviar: la data de l'última recol·lecció, els sets, les metadades i patrons de recol·lecció. 4. L'administrador guarda el fitxer	
Flux alternatiu: cap	
Postcondicions: Cap	

Nom:	3.5 Apagar/encendre índex
Autor:	Oriol Olivé
Data:	24/01/12
Descripció: L'administrador pot apagar o encendre l'índex	
Actors: Administrador	
Precondicions: L'usuari té privilegis d'administrador	
Flux normal: 1. El sistema mostra l'estat de l'índex i un botó 2. L'administrador clica encendre/apagar 3. El sistema commuta l'estat de l'índex.	
Flux alternatiu: cap	
Postcondicions: Cap	

Nom:	3.6 Consultar estadístiques
Autor:	Oriol Olivé
Data:	24/01/12
Descripció:	L'administrador pot consultar les estadístiques totals, mensuals o diàries, així com les estadístiques per ip, ítem, autor o matèria.
Actors:	Administrador
Precondicions:	L'usuari té privilegis d'administrador i hi ha estadístiques generades
Flux normal:	1. El sistema mostra els llistats totals per ítems, autors, matèries i ips. 2. L'administrador pot filtrar les dades.
Flux alternatiu:	cap
Postcondicions:	Cap

Nom:	3.7 consultar logs
Autor:	Oriol Olivé
Data:	24/01/12
Descripció:	L'administrador consulta els logs
Actors:	Administrador
Precondicions:	L'usuari té privilegis d'administrador i hi ha logs genrats
Flux normal:	1. El sistema mostra un llistat dels logs totals i repartits en informació, avisos i errors. 2. L'administrador pot filtrar els resultats així com mirar logs més antics
Flux alternatiu:	cap
Postcondicions:	Cap

Nom:	3.8 Acceptar autoritats
Autor:	Oriol Olivé
Data:	24/01/12
Descripció:	

L'administrador accepta i/o denega peticions d'autoritats noves o existents
Actors: Administrador
Precondicions: L'usuari té privilegis d'administrador i hi ha autoritats pendents de validar
Flux normal: 1. El sistema mostra un llistat amb els ítems que tenen autoritats per validar 2. L'administrador clica un ítem 3. El sistema mostra les dades de l'autoritat amb els autors que s'hi assemblen prou i el factor de semblança. L'administrador pot consultar l'ítem 4. L'administrador pot acceptar o denegar l'autoritat 5. L'ítem es guarda a l'índex modificat
Flux alternatiu: 5. En cas de denegar l'autoritat no cal fer res perquè l'ítem ja existeix a l'índex
Postcondicions: Cap

Nom:	3.9 Modificar sinònims
Autor:	Oriol Olivé
Data:	07/02/12
Descripció: L'administrador modifica el fitxer de sinònims de l'índex	
Actors: Administrador	
Precondicions: L'usuari té privilegis d'administrador	
Flux normal: 1. El sistema mostra el fitxer de sinònims 2. L'administrador el modifica 3. El sistema guarda el fitxer de sinònims 4. El sistema reinicia l'índex	
Flux alternatiu: Cap	
Postcondicions: Cap	

8.1.3 Diagrames d'activitat

Els diagrames d'activitat ens permetran fer-nos una idea de les activitats que es duran a terme en cada cas d'ús, i veure on es desenvolupa.

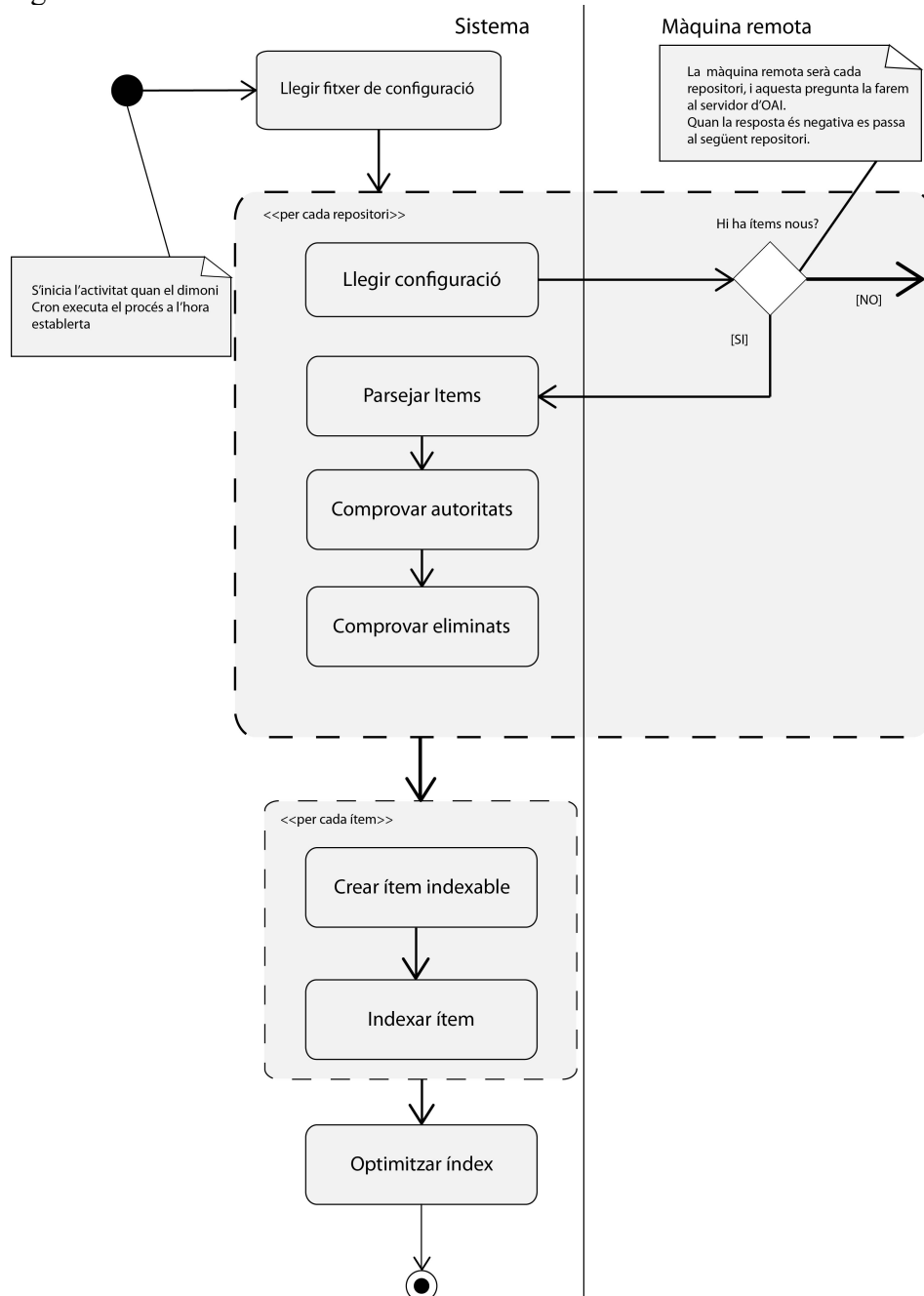
En aquest apartat mostrarem els diagrames d'activitat més interessants:

- Recol·lectar i indexar
- Crear estadístiques
- Consultar ciutats
- Eliminar ítems
- Acceptar autoritats
- Cercar ítems i veure ítem
- Llistar autors

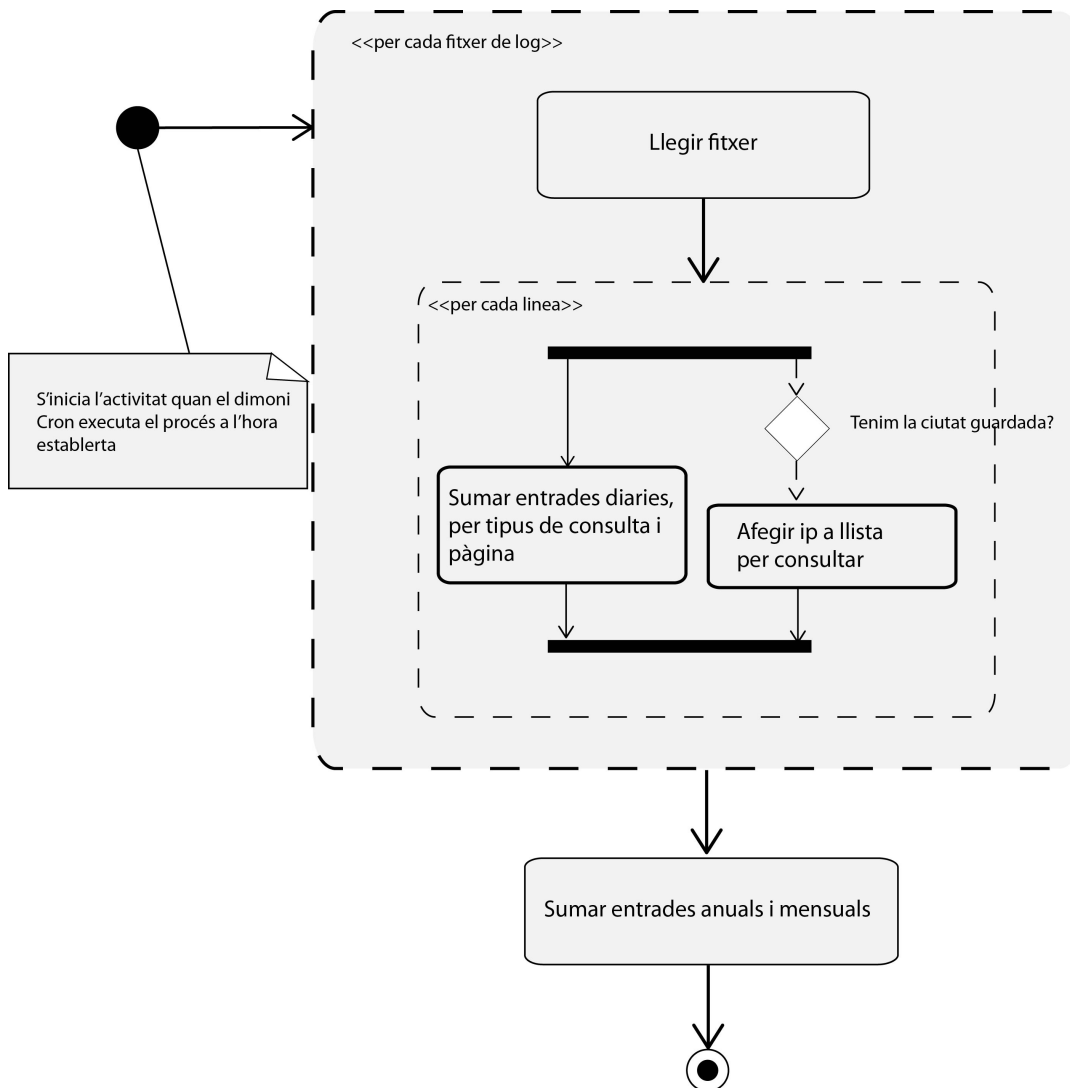
Diagrames d'activitat del sistema:

Aquests diagrames estan fets amb la metodologia UML.

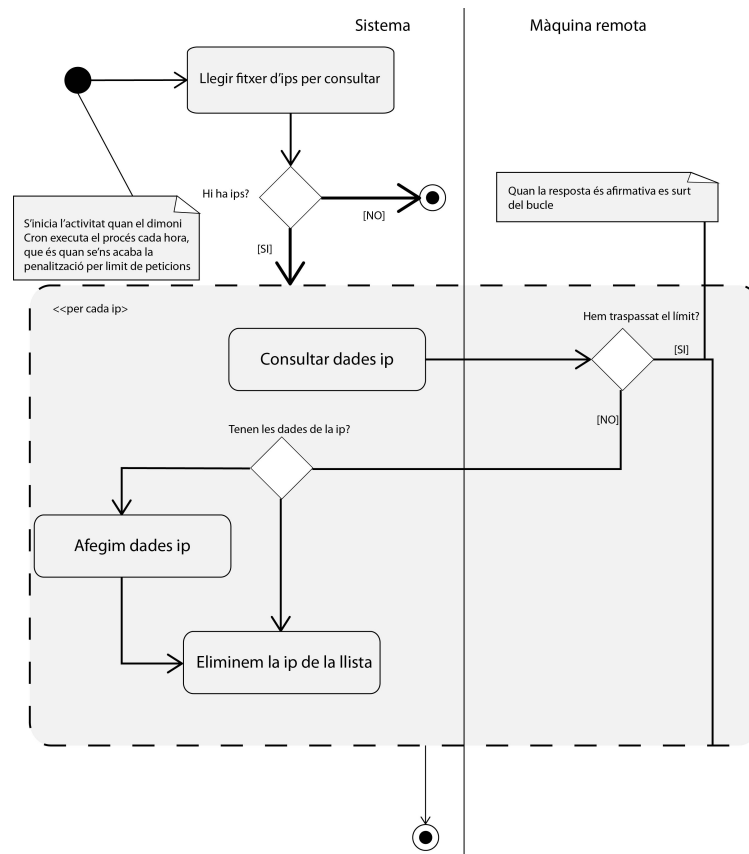
El diagrama següent mostra l'activitat de recol·lectar i indexar.



El diagrama següent mostra l'activitat de crear les estadístiques. És una activitat que es farà periòdicament i automàticament.



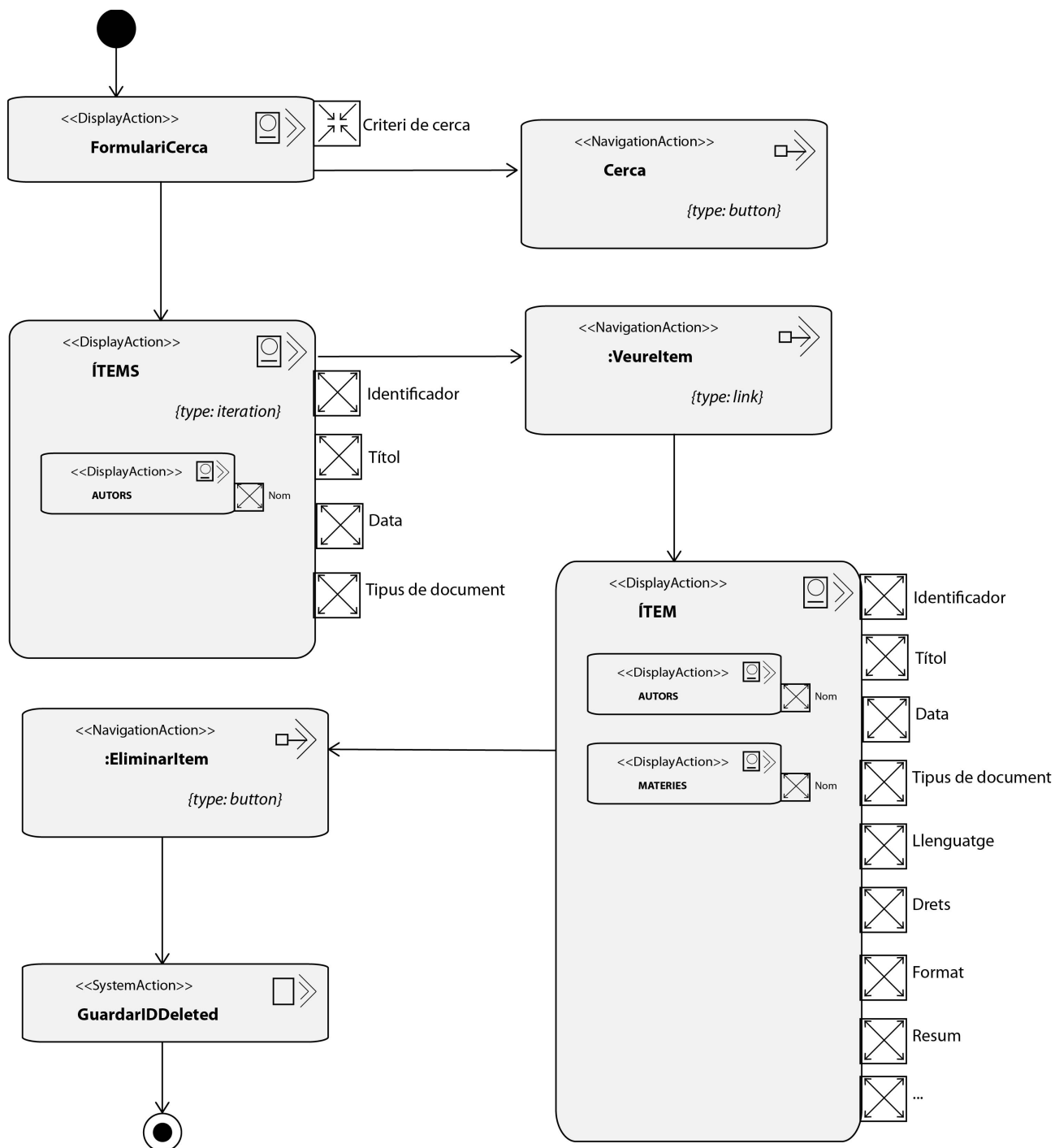
Aquest diagrama representa l'activitat de consultar les dades d'una ip. L'activitat es farà automàticament cada hora, ja que la API utilitzada ens permet consultar-ne 20 cada hora.



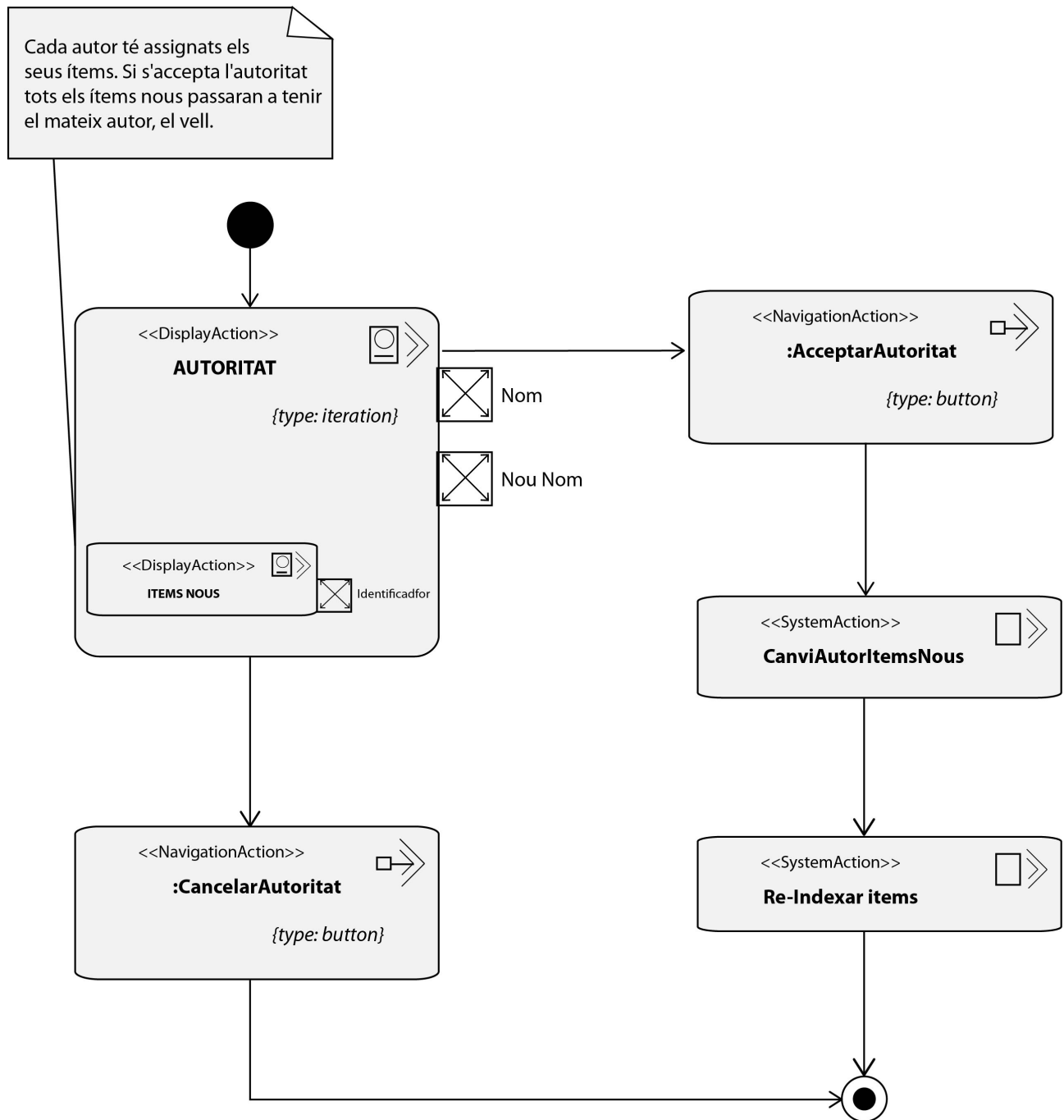
Diagrames d'activitat de l'usuari i l'administrador

La resta de diagrames d'activitat estan fets utilitzant la metodologia UWE, ja que són processos o activitats de navegació.

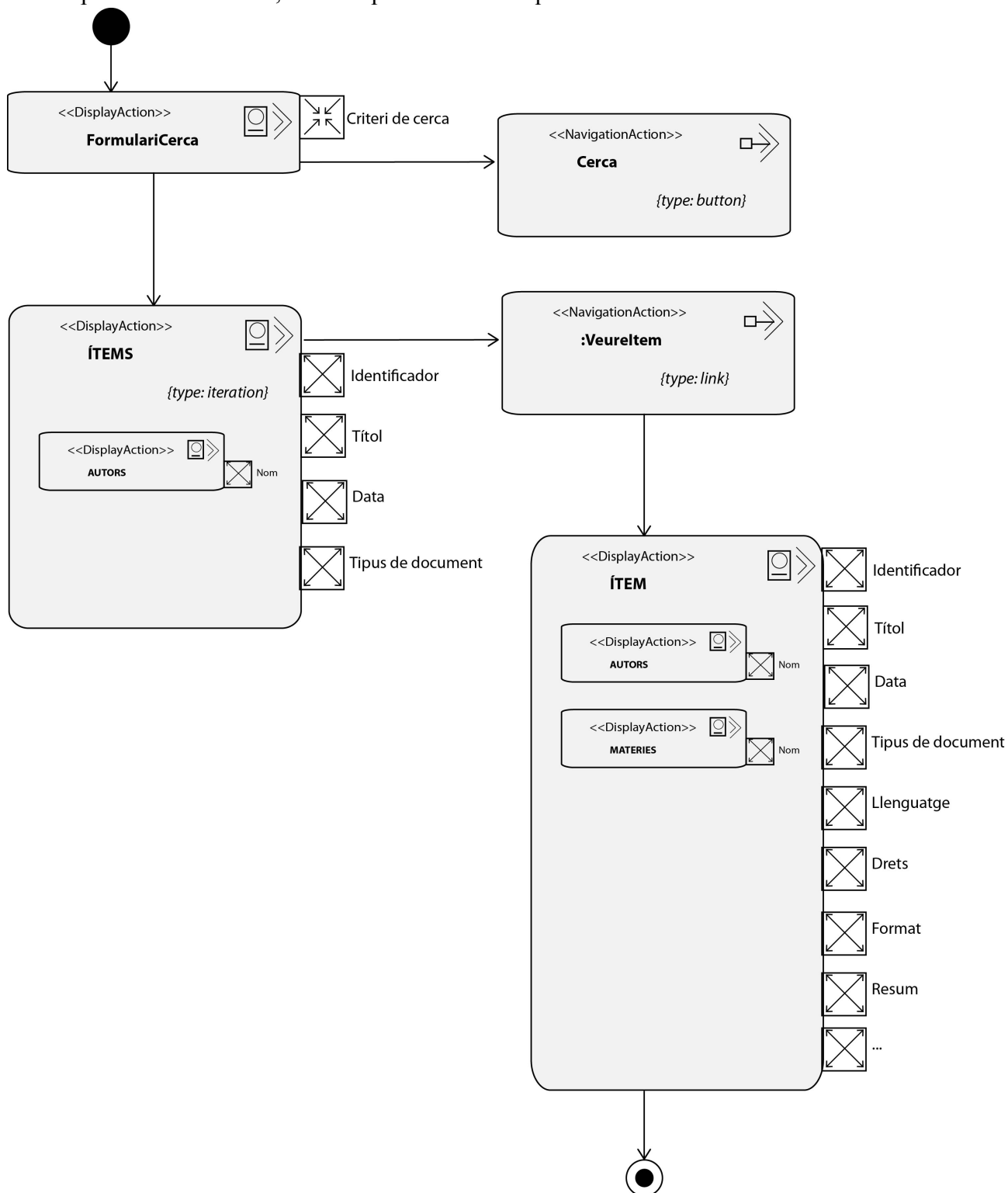
El diagrama següent representa l'activitat d'eliminar un ítem. Hem afegit l'activitat de **GuardarIDDeleted** que ens permetrà comprovar que els ítems entrats a posterior no hagin estat eliminats.



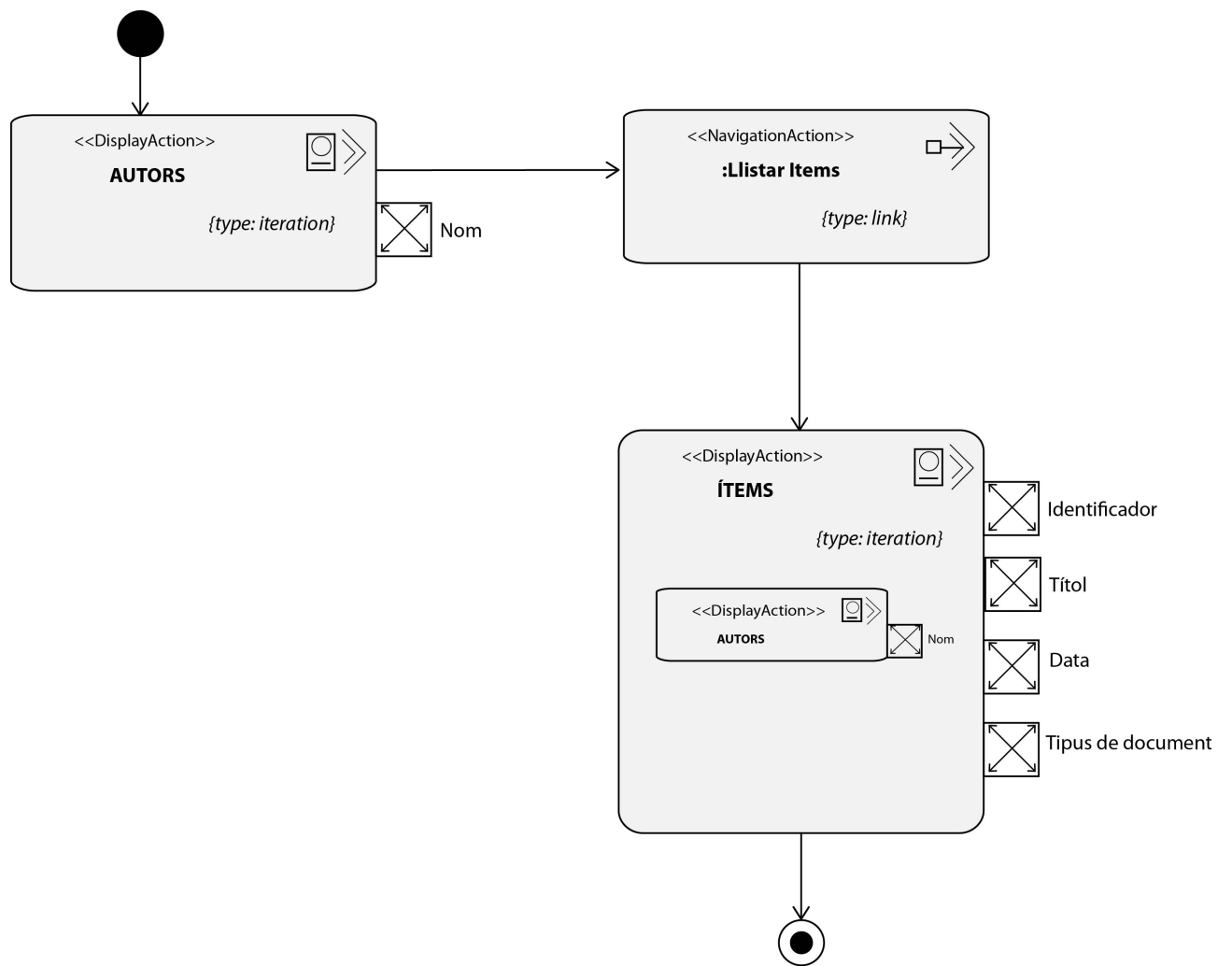
El diagrama següent ens mostra l'activitat de confirmar una autoritat. El sistema ens proporciona un llistat que consisteix en dos noms d'autors que ell ha relacionat i l'ítem o ítems que estan assignats a l'autor nou. L'administrador ha de decidir si la relació que ha fet el sistema és correcta, i aquells dos autors són la mateixa persona. En aquest cas, els ítems de l'autor nou, que és l'incorrecte, passen a tenir l'autor correcte.



El diagrama següent ens mostra l'activitat de cercar i mostrar un ítem. Aquest diagrama seria el mateix per al llistat d'ítems, sense la part de la cerca que tenim a l'inici.



Aquest últim diagrama ens mostra el cas dels autors, que seria exactament igual per a les matèries. Un cop al llistat d'ítems, tindrem la possibilitat de veure l'ítem detalladament. En aquest diagrama no ho hem mostrat ja que seria idèntic al diagrama mostrat anteriorment.

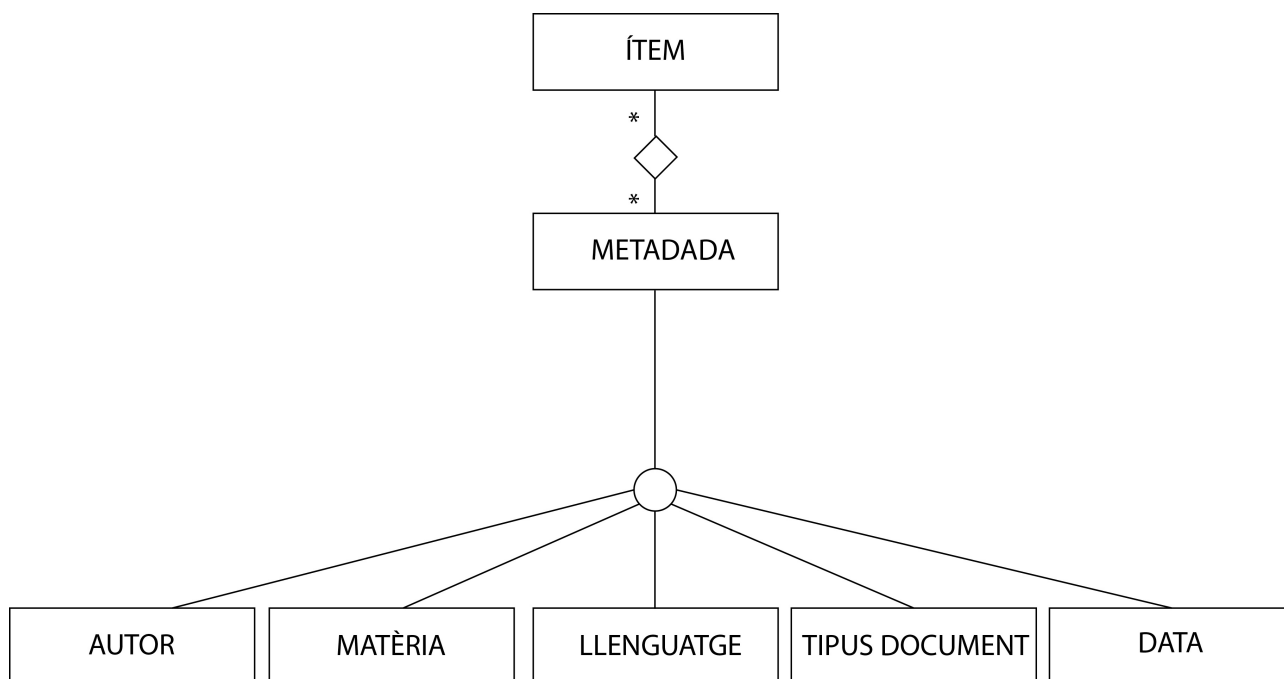


8.1.4 Model de dades

En el nostre projecte només tenim l'objecte Ítem.

Un ítem és el conjunt que agrupa la URL d'uns documents (poden ser més d'un), el seu títol, els seus autors, el seu format, tipus, paraules clau, descripció, etc. No té sentit parlar d'autors si no parlem d'un ítem. Ni té sentit parlar de les paraules clau si no fan referència a un ítem.

Així doncs, en un model de dades entitat relació podríem tenir una taula ítem relacionada amb una taula metadada (*:*) i heretant d'aquesta, totes les metadades que volem.



Tanmateix, amb una única taula ítem que contingués un títol, un vector d'autors, un vector de paraules clau, descripció, un vector de URL's de documents, etc, en fariem prou.

ÍTEM
- Títol
- Autors *
- Matèries *
- Resum
- Documents *
...

Però nosaltres el que volem és tenir moltes oportunitats de trobar el que busquem, així doncs, el que ens interessaria seria una cosa més tipus això.

ÍTEM
- Títol
- Títol (<i>minúscules</i>)
- Títol (<i>Capitalitzat</i>)
- Títol (<i>Tokenitzat</i>)
- Títol (<i>Sense Stop Words</i>)
- Títol (<i>Sense accents</i>)
...

Per crear l'esquema de dades al SOLR, hem definit els següents camps:

A l'ANNEX 2 sobre el SOLR s'hi pot trobar més informació sobre els camps i els tipus de camp de solr.

Camps principals:

Nom	Tipus	Guardat	Indexat	Multi- valuat	Requerit	Descripció
dc.identifier.uri	String	Si	Si	No	Si	Aquest camp serà l'identificador únic de l'ítem. Tindran forma d'URL, ja que utilitzarem el handle dels repositoris.
dc.identifier.issn	String	Si	Si	Si	No	L'ISSN de l'ítem.
dc.identifier.isbn	String	Si	Si	Si	No	L'ISBN de l'ítem.
dc.identifier.citation	String	Si	Si	No	No	Cita de l'ítem
dc.identifier	String	Si	Si	Si	No	Altres identificadors
dc.identifier.other	String	Si	Si	Si	No	Altres identificadors
dc.creator	Text	Si	No	Si	No	Autor
dc.contributor.other	Text	Si	No	Si	No	Altres autors, directors...
dc.contributor	Text	Si	No	Si	No	Altres autors
dc.title	Text	Si	Si	Si	Si	Títol de l'ítem
dc.subject	Text	Si	No	Si	No	Matèries de l'ítem
dc.description.abstract	Text	Si	No	Si	No	Resum de l'ítem
dc.description	Text	Si	No	Si	No	Altres descriptors
dc.description.Sponsorship	Text	Si	No	No	No	Sponsor de l'ítem
dc.publisher	Text	Si	Si	Si	No	Editors de l'ítem
dc.format	String	Si	No	Si	No	Format de l'ítem
dc.format.mimetype	String	Si	No	Si	No	Mimetype de l'ítem
dc.date	String	Si	No	No	No	Data del document
dc.rights	Text	Si	No	No	No	Drets del document
dc.rights.uri	String	Si	No	No	No	URL dels drets
dc.rights.accessRights	Text	Si	No	No	No	Tipus d'accés d'alguns ítems de projectes Europeus
dc.type	String	Si	No	No	No	Tipus de document de l'ítem
dc.type.version	String	Si	No	No	No	Versió del document
dc.language	String	Si	No	Si	No	Llenguatge de l'ítem
dc.relation.uri	String	Si	Si	Si	No	URL d'ítems relacionats

dc.relation	String	Si	No	Si	No	Relacions
dc.relation.ispartof	String	Si	No	Si	No	Si forma part d'algun tot més gran
dc.relation.isformatof	String	Si	No	Si	No	Si és un mateix ítem en un altre format
dc.coverage.spatial	String	Si	No	No	No	Abast del contingut del recurs
dc.source	String	Si	No	No	No	Font del document
repository	Text	Si	Si	No	Si	Repositori del qual prové
files	String	Si	No	Si	No	Fitxers de l'ítem

Camps auxiliars:

Nom	Tipus	Guardat	Indexat	Multi- valuat	Requerit	Descripció
text	Text	No	Si	Si	No	La cerca per defecte es farà a aquest camp
data_aux	String	No	Si	No	No	Data auxiliar per a facetes i ordenació
text_autors	String	No	Si	Si	No	Camp auxiliar per a les facetes
text_subject	String	No	Si	Si	No	Camp auxiliar per a les facetes
text_lang	String	No	Si	Si	No	Camp auxiliar per a les facetes
test_type	String	No	Si	Si	No	Camp auxiliar per a les facetes
text_title	text_title	No	Si	No	No	Camp auxiliar per a la ordenació

Els camps auxiliars, contindran una còpia dels camps principals, però amb un altre tipus de camp. Al camp text hi copiarem els següents camps per a la cerca per defecte:

- dc.identifier.uri
- dc.title
- dc.identifier.citation
- dc.description.abstract
- dc.publisher
- dc.creator
- dc.contributor
- dc.contributor.other
- dc.subject
- dc.language
- dc.date

Els camps al SOLR poden ser només *indexats*, *indexats i guardats* o només *guardats*.

Segons la multiplicitat: *simples* o *multi-valuats*. Cal tenir en compte que en el cas de no tenir un camp indexat, la cerca serà més lenta.

Els tipus de camps que hem fet servir, han sigut *text*, *string* i *text_title*.

El tipus de camp text porta un tokenitzador per espais, un filtre per convertir-ho tot a minúscules, un filtre de sinònims, un filtre per treure les paraules buides, un filtre per els accents, i un filtre per lematitzar sense tenir en compte l'idioma.

El tipus de camp *string* el necessitem per tots els camps on no s'ha d'aplicar cap filtre ni tokenitzar. I el tipus de camp *text title*, no tokenitza la cadena, però aplica un filtre per treure les paraules buides del principi del títol, ja que no s'han de tenir en compte per ordenar. El camp **text** és el camp on l'índex cercarà per defecte.

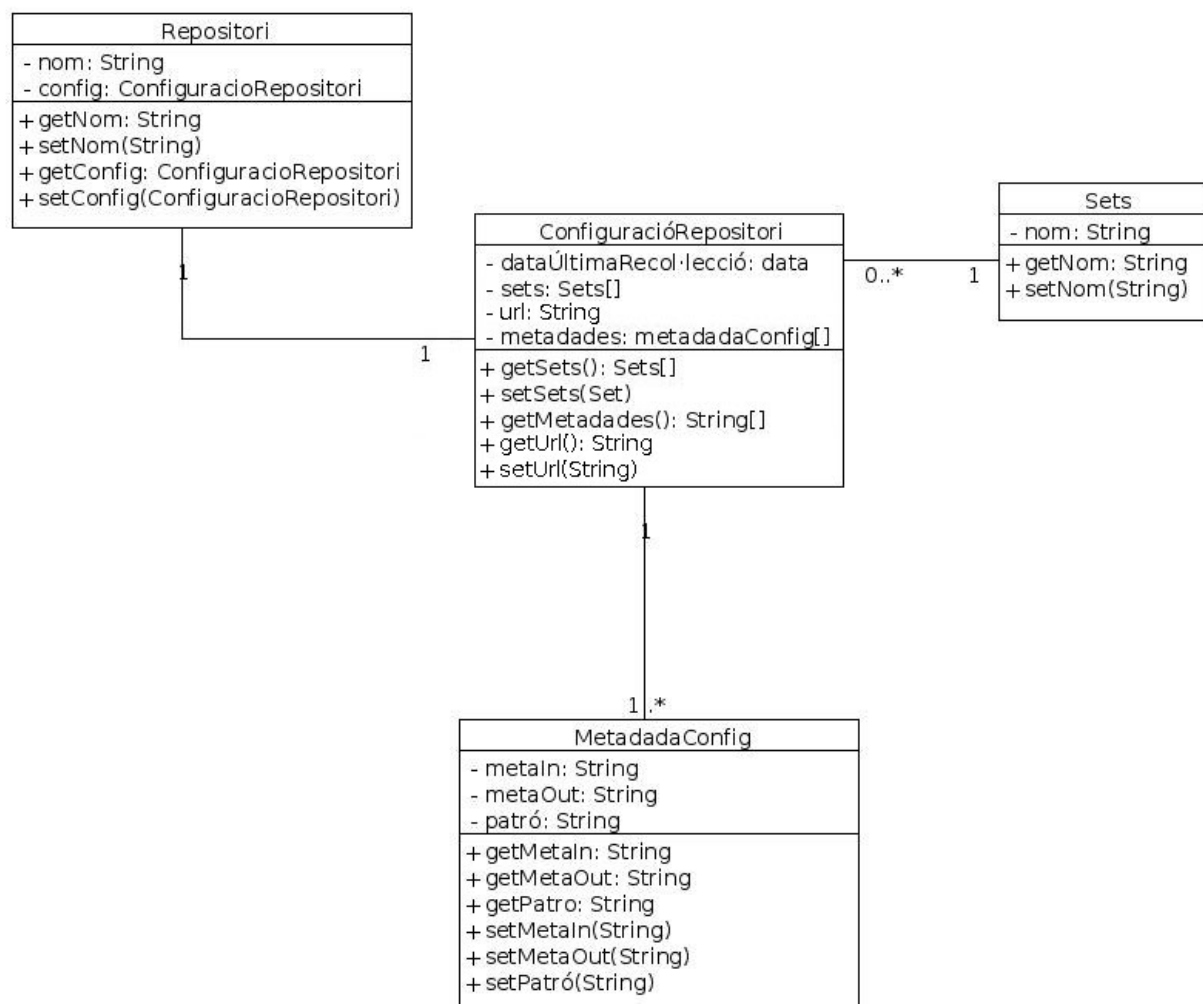
8.2 Disseny

Els diagrames de disseny ens permetran acabar de modelar més detalladament el projecte. En aquest punt també separarem els diagrames de classes de la part del sistema, dels d'administració i consulta.

8.2.1 Diagrames de classes de la part del sistema

Vegem primer la lògica de la part de sistema.

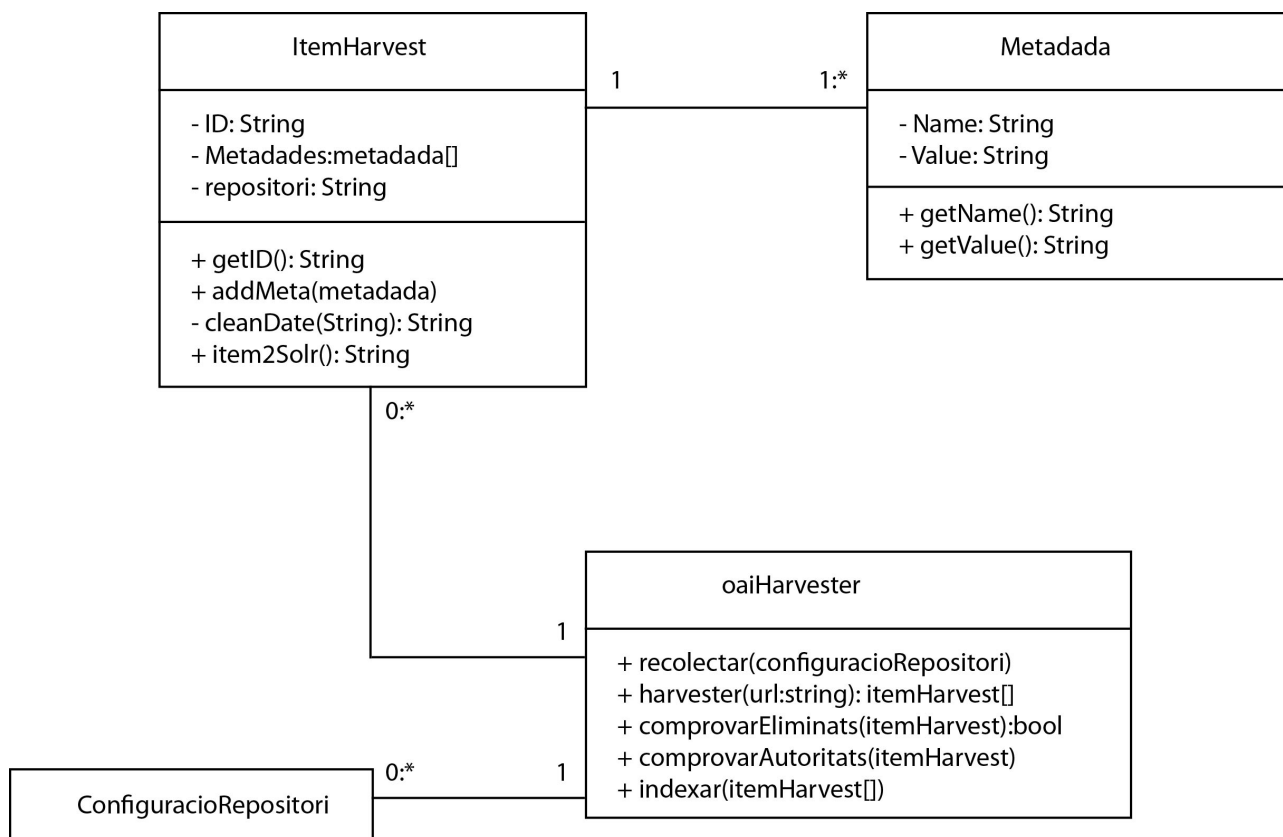
L'objecte **configuracióRepositori** encapsula els sets i les metadades de configuració d'un repositori. L'objecte **metadataConfig** contindrà els mètodes i atributs per permetre identificar les metadades i endevinar el qualificador perdut en la recollida.



El diagrama següent ens mostra la relació de la **configuracioRepositori** vista anteriorment amb la classe encarregada de recollir aquests repositoris. Aquesta es relaciona amb **itemHarvest** que

està format de metadades. **ItemHarvest** conté els mètodes necessaris per preparar l'ítem per importar-lo a l'índex SOLR.

El mètode **cleanDate** ens permetrà netejar la data per crear l'atribut **dataAux** de la classe ítem del sistema de consulta.



Per la part de la consulta, el nostre sistema és basa quasi al 100% en l'objecte “**ítem**”. Realment només mostrem ítems i els llistem agrupats segons diferents atributs com seran autors, matèries, data...

Així doncs, a la nostra classe ítem hi necessitem un mètode que ens retorni tots els atributs, que en aquest cas seran metadades que poden ser úniques o repetides, i uns altres que ens permetin afegir aquestes metadades.

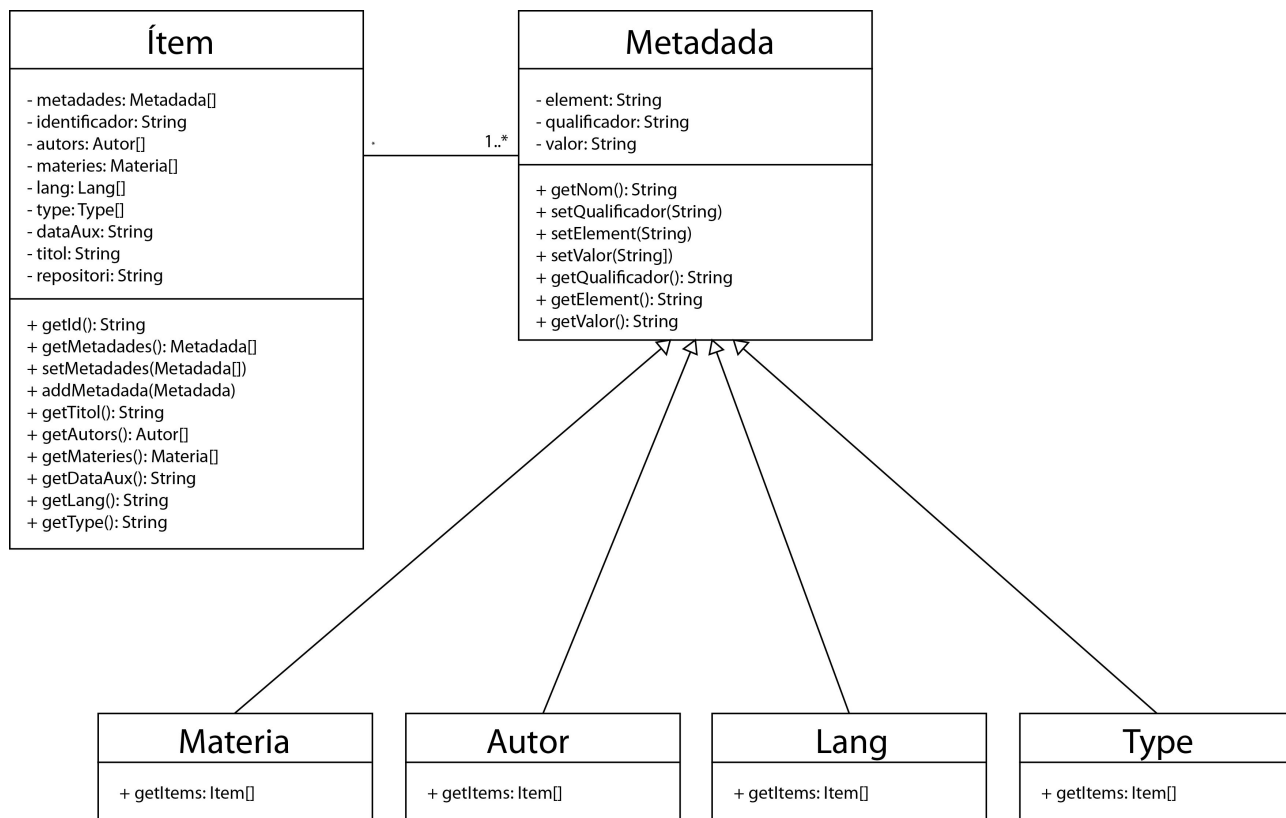
En el cas d'atributs tant importants com serien l'autor, la matèria, el títol, l'identificador i la data, podem crear uns mètodes que ens retornin i ens creïn aquestes metadades.

El diagrama següent podem veure que **Matèria** i **Autor** hereten de **Metadada** perquè necessitem obtenir els ítems que tenen aquesta metadada. És important fer notar que la classe **metadada** no és la mateixa utilitzada en el moment de *harvestejar*.

També cal destacar el cas de l'atribut **dataAux**, que és de tipus *String* i no és tractat com una metadada ni com una data perquè quan no se sap la data, el món bibliotecari tendeix a posar-hi valors del tipus *[s.d.]* o *19??*¹⁸.

La classe **Ítem** no conté cap mètode per afegir una data ni l'identificador perquè es considera que aquests atributs s'afegeixen en el moment de creació de l'ítem, i que són immutables. El mateix passa amb el nom de la classe metadada.

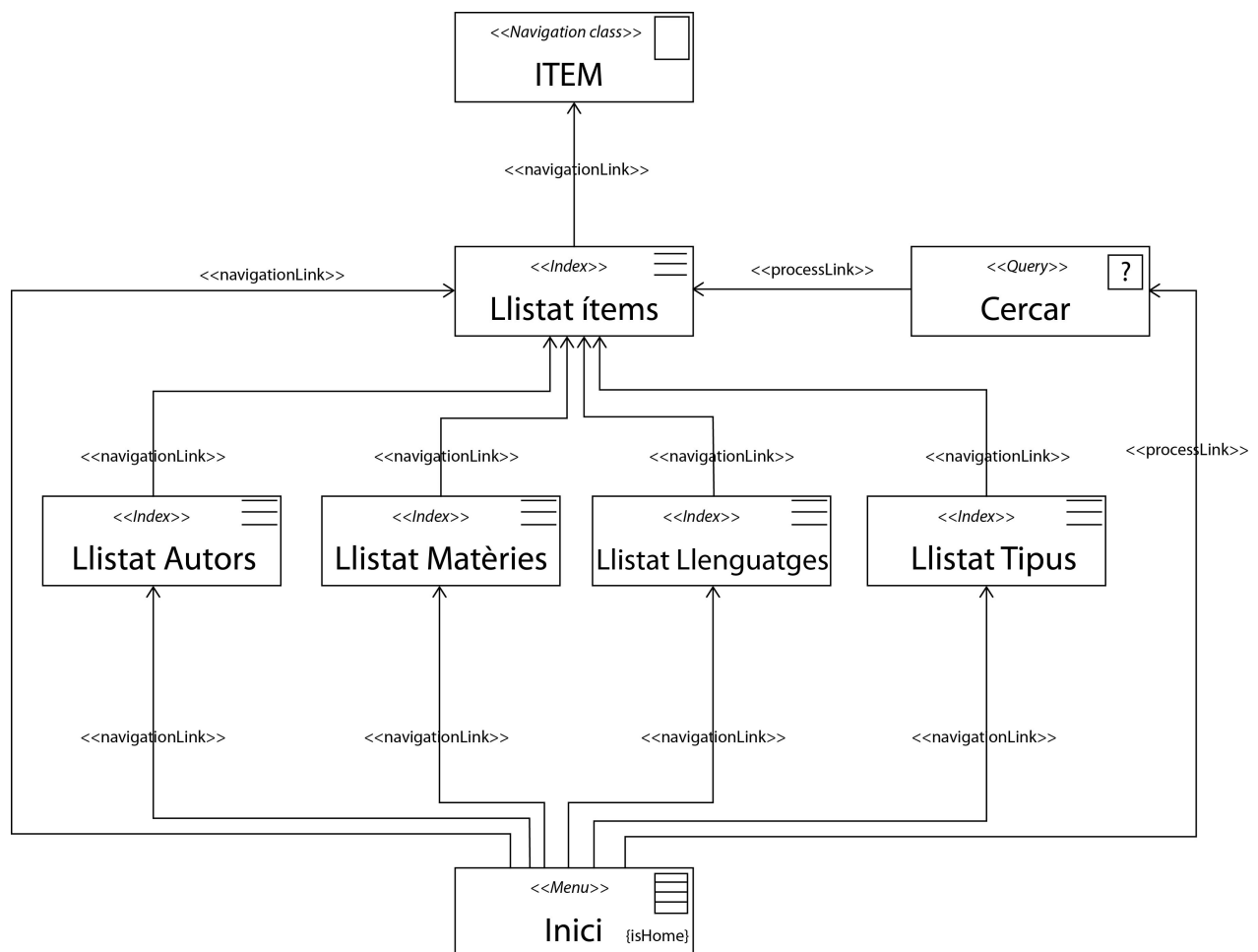
¹⁸ [s.d.] es refereix a sine data, i s'utilitza quan no en coneixem la dada. En l'exemple de 19?? entenem que no sabem l'any exacte, però estem segurs que és del segle XX.



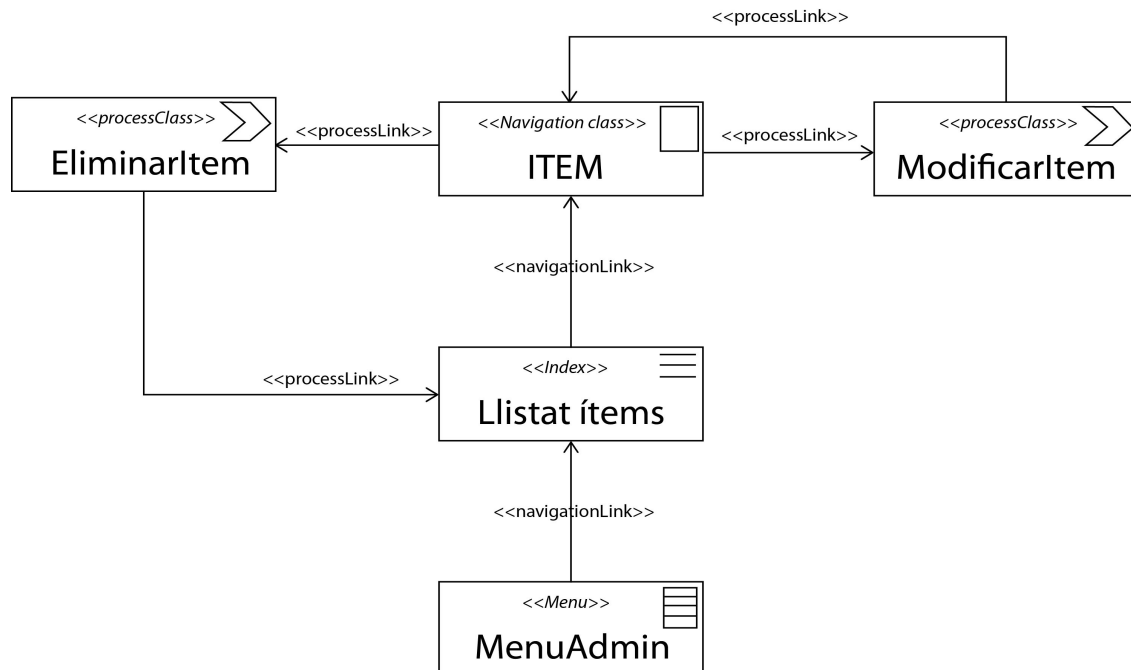
8.2.2 Model de navegació

Els diagrames de navegació ens permetran saber com estan enllaçades les pàgines de la nostra web. Per crear-los, partim del model de contingut. Més concretament del diagrama de classes de la part de consulta, ja que aquest diagrama només té sentit en la part visible del sistema.

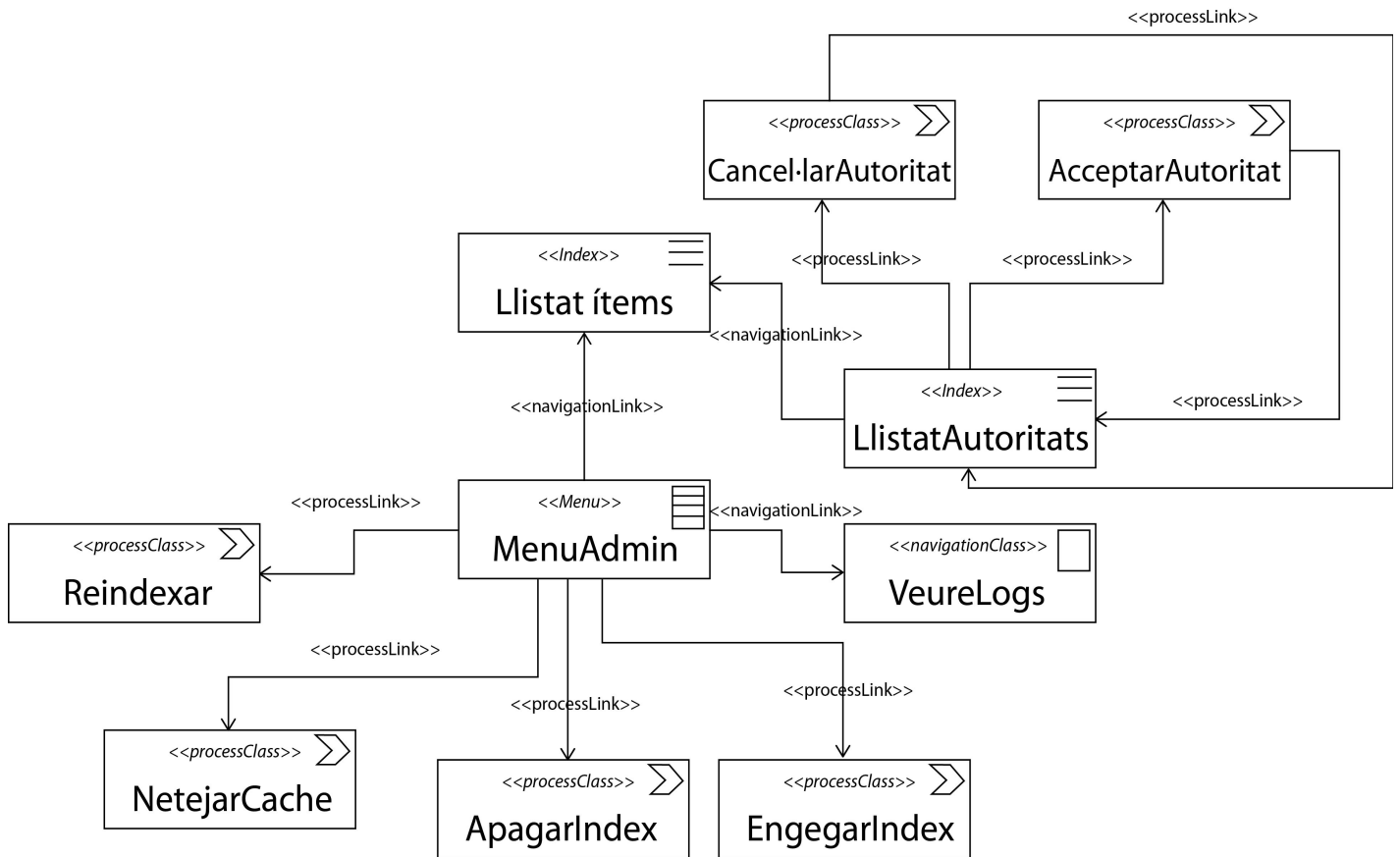
El diagrama següent mostra la navegació de l'usuari sense privilegis, podem veure com a partir del menú principal, podrà llistar els ítems directament, llistar els autors, matèries, llenguatges i tipus, i després llistar els ítems que tenen aquestes metadades, o cercar ítems, que també condueix al llistat d'ítems des d'on es pot visualitzar un ítem concret.



Aquest altre diagrama mostra la navegació de l'administrador. Aquest mostra la navegació respecte a l'objecte **Item**. S'ha de tenir en compte que tant l'usuari com l'administrador comparteixen tota la navegació mostrada en el diagrama anterior. I aquest diagrama només mostra la referent a l'ítem per simplificar.



Aquest altre diagrama mostra la resta de navegació de l'administrador. Des del llistat d'autoritats es poden acceptar o cancel·lar, i en ambdós casos es torna al llistat. També es dona la possibilitat de consultar els ítems assignats a cada autoritat.



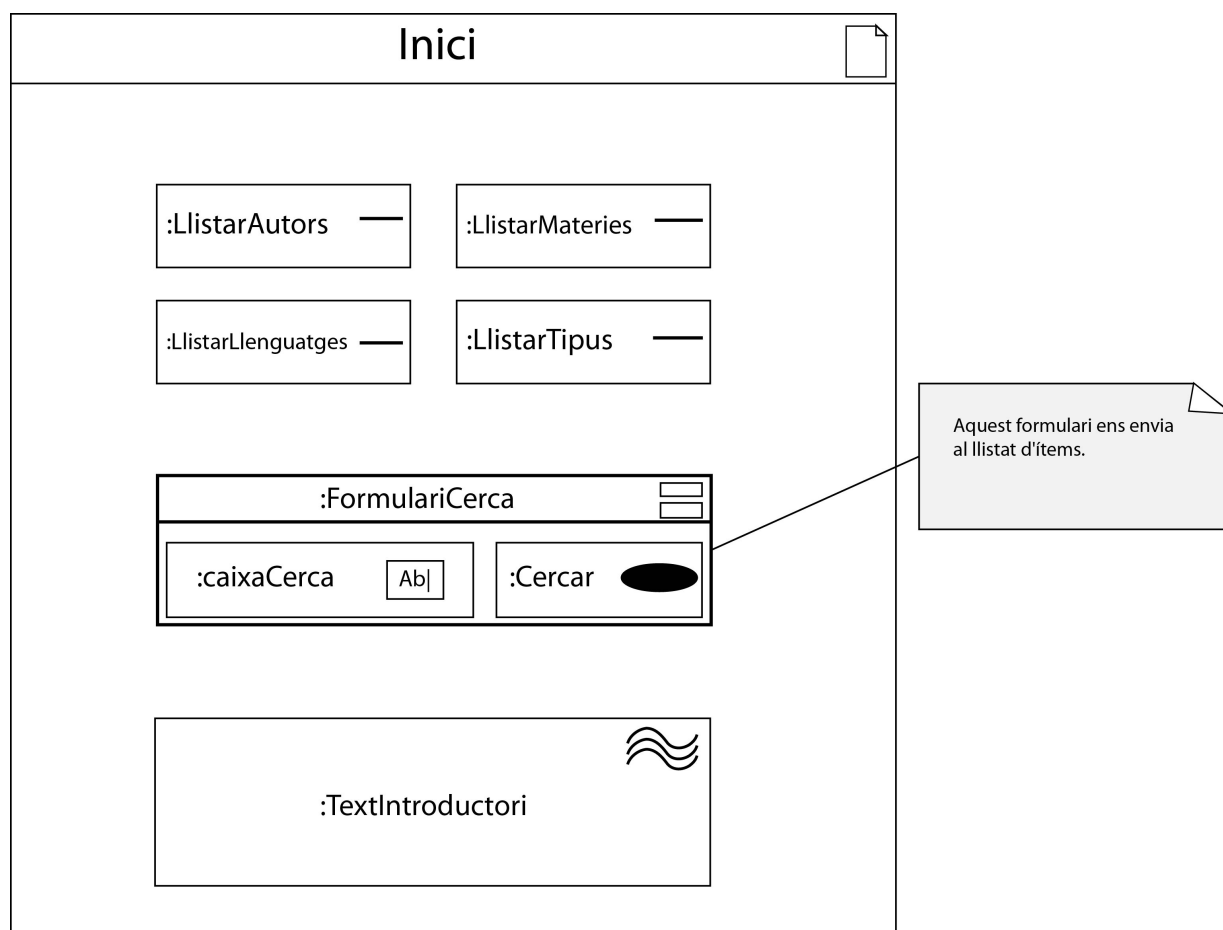
8.2.3 Model de presentació

Els diagrames de presentació ens permeten expressar quin element està contingut dins d'una pàgina. Ens poden ser molt útils per estructurar la informació i anar un pas més enllà que els diagrames de navegació, que només ens permetien dissenyar quines pàgines ens porten a quines altres.

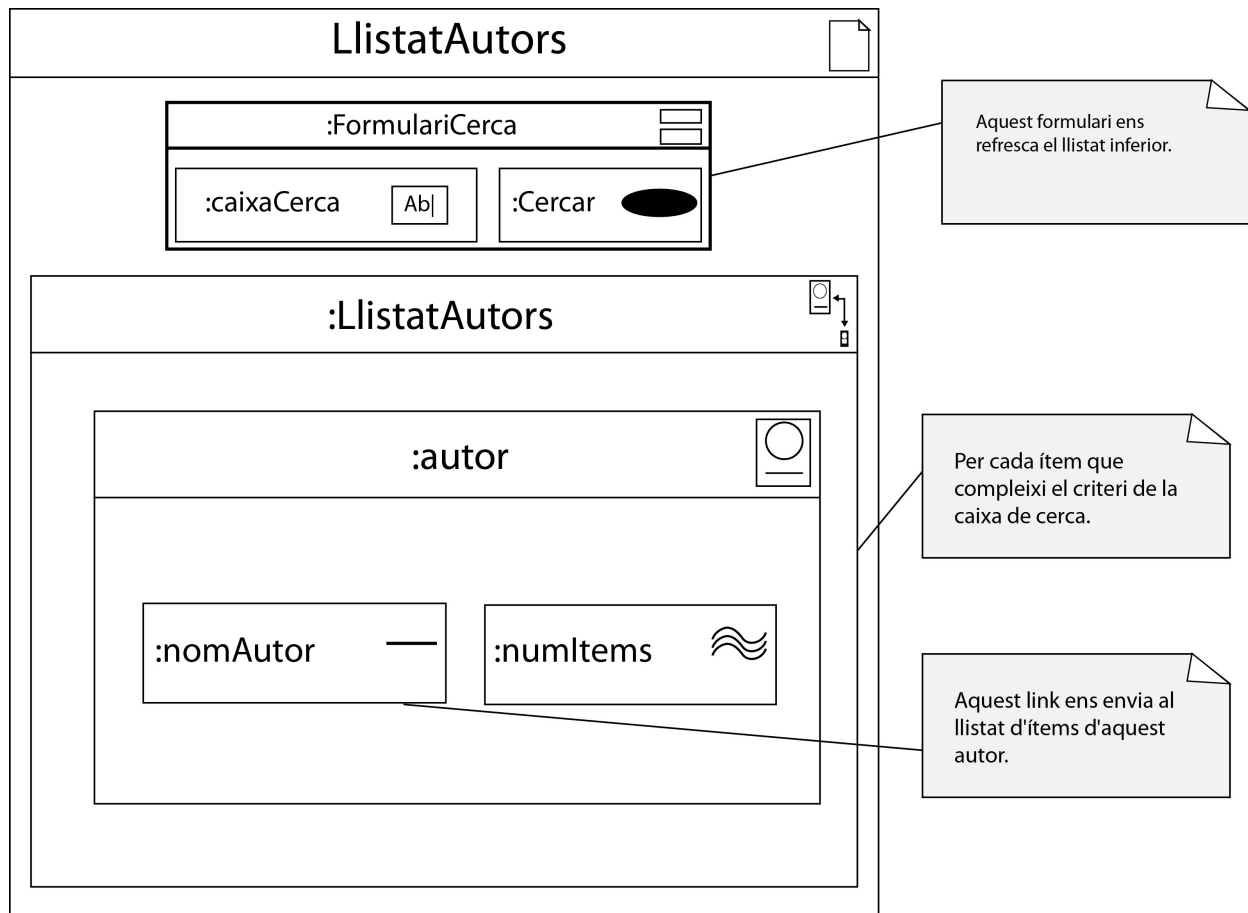
En mostrarem els més rellevants i representatius:

- Inici
- Llistat d'autors
Que també representa el llistat de matèries, el de llenguatges i el de tipus de document.
- Llistat d'ítems
- Llistat d'autoritats

Aquest diagrama següent és molt senzill. A la pàgina només i tindrem els links cap als llistats d'autors, matèries, llenguatges i tipus, que formaran el menú principal, el formulari de cerca i el text de benvinguda i d'introducció al DUGi. El formulari de cerca ens portarà a la pàgina del llistat d'ítems.



Aquesta serà la pàgina tipus dels llistats d'autors, matèries, llenguatges i tipus de documents. Hi tenim un formulari de cerca que ens permet filtrar els autors anant afegint les lletres per les quals comencen. De cada autor hi podem veure el nom i en nombre d'ítems que el contenen. El nom de l'autor en fa de d'enllaç i ens porta al llistat d'ítems d'aquest autor.

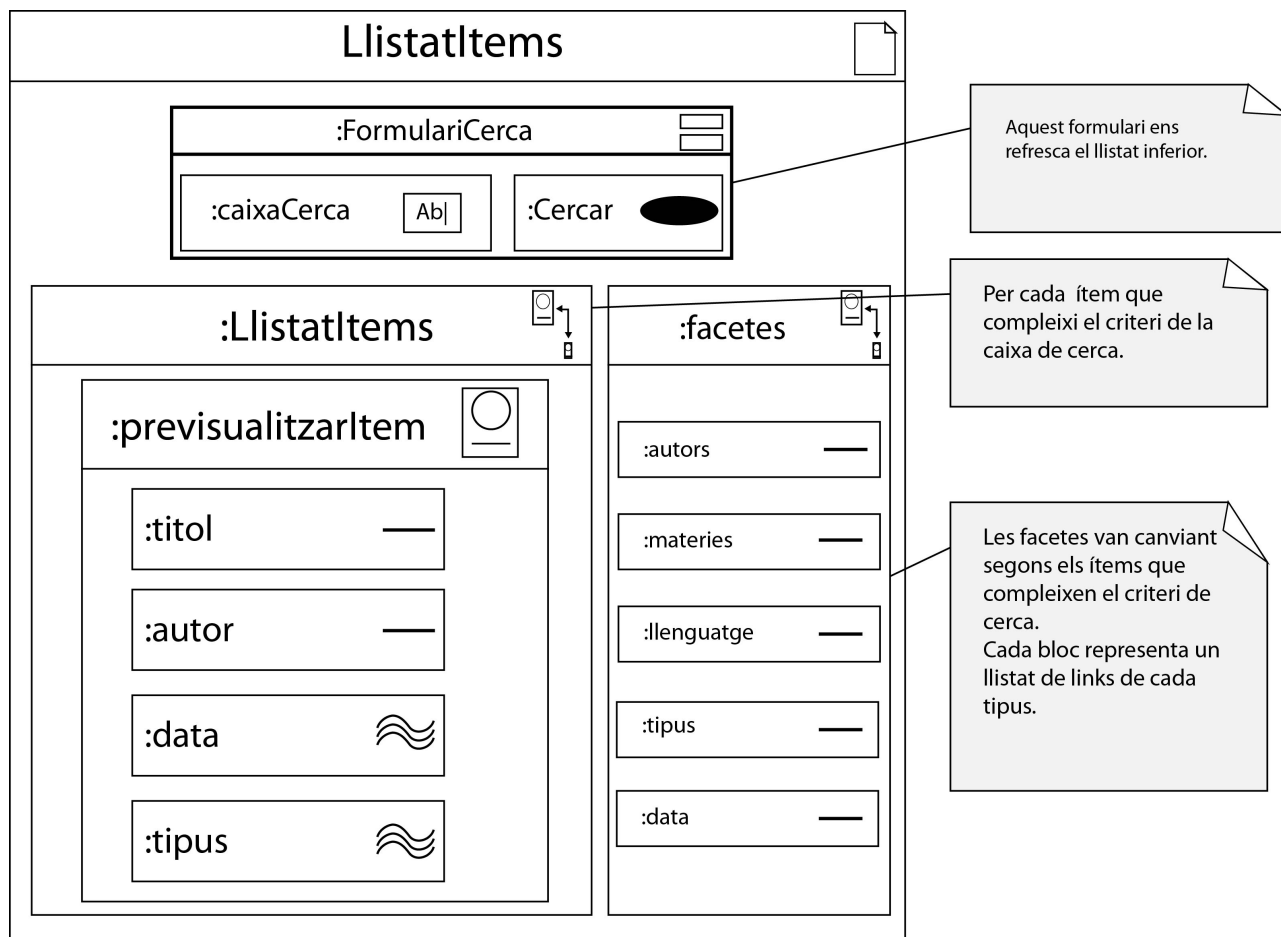


La pàgina del llistat d'ítems és la més important del DUGi, com hem vist en els diagrames de navegació, quasi totes les pàgines ens porten a ella.

Conté un formulari de cerca, que és el mateix de la pàgina principal i ens permet anar refinant la cerca. A més a més, tenim el panell de facetes que encara ens permet refinar més la cerca, podent filtrar per autor, matèria, llenguatge, tipus de document i data.

La previsualització de cada ítem només ens ensenyarà les dades bàsiques. Per exemple el títol, l'autor, la data i el tipus de document. Per veure totes les dades i poder accedir al document, es pot seguir l'enllaç del títol de l'ítem que ens mostrarà la visualització completa de l'ítem.

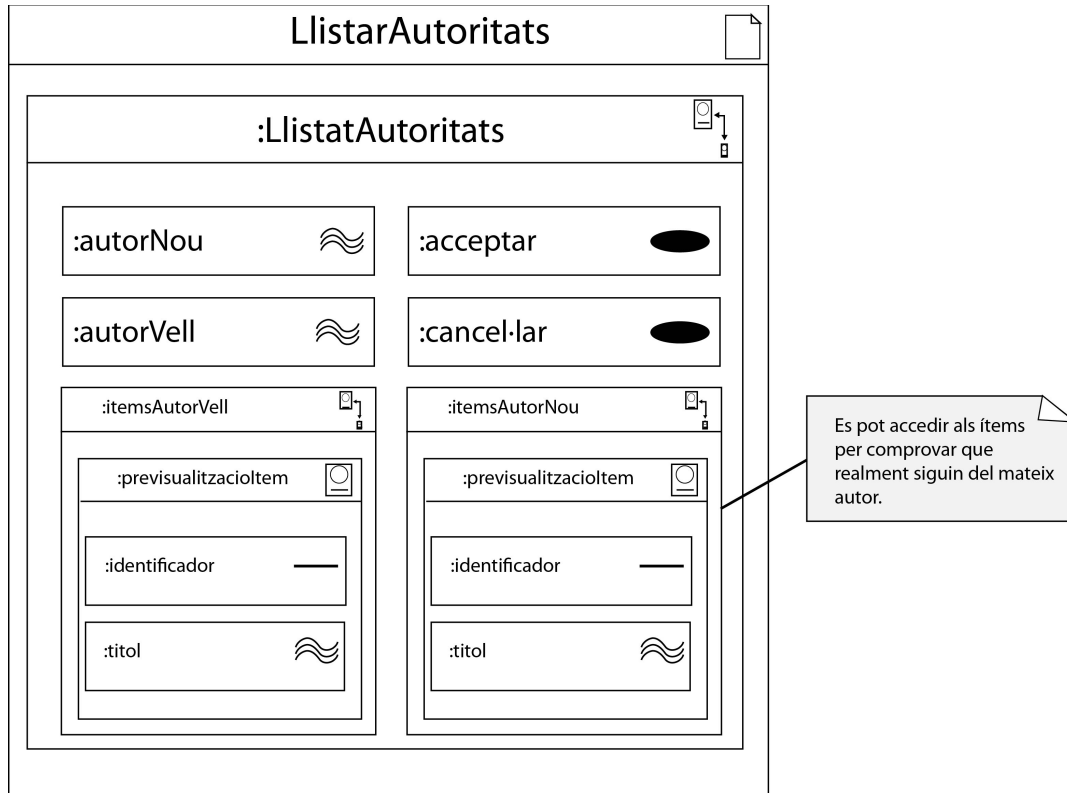
Si es segueix l'enllaç de l'autor, s'anirà al llistat d'ítems d'aquest autor.



La pàgina de llistar autoritats, ens mostrarà els noms dels dos autors, el vell i el nou, els quals el sistema ens proposa d'unificar, els botons d'acceptar o cancel·lar la proposta d'unificació, i un llistat d'ítems.

En altres diagrames s'ha proposat mostrar només el llistat dels ítems nous, però potser serà interessant mostrar també els ítems vells, per poder determinar si és realment el mateix autor.

Aquesta decisió s'acabarà de prendre a la fase de proves.



8.2.4 Model de procés

El model de procés ens permet modelar les accions que es duen a terme quan s'executa un procés. Aquest model es divideix en dues parts molt diferenciades:

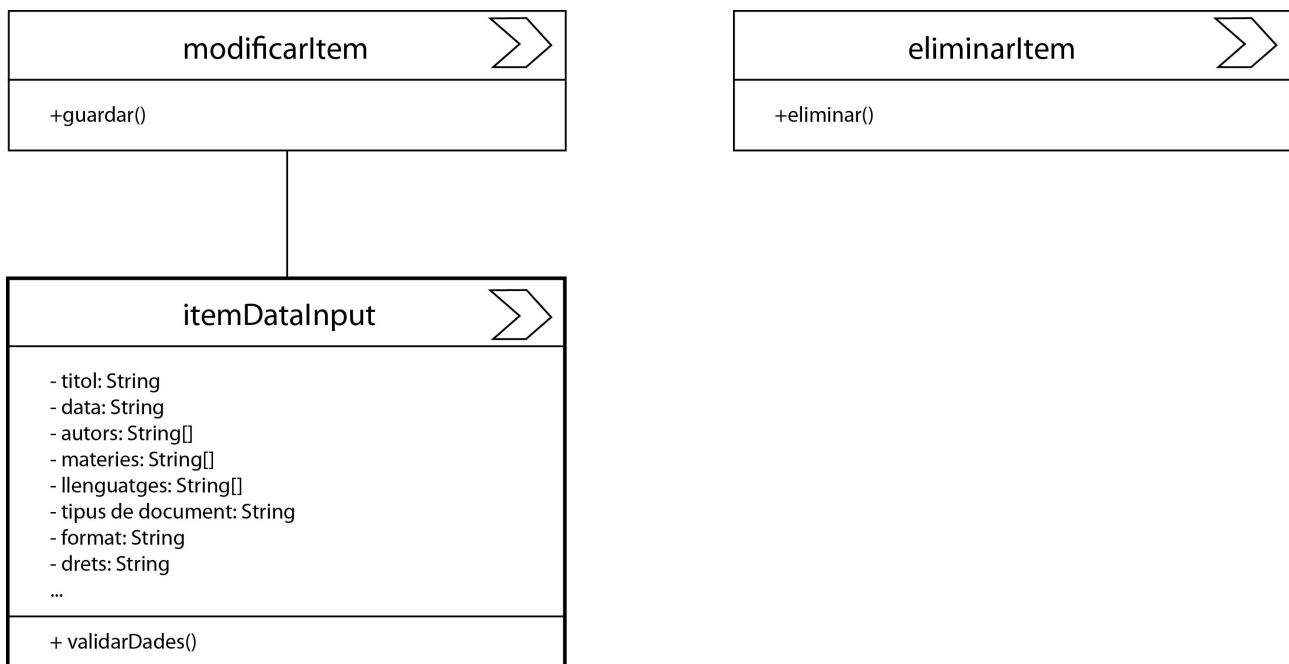
1. El model de l'estructura del procés
2. El model de flux del procés

El primer ens permetrà descriure les relacions entre les diferents classes que entren en el procés, i per tant aquest torna a ser un diagrama de classes.

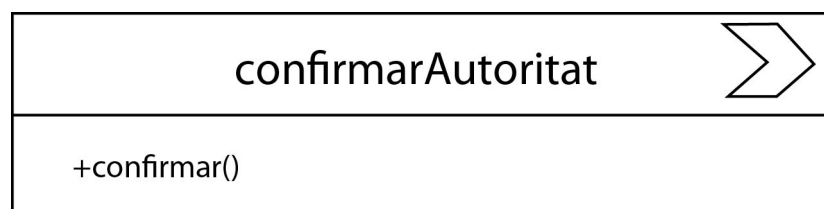
El segon, és representat com un diagrama d'activitats, i ens mostrarà el comportament de cada classe de procés modelada en el diagrama anterior.

Model d'estructura del procés

En el diagrama següent descriu l'estructura dels processos de modificar i eliminar ítem. Podem veure com **eliminarItem** no està relacionat amb cap classe directament. En canvi la classe **modificarItem** està relacionada amb **itemDataInput**, que ens permetrà entrar les dades d'un ítem i validar-les.

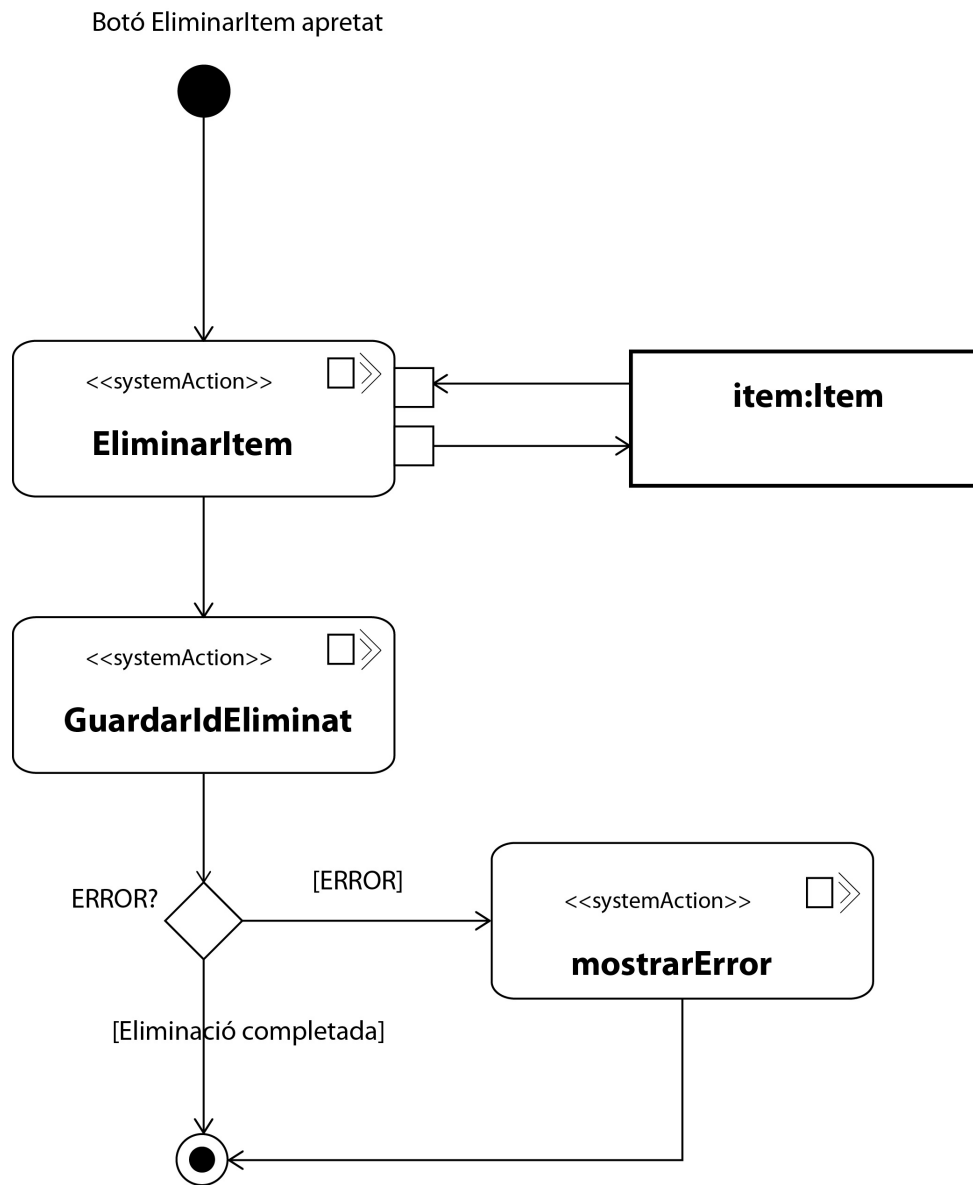


En el cas de confirmar una autoritat, tampoc hi haurà cap relació, ja que el cas és bastant similar al d'eliminar un ítem.

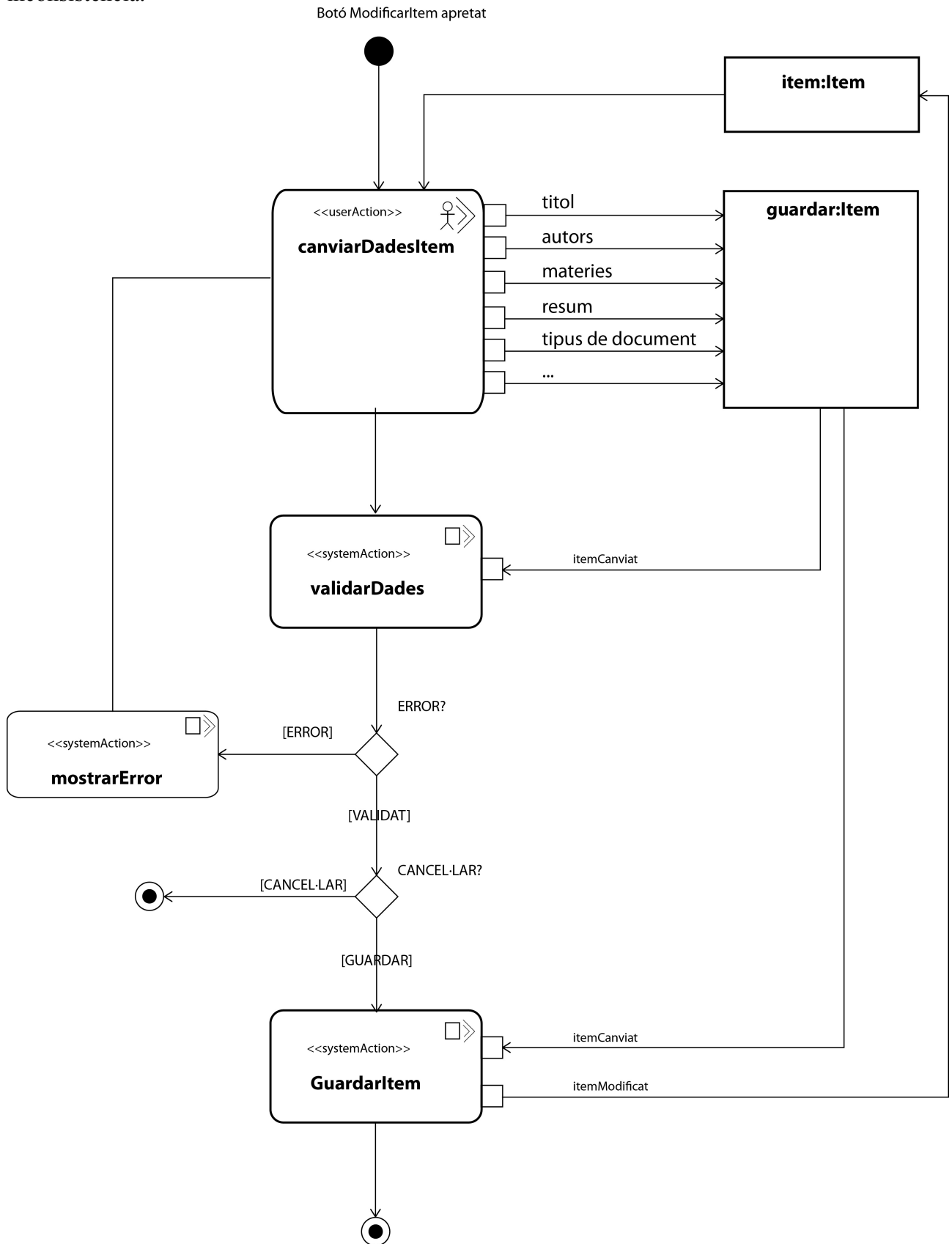


Model de flux del procés

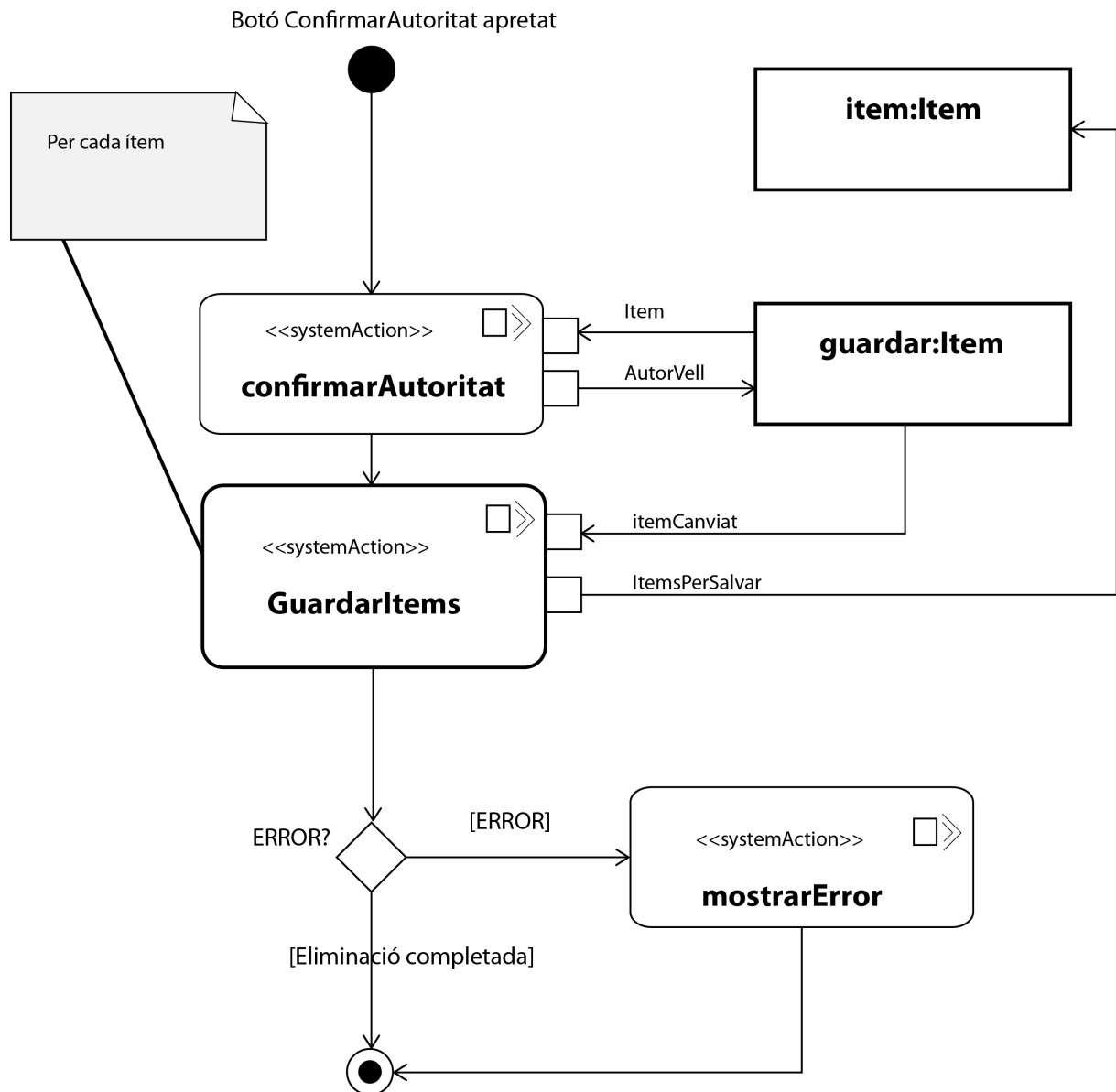
En el cas d'eliminar un ítem, només s'ha de localitzar i posar-lo a la llista negra perquè no es torni a recol·lectar cap altre vegada.



En el cas de modificar l'ítem, s'ha de validar les dades per assegurar-nos que no hi ha cap inconsistència.



Per confirmar les autoritats, cal modificar l'autor de cada ítem dels que pertanyen a l'autor nou, que resulta ser el mateix que el que el sistema ens ha relacionat.

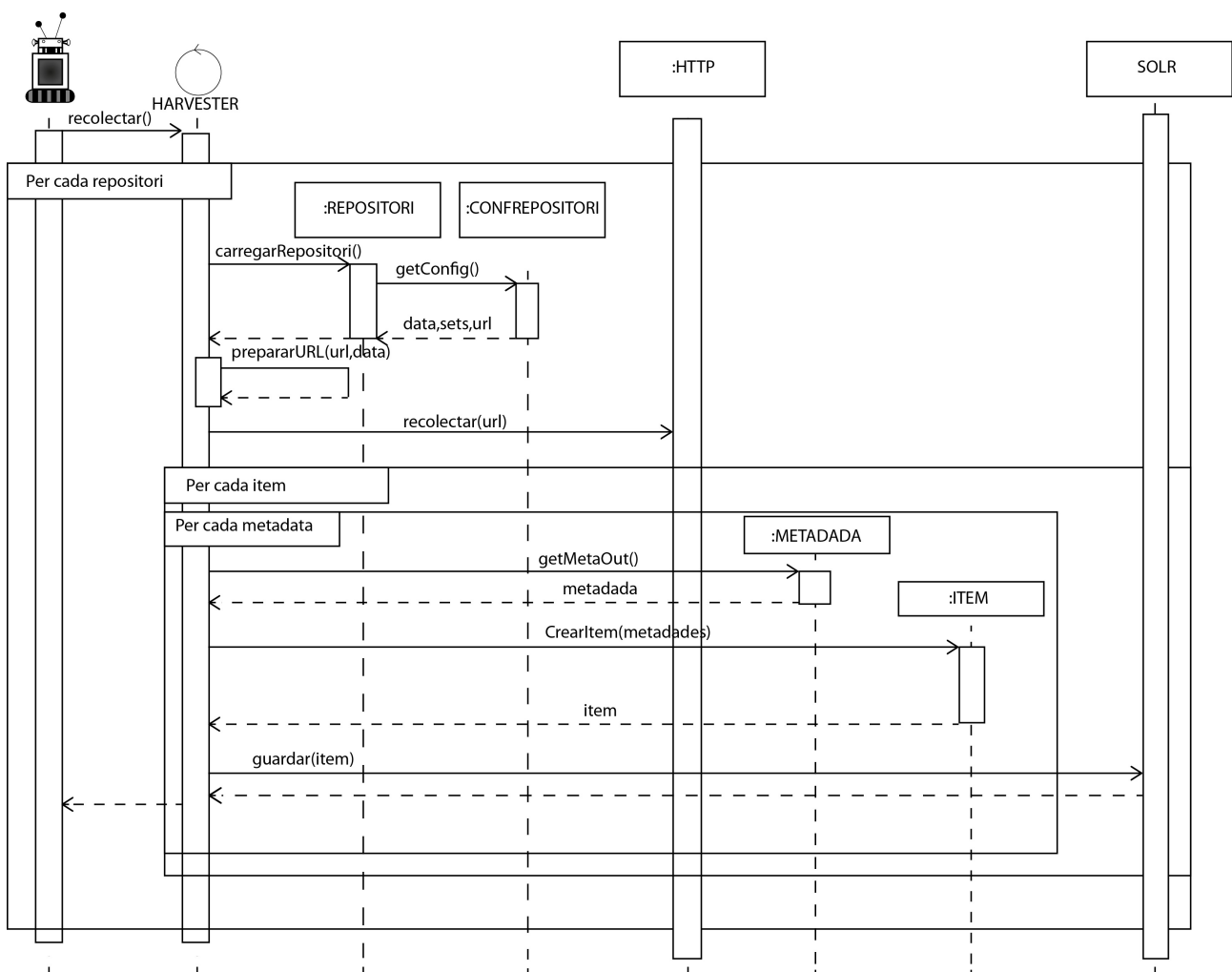


8.2.5 Diagrames de seqüència

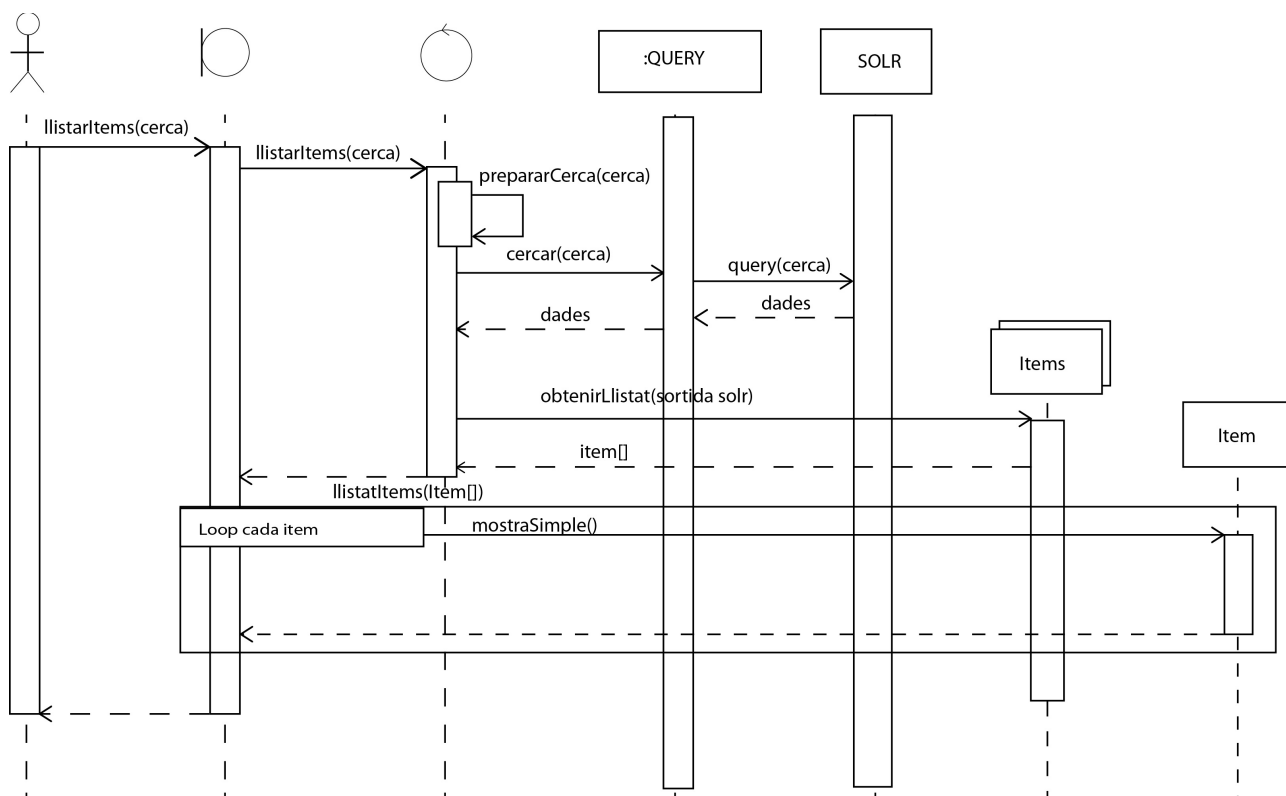
Els diagrames de seqüència i els de col·laboració sempre són interessants perquè ens aporten un punt de vista diferent del dels diagrames de classe. Ens permeten veure els missatges que s'envien els objectes entre ells i apreciar-ne l'ordre. Però si hem d'escollir entre un d'aquests diagrames, potser ens quedàriem amb els de seqüència, que ens permeten fer-nos una idea del temps de resposta de les peticions.

En veurem un parell, el de recol·lecció, i el d'una petició de cerca d'ítems.

En el cas del diagrama de seqüència del procés de recol·lecció, hem de destacar que aquest és executat per el sistema, després és carrega cada repositori, i es fa la petició al repositori, amb l'objecte HTTP, i amb les dades recol·lectades, per cada ítem es transformen les metadades amb el mètode **getMetaOut**, i amb elles es creen els ítems que es guarden a l'índex.



En el cas del procés de llistar els ítems, es pot apreciar el sistema vista-controlador, i al igual que en el diagrama anterior, tant l'objecte **QUERY** com el **SOLR**, ja estan creats, i només els hi fem una petició.

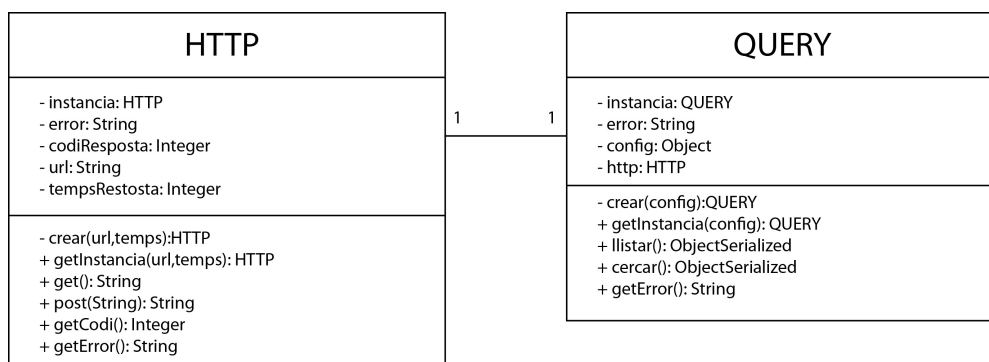


8.2.6 Patrons

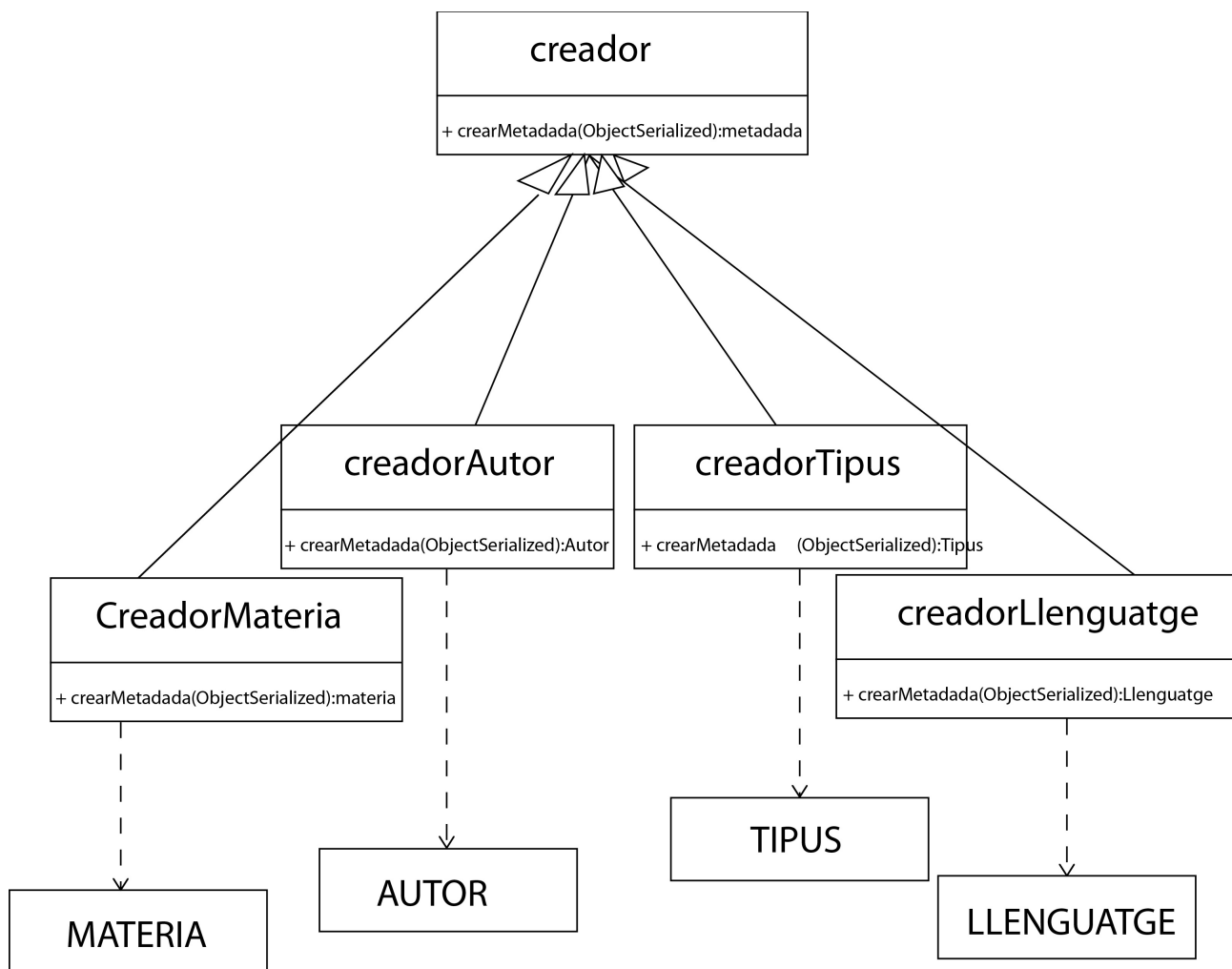
Hem aplicat alguns patrons per tal de solucionar problemes ja coneguts d'una manera que ja es sap que funciona, que és eficient i ens permet reutilitzar els recursos. Hem utilitzat els següents patrons:

- Singleton
- Factory Method
- Facade

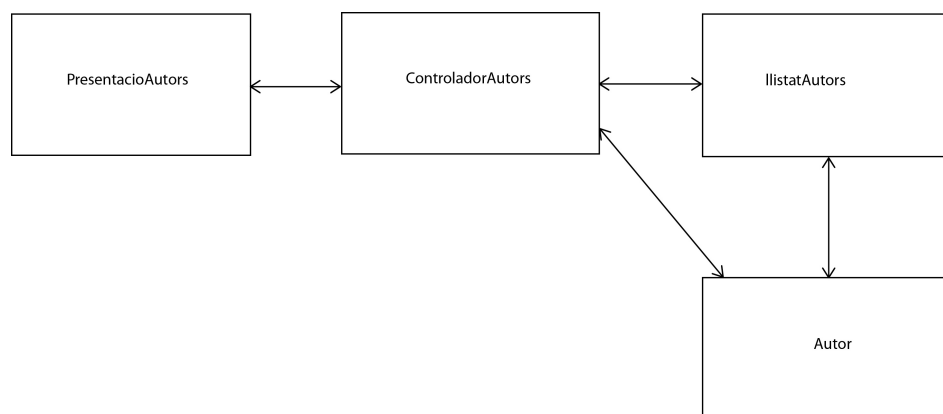
En el cas dels objectes HTTP i QUERY vistos en els diagrames de seqüència, s'ha utilitzat el patró Singleton, ja que només necessitem una instància d'aquests objectes i crear-ne un cada vegada seria poc apropiat. És interessant veure que el SOLR ens retornarà un l'objecte PHP serialitzat, que conté els autors, matèries, etc i el nombre d'ítems, entre altres coses.



Les classes creadorXXX hereten de la classe creador, i redefeixen el mètode **factoryMethod**, retornant-nos l'objecte concret. En aquest cas, l'entrada serà l'objecte *serialitzat* que ens retorna la classe **QUERY** després de fer-li la petició al SOLR.



Com es pot veure en el diagrama següent, hem utilitzat el patró *Facade* utilitzant un controlador per separar les capes de presentació de les lògiques de negoci. Aquesta classe **controladorAutors** és la que es relaciona amb la capa de presentació, i crea i utilitza els objectes **llistatAutors** i **Autor**. Utilitzem aquest patró per totes les pàgines del DUGi, com per exemple el llistat d'ítems, el d'autors, el de matèries, etc. Aquest patró també és recomanat per GRASP (object-oriented design General Responsibility Assignment Software Patterns).



8.2.7 Interfícies d'usuari

En aquest apartat mostrarem les interfícies d'usuari que hem dissenyat per al DUGi.

Les escollides són:

- Llistat d'ítems
- Llistat de matèries
- Panell d'administració
- Panell de gestió d'ítems
- Panell de recollida

El llistat d'ítems serà una taula amb els valors de data, títol i autors. Si es pot es posarà una icona representativa del tipus de document. A l'esquerra hi haurà els enllaços cap als repositoris, en forma de pastilles, a la dreta un filtre per autor i per matèria. Si es pot es posarà el filtre de totes les facetes que s'han escollit: autor, matèria, llenguatge, tipus de document i data.

La caixa de cerca permetrà cercar o filtrar aquests ítems o llistar-los tots en cas de no fer-la servir.

El següent prototip mostra una mica com serà aquesta interfície. El resultat real es mostra al punt 10.


Portal de cerca conjunta als documents digitals de la UdG

[Inici](#) [Llistar Ítems](#) [Llistar Matèries](#) [Llistar Autors](#)

[DUGiDocs](#)
[DUGiMedia](#)
[DUGiFE](#)
[RACO](#)
[TDX](#)
[MDX](#)
[Recercat](#)
[MDC](#)

Cercar

Llistar Ítems
S'han trobat 13500 ítems.
Mostrant ítems des de 0 a 20.

◀ 1 2 3 4 5 ▶▶

Data	Títol	Autors
15-07-1982	Blablabla	Blablabla
15-07-1982	Blablabla	Blablabla
15-07-1982	Blablabla	Blablabla
15-07-1982	Blablabla	Blablabla
15-07-1982	Blablabla	Blablabla
15-07-1982	Blablabla	Blablabla
15-07-1982	Blablabla	Blablabla
15-07-1982	Blablabla	Blablabla
15-07-1982	Blablabla	Blablabla
15-07-1982	Blablabla	Blablabla
15-07-1982	Blablabla	Blablabla
15-07-1982	Blablabla	Blablabla

◀ 1 2 3 4 5 ▶▶

Autors
Blablabla
Blablabla
Blablabla
Blablabla
Blablabla
Blablabla
Blablabla
Blablabla

Matèries
Blablabla
Blablabla
Blablabla
Blablabla
Blablabla
Blablabla
Blablabla
Blablabla

© 2011 Universitat de Girona | Biblioteca UdG

[Ajuda](#) | [Avís Legal](#) | [Mapa del Web](#) | [Comentaris](#)

En el cas dels llistats, es mostraran els resultats agrupats per les lletres d'inici. I amb la caixa de cerca es permetrà filtrar els resultats per aquestes lletres inicials.



Portal de cerca conjunta als documents digitals de la UdG

Inici

Llistar Ítems

Llistar Matèries

Llistar Autors

Llistar Matèries

Comença per:

S'han trobat 13500 matèries.
Mostrant de 0 a 4000.

[següent >](#)

A

Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)

B

Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)

C

Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)

D

Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)

E

Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)

F

Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)

G

Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)

H

Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)

I

Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)

J

Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)

K

Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)

L

Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)
Blabla (10)

Autors

Blablabla
Blablabla
Blablabla
Blablabla
Blablabla
Blablabla
Blablabla
Blablabla

Matèries

Blablabla
Blablabla
Blablabla
Blablabla
Blablabla
Blablabla
Blablabla
Blablabla

DUGiDocs

DUGiMedia

DUGiFE

RACO

TDX

MDX

Recercat

MDC

A la interfície d'administració hi ha d'haver els panells de revisar autoritats, veure els nous ítems recollits, l'estat de l'índex, i la opció de netejar la cache. També hi ha d'haver una vista prèvia de les últimes consultes, que serà molt més acurada al panell de log i/o al d'estadístiques.



Portal de cerca conjunta als documents digitals de la UdG

logout

DUGi: Administració

Administració

Recol·lecció

Gestió d'ítems

Estadístiques

Logs

Autoritats per revisar

Olive Comadira, Oriol -> Olivé Comadira, Oriol
 Olive Comadira, Oriol -> Olivé Comadira, Oriol
 Olive Comadira, Oriol -> Olivé Comadira, Oriol
 Olive Comadira, Oriol -> Olivé Comadira, Oriol
 Olive Comadira, Oriol -> Olivé Comadira, Oriol

Nous ítems

http://hdl.handle.net/10256.1/3003
 http://hdl.handle.net/10256.1/3002
 http://hdl.handle.net/10256.1/3001
 http://hdl.handle.net/10256.1/3000
 http://hdl.handle.net/10256.1/2999

Estat Índex

Actiu

[Reiniciar](#)

Cache

[Netejar cache](#)

Últimes consultes

Data	IP	URL
2013-04-20	84.88.136.42	subject/Superf%C3%ADcle%20d%E2%80%99equilibri
2013-04-20	84.88.136.42	author/Pulderrajols,%20P.
2013-04-20	84.88.136.42	item/http:@:@@hdl.handle.net@@10256@@1975
2013-04-20	84.88.136.42	subject/Superf%C3%ADcle%20d%E2%80%99equilibri
2013-04-20	84.88.136.42	subject/Superf%C3%ADcle%20d%E2%80%99equilibri

© 2011 Universitat de Girona | Biblioteca UdG

En el panell de gestió d'ítem ens hi cal una caixa de cerca per poder trobar l'ítem, un panell amb els resultats de la cerca per escollir l'ítem a modificar o eliminar, i un altre amb l'ítem per modificar. Al panell de modificació de l'ítem, hem de poder afegir valors a les metadades multivaluades, guardar l'ítem, i afegir metadades noves.



Portal de cerca conjunta als documents digitals de la UdG

logout

DUGi: Gestió d'Ítems

Cercar



Ítems

Query: Ezquerra

Ítems trobats: 6

<http://hdl.handle.net/10256.1/3001> Eliminar

<http://hdl.handle.net/10256.1/3000> Eliminar

<http://hdl.handle.net/10256.1/2999> Eliminar

Editar

Editant ítem: <http://hdl.handle.net/10256/1234>

Recorda que per eliminar un valor cal deixar la casella en blanc

Títol	
Data	
Autors	+
Matèries	+
Resum	
Tipus doc	
Format	
Llenguatge	+

[Afegir metadada](#)

El panell de recol·lecció, es divideix en 4 parts. La primera és l'edició dels sinònims de l'índex. Si es cregués necessari, també es podria afegir un altre panell igual per a la gestió de les paraules buides. Les altres tres estan relacionades, i són les referents a la recol·lecció realment. La primera és la que ens permetrà seleccionar el repositori, a la segona podrem canviar els sets, la data a partir de la qual volem recol·lectar, i la url de recol·lecció del servidor. La part més interessant és la última. En aquesta tenim una taula amb les metadades d'entrada, les de sortida, el patró i les opcions. Com que per poder aplicar bé els patrons necessitem tenir en compte l'ordre de les metadades, el botó “m” ens permetrà canviar l'ordre de la fila metre el mantinguem premut. El botó amb la “x” eliminarà la fila.


Portal de cerca conjunta als documents digitals de la UdG

logout

Administració

Recol·lecció

Gestió d'ítems

Estadístiques

Logs

DUGi: Administració > Recol·lecció

Sinònims:

udg => universitat de girona
biblio => biblioteca
xavi => xavier
pep => josep
marissa => maria lluisa

Guardar

Servidors:

DUGiDocs
DUGiMedia
DUGiFonsEspecials
RACO
MDX

Valors:

URL:

SETS:

DATA:

Metadades

Metadada Entrada	Metadada Sortida	Patró	Opcions
dc.identifier	dc.identifier.uri ▼	http://hdl.handle.net/10256V	m x
dc.identifier	dc.identifier.other ▼	http://	m x
dc.identifier	dc.identifier.citation ▼		m x
dc.creator	dc.creator ▼		m x
dc.contributor	dc.contributor.other ▼	Universitat	m x
dc.contributor	dc.contributor ▼		m x
dc.rights	dc.rights.uri ▼	http://	m x
dc.rights	dc.rights ▼		m x
dc.subject	dc.subject ▼		m x
dc.title	dc.title ▼		m x
dc.date	dc.date ▼	d{4}(?:-d{2})?(?:-d{2})?\$	m x
dc.date	dc.date ▼	[.+\\]	m x
dc.language	dc.language ▼	w{3}	m x
Metadada Entrada	Metadada Sortida	Patró	Opcions

9. Implementació i proves

9.1 Implementació

En aquest punt parlarem dels principals problemes que ens hem trobat en la implementació del projecte, i de les solucions que hi hem aplicat.

Els problemes han sigut els següents:

- Detectar les autoritats conflictives
- Llistar els autors i les matèries ordenats alfabèticament tenint en compte els accents
- Dates
- Estadístiques sense base de dades
- Cache
- Multi-llenguatge
- Logs
- Seguretat
- Obtenció de la url dels documents
- SEO
- Optimització

Per implementar aquest projecte, primer s'ha resolt els problemes de detectar autoritats, de llistar els autors ordenats i els problemes amb les dates a petita escala. S'han preparat unes proves que cobrissin un espectre prou ampli de possibilitats per anar refinant els algoritmes. I s'ha implementat les proves per a fer-se des de la línia de consola amb PHP-CLI¹⁹ per a poder re-aprofitar el codi.

Per a la implementació general del projecte, s'ha anat implementant les parts independents, provant-ne el funcionament, i després s'ha muntat l'aplicació en un model vista-controlador.

Finalment s'ha refinat la presentació.

9.1.1 Detectar les autoritats conflictives

Aquest era sens dubte un dels punts més interessants del projecte. Des de fan molt de temps, dins del món dels repositoris, es busca una solució a aquest problema. Fins ara és una feina que es fa molt meticulosament i no es pot automatitzar.

Fins ara la tècnica consisteix en assegurar-se bé que la forma del nom és la correcta en el moment d'entrada en el repositori, i en anar fent escombrats periòdicament per detectar-ne els possibles conflictes.

La nostra solució

En primer moment es va estudiar la possibilitat de fer servir l'algoritme *Distància de Levenshtein*²⁰ directament per calcular com de diferents eren dues cadenes de text.

La idea semblava bona, però no era prou acurada en el nostre cas.

En el cas de comparar **Olive**, **Oriol** amb **Olivé**, **Oriol** els resultats eren bons, però al comparar **Olive**, **Oriol** amb **Olivé Comadira**, **Oriol** o amb **Olivé**, **O**. els resultats no eren acceptables.

Per a la nostra aproximació a un resultat acceptable, vam idear una funció heurística que mitjançant pesos ens retorna un valor d'igualtat entre dos autors.

El més important és que coincideixi el primer cognom en un tant per cent molt elevat. Després que

¹⁹ PHP command line usage. Permet executar scripts de php des de la línia de consola.

²⁰ Algoritme que busca el nombre mínim d'operacions per transformar una cadena en una altra. Rep el nom en honor del científic rus Vladimir Levenshtein que fa idear l'algoritme al 1965.

ens coincideixi la inicial del nom, el segon cognom i la resta del nom.

Així doncs hem proposat un 70% per al primer cognom, un 15% per la inicial del nom, un 10% per el segon cognom i un 5% per a la resta del nom.

La nostra funció heurística és una cosa així:

$$Dist = dist_cognom1*0.7 + dist_inicial*0.15 + dist_cognom2*0.1 + dist_nom*0.05$$

Llavors amb un valor llindar recollirem els possibles candidats.

L'algoritme en pseudocodi és el següent:

*Hi pot haver casos de cognoms amb més d'una paraula com per exemple els cognoms que comencen amb **de**, **van**, **von**, **el**, etc, que no tenim en compte en aquest pseudocodi.*

```
// busquem el primer cognom sense accents
patro=partir(/[ ,]/,ignorar_accents(autor))[0];
// el prefix serà fins la diferència entre els dos noms
prefix=(autor,patro);

// consultar a l'índex per recollir candidats que comencin per el prefix
possibles=llistat(prefix);

trobat='';
candidats=array();
i=0;
mentre(! trobat){
    si (autor = possibles[i]){
        trobat=possibles[i]
    }
    altrament{
        // apliquem la funció heurística
        factor=distancia(autor, possibles[i])*100;
        // comparem amb el nostre valor llindar [0-100]
        si(factor>=VALOR_LLINDAR){
            push(candidats,possibles[i]);
        }
    }
}

// funció heurística
funcio distancia(autor_a, autor_b){
    fact=0;
    autor_a_partit=partir(/[ ,]/,autor_a);
    autor_b_partit=partir(/[ ,]/,autor_b);

    // el valor de sortida de l'algoritme de Levenshtein és més gran com més allunyat
    // estigui per això es resta a 10
    fact_cognom1=(10-levenshtein(autor_a_partit[0],autor_b_partit[0]))*0.07;

    fact+=fact_cognom1;

    // tenim segon cognom?
    si (autor_a_partit > 2 i autor_b_partit > 2){
        fact_cognom2=(10-levenshtein(autor_a_partit[1],autor_b_partit[1]))*0.01;
        fact+=fact_cognom2;
    }

    // inicial i nom
```

```

        si(substring(autor_a_partit[2],0,1),substring(autor_b_partit[2],0,1)){
            fact+=0.15;
        }
        fact_nom=(10-levenshtein(autor_a_partit[2],autor_b_partit[2]))*0.005;
        fact+=fact_nom;
        retorna fact;
    }

```

Finalment el sistema ha de guardar els possibles candidats en un fitxer de text, i la forma més útil és guardar l'objecte de php serialitzat.

Aquest algoritme dona resultats bastant acceptables i en un temps força raonable. És interessant fer notar que es pot anar acotant el valor llindar per aconseguir depurar més els resultats, i que com que els resultats sempre seran repassats per l'administrador, aquest detectarà molt fàcilment els falsos-positius.

En el cas d'un autor que el nom li comenci per accent, el prefix que es trobarà serà buit, i llavors caldrà repassar tots els autors del DUGi, fent d'aquest el cas menys eficient de l'algoritme.

Al fer-se el procés en hores de poques consultes i amb la gestió dels recursos del sistema, aquests casos no han de crear cap problema de rendiment.

De totes maneres, tot i que els resultats són molt satisfactoris, cal seguir treballant-hi per millorar-lo.

9.1.2 Llistar autors i matèries alfabèticament

El cas dels accents en el software sempre és conflictiu, ja que en anglès no n'hi ha i aquest és el llenguatge per referència en aquest món.

Així doncs, tot i que l'índex Solr permet cercar sense tenir en compte els accents, a l'hora de treure els llistats per a les facetes, Solr no ordena bé. O si més no, no ordena de la forma que a nosaltres en interessa.

El problema té a veure amb l'ordre dels caràcters en UTF-8²¹, que al igual que en altres estàndards, codifica els 128 caràcters de US-ASCII²² tots seguits en un sol byte per caràcter, en un rang que va de U+0000 a U+007F.

Llavors els caràcters accentuats es formen a partir de dos bytes que es troben a partir de la posició U+00A0.

Com hem vist al punt 8.1.4, els camps de Solr utilitzats per a fer les facetes, no tenen cap filtre aplicat per a treure accents, ni paraules buides, ni cap tokenitzador. Això era necessari per fer-los servir a les facetes, però ens perjudica en el moment de llistar-los alfabèticament.

Vegem els següents autors:

1. Álvarez, Manolo
2. Moré, Laura
3. Olivé, Oriol

Ara estan ordenats correctament, però al treure un llistat ordenat alfabèticament amb el Solr ens els retornaria en aquest ordre:

1. Moré, Laura
2. Olivé, Oriol
3. Álvarez, Manolo

21 **UTF-8** (8-bit Unicode Transformation Format): Estàndard de codificació de caràcters que permet codificar els caràcters de totes les llengües utilitzant d'un a quatre bytes.

22 **US-ASCII** (American Standard Code for Information Interchange): Estàndard de codificació Americà que permet codificar els caràcters de lletres, nombres i puntuació amb 7 bits, i no permet codificar els caràcters accentuats ni especials sense una extensió.

A més a més, Solr tampoc ens retornarà els autors que comencin per una lletra accentuada en el cas d'utilitzar-la de prefix.

La nostra solució

Per al nostre llistat, necessitem ordenar els valors sense tenir en compte els accents, i quan ens vingui un prefix, fer la petició les vegades que calgui per tots els possibles accents del caràcter.

Vegem-ne l'algoritme en pseudo-codi:

```
autors=array();
// si tenim prefix
si(prefix != ''){
    // calculem les possibles variants del prefix
    possibles=array();
    possibles=crear_variants_accentuades(prefix);

    per cada possible dins de possibles{
        // tots els autors que comencen pel prefix possible
        push(autors,llistar(possible));
    }
}
altrament{
    autors=llistar();
}

// ordenació a la que indiquem la funció de comparació que volem
autors=ordenar(autors,comparar_sense_accents);

funció comparar_sense_accents(autor_a, autor_b){
    a=convertir_sense_accents(autor_a);
    b=convertir_sense_accents(autor_b);

    // comparar sense tenir en compte les majúscules i minúscules
    retorna comparar_sense_majus(a,b);
}
```

Amb aquest algoritme solucionem el problema de l'ordre, i del prefix. En el cas del prefix la solució no és gaire bona, però és acceptable si tenim en compte que amb 3 o 4 lletres l'usuari ja ha filtrat suficient com per veure el seu autor.

S'haurà d'estudiar una manera de millorar la solució en el cas dels prefixos accentuats.

9.1.3 Dates

Aquest és un problema que sorgeix de la necessitat dels bibliotecaris de posar dates indefinides.

Per exemple, en el cas de no tenir la data, mai es deixa buida, sempre es posa el valor “[s.d.]” que vol dir *sine data* que en llatí vol dir *sense data*.

Els claudàtors “[]” s'utilitzen sempre que no es té la seguretat total de que la data és la correcta. Podem trobar, per exemple, valors de dates tipus [2012-05-12].

En el cas de no tenir clar l'any, el mes o el dia, s'utilitza per conveniència el comodí “?”.

Per exemple, en cas de saber que l'ítem és del gener de l'any 1965, el valor que posaríem seria el següent: 1965-01-??, i en el cas de saber que va ser al segle XX, posaríem 19?? i prou.

Com es pot imaginar, això ens suposa un problema, sobretot a l'hora d'ordenar els ítems per data.

La nostra solució

En el nostre cas, necessitem que l'usuari pugui cercar els ítems que no tenen data utilitzant el valor

[s.d] o s.d. directament, per tant, hem de mantenir el valor antic de la data, que a més a més serà el valor que es mostrarà en el cas de veure un ítem.

El que proposem per a complir el requisit de poder cercar en *mode bibliotecari* i poder ordenar per data, és crear un camp de data auxiliar, que serà el que utilitzarem per fer els llistats de facetes i per ordenar els ítems, i en el cas de cercar per data, utilitzar un tercer camp de còpia, que contindrà els valors tant de les dates que ens arriben per defecte, com de les dates auxiliars.

El camp auxiliar contindrà els valors del camp data filtrats en el moment de la creació de l'ítem, i hi traurem tots els valors que puguin ser problemàtics a l'hora d'ordenar com els **interrogants** i els **[s.d.]**.

En el cas de les dates incompletes i amb interrogants, el sistema generarà una data entera amb la forma **AAAA-MM-DD**, substituint els comodins per **01**. La idea és aconseguir el primer dia possible sense comodins que estigui dins del rang esperat.

I en el cas dels **[s.d.]**, es substituirà per **0**.

Amb el tercer camp aconseguim que es pugui cercar per **[s.d.]**, ja que el SOLR permet el comodí de cadena “*” i el comodí de caràcter “?”. Així doncs, per el cas de les dates amb comodí, no ens faria falta.

Aquest tercer camp és de tipus còpia, i el SOLR el generarà automàticament.

Aquesta solució ens permet obtenir una ordenació correcta dels ítems per data, i no és gens costosa per a l'índex. No creiem que necessiti cap millora.

9.1.4 Estadístiques sense Base de Dades

Al no fer servir una base de dades, i no voler fer-la servir, tenim dues opcions, o fem servir el SOLR, i creem una col·lecció per les estadístiques, o muntem un sistema amb fitxers.

Les nostres estadístiques han de permetre fer un recompte de les consultes a un ítem, a un autor, matèria així com en quin llenguatge en han consultat, quines cerques han fet i des d'on venien anteriorment per a poder fer estudis de SEO²³. A més a més, ha de permetre determinar des de quina ciutat ens han consultat.

La nostra solució. Nosaltres hem escollit la opció del sistema de fitxers.

El sistema es basa en un procés que s'executa cada nit, i repassa totes les peticions fetes aquest dia. Això es fa parsejant el fitxer de log de les consultes.

De cada línia ens interessen la IP, el referer²⁴, la url i el llenguatge utilitzat.

Un cop tenim la url, necessitem distribuir la consulta entre llistat d'ítems, d'autors, de matèries, vista d'ítem, o altres.

Finalment fem un sumatori per tipus de consulta i per ip i per dia, mes i any.

Les dades de les ips es guarden formant una estructura d'arbre.

Per exemple la ip *38.101.148.12* es guardarà al directori *ips/38/101/* que contindrà els fitxers *38.101.148.geo* i *38.101.148.stats*. Dins d'aquests fitxers tindrem les dades d'aquestes ips serialitzades. Així podrem actualitzar-les quan ens convingui, i mantenir els sumatoris al dia.

Tot i que la solució funciona bé, en un futur s'hauria de millorar utilitzant algun tipus de base de dades nosql com per exemple redis, que permet definir una estructura de hash i anar incrementant els valors numèrics, el que ens podria servir per fer els sumatoris.

²³ Sigles en anglès de Search engine Optimization. Es refereix a l'optimització del contingut per a la bona indexació per part dels cercadors web.

²⁴ Capçalera del protocol HTTP que ens permet saber de quina url prové el client.

9.1.5 Cache

Un sistema de memòria cau pot ser molt interessant en el cas de que els usuaris cerquin sovint el mateix.

En el nostre sistema, necessitem controlar nosaltres totalment aquesta cache, hi ha moments en els que ens interessarà no guardar ni servir cache, ja que necessitem que s'executi el codi php darrera de la cerca.

La vida de la cache ha de ser com a màxim d'un dia, ja que cada nit, es recol·lecten els nous ítems. També s'hauria de netejar en el cas de fer alguna modificació en el codi del sistema. Per això al panell d'administració hi ha un botó per netejar-la.

La nostra solució

El llenguatge php proporciona una funcionalitat que permet que es dispari un script en el moment de rebre una petició, i també al final de servir-la. Aquesta funcionalitat és la que ens permetrà crear un sistema de cache. A més a més, per generar els fitxers de cache, necessitem controlar el buffer de sortida. Al PHP es fa amb les funcions **ob_start**, **ob_end_flush** i **ob_get_contents** o **ob_get_clean**. En el nostre pseudocodi utilitzarem **cache_comença**, **cache_contingut** i **cache_acaba**.

La funcionalitat s'activa al fitxer *.htaccess* de la següent forma:

```
php_value auto_prepend_file start_cache.php
php_value auto_append_file end_cache.php
```

Llavors creem dos fitxers amb el nom que els hi hem posat i el nostre pseudocodi serà el següent:

start_cache.php

```
// carreguem el fitxer de configuració que conté les variables de sistema
//requerir fitxer de configuració;

// recollim la url de base
url_base=SERVIDOR_BASE;

// necessitem muntar tots els paràmetres tant GET com POST per crear l'adreça
//única
$url = muntar_url();

// sistema de log
log = nou log(FITXER_QUERY_LOG);

// creem la ruta del fitxer
fitxer_cache= PATH_LOG.md5(url).'.html';

// gaudem la informació de la petició i d'on venim
log->info(url, referer, fitxer_cache);

// carreguem els patrons de les urls sense cache
patrons_no_cache=partir('|',PATRONS_NO_CACHE);

// cal fer cache o carregar el fitxer?
si(existeix_fitxer(fitxer_cache) i no preg_match(/patrons_no_cache/,url_base)){
    si(temps_fitxer(fitxer_cache) < TEMPS_MAX_CACHE){
        // mostrar el fitxer de cache
        mostrar_fitxer(fitxer_cache);
        // matem la petició perquè ja hem mostrat el contingut
```

```

        mor();
    }
    cache_comença();
}

```

end_cache.php

```

// obrim el fitxer de la cache
f=obrir_fitxer(fitxer_cache);
// guardem i tanquem
escriure_fitxer(f,cache_contingut());
tancar_fitxer(f);
cache_acaba();

```

La solució mostrada és satisfactòria i ens està donant bons resultats en quan a optimització i rendiment, ja que disminueixen considerablement les peticions a l'índex, i tot el sistema rendeix molt per sota dels seus màxims.

9.1.6 Multi-llenguatge

En el nostre sistema necessitem un sistema de multi-llenguatge. No només cal traduir els textos, si no que a més a més, tenim que les dates en anglès s'escriuen a la inversa que en català.

La nostra solució

En alguns casos, com per exemple en l'inici o l'ajuda, ens val la pena carregar una plantilla per cada llenguatge, ja que ens serà molt més pràctic per editar el text.

Per la resta de casos, sobretot el llistat d'ítems, tenim una classe llenguatge que ens permet escriure la data en format humà segons la llengua, ens permetrà també escriure els noms de les metadades en l'idioma pertinent i substituir valors dins d'un text.

Llenguatge
- locale: String - data_pat: String - mesos: String[] - valors: hash
+ constructor(lang:String='ca') + setLocale(lang:string) + setDataPat(String) + setValors(hash) + load() + getValor(key:string) + getSustValor(key:String,pat:String[],val:String[]) + normData(data:string) + getKey(val:string)

Els valors, el patró de data i els mesos es guarden en un fitxer en format YAML per cada llenguatge suportat.

Vegem-ne un tall del català:

```

data:
  pat: ${dia} ${mes} ${any}
  mesos: [gener, febrer, març, abril, maig, juny, juliol, agost, setembre, octubre,
novembre, desembre]

```

```

valors:
  header.head.title: Repositori digital de la Universitat de Girona

  header.header.slogan: Portal d'accés als documents digitals de la UdG

  header.tabs.inici: Inici
  header.tabs.cercar: Cercar
  header.tabs.items: Llistar Ítems
  header.tabs.autors: Llistar Autors
  header.tabs.materies: Llistar Matèries

  titols.num-found: S'han trobat ${num} ítems
  titols.desde-a: Llistant ítems des de ${start} a ${max}
  titols.one-found: S'ha trobat un ítem
  titols.no-found: No s'ha trobat cap ítem
  titols.first: Primera
  titols.last: Última
  titols.autor: Llistar Títols per Autor "${val}"
  titols.materia: Llistar Títols per Matèria "${val}"

  item.title: Ítem
  item.not-found: No s'ha trobat l'ítem identificat com "${id}"

  autors.num-found: S'han trobat ${num} autors
  autors.pag: Llistant autors des del ${inici} fins al ${fi}
  materies.num-found: S'han trobat ${num} matèries
  materies.pag: Llistant matèries des de la ${inici} fins a la ${fi}

```

Podem veure que en català el patró utilitzat és **\${dia} \${mes} \${any}**, igual que en castellà, però en anglès el patró serà el següent: **\${any} \${mes} \${dia}**.

L'array de mesos, ens permetrà escriure les dates en format humà, fent servir el patró anterior.

Una data com per exemple 15-07-1982 en català s'escriurà com a 15 juliol 1982, i en anglès 1982 15 july. Si es volgués aconseguir una data més humana encara, com per exemple 15 de juliol del 1982, només caldria afegir les preposicions al patró.

També és interessant la opció de recuperar un valor amb substitució. En aquests casos, substituïrem les variables marcades amb **\${}** en el text de sortida utilitzant els dos arrays **pat** i **val** que es poden veure com a paràmetres del mètode **getSustValor**. Si volguéssim substituir el text amb clau *materies.pag*, el primer array contindrà els valors *inici* i *fi*, i l'array *val* contindrà els *dos nombres*.

Finalment també tenim el mètode per recuperar una clau a partir del seu valor. Aquest mètode és pot segur, ja que les claus estan controlades per ser úniques, i els valors no, però l'hem implementat perquè ens permetrà fer una cerca més humana.

Imaginem que un usuari vol cercar els ítems que contenen la paraula *dugi* al *títol*. Com que el nom del camp és intern i l'usuari no té perquè saber-lo, segurament buscaria *títol: dugi*, i si fos un usuari de parla anglesa, possiblement buscaria *title: dugi*. En aquest cas, recuperar la clau a partir del valor, ens pot ajudar a permetre aquesta opció.

De totes maneres, s'ha d'explicar a l'ajuda com fer aquestes cerques per camps, i utilitzar exactament el mateix valor que utilitzem per escriure els camps en el la vista d'ítems.

A la classe llenguatge hi hem aplicat un *patró singleton* per no haver de carregar cada vegada els valors d'un locale en el cas d'un client que ens fa més d'una petició.

Aquesta solució ens permet complir tots els requeriments en aquest camp, i ens ha donat uns resultats molt satisfactoris. De totes maneres, com hem explicat al principi del punt, en alguns casos on teníem textos molt llargs, ens ha resultat més còmode fer diferents plantilles per als idiomes.

9.1.7 Sistema de logs

En aquest cas, necessitem un sistema de logs que ens permeti 3 nivells de missatges. El primer ha de ser el nivell d'informació, el segon el nivell d'alerta i el tercer el nivell d'error.

El primer nivell és només per a les estadístiques i per guardar un inventari de les consultes i operacions que va fent el sistema, com per exemple la recoll·lecció, el procés de detecció d'autoritats, etc.

El segon nivell s'ha de fer servir en el cas d'errors que no són preocupants per al funcionament del sistema, però que cal tenir-los en compte per arreglar-los quan es pugui. Aquests podrien ser, per exemple, problemes amb metadades o amb ítems que no s'han pogut entrar bé en el moment de recoll·lecció per repetir algunes metadades que han de ser úniques, etc. En aquest cas caldrà guardar-ne registre, i enviar un email amb totes les del dia a l'administrador, o mostrar-les totes quan aquest entri al panell d'administració.

El tercer nivell és el més urgent. En aquest cas caldrà enviar un email a l'administrador. Aquests seran errors de comunicació amb l'índex, problemes d'escriptura al disc, i altres errors que es cregui que requereixen assistència urgent.

La nostra solució

S'ha creat una classe que ens permetrà controlar els nivells de log, i que escriurà els logs i enviarà els missatges utilitzant la funció **error_log** del PHP. Aquesta funció permet escriure logs en un fitxer nostre o utilitzant la funció de log del sistema, i enviar emails utilitzant la configuració associada al fitxer php.ini.

Logs
- mail: string - fitxer: string
+ constructor(fitxer: String, mail: String) + INFO(msg: String) + WARN(msg: String) + ERROR(msg: String)

Aquesta classe ens permet crear diferents objectes log segons les nostres necessitats, i que escriuin en diferents fitxers de log. Això ens serà molt útil, sobretot, per crear les estadístiques.

Els missatges que s'escriuran als fitxers de log tindran la següent forma:

Nivell INFO:

[data] INFO [ip; url] msg

Nivell WARN:

[data] WARN [ip; url] msg

Els email en aquest cas seran enviats amb el tema WARNINGS DUGi per l'adreça warnings@dugi.udg.edu

Nivell ERROR:

[data] ERROR [ip; url] msg

Al mail s'enviarà a més a més de l'error retornat pel PHP l'stack trace del sistema. I el tema serà ERROR DUGi enviat per l'adreça error@dugi.udg.edu

A més a més, per controlar que els logs no creixin massa, s'ha creat una utilitat de rotació de logs que ens permetrà anar guardant els logs ordenats per data i comprimits amb gzip. Cada nit el sistema cridarà a aquesta utilitat.

Vegem-ne el pseudocodi:

```
// carreguem el fitxer de configuració que conté les variables de sistema
requerir fitxer de configuració;

dia=data("Y-m-d");

per cada fitxer amb la forma *.log a la ruta LOG_PATH{
    dades=contingut_fitxer(fitxer);
    dades_comprimides=comprimir(dades);
    fitxer_comprimit=reemplaça('/.log$/', '', fitxer).'_'.dia.'.log.gz';
    fitxer_guardar_contingut(fitxer_comprimit, dades_comprimides);
    // netejar fitxer
    fitxer_guardar_contingut(fitxer, '');
}
```

La solució aplicada també és satisfactòria en aquest cas, tot i que amb les opcions que ens dona la funció `error_log` del PHP, podríem arribar a lligar-ho amb algun gestor de logs del sistema, i aprofitar el seu sistema de rotació de logs.

9.1.7 Seguretat

La seguretat és molt important ja que el nostre sistema podria ser una porta d'entrada a tota la infraestructura informàtica de la Universitat.

En general s'ha tingut molt en compte el tema de les inclusions remotes i locals. En el cas de no controlar totalment les variables dels mètodes de php que ens permeten incloure fitxers, es podria permetre a un atacant executar codi en el nostre servidor, o obtenir informació sensible del sistema.

No ens hem de preocupar per les vulnerabilitats del tipus injecció de sql, ldap o xml, ja que no utilitzem cap d'aquests serveis per a emmagatzemar les nostres dades.

El que si que hem de tenir en compte, són les vulnerabilitats del tipus XSS i CSRF, que són injeccions de codi html i javascript, que ens permeten en el primer cas aconseguir manipular la informació mostrada a la pàgina, i per exemple robar la cookie de l'usuari, i en el segon cas, fer executar comandes del sistema a un usuari amb privilegis. En el segon cas, el de les injeccions CRSF, l'atacant necessitaria conèixer el sistema per a poder executar alguna comanda interessant. Aquest tipus d'atacs es podem mitigar mitjançant l'addició de tokens a les peticions per executar comandes d'administració. De totes maneres, com hem vist en el punt 8, l'administrador no fa acció que pugui ser comprometedora, i en el cas d'aconseguir escriure codi en un fitxer com el de logs o el fitxer de sinònims, no podria executar-lo.

Així doncs, no creiem necessari eliminar la potencialitat d'aquesta vulnerabilitat afegint tokens a les peticions d'administració.

En el cas de les vulnerabilitats XSS, s'ha de filtrar totes les entrades de text per part de l'usuari que puguin escriure al codi final de la pàgina. Aquestes vulnerabilitats són molt complicades de filtrar, perquè hi ha moltíssimes variants, i la seva execució sol dependre del navegador.

Com comentàvem, en cas de un atacant aconseguís escriure codi en algun fitxer, segurament seria a les carpetes de l'índex, a les de cache o a la de log. En aquest cas, l'atacant tindria problemes per executar aquest codi, ja que aquestes carpetes estan fora dels dominis del servidor web.

En el cas de l'índex SOLR, hem de filtrar les ips que puguin contactar amb ell, ja que fins que no permeti fer les actualitzacions i les insercions dels ítems amb un usuari i password, no hem de permetre fer aquestes accions a cap ip que no sigui fiable.

Per permetre l'accés només a certes ips, amb Jetty²⁵, necessitem crear un nou context que són com els hosts virtuals d'Apache.

Així podrem indicar al jetty quin és el nostre directori de context.

Primer necessitem desconnectar la opció del jetty de carregar tots els contextos que trobi al directori **webapps**.

Per fer-ho editem el fitxer **etc/jetty.xml** i comentem el següent bloc:

```
<!-- ===== -->
<!-- Configure the webapp deployer. -->
<!-- A webapp deployer will deploy standard webapps discovered -->
<!-- in a directory at startup, without the need for additional -->
<!-- configuration files. It does not support hot deploy or -->
<!-- non standard contexts (see ContextDeployer above). -->
<!-- -->
<!-- This deployer is configured to deploy webapps from the -->
<!-- $JETTY_HOME/webapps directory -->
<!-- -->
<!-- Normally only one type of deployer need be used. -->
<!-- -->
<!-- ===== -->

<!-- oriol per no carregar tots els contextos del directory webapps
<Call name="addLifeCycle">
  <Arg>
    <New class="org.mortbay.jetty.deployer.WebAppDeployer">
      <Set name="contexts"><Ref id="Contexts"/></Set>
      <Set name="webAppDir"><SystemProperty name="jetty.home"
default="."/>/webapps</Set>
      <Set name="parentLoaderPriority">false</Set>
      <Set name="extract">true</Set>
      <Set name="allowDuplicates">false</Set>
      <Set name="defaultsDescriptor"><SystemProperty name="jetty.home"
default="."/>/etc/webdefault.xml</Set>
    </New>
  </Arg>
</Call>
-->
```

Un cop tenim desconnectats els contextos necessitem crear el directori **contexts** i el fitxer **solr.xml**.

```
<Configure id="solr" class="org.mortbay.jetty.webapp.WebAppContext">
  <Set name="resourceBase"><SystemProperty name="jetty.home"
default="."/>/webapps/solr/</Set>
  <Set name="contextPath">/solr</Set>
  <Call name="setSecurityHandler">
    <Arg>
      <New class="org.mortbay.jetty.security.HTAccessHandler">
        <Set name="protegee">
          <Ref id="solr"/>
        </Set>
      </New>
    </Arg>
  </Call>
</Configure>
```

²⁵ És el servidor web que utilitza el SOLR per defecte. És molt poc pesat i portable. El SOLR també ens permet utilitzar el Tomcat com a servidor web de servlets. Per a més informació sobre l'estructura del SOLR, veure l'annex 14.2 sobre el SOLR.

```

        </Set>
    </New>
</Arg>
</Call>
<Set name="defaultsDescriptor"><SystemProperty name="jetty.home"
default="."/>etc/webdefault.xml</Set>
</Configure>

```

Al fitxer `solr.xml` necessitem indicar-li on està el directori de l'aplicació, i sobretot que ens permeti assignar-li un fitxer `.htaccess` com els d'apache per a restringir l'accés. Aquest fitxer `.htaccess` no funciona exactament igual com en apache per denegar l'accés a una carpeta.

Així doncs, al fitxer `.htaccess` del directori `webapps/solr`, que li hem indicat prèviament, hi posem el següent bloc XML.

```

<Limit>
    satisfy all
    order deny,allow
    deny from all
    allow from 127.0.0.1 xxx.xxx.xxx.xxx yyy.yyy.yyy.yyy
</Limit>

```

Permetem així que hi connectin el localhost, que serà el nostre PHP, i les ips `xxx.xxx.xxx.xxx` i `yyy.yyy.yyy.yyy` en aquest cas.

Amb aquesta simple configuració, tanquem la principal debilitat que tenia el nostre sistema. Si a més a més estem alerta de les possibles actualitzacions de seguretat de tot el programari utilitzat, podem estar tranquils.

9.1.8 Obtenció de la url dels documents

Tot i que l'accés directe als documents, que formen part dels ítems, el deixarem com a treball futur perquè volem que quedi ben integrat al DUGi segons el tipus de contingut, volem recuperar-ne les dades per no haver de refer l'índex al moment d'acabar d'implementar aquesta funcionalitat.

Com hem vist al punt 7.3.3 quan parlàvem dels formats de metadades suportats per OAI-PMH, el format METS era el més complet, però era 5 o 6 vegades més extens que l'OAI-DC. Això ens feia optar per l'OAI-DC a l'hora de recol·lectar els ítems fent servir el verb **ListRecords**, ja que aquest mostra 200 ítems de cop, i en el cas de fer servir METS rebíem un error de *timeout* del servidor.

Així doncs, la solució ha sigut la següent:

1. Petició `ListRecords` amb el format OAI-DC
2. Parsejar cada ítem
3. Per cada ítem per una altra petició `GetRecord` amb el format de dades METS
4. Parsejar la zona de fitxers del METS per obtenir tots els fitxers i els thumbnails, i identificar el fitxer principal.

S'ha de guardar els thumbnails lligats amb els fitxers per a poder-los mostrar correctament.

Tot i fer una petició per cada ítem més, al no enviar tantes dades de cop, el rendiment és molt superior que en el cas de rebre-ho tot de cop, i a l'hora de parsejar també ens és molt més eficient.

9.1.9 SEO (Optimització per cercadors)

L'optimització del contingut per a permetre als cercadors una millor indexació és un tema que ens preocupa. Per als nostres autors és molt important la visibilitat del seu contingut.

No ens serveix de res tenir molt contingut de qualitat si els usuaris no el podran trobar.

Per tal de millorar aquest punt apliquem les següents accions:

- Hem de crear un mapa del lloc per permetre als cercadors conèixer l'estructura de la nostra pàgina.
- Hem de fer servir el mòdul d'apache **mod_rewrite** per crear unes adreces que defineixin el contingut que hi ha al darrera. Les urls seran les següents:
 - list-items
 - list-authors
 - list-subjects
 - item/xxxx
 - author/xxx
 - subject/xxx
- Utilitzarem el símbol '-' per separar paraules en una url, ja que els cercadors no tenen en compte el símbol '_' i no indexarien el text com a dues paraules separades.
- Separar el contingut de l'estil. Mai utilitzarem tags html com **font**, **blink**, **br**, **b**, **i**. Sempre utilitzarem tags amb contingut semàntic com **em**, **strong**.
- No utilitzarem taules per crear l'estructura de la pàgina. Utilitzarem **div** i **css** per fer-ho.
- Els tags **title**, **description** i **keywords** canviaran a cada pàgina dinàmica per definir el contingut que mostraran.
- Sempre farem servir el l'atribut **title** en els enllaços, per definir millor la pàgina on ens portaran.
- Amb el **mod_rewrite** redirigirem el contingut del DUGi antic al nou, utilitzant el **codi HTTP 301**: mogut permanentment.

Els resultats d'aplicar aquestes premisses es veuen al cap dels mesos, de totes maneres estem segurs que ens donaran el resultat esperat. Per optimitzar encara més el nostre web, el pròxim nivell seria fer servir els tags d'**HTML5 header**, **section**, **aside**, **footer**, així com els atributs del tag **input**, per expressar dates i cerca. També serà interessant crear un **sitemap** en format xml per indicar als cercadors tots les urls de tots els ítems que tenim.

9.1.10 Optimització

Creiem que és important oferir les pàgines de la nostra aplicació en un temps raonable. Per a aconseguir-ho aplicarem les següents accions:

- Agrupar i minimitzar tots els javascripts en un sol fitxer.
- Agrupar i minimitzar els fitxers de css.
- Incrustar les imatges i fonts que formen part de l'estil al fitxer de css en codificats en base64.
- Comprimir les pàgines, el javascript i el css utilitzant el modul d'apache **mod_deflate**.
- Activar el **mod_expires** d'apache per indicar als navegadors que el contingut no expira en menys d'un dia.

Al aplicar totes aquestes premisses, aconseguim que els navegadors carreguin les pàgines força més ràpid, ja que només han de fer 3 peticions per carregar l'HTML, el JS i el CSS. A més a més, en el

cas de pàgines llargues amb molt de contingut com els llistats, al estar comprimit el contingut, s'enviarà molt més ràpid, i es permetrà que inclús els usuaris amb una connexió lenta, poden tenir un temps de resposta força satisfactori.

9.2 Proves

Per tal d'anar testejant les possibles solucions als problemes, hem anat fent les següents proves unitàries.

9.2.1 Detecció d'autoritats

Com explicàvem al punt 9.1.1, la primera prova va ser utilitzar directament l'algoritme de Levenshtein.

```
dugi@dugi:/dades/parser$ php -r "echo levenshtein('Olivé, O.','Olivé, Oriol').\\\"\\n\\\";"
4
dugi@dugi:/dades/parser$ php -r "echo levenshtein('Olivé, O.','Olivé, O').\\\"\\n\\\";"
1
dugi@dugi:/dades/parser$ php -r "echo levenshtein('Olivé, O.','Olivé, Oriol').\\\"\\n\\\";"
4
dugi@dugi:/dades/parser$ php -r "echo levenshtein('Olivé Comadira, Oriol','Olivé, Oriol').\\\"\\n\\\";"
9
```

Les proves anteriors estan fetes utilitzant la funció levenshtein que proporciona el php. I executant-les des de PHP-CLI. Aquests resultats no ens aportaven massa. La cosa canvia quan es tracta de cadenes amb la mateixa longitud.

```
dugi@dugi:/dades/parser$ php -r "echo levenshtein('Esquerra','Ezquerra').\\\"\\n\\\";"
1
```

Després es van anar implementant diferents algorismes que van ser diferents aproximacions a la solució final. Aquestes proves, es van fer entrant valors d'autors reals al nostre índex.

```
dugi@dugi:/dades/parser$ php pattern_match3.php Nonó, B.
Nonó, B. -> Non
Array
(
    [0] => Nonó Rius, Brigit
    [1] => Nonó Rius, Brígida
    [2] => Nonó i Rius, Brígida
    [3] => Nonoyama, Michiko
)
0..1..2..3..
Nonó Rius, Brigit -> 87.5
Nonó Rius, Brígida -> 87.5
Nonó i Rius, Brígida -> 87.5
```

Els dos primers algorismes es van descartar perquè la funció heurística no donava els valors esperats.

L'algoritme de pattern_match3.php és el mateix que l'escollit finalment amb la diferència que enlloc d'utilitzar l'algoritme de Levenshtein per calcular les distàncies, en fèiem servir un desenvolupat per nosaltres. El nostre algoritme no considerava més pròximes unes cadenes si tenien una mida més semblant.

```
dugi@dugi:/dades/parser$ php pattern_match4.php Nonó, B.
Nonó, B. -> Non
Array
(
    [0] => Nonó Rius, Brigit
    [1] => Nonó Rius, Brígida
    [2] => Nonó i Rius, Brígida
    [3] => Nonoyama, Michiko
)
0..1..2..3..
Nonó Rius, Brigit -> 87.5
Nonó Rius, Brígida -> 87
Nonó i Rius, Brígida -> 87
```

En aquest cas, utilitzem l'algoritme final que usa la funció Levenshtein de PHP. Podem veure que els resultats són molt pròxims als de l'algoritme de pattern_match3.php, però en aquest cas es valora la diferència o semblança de mida també.

```
dugi@dugi:/dades/parser$ php pattern_match4.php Nonó Rius, Brigit
Nonó Rius, Brigit -> Non
Array
(
    [0] => Nonó Rius, Brigit
    [1] => Nonó Rius, Brígida
    [2] => Nonó i Rius, Brígida
    [3] => Nonoyama, Michiko
)
0..Trobat: Nonó Rius, Brigit
```

En aquesta prova, es pot veure com l'algoritme ha trobat l'autoritat exacta, per tant no ens caldria guardar-la a la llista.

```
dugi@dugi:/dades/parser$ php pattern_match4.php Ezquerra, X.
Ezquerra, X. -> Ezquerra
Array
(
    [0] => Ezquerra Elias, Xavier
    [1] => Ezquerra, Anabel
)
0..1..
Ezquerra Elias, Xavier -> 87.5

dugi@dugi:/dades/parser$ php pattern_match4.php Ezquerra, Xavier
Ezquerra, Xavier -> Ezquerra
Array
(
    [0] => Ezquerra Elias, Xavier
    [1] => Ezquerra, Anabel
)
0..1..
Ezquerra Elias, Xavier -> 90
```

En aquestes altres proves podem veure com funciona l'algoritme amb diferents entrades.

```

dugi@dugi:/dades/parser$ php pattern_match4.php Àlvarez, J.|more
Àlvarez, J. ->
Array
(
    [0] => Anònim
    [1] => Ferrater Mora, José
    [2] => Pallí i Buxó, Lluís
    [3] => Universitat de Girona. Escola Politècnica Superior
    [4] => Universitat de Girona. Servei de Publicacions
    [5] => Universitat de Girona. Facultat de Ciències
    [6] => Vicens Vives, Jaume
    [7] => Ferrater Mora, Mercè
    [8] => Universitat de Girona. Departament d'Informàtica i Matemàtica Aplicada
    [9] => Aragó, Narcís-Jordi
    [10] => Santaló, Lluís

    ...
    ...13639..13640..13641..13642..13643..13644..13645..13646..13647..13648..13649..136
50..13651..13652..13653..13654..13655..13656..13657..13658..13659..13660..13661..13662.
.13663..13664..13665..13666..13667..13668..13669..13670..13671..13672..13673..13674..13
675..13676..13677..13678..13679..13680..13681..13682..13683..13684..13685..13686..13687
..13688..13689..13690..13691..13692..13693..13694..13695..13696..13697..13698..13699..1
3700..13701..13702..13703..13704..13705..
    Álvarez Rodríguez, Javier -> 87.5
    Álvarez de la Puente, J. M. -> 88.5

```

Aquesta prova anterior, mostra el problema que comentàvem en el punt 9.1.1. L'algoritme no fa servir cap prefix, ja que aquest es descarta per l'accent. De totes maneres, el resultat és el correcte, i el temps d'execució amb 13705 autors és inferior a un segon.

```

dugi@dugi:/dades/parser$ php pattern_match4.php Àlvarez, J.
...
Álvarez Rodríguez, Javier -> 87.5
Álvarez de la Puente, J. M. -> 88.5

```

```

temps d'execucio: 0.823423862457

```

```

dugi@dugi:/dades/parser$ php pattern_match5.php Nonó, B.
Nonó, B. -> Non

```

```

Array
(
    [0] => Nonó Rius, Brigit
    [1] => Nonó Rius, Brígida
    [2] => Nonó i Rius, Brígida
    [3] => Nonoyama, Michiko
)

```

```

0..1..2..3..

```

```

Nonó Rius, Brigit -> 87.5

```

```

Nonó Rius, Brígida -> 87

```

```

Nonó i Rius, Brígida -> 87

```

```

temps d'execucio: 0.00681114196777

```

La prova anterior mostra una comparativa entre aquests temps d'execució.

9.2.2 Ordenació d'autors i matèries

Per a resoldre el problema de l'ordenació d'autors i matèries, es va crear el següent entorn de proves. Es va muntar un índex SOLR en local amb un esquema amb dos camps:

- ID
- Nom

La idea era anar provant els tipus de camp que donaven millor resultat en la ordenació i en les facetes.

Per fer-ho vam afegir el següent camp al fitxer schema.xml de la configuració del SOLR:

```
<!-- oriol proves -->
<field name="text_autors" type="strautor" indexed="true" stored="true"
multiValued="false"/>
```

En aquesta prova, el camp no serà multivaluat, ja que si no, no ens permetrà ordenar per el camp. Així, aquestes proves també ens serviran per al camp títol, que si que l'haurem de fer servir per ordenar sense tenir en compte els accents.

També hi vam afegir la definició del tipus de camp *strautor*.

En aquesta definició es va anar canviant els filtres per tal de trobar els que ens podien permetre resoldre el problema.

Vam entrar a l'índex els següents autors:

```
<add>
<doc>
  <field name="id">1</field>
  <field name="text_autors">Alvarez, Antonio</field>
</doc>
<doc>
  <field name="id">2</field>
  <field name="text_autors">Àlvarez, Antonio</field>
</doc>
<doc>
  <field name="id">3</field>
  <field name="text_autors">Álvarez, Antonio</field>
</doc>
<doc>
  <field name="id">4</field>
  <field name="text_autors">Olivé Comadira, Oriol</field>
</doc>
<doc>
  <field name="id">5</field>
  <field name="text_autors">Olive, Oriol</field>
</doc>
<doc>
  <field name="id">6</field>
  <field name="text_autors">Oliver, Oriol</field>
</doc>
</add>
```

```
C:\Users\ori\projecte\solr-4.2.0\example\exampledocs>java -jar post.jar ori.xml
SimplePostTool version 1.5
Posting files to base url http://localhost:8983/solr/update using content-type a
```



```
pplication/xml..  
POSTing file ori.xml  
1 files indexed.  
COMMITting Solr index changes to http://localhost:8983/solr/update..  
Time spent: 0:00:00.636
```

Vegem les proves que es van fer:

```
<!-- oriol prova autors -->  
  <fieldType name="strautor" class="solr.TextField" omitNorms="true">  
    <analyzer type="index">  
      <!-- KeywordTokenizer does no actual tokenizing, so the entire input string is  
preserved as a single token -->  
      <tokenizer class="solr.KeywordTokenizerFactory" />  
    <!-- The TrimFilter removes any leading or trailing whitespace -->  
    <filter class="solr.TrimFilterFactory" />  
  </analyzer>  
  <analyzer type="query">  
    <tokenizer class="solr.KeywordTokenizerFactory" />  
  </analyzer>  
</fieldType>
```

Aquest tipus de contingut ens permet tractar el text com si fos un String, sense aplicar-hi cap més filtre que el de treure els caràcters blancs de davant i darrera.

Si fem una cerca ordenada per autor, el resultat amb aquest tipus de camp és el següent:

http://localhost:8983/solr/collection1/select/?q=*&sort=text_autors%20asc

```
<result name="response" numFound="6" start="0">  
  <doc>  
    <str name="text_autors">Alvarez, Antonio</str>  
  </doc>  
  <doc>  
    <str name="text_autors">Olive, Oriol</str>  
  </doc>  
  <doc>  
    <str name="text_autors">Oliver, Oriol</str>  
  </doc>  
  <doc>  
    <str name="text_autors">Olivé Comadira, Oriol</str>  
  </doc>  
  <doc>  
    <str name="text_autors">Àlvarez, Antonio</str>  
  </doc>  
  <doc>  
    <str name="text_autors">Álvarez, Antonio</str>  
  </doc>  
</result>
```

Podem veure que la ordenació en aquest cas ha sortit una mica malament.

http://localhost:8983/solr/collection1/select/?q=*&sort=text_autors%20asc&fl=text_autors&facet=true&facet.field=text_autors

```
<lst name="text_autors">  
  <int name="Alvarez, Antonio">1</int>  
  <int name="Olive, Oriol">1</int>
```

```

<int name="Oliver, Oriol">1</int>
<int name="Olivé Comadira, Oriol">1</int>
<int name="Àlvarez, Antonio">1</int>
<int name="Álvarez, Antonio">1</int>
</lst>

```

En el cas del llistat de facetes, els valors es mostren bé, però no ens ordena bé.

En la següent prova, s'ha intentat de treure els accents i posar-ho en minúscules a l'hora d'indexar. Per a poder fer la prova, s'ha de tancar l'índex, eliminar les dades, tornar a importar els autors, i tornar a engegar-lo.

```

<!-- oriol prova autors -->
<fieldType name="strautor" class="solr.TextField" omitNorms="true">
  <analyzer type="index">
    <!-- KeywordTokenizer does no actual tokenizing, so the entire input string is
preserved as a single token -->
    <tokenizer class="solr.KeywordTokenizerFactory" />
    <filter class="solr.LowerCaseFilterFactory" />
    <!-- oriol accents -->
    <charFilter class="solr.MappingCharFilterFactory" mapping="mapping-
ISOLatin1Accent.txt"/>
    <!-- The TrimFilter removes any leading or trailing whitespace -->
    <filter class="solr.TrimFilterFactory" />
  </analyzer>
  <analyzer type="query">
    <tokenizer class="solr.KeywordTokenizerFactory" />
    <filter class="solr.LowerCaseFilterFactory" />
    <!-- oriol accents -->
    <charFilter class="solr.MappingCharFilterFactory" mapping="mapping-
ISOLatin1Accent.txt"/>
  </analyzer>
</fieldType>

```

Els resultats són els següents:

http://localhost:8983/solr/collection1/select/?q=*&sort=text_autors%20asc

```

<result name="response" numFound="6" start="0">
  <doc>
    <str name="text_autors">Alvarez, Antonio</str>
  </doc>
  <doc>
    <str name="text_autors">Àlvarez, Antonio</str>
  </doc>
  <doc>
    <str name="text_autors">Álvarez, Antonio</str>
  </doc>
  <doc>
    <str name="text_autors">Olivé Comadira, Oriol</str>
  </doc>
  <doc>
    <str name="text_autors">Olive, Oriol</str>
  </doc>
  <doc>
    <str name="text_autors">Oliver, Oriol</str>
  </doc>
</result>

```

En aquesta prova hem aconseguit ordenar els autors correctament.

El cas de les facetes:

http://localhost:8983/solr/collection1/select/?q=*&sort=text_autors%20asc&fl=text_autors&facet=true&facet.field=text_autors

```
<lst name="text_autors">
<int name="alvarez, antonio">3</int>
<int name="olive comadira, oriol">1</int>
<int name="olive, oriol">1</int>
<int name="oliver, oriol">1</int>
</lst>
```

La prova no ha sortit tant bé, els autors surten sense majúscules ni accents, tots agrupats i ben ordenats.

Així docs, aquesta prova ens dona el tipus de camp que hem de fer servir per ordenar per títol, però no gaire res més.

Després de fer aquestes proves, ens en donem compte de la necessitat de tractar la sortida de les facetes amb el PHP, i fer l'ordenació nosaltres.

```
dugi@DUGip:/dades/dugi/old/proves$ php autors.php
Array
(
    [[Alexander, Luis]] => 2
    [[B. de Roch, Montserrat]] => 1
    [[Loros,?] José del] => 1
    [[Plana], Joaquina] => 1
    [[Tom]] => 1
    [Abad Carretero, Luís] => 4
    [Abad Power, Paloma] => 2
    [Abad, José M.] => 1
    [Abadal [i Calderó], Ramon d'] => 1
    [Abadal i de Vinyals, Ramon d'] => 3
    [Abadal, Ernest] => 1
    [Abadal, Joan de la Creu d'] => 3
    [Abadal, Maria] => 6
    [Abarca, Carmen] => 1
    [Abargues Casanova, Carlos] => 1
    [Abarquero Zorrilla, Fernando] => 1
    [Abart, Rainer] => 1
    [Abascal Azpeitia, Estibaliz] => 1
    [Abate, Charles J.] => 1
    [Abattouy, Naima] => 1
    [Abecasis, Rhoda] => 1
    [Abel, Sherry] => 1
    [Abellà Ametller, Carles] => 27
    [Abella García, Alberto] => 3
    [Abella Villar, Pablo] => 1
    [Abella, Carlos] => 1
    [Abellán, José Luis] => 12
```

Aquesta és la prova de l'execució de l'algoritme mostrat al punt 9.1.2. Es pot comprovar com funciona perfectament si ens fixem en l'ordre d'**Abellà Ametller** abans d'**Abella García**.

```
[Aguilar, Eva] => 1
[Aguilar, Francisco] => 9
[Aguilar, Helena] => 1
```

```
[Aguilar, Lucía] => 1
)
Temps d'execució: 1.11682200432
```

En aquest cas, es pot veure el temps d'execució de l'algoritme. Tot i que l'algoritme només mostra els 100 primers autors, al haver d'ordenar el llistat d'autors complet, el temps sempre serà similar a aquest.

S'ha optat per trossejar la sortida perquè llistar tots els autors o les matèries de cop, pot ser una mica costós tant per la mida de la sortida com per la interpretació per part del navegador.

A més a més, s'ha de tenir en compte, que en aquests casos si que serà molt important la cache.

```
dugi@DUGip:/dades/dugi/old/proves$ php autors.php bal
Array
(
    [Balaguer i Condom, Maria Dolors] => 11
    [Balaguer, V.] => 1
    [Balart Gimeno, Rafael Antoni] => 2
    [Balasch Martín, Jordi] => 1
    [Balasch Solé, Núria] => 1
    [Balbás Alonso, David] => 1
    [Balbin, Carlos Francisco] => 1
    [Balcells Rocamora, Enrique] => 1
    [Balcells, Josep Maria] => 4
    [Balcells, Mercè] => 1
    [Baldiris Navarro, Sílvia M.] => 5
    [Baldiris Navarro, Silvia Margarita] => 1
    [Baleato González, Sandra] => 1
    [Balibrea Pérez, Marta] => 1
    [Ball-llosera, Dolors] => 1
    [Ballart Junyer, Luis] => 1
    [Ballart, L.] => 2
    [Ballespí, Sergi] => 1
    [Ballestar Tarín, Maria Luisa] => 2
    [Ballesté Martorell, David] => 1
    [Ballesté Pau, Marta] => 1
    [Ballester Ferrando, David] => 24
    [Ballester, Ferran] => 3
    [Ballester, Ramon] => 1
    [Ballesteros Beretta, Antonio] => 1
    [Ballesteros i Segarra, Enric] => 7
    [Ballesteros Tribaldo, Evaristo Antonio] => 1
    [Ballestín Navarro, Xavier] => 1
    [Ballestrero, Maria Vittoria] => 2
    [Ballonga Serra, Cristina] => 1
    [Ballús i Edo, Ignasi] => 1
    [Ballús, Ignasi] => 1
    [Ballvé, José L.] => 1
    [Balmaceda Ramos, Carmen] => 5
    [Balmaña, Noemí] => 1
    [Balny, Claude] => 1
    [Baló, Agnes] => 1
    [Balsabre, Armand] => 1
    [Balsach, Maria-Josep] => 13
    [Balsalobre, Nadeia] => 2
    [Baltasar Bagué, Alícia] => 7
    [Baltasar, Basilio] => 1
    [Balty, Jean Ch.] => 1
)
Temps d'execució: 0.00803399085999
```

En el cas d'introduir un prefix, el temps de resposta és molt inferior, ja que només hem d'ordenar els autors que comencen per el prefix, i al SOLR també li costa molt menys recuperar-los.

```
dugi@DUGip:/dades/dugi/old/proves$ php autors.php alv
Array
(
    [Alvar, Carlos] => 1
    [Álvarez Álvarez, Manuela] => 2
    [Álvarez Bolado, Alfonso] => 1
    [Álvarez de la Puente, J. M.] => 1
    [Álvarez Etxeberria, Igor] => 1
    [Álvarez Huacón, Vicente] => 1
    [Álvarez Leiva, Sergio] => 1
    [Álvarez Puig, Agustí] => 1
    [Álvarez Rodríguez, Javier] => 1
    [Álvarez Rodríguez, Lola] => 1
    [Álvarez, Fèlix] => 1
    [Álvarez, Oscar D.] => 1
    [Álvarez, Pablo G.] => 1
    [Álvarez, Ramon] => 1
    [Álvarez, Román] => 1
    [Álvarez-Dardet, Carlos] => 1
    [Álvarez-Uría, Fernando] => 2
    [Alves, Nair] => 1
    [Alvis Arrieta, Jorge Luis] => 1
    [Alvite Díaz, Ramiro] => 1
)
Temps d'execució: 0.0253221988678
```

Aquest és un dels casos que comentàvem al punt 9.1.2, el de possibles accents en el prefix. En aquest cas l'algoritme fa 3 cerques al SOLR i les concatena per finalment ordenar l'array enter. Es pot veure que el temps no és una mica superior al del prefix *bal*, però és un temps acceptable i força raonable.

9.2.3 Dates

Les proves amb les dates, s'han fet directament sobre les dades recol·lectades. Això si, abans es va provar bé l'algoritme de neteja de les dates.

```
dugi@DUGip:/dades/dugi/old/proves$ php dates.php [s.d.]
Data entrada: [s.d.]
Data sortida: 0000
dugi@DUGip:/dades/dugi/old/proves$ php dates.php 19??
Data entrada: 19??
Data sortida: 1900-01-01
dugi@DUGip:/dades/dugi/old/proves$ php dates.php 19??-01-??
Data entrada: 19??-01-??
Data sortida: 1900-01-01
dugi@DUGip:/dades/dugi/old/proves$ php dates.php 196?-01-??
Data entrada: 196?-01-??
Data sortida: 1960-01-01
dugi@DUGip:/dades/dugi/old/proves$ php dates.php 196?-01
Data entrada: 196?-01
Data sortida: 1960-01-01
dugi@DUGip:/dades/dugi/old/proves$ php dates.php 1988?
```

```
Data entrada: 1988?
Data sortida: 1988-01-01
dugi@DUGip:/dades/dugi/old/proves$ php dates.php 1988?-01-01
Data entrada: 1988?-01-01
Data sortida: 1988-01-01
```

En la prova anterior es pot observar com l'algoritme neteja bé les dates que s'entraran al camp `data_aux`.

Ara vegem un exemple d'ordenació per data al SOLR.

```
dugi@DUGip:/dades/dugi/old/proves$ curl "http://localhost:8983/solr/select/?
q=*&fl=dc.date,data_aux&wt=json&indent=true"
```

```
{
  "responseHeader":{
    "status":0,
    "QTime":0,
    "params":{
      "fl":"dc.date,data_aux",
      "indent":"true",
      "wt":"json",
      "q":"*:*"
    }
  },
  "response":{"numFound":19097,"start":0,"docs":[
    {
      "data_aux":"2012-11-30",
      "dc.date":["2012-11-30"]
    },
    {
      "data_aux":"2009-11-11",
      "dc.date":["2009-11-11"]
    },
    {
      "data_aux":"1989-01-01",
      "dc.date":["1989"]
    },
    {
      "data_aux":"1991-01-01",
      "dc.date":["1991"]
    },
    {
      "data_aux":"2004-09-15",
      "dc.date":["2004-09-15"]
    },
    {
      "data_aux":"1991-01-01",
      "dc.date":["1991"]
    },
    {
      "data_aux":"1991-01-01",
      "dc.date":["1991"]
    },
    {
      "data_aux":"1991-11-01",
      "dc.date":["1991-11"]
    },
    {
      "data_aux":"1992-01-01",
      "dc.date":["1992"]
    },
    {
      "data_aux":"1990-11-08",
      "dc.date":["1990-11-08"]
    }
  ]}
```

En aquest cas podem veure els 10 primers resultats sense ordenar-los. El format de sortida escollit ha estat el JSON.

[illegible]

Es pot observar a la prova anterior que tots els valors són [s.d.], i per això el valor de la data_aux és 0000.

```
dugi@DUGip:/dades/dugi/old/proves$ curl "http://localhost:8983/solr/select/?q=*&fl=dc.date,data_aux&sort=data_aux%20desc&wt=json&indent=true"
{
  "responseHeader":{
    "status":0,
    "QTime":0,
    "params":{
      "fl":"dc.date,data_aux",
      "sort":"data_aux desc",
      "indent":"true",
      "q":"*:*:*",
      "wt":"json"}}},
```

```
"response":{"numFound":19097,"start":0,"docs":[
  {
    "data_aux":"2013-03-20",
    "dc.date":["2013-03-20"]},
  {
    "data_aux":"2013-03-19",
    "dc.date":["2013-03-19"]},
  {
    "data_aux":"2013-03-19",
    "dc.date":["2013-03-19"]},
  {
    "data_aux":"2013-03-19",
    "dc.date":["2013-03-19"]},
  {
    "data_aux":"2013-03-19",
    "dc.date":["2013-03-19"]},
  {
    "data_aux":"2013-03-15",
    "dc.date":["2013-03-15"]},
  {
    "data_aux":"2013-03-12",
    "dc.date":["2013-03-12"]},
  {
    "data_aux":"2013-03-08",
    "dc.date":["2013-03-08"]},
  {
    "data_aux":"2013-03-08",
    "dc.date":["2013-03-08"]},
  {
    "data_aux":"2013-03-08",
    "dc.date":["2013-03-08"]}
]
```

En aquesta prova hem ordenat de manera descendent per el camp data_aux.

9.2.3 Recol·lecció

Per comprovar que la recol·lecció funcionava correctament vam fer les següents proves.

Primer fem una recol·lecció a dia 1 de maig de 2013. Si tot ha anat bé, ens ha de recol·lectar els ítems entrats o modificats a data d'ahir o avui, ja que se suposa que va funcionant correctament.

```
dugi@DUGip:/dades/parser$ date
dc mai 1 19:37:35 CEST 2013
dugi@DUGip:/dades/parser$ php recolectar.php

ID: http://hdl.handle.net/10256/7759
ID: http://hdl.handle.net/10256/7760
ID: http://hdl.handle.net/10256.2/10935
ID: http://hdl.handle.net/10256.2/13698
```

El sistema ha recol·lectat els 4 ítems anteriors.
Comprovem que realment es van modificar ahir o avui.


```

    dc.identifier: { pattern: '^\\d{9}[\\dXx]{1}', to_meta: dc.identifier.isbn }
-
    dc.identifier: { pattern: null, to_meta: dc.identifier.citation }
-
    dc.creator: { pattern: null, to_meta: dc.creator }
-
    dc.contributor: { pattern: '^Universitat de Girona', to_meta:
dc.contributor.other }
-
    dc.description: { pattern: '^.{40}', to_meta: dc.description.abstract }
-
    dc.rights: { pattern: '^http:\\/\\/', to_meta: dc.rights.uri }
-
    dc.rights: { pattern: null, to_meta: dc.rights }
-
    dc.title: { pattern: null, to_meta: dc.title }
-
    dc.type: { pattern: null, to_meta: dc.type }
-
    dc.date: { pattern: '^\\d{4}(?:-\\d{2})?(?:-\\d{2})?$', to_meta: dc.date }
-
    dc.date: { pattern: '^\\d{4}-\\d{2})-\\?\\?$', to_meta: dc.date }
-
    dc.date: { pattern: '^\\d{4})-\\?\\?$', to_meta: dc.date }
-
    dc.date: { pattern: '^\\d{4})-\\d{4}$', to_meta: dc.date }
-
    dc.date: { pattern: '\\[.+]\\', to_meta: dc.date }
-
    dc.format: { pattern: '\\/', to_meta: dc.format.mimetype }
-
    dc.format: { pattern: null, to_meta: dc.format }
-
    dc.language: { pattern: '\\w{3}', to_meta: dc.language }
-
    dc.publisher: { pattern: null, to_meta: dc.publisher }
-
    dc.relation: { pattern: '^http:\\/\\/', to_meta: dc.relation.uri }

```

En aquest tros del fitxer de configuració, en format **YAML**²⁶, conté la configuració del DUGiFonsEspecials i hi podem apreciar el camp *last*, que conté la data d'avui.

A més a més, també hi podem apreciar les metadades d'entrada, les de sortida i el patró que ha de fer servir per saber a quina metadada de sortida ha d'anar el valor.

Anem a comprovar que s'ha recollert bé l'ítem.

```

dugi@DUGip:/dades/parser$ curl "http://localhost:8983/solr/select/?
q=dc.identifier.uri:%22http://hdl.handle.net/10256.2/13698%22&wt=json&indent=true"
{
  "responseHeader":{
    "status":0,
    "QTime":0,
    "params":{
      "indent":"true",
      "wt":"json",
      "q":"dc.identifier.uri:\\\"http://hdl.handle.net/10256.2/13698\\\""},
    "response":{"numFound":1,"start":0,"docs":[

```

²⁶ Acronim de "YAML Ain't Another Markup Language", és un format de serialització de dades semblant a JSON. És forma menys pesat que l'XML perquè no requereix obrir i tancar el tag. Per més informació es pot visitar: <http://www.yaml.org/>

```

{
  "dc.title":"El Conde-duque: el hombre y su política",
  "data_aux":"0000",
  "dc.type":"text",
  "dc.identifier.uri":"http://hdl.handle.net/10256.2/13698",
  "dc.rights":"Tots els drets reservats",
  "repository":"DUGiFonsEspecials",
  "dc.creator":[
    "Vicens Vives, Jaume"],
  "files":
["http://dugifonsespecials.udg.edu/bitstream/10256.2/13698/4/MSVV24_054.pdf.jpg|
http://dugifonsespecials.udg.edu/bitstream/10256.2/13698/3/MSVV24_054.pdf"],
  "dc.description.abstract":[
    "Apunts sobre la política del compte duc d'Olivares durant el regnat de
Felip IV"],
  "dc.date":["[s.d.]"],
  "dc.language":["spa"],
  "dc.publisher":[
    "Universitat de Girona. Biblioteca. Fons Jaume Vicens Vives"],
  "dc.identifier.other":["MSVV24_054"],
  "dc.format":["8 p."],
  "dc.format.mimetype":["application/pdf"]}
}

```

Podem comprovar com tots els valors han anat a les metadades que tocaven segons la configuració anterior. També podem apreciar que la recollida de les urls dels documents també s'ha fet correctament, i que cada thumbnail està lligat amb el fitxer que representa.

9.2.4 Mod_rewrite

Comprovar que les redireccions de les urls funcionin bé és molt important. Per això hem anat comprovant totes les expressions regulars del nostre fitxer *.htaccess*.

Per poder-les provar, hem creat el següent codi php al fitxer *index.php*:

```

<?php
    print_r($_GET);
?>

```

Així, quan per exemple volíem comprovar les següents regles, les posem al fitxer *.htaccess*:

```

RewriteEngine On
# autors
RewriteRule ^browse-authors/(.+)/(.*)$ index.php?pag=auts&q=$1&p=$2
RewriteRule ^browse-authors/(.*)$ index.php?pag=auts&q=$1
RewriteRule ^browse-authors/$ index.php?pag=auts
RewriteRule ^browse-authors$ index.php?pag=auts

```

Després fem peticions per anar comprovant que funcionin totes les regles

```

$ curl "http://localhost/browse-authors"
Array
(
    [pag] => auts
)

```

```
$ curl "http://localhost/browse-authors/"
Array
(
    [pag] => auts
)
$ curl "http://localhost/browse-authors/prova"
Array
(
    [pag] => auts
    [q] => prova
)
$ curl "http://localhost/browse-authors/prova/A"
Array
(
    [pag] => auts
    [q] => prova
    [p] => A
)
```

9.2.5 SSL a l'administració

Hem de comprovar que es faci la redirecció cap al protocol https. Per fer-ho fem una petició amb el paràmetre -I de curl, que ens permet veure les capçaleres HTTP.

Hem afegit una regla de **mod_rewrite** al fitxer *.htaccess*:

```
# dugiadmin secure
RewriteCond %{HTTPS} off
RewriteRule dugiadmin https://%{HTTP_HOST}%{REQUEST_URI}
```

Comprovem que funciona:

```
curl -I "http://localhost/dugadmin/"
HTTP/1.1 302 Found
Date: Sun, 05 May 2013 11:28:47 GMT
Server: Apache
Location: https://localhost/dugiadmin/
Cache-Control: max-age=-1054290
Expires: Tue, 23 Apr 2013 06:37:17 GMT
Vary: Accept-Encoding
Content-Type: text/html; charset=iso-8859-1
```

Podem veure la capçalera **Location**, que ens indica que el servidor ens redirigeix.

Ara hem de comprovar que funciona el protocol https.

```
$ curl "https://localhost/dugiadmin/"
curl: (77) error setting certificate verify locations:
CAfile: /etc/ssl/certs/ca-certificates.crt
CApath: none
```

Veiem que ens demana el certificat.

Provem en mode insegur.

```
$ curl "https://localhost/dugiadmin/" --insecure | head -n 5
% Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
           %         0         0     0    0    0    0    0    0
           %         0         0     0    0    0    0    0    0
           %         0         0     0    0    0    0    0    0
           %         0         0     0    0    0    0    0    0
           %         0         0     0    0    0    0    0    0
```

```

100 8396 0 8396 0 0 216k 0 --:--:-- --:--:-- --:--:-- 911k
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">

<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="ca" lang="ca">

<head>

<meta http-equiv="Content-Type" content="text/html; charset=UTF-8" />

<meta name="Distribution" content="Global" />
<meta name="Author" content="Biblioteca UdG - biblioteca.projectes@udg.edu" />

```

Funciona correctament.

9.2.6 Compressió de la sortida

Fem la prova del **mod_delfate**, per comprimir els fitxers. Posem les següents directives al fitxer *.htaccess*.

```

# tema gzip
SetOutputFilter DEFLATE
AddOutputFilterByType DEFLATE text/html text/css text/plain text/xml text/css
text/js application/javascript application/x-javascript

```

Fem la comprovació amb el navegador.

```

Request URL: http://dugi.udg.edu/
Request Method: GET
Status Code: 200 OK
▼ Request Headers view source
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Charset: ISO-8859-1,utf-8;q=0.7,*;q=0.3
Accept-Encoding: gzip,deflate,sdch
Accept-Language: ca-ES,ca;q=0.8
Cache-Control: max-age=0
Connection: keep-alive

```

Vegem ara, que la sortida realment és comprimida:

```

▼ Response Headers view source
Cache-Control: no-store, no-cache, must-revalidate, post-check=0, pre-check=0
Connection: Keep-Alive
Content-Encoding: gzip
Content-Length: 6558
Content-Type: text/html
Date: Sun, 05 May 2013 11:46:14 GMT
Expires: Thu, 19 Nov 1981 08:52:00 GMT
Keep-Alive: timeout=15, max=98
Pragma: no-cache
Server: Apache
Vary: Accept-Encoding

```

Per fer la comprovació amb el **curl**, hem d'afegir-li la capçalera *accept-encoding: gzip*.

```

$ curl -I "http://localhost/" --header "Accept-Encoding: gzip" HTTP/1.1 200 OK
Date: Sun, 05 May 2013 11:58:02 GMT

```

```

Server: Apache
Set-Cookie: PHPSESSID=e3b6d42601def3cb8d097266f25f391b; path=/
Expires: Thu, 19 Nov 1981 08:52:00 GMT
Cache-Control: no-store, no-cache, must-revalidate, post-check=0, pre-check=0
Pragma: no-cache
Vary: Accept-Encoding
Content-Encoding: gzip
Content-Length: 20
Content-Type: text/html

```

Podem veure com el servidor ens contesta que el contingut és comprimit amb gzip.
La sortida del contingut és la següent

```

curl "http://localhost/" --header "Accept-Encoding: gzip"
tKk'FnV_Hd&6H0)JB:e^#E n*v>l D h=;;s n"w t v z @ r>
> c 99 c h@|3 Udm/B?
JA.D' c G 6 3p g &K &ByWF 4 GY yùv0 6y 0 Aa"*M J c t v n{
^c`f 9 Xa .rq xK zú? |\c Q> bH &j" B; |p YB:d Q,^201 i j'

```

Realment el contingut surt comprimit.

9.2.7 Mod_expires

Fem la prova del funcionament del *mod_expires*, perquè els navegadors guardin en cache les imatges.

Afegim les següents línies al fitxer *.htaccess*

```

#mod expires
### activate mod_expires
ExpiresActive On
### Expire .gif's 1 month from when they're accessed
ExpiresByType image/gif A2592000
ExpiresByType image/jpeg A2592000
ExpiresByType image/png A2592000
### Expire everything else 1 day from when it's last modified
### (this uses the Alternative syntax)
ExpiresDefault "modification plus 1 day"

```

Comprovem que una imatge estigui en cache 1 mes

```

$ curl -I "http://dugi.udg.edu/images/up.png"
HTTP/1.1 200 OK
Date: Sun, 05 May 2013 12:00:59 GMT
Server: Apache
Last-Modified: Tue, 01 Mar 2011 10:48:49 GMT
ETag: "18059-10a-49d698a952640"
Accept-Ranges: bytes
Content-Length: 266
Cache-Control: max-age=2592000
Expires: Tue, 04 Jun 2013 12:00:59 GMT

```

Vary: Accept-Encoding
Content-Type: image/png

Podem veure que funciona perfectament. La cache d'aquesta imatge caduca el dia 4 de juny. D'aquí a 30 dies exactament.

També és important veure que tenim la capçalera ETag, que permet al navegador saber quina versió del document està consultant, i per tant, si s'ha modificat des de l'última vegada que es va consultar.

10. Implantació i resultats

Per a implantar el sistema en producció, la instal·lació dels serveis no ha de ser la mateixa que al servidor de proves. No hem de donar pistes dels errors, i s'ha de cuidar tots els detalls.

Un cop instal·lats els serveis necessaris, implantació s'ha fet en dues parts:

- La primera part ha sigut la implantació del sistema bàsic, que consta del sistema de recollida i el sistema de consulta.
- La segona part, ha sigut afegir tota la part d'administració.

Evidentment, abans de fer aquesta implantació, hem fet un testeig del sistema a la màquina de proves.

10.1. Instal·lació dels serveis

La instal·lació de la màquina de producció es va fer clonant la màquina de proves abans de començar a posar-hi el codi. Després es va acabar de polir la instal·lació amagant tots els missatges d'error i de versions dels serveis.

De totes maneres, en aquest punt mostrarem la instal·lació completa dels serveis.

La base

El DUGi s'ha instal·lat sobre un linux amb Ubuntu Server 10.04 LTS.

```
$ lsb_release -a
No LSB modules are available.
Distributor ID: Ubuntu
Description:   Ubuntu 10.04.4 LTS
Release:      10.04
Codename:     lucid
```

Aquesta versió està pensada per a ser utilitzada en servidors, i és mantinguda molt temps per la gent d'Ubuntu. Així doncs, ens assegurem que hi haurà actualitzacions dels paquets durant molt de temps, abans de que ens decidim per pujar la versió del sistema operatiu.

Creació de l'usuari dugi

L'usuari dugi serà el que tindrà permisos sobre els fitxers que servirà l'apache.

```
$ sudo adduser dugi
```

Contestem les preguntes que ens fa el sistema i ja tenim l'usuari creat.

```
$ cat /etc/passwd|grep dugi
dugi:x:1001:33:,,,:/home/dugi:/bin/bash
```

Instal·lar servidor de correu

Hem escollit el servidor de correu **ssmtp** per la seva simplicitat, i perquè podem fer-lo servir d'esclau d'un altre servidor de correu, en aquest cas el de la UdG, el **manol**.

Abans d'instal·lar res, actualitzem les fonts de programari.

```
$ apt-get update
```

Ara ja podem instal·lar l'ssmtp.

```
$ sudo apt-get install ssmtp
```

Un cop instal·lat, hem de configurar-lo perquè actuï de client del servidor de correu **manol.udg.edu**. Per fer-ho, editem amb permisos de super-usuari el fitxer */etc/ssmtp/ssmtp.conf*.

```
# Config file for sSMTP sendmail
#
# The person who gets all mail for userids < 1000
# Make this empty to disable rewriting.
#root=postmaster
# oriol

# The place where the mail goes. The actual machine name is required no
# MX records are consulted. Commonly mailhosts are named mail.domain.com
#mailhub=mail
# oriol
mailhub=manol.udg.edu:25

# Where will the mail seem to come from?
#rewriteDomain=

# The full hostname
#hostname=dugi.udg.edu
# oriol
hostname=manol.udg.edu:25

# Are users allowed to set their own From: address?
# YES - Allow the user to specify their own From: address
# NO - Use the system generated From: address
#FromLineOverride=YES
# oriol
FromLineOverride=YES
```

Amb aquesta configuració fem que els mails els envii **manol.udg.edu**, i que podem editar qui envia el mail.

Instal·lar Apache2

Instal·lem l'**Apache** amb el mòdul de **ssl** per a la zona d'administració del DUGi.

```
$ sudo apt-get install apache2 libapache-mod-ssl
```

Un cop instal·lat, hem de configurar-lo.

Comencem indicant-li que volem que el servei s'engegui amb l'usuari dugi.

Per fer-ho, editem el fitxer */etc/apache2/envvars*.

```
# envvars - default environment variables for apache2ctl

# Since there is no sane way to get the parsed apache2 config in scripts, some
# settings are defined via environment variables and then used in apache2ctl,
# /etc/init.d/apache2, /etc/logrotate.d/apache2, etc.
# oriol
#export APACHE_RUN_USER=www-data
export APACHE_RUN_USER=dugi
```

```
export APACHE_RUN_GROUP=www-data
export APACHE_PID_FILE=/var/run/apache2.pid
```

Seguidament, podem activar els mòduls d'**Apache** que necessitem. Utilitzarem la utilitat d'**apache2 a2enmod**, que l'únic que fa es fer un enllaç dels fitxers d'/etc/apache2/mods-available/ a /etc/apache2/mods-enabled.

```
$ sudo a2enmod rewrite
$ sudo a2enmod deflate
$ sudo a2enmod expires
$ sudo a2enmod ssl
```

Ara afegim un filtre per no deixar veure els fitxers de backup. Això ho farem al fitxer /etc/apache2/apache2.conf.

```
AccessFileName .htaccess

#
# The following lines prevent .htaccess and .htpasswd files from being
# viewed by Web clients.
#
<Files ~ "^\.ht">
    Order allow,deny
    Deny from all
</Files>
# Oriol
<Files ~ "\.bak$">
    Order allow,deny
    Deny from all
</Files>
<Files ~ "~$">
    Order allow,deny
    Deny from all
</Files>
```

Al mateix fitxer, una mica més avall, desactivem les signatures de l'Apache.

```
#ServerTokens Full
# oriol
# amagant info
ServerTokens Prod
ServerSignature Off
```

Activem que es mostri les nostres pàgines quan es produeixin els següents errors:

```
# oriol error
ErrorDocument 503 "/error/503.html"
ErrorDocument 403 "/error/403.html"
ErrorDocument 500 "/error/500.html"
```

Comprovem que l'**Apache2**, per defecte, mostra informació de la versió i el sistema. Utilitzarem el paràmetre **-I** del **curl**, que ens permet mostrar només les capçaleres HTTP.

```
$ curl -I "dugi.udg.edu"
HTTP/1.1 200 OK
Date: Sun, 05 May 2013 09:28:41 GMT
Server: Apache/2.2.8 (Ubuntu) mod_ssl/2.2.8 OpenSSL/0.9.8g
```

```
Expires: Thu, 19 Nov 1981 08:52:00 GMT
Cache-Control: no-store, no-cache, must-revalidate, post-check=0, pre-check=0
Pragma: no-cache
Vary: Accept-Encoding
Content-Type: text/html
```

Com es pot apreciar, l'Apache dóna informació massa interessant per algun atacant, inclús dóna la versió d'alguns mòduls instal·lats.

Després de reiniciar l'Apache2, les capçaleres són les següents:

```
$ curl -I "dugi.udg.edu"
HTTP/1.1 200 OK
Date: Sun, 05 May 2013 09:24:07 GMT
Server: Apache
Expires: Thu, 19 Nov 1981 08:52:00 GMT
Cache-Control: no-store, no-cache, must-revalidate, post-check=0, pre-check=0
Pragma: no-cache
Vary: Accept-Encoding
Content-Type: text/html
```

Creació del virtual host per defecte

Editem el fitxer `/etc/apache2/sites-available/default` per configurar el host principal.

Deixem el virtual host a tots els ports, perquè ja sobreescrivem el port **443** per al host **ssl**. El que si que canviarem serà l'arrel de l'apache, així, tota la resta del sistema la posarem fora d'aquesta base, i hi accedirà el php.

```
# oriol
#       DocumentRoot /var/www/
#       DocumentRoot /dades/dugi/
```

Bloquegem el llistat dels documents als directoris, però permetem que es pugui sobreescrivre la configuració amb el fitxer `.htaccess`. Això es fa amb la directiva *AllowOverride All*, i ens permetrà configurar el **mod_rewrite** a l'`.htaccess` del nostre directori base.

```
#       <Directory /var/www/>
<Directory /dades/dugi/>
#           Options Indexes FollowSymLinks MultiViews
Options FollowSymLinks MultiViews
# oriol
#           AllowOverride None
#           AllowOverride All

#           Order allow,deny
#           allow from all

</Directory>
```

Instal·lació del virtual host d'administració

Aquest virtual host, correrà al port 443 amb ssl. Tot i que hem instal·lat el mòdul d'apache, instal·larem una utilitat per generar un certificat ssl.

```
$ sudo apt-get install ssl-cert
```

Amb aquesta utilitat generem el certificat, contestant les preguntes que ens demani. Guardarem el certificat a la carpeta `ssl` del directori `/etc/apache2`.

```
$ make-ssl-cert /usr/share/ssl-cert/ssleay.cnf /etc/apache2/ssl/apache.pem
```

Després editem el fitxer */etc/apache2/ports.conf* per indicar a l'**Apache** que escolti pel port **443**.

```
Listen 80

<IfModule mod_ssl.c>
    Listen 443
</IfModule>
```

Després copiem el fitxer */etc/apache2/sites-avaible/default* a */etc/apache2/sites-avaiable/ssl* i hi editem les primeres línies.

```
NameVirtualHost *:443
<VirtualHost *:443>
    ServerAdmin oriol.olive@udg.edu

    SSLEngine On
    SSLCertificateFile /etc/apache2/ssl/apache.pem
```

Ara ja podem activar els virtual hosts.

Activació dels virtual hosts

Per activar els virtual hosts, utilitzarem la utilitat d'**apache2 a2ensite**, que ja ens farà l'enllaç al directori */etc/apache2/sites-enabled/*.

```
$ sudo a2ensite default
$ sudo a2ensite ssl
```

Després reiniciem l'apache perquè al haver activat nous mòduls, agafi els canvis.

```
$ sudo /etc/init.d/apache2 restart
```

Ara és el moment de comprovar que el nostre servidor respon correctament tant a l'adreça <http://dugi.udg.edu> com a l'adreça <https://dugi.udg.edu/>, i ens contesta amb el fitxer de prova que hi hem posat.

Instal·lar el PHP5

Ara que ja hem comprovat que tot funciona, ja podem instal·lar el php5.

```
$ sudo apt-get install php5
```

Un cop instal·lat, necessitem configurar-lo. Com que el php es pot executar tant com a un mòdul de l'apache com des de la línia de comanda, ens porta dos fitxers de configuració independents per a cada us. Així doncs, podem assignar més memòria al **PHP CLI** que al **PHP** que corre sota apache.

Configurar el PHP sota Apache2

Al PHP que executarà l'Apache, no ens interessa, per exemple, que es mostri la versió d'aquest. Per configurar-lo, editem el fitxer */etc/php5/apache2/php.ini*.

```
;expose_php = On
; oriol
```

```
expose_php = Off
```

Abans d'amagar la signatura del PHP, anem a veure què mostra.

```
$ curl -I "dugi.udg.edu"
HTTP/1.1 200 OK
Date: Sun, 05 May 2013 09:27:07 GMT
Server: Apache
X-Powered-By: PHP/5.2.4-2ubuntu5.14
Set-Cookie: PHPSESSID=4b196287c5c1900b9d63b9d2dec6c98; path=/
Expires: Thu, 19 Nov 1981 08:52:00 GMT
Cache-Control: no-store, no-cache, must-revalidate, post-check=0, pre-check=0
Pragma: no-cache
Vary: Accept-Encoding
Content-Type: text/html
```

Igual que en el cas de l'Apache, la versió exacta del PHP és una informació sensible. Per comprovar-ho hem de reiniciar l'apache, fer una petició i fixar-nos en les capçaleres de l'HTTP.

```
$ curl -I "dugi.udg.edu"
HTTP/1.1 200 OK
Date: Sun, 05 May 2013 09:24:07 GMT
Server: Apache
Set-Cookie: PHPSESSID=bf82a40e72e06b15e666833e106b9444; path=/
Expires: Thu, 19 Nov 1981 08:52:00 GMT
Cache-Control: no-store, no-cache, must-revalidate, post-check=0, pre-check=0
Pragma: no-cache
Vary: Accept-Encoding
Content-Type: text/html
```

Seguidament, anem a amagar els errors. Els errors del PHP també poden donar informació sensible, com per exemple rutes de fitxers i directoris, usuaris amb permisos per executar codi php, llibreries i funcions que utilitzem, etc.

```
; oriol
;error_reporting = E_ALL & ~E_NOTICE
error_reporting = 0

; oriol
;display_errors = On
display_errors = Off

; Even when display_errors is on, errors that occur during PHP's startup
; sequence are not displayed. It's strongly recommended to keep
; display_startup_errors off, except for when debugging.
display_startup_errors = Off
```

Com que tenim una classe que controla els logs del nostre sistema, aquesta ja s'encarregarà d'avisar-nos si hi ha algun problema greu. No necessitem que l'usuari tingui informació del que ha passat. En el cas del **PHP CLI**, no necessitem amagar els errors, ja que no estaran exposats via HTTP.

Ampliem una mica el límit de consum màxim de memòria del PHP sobre Apache.

```
; oriol
;memory_limit = 16M          ; Maximum amount of memory a script may consume (16MB)
memory_limit = 128M
```

Activem els logs del PHP per al nivell de log d'ERROR.

```
; oriol activar log
;log_errors = Off
log_errors = On
```

A més a més, especifiquem el fitxer on volem que vagin a parar aquests logs, i desactivem l'opció del **syslog**.

```
; Log errors to specified file.
; oriol activar logs
;error_log = filename
error_log = /var/log/php/dugi.log

; Log errors to syslog (Event Log on NT, not valid in Windows 95).
; oriol activar logs
;error_log = syslog
```

Finalment, lliguem el sistema d'enviament de mails del PHP amb el servidor **smtp** que hem instal·lat anteriorment.

```
; For Unix only. You may supply arguments as well (default: "sendmail -t -i").
;sendmail_path =
; oriol
sendmail_path = smtp
```

És importantíssim vigilar que la directiva *allow_url_include* estigui desactivada, ja que és la que permet fer incusions remotes de codi php. Així, en el cas que haguéssim comés algun error i permetéssim que l'usuari pogués passar per paràmetre el nom, o part del nom, del fitxer que s'inclou o es requereix, aquest només podria fer una inclusió local.

```
; Whether to allow include/require to open URLs (like http:// or ftp://) as files.
allow_url_include = Off
```

També és molt important que la directiva *safe_mode_allowed_env_vars* contingui un prefix, ja que de no ser així permetria a l'usuari modificar les variables d'entorn del php sense voler, si poses el mateix nom a una variable pròpia.

```
; Note: If this directive is empty, PHP will let the user modify ANY
; environment variable!
safe_mode_allowed_env_vars = PHP_
```

Una altra directiva perillosa és la de *register_globals*. També ens n'hem d'assegurar que estigui desactivada, ja que si no, un usuari podria modificar el valor de les nostres variables des de la url en cas de haver-les inicialitzat, i saltar-se així restriccions. Aquest és un cas semblant al de la inclusió de codi, en teoria, si s'ha programat correctament no hi ha d'haver opció per a aquests tipus d'atacs, però mai està de més desactivar les directives perilloses.

```
; Whether or not to register the EGPCS variables as global variables. You may
; want to turn this off if you don't want to clutter your scripts' global scope
; with user data. This makes most sense when coupled with track_vars - in which
; case you can access all of the GPC variables through the $HTTP_*_VARS[],
; variables.
;
; You should do your best to write your scripts so that they do not require
```

```

; register_globals to be on; Using form variables as globals can easily lead
; to possible security problems, if the code is not very well thought of.

; NOTE: applications relying on this feature will not receive full
; support by the security team. For more information please
; see /usr/share/doc/php5-common/README.Debian.security
;
register_globals = Off

```

A partir de la versió 4.1.10 del php, es recomana fer servir les variables `$_GET` i `$_POST`, entre d'altres, enlloc de `$HTTP_GET_VARS` i `$HTTP_POST_VARS`. Per lo que no necessitem que la directiva `register_log_arrays` estigui activada.

```

; Whether or not to register the old-style input arrays, HTTP_GET_VARS
; and friends. If you're not using them, it's recommended to turn them off,
; for performance reasons.
;register_long_arrays = On
; oriol
register_long_arrays = Off

```

Tampoc necessitem que es lligui les variables `$argv` i `$argc` contiguin els valors de `$_GET`, ja que al programar orientat a objectes, no s'executaria el mateix *main* des de l'apache que des de la línia de comanda.

```

; This directive tells PHP whether to declare the argv&argc variables (that
; would contain the GET information). If you don't use these variables, you
; should turn it off for increased performance.
;register_argc_argv = On
; oriol
register_argc_argv = Off

```

Configurar el PHP CLI

Posarem totes aquestes modificacions al fitxer `/etc/php5/cli/php.ini`, menys les referents a mostrar errors. La directiva `expose_php` no té cap sentit aquí, i el valor que hi deixem és del tot indiferent. El que sí que canviarem és el límit de memòria, ja que potser alguns processos en requereixen més, i el temps d'execució no ens és tant important. Tampoc és important en aquest cas, que la directiva `allow_url_include` estigui desactivada, ja que un atacant no té manera de fer incloure un fitxer. Tampoc ho són, per la mateixa raó, les altres directives que hem comentat.

```

;memory_limit = 32M      ; Maximum amount of memory a script may consume (32MB)
; oriol + memory
memory_limit = 512M

```

Instal·lar les llibreries de PHP

Necessitem les següents llibreries de PHP:

- Llibreria YAML de symfony2: Són els principals promotors del YAML, i crec que tenen la millor llibreria de yaml per php.
- Llibreria de Savant3: Per utilitzar el sistema de plantilles Savant.
- Llibreria CURL: Per fer peticions HTTP.

YAML

Per instal·lar la llibreria YAML de Symfony no cal instal·lar tot el Symfony enter. Només cal descarregar-la i posar-la a la carpeta que volem.

La descarreguem de la següent url:

```
https://github.com/symfony/Yaml
```

Quan tinguem creat tot el sistema de carpetes del DUGi, la posarem a la carpeta **lib**.

Savant3

Per instal·lar les llibreries de Savant3, la millor manera és instal·lant el paquet **pear** de PHP.

```
$ sudo apt-get install php-pear
```

Un cop instal·lat el pear, instal·lem el Savant3 amb les següents comandes:

```
$ pear channel-discover phpsavant.com
$ pear install savant/Savant3
```

CURL

Tot i que amb la directiva *allow_url_fopen* activada el php ens permet fer peticions web amb les funcions **fopen** i **file_get_contents**, hem optat per fer servir la llibreria **CURL**, que ens dóna moltes més opcions.

La instal·lem de la següent manera:

```
$ apt-get install php5-curl
```

Un cop instal·lada la llibreria CURL, ja tenim la base preparada per a instal·lar-hi el DUGi.

10.2 Instal·lació del DUGi

Per instal·lar el DUGi només ens cal copiar la carpeta */dades* del servidor de proves al servidor de producció. Ho fem per exemple, des del servidor de proves amb la comanda `scp` i amb el paràmetre `-r` que ens permet fer-ho recursivament.

```
$ scp -r /dades/* dugi@dugi.udg.edu:/dades/
```

L'estructura del directori */dades* queda així:

```
/dades/autoritats
/dades/dugi
/dades/config
/dades/parser
/dades/stats
/dades/lib
/dades/index
/dades/cache
/dades/log
```


El directori **autoritats** conté les autoritats que falten per fer. El directori **dugi** conté el codi de l'aplicació web. El directori **config** conté els fitxers de configuració que utilitzen els sistemes, i també conté els fitxers de multi-llenguatge. El directori **parser** conté el codi de la recollida. El directori **stats** conté les estadístiques. El directori **lib** conté les llibreries de php, tant les de tercers com les classes creades per nosaltres que utilitzen més d'una part del sistema. El directori **index** conté l'index del DUGi. El directori **cache** conté la cache del sistema. I el directori **log** conté els logs del sistema.

El directori dugi està format per els següents directoris:

```
/dades/dugi/controllers
/dades/dugi/css
/dades/dugi/dugiadmin
/dades/dugi/images
/dades/dugi/js
/dades/dugi/templates
```

El directori **controllers** conté les classes controlador de la web, i el directori **templates** conté les plantilles de la vista.

Un cop copiat el DUGi, cal netejar els directoris de les estadístiques, les autoritats, l'index, la cache i els logs.

```
$ rm -rf /dades/stats/*
$ rm /dades/autoritats/*
$ rm -rf /dades/index/data/*
$ rm /dades/cache/*
$ rm /dades/log/*
```

Un cop nets aquests directoris, hem d'acabar d'instal·lar l'index.

Creem el fitxer `/etc/init.d/solr_index` amb el següent codi:

```
#!/bin/sh -e
#####
#
# Oriol Olivé
# index up script
#
#####

ENV="env -i LANG=C PATH=/usr/local/bin:/usr/bin:/bin"
set -e
. /lib/lsb/init-functions

INDEX_DIR="/dades/index/"
INDEX_USER="dugi"
case $1 in
    start)
        PID=`pgrep -f 'java -jar start.jar'`||PID=""
        if [ "$PID" != "" ] ; then
            log_warning_msg "Warning service solr_index already started"
            log_daemon_msg "Starting service " "solr_index"
            log_end_msg 1
        else
            cd $INDEX_DIR
```

```

        sudo -u $INDEX_USER java -jar start.jar &
        log_daemon_msg "Starting service " "solr_index"
        log_end_msg 0
    fi
;;
stop)
    log_daemon_msg "Stoping service " "solr_index"
    pkill -f 'java -jar start.jar'
    rm $INDEX_DIR/nohup.out
    log_end_msg 0
;;
*)
    log_success_msg "Usage: /etc/init.d/solr_index {start|stop}"
    exit 1
;;

esac

```

El fem executable:

```
$ sudo chmod +x /etc/init.d/solr_index
```

Ara l'hem d'instal·lar als runlevels pertinents, perquè s'engegui el servei al engegar la màquina. L'apache s'engega amb el valor 91, nosaltres engegarem l'índex al 92, doncs.

```
$ sudo update-rc.d solr_index defaults 92 08
```

Finalment, activem els processos que s'han de fer periòdicament al crontab de l'usuari dugi.

```

# m h dom mon dow   command
# stats
10 * * * * /usr/bin/php /dades/lib/log/ip_geo.php
55 23 * * * /usr/bin/php /dades/lib/log/stats.php
# logrotate
59 23 * * * /usr/bin/php /dades/lib/log/logrotate.php
# harvesting
00 1 * * * cd /dades/parser && /usr/bin/php /dades/parser/recolectar.php

```

Cada dia a la una de la matinada s'ha d'executar la recol·lecció. Cap a mitjanit, s'executarà el generador d'estadístiques, i just a la mitjanit, es farà la rotació dels logs. Cada hora es mirarà si hi ha ips pendents de processar.

Un cop fet això, el DUGi ja estarà instal·lat completament i a punt per funcionar a tot rendiment. És el moment d'aixecar els serveis.

```

$ /etc/init.d/apache2 start
$ /etc/init.d/solr_index start

```

10.3 Resultats

Dins dels resultats, diferenciarem entre els resultats de la implantació i el grau d'assoliment dels objectius.

10.3.1 Resultats de la implantació

Pel que fa a l'ús que està tenint el DUGi, no n'acabem d'estar del tot satisfets. Veiem que hi ha moltes persones que, potser perquè tenen l'adreça guardada als marcadors, consulten directament el DUGiDocs o el DUGiMedia.

Diàriament tenim unes 500 visites humanes i unes 3000 de robots de mitjana. D'aquestes 500, unes 130 són en català, 150 en castellà i 320 en anglès. Al cap de setmana aquestes 500 baixen notablement a unes 350 visites humanes.

Potser s'hauria d'haver fet una campanya de difusió a més a més de posar l'enllaça a la web de la Biblioteca de la UdG.

El nombre de visites diàries dels robots, són força satisfactòries. Els *crawlers*²⁷ dels cercadors troben contingut nou i interessant, i ens indexen més pàgines.

El google, per exemple, ens té indexades 138.000 pàgines. Això es pot veure fent la cerca **site:dugi.udg.edu**.

De totes maneres, caldrà seguir treballant per aconseguir que molts més resultats del DUGi surtin a la primera pàgina de resultats dels cercadors.

Pel que fa al canvi del sistema vell per al nou, el canvi es va fer d'un dia per l'altre i no va crear cap problema. Es va aixecar la màquina nova, i es va tancar l'antiga. Potser algun usuari es va trobar redirigit al clicar un link, però tot i canviar la pàgina, el contingut va ser el mateix.

Per altra banda, els usuaris del departament de Projectes n'estan força contents perquè ajuda a detectar problemes i permet fer cerques que afecten a tots els repositoris alhora.

10.3.2 Grau d'assoliment dels objectius

L'objectiu principal del projecte era el de desenvolupar un portal web amb un sistema d'indexació que permeti la cerca per diferents criteris. Aquest objectiu principal s'ha complert, com es pot veure a les captures de pantalla següents.

27 Els robots dels cercadors web, que rastregen el contingut de les pàgines web per a indexar-lo. També són coneguts com a aranyes.

El següent gràfic és el llistat d'ítems per títol. Podem veure com es pot anar passant les pàgines d'una a una, anar al final, etc.

DUGiDocs

DUGiMedia

DUGiFonsEspecials

RACO

TDX

MDX

Recercat

MDC

DUGi

Portal de cerca conjunta als documents digitals de la UdG

Inici

Llistar Títols

Llistar Autors

Llistar Matèries

Cercar:

Llistar Títols

Shan trobat 32944 ítems

Llistant ítems des de 0 a 25:

<

1

2

3

4

5

6

>

Format	Data	Títol	Autor
	[s.d.]	104. Delimitación del período: de 1871 a 1974	Vicens Vives, Jaume
	[s.d.]	105. La crisis de julio de 1914	Vicens Vives, Jaume
	[s.d.]	107	Vicens Vives, Jaume
	23 setembre 2008	10 anys de l'Escola Samba Kubally	López, Anna ; Besalú i Costa, Xavier
	23 setembre 2008	10 anys de la UdG. Una mirada externa	Anònim

Matèries -

Open source software -- Congresses

Visualització tridimensional (Informàtica) Ensenyament

universitari -- Metodologia -- Congressos Tecnologia

educativa Computer algorithms Internet in higher

education Density functionals

Autors -

Panyella Aldrúfeu, Josep Maria Desconegut Duran i

Portas, Miquel Barceló i Vidal, Carles Nolla, Josep

Maria Sala, Joan Rosa, Josep Lluís de la

El següent gràfic mostra el llistat d'autors. Tenim un caixetí que ens permet filtrar pel començament del nom dels autors. Els resultats estan organitzats per el prefix més la següent lletra, així doncs, si posem que comença per **ba**, els resultats estaran ordenats per **baa**, **bab**, **bac**, **bad**, etc

DUGiDocs

DUGiMedia

DUGiFonsEspecials

RACO

TDX

MDX

Recercat

MDC

DUGi

Portal de cerca conjunta als documents digitals de la UdG

Inici

Llistar Títols

Llistar Autors

Llistar Matèries

Llistar Autors

Shan trobat 360 autors

Comença per:

Llistant autors des del 0 fins al 360

BAB		BAC		BAD	
Babault, Julien	[1]	Bacardí, Montserrat	[1]	Badell Palau, Cristina	[2]
Babini, J.	[1]	Bacardit Peñarroya, Monts...	[1]	Badia Brea, Maria Elena	[2]
Babot, Montse	[1]	Bacca Cortés, Eval Bladim...	[1]	Badia i Gorchs, Marc	[3]
Baburès, Joan	[1]	Bacci, Michele	[2]	Badia i Homs, Joan	[18]
		Baceta, Juan Ignacio	[1]	Badia i Homs, Lola	[2]
		Bach Ariza, Àlex	[2]	Badia i Margarit, Antoni M...	[2]
		Bach i Plaza, Joan	[5]	Badia Roig, Carmina	[1]
		Bach Plaza, Joan	[1]	Badia, E.	[1]
		Bach, A.	[3]	Badia, Enrique	[1]
		...	[+]	...	[+]

BAE		BAF		BAG	
Baena Muñoz, Marta	[2]	Bafaluy Uruel, David	[2]	Baglaenko, Nina	[2]

Matèries -

Espai Europeu d'Educació Superior

Imatges -- Processament Ciències del medi ambient

Anàlisi multivariable Química quàntica -- Congressos

Algorismes computacionals MPLS (Norma)

Autors -

Vicens Vives, Jaume Muguerza, Javier Martín

Fernández, Josep Antoni Dalmau, Jordi Pablo, Moisès

de Sala, Joan Nadal, Josep M.

Per cercar ítems hem de fer servir el caixetí de cerca. En l'exemple següent es cerca ítems que tinguin la paraula *catàleg* al **títol**. Com es pot veure a la captura, la cerca no té en compte les majúscules ni els accents.

The screenshot shows the DUGi portal interface. At the top, there's a search bar with 'catàleg' entered. Below it, a table lists search results under the heading 'Llistar Títols'. The table has columns for Format, Data, Títol, and Autor. The results show four items related to catalogs and library systems. On the right side, there's a sidebar with 'Matèries' (Subjects) and 'Autors' (Authors) lists.

Format	Data	Títol	Autor
[Icon]	1 gener 1985	Addenda al catàleg de peixos del mar català de les col·leccions de la Universitat Autònoma de Barcelona	García-Berthou, Emili ; Moreno Amich, Ramon
[Icon]	1985	Addenda al catàleg de peixos del mar català de les col·leccions de la Universitat Autònoma de Barcelona	García-Berthou, Emili ; Moreno Amich, Ramon
[Icon]	14 febrer 2008	Canvi de sistema del catàleg de la Biblioteca de la UdG: introducció a Millennium	Soler Pujol, Amàlia
[Icon]	1 març 1988	Catàleg de 1988-03-01 a José Ferrater Mora des de Madrid (Espanya)	Sánchez Cuesta, León

L'aplicació permet anar obrint els ítems com si fossin pop-ups per a comparar-los, o obrir-los en una nova pestanya. Si s'han obert com a pop-ups, es poden maximitzar, minimitzar i escalar a voluntat per una millor consulta.

This screenshot shows the same search results table, but with two items opened in pop-up windows for comparison. The first pop-up is for item 105, 'LA CRISIS DE JULIO DE 1914', and the second is for item 104, 'DELIMITACIÓN DEL PERÍODO: DE 1871 A 1974'. Both pop-ups display detailed metadata for each item, including author, date, summary, pages, format, citation, document URL, language, editor, rights, title, type, and repository.

Format	Data	Títol	Autor
[Icon]	[s.d.]	104. Delimitación del período: de 1871 a 1974	
[Icon]	[s.d.]	105. LA CRISIS DE JULIO DE 1914	

El disseny del portal, es pot millorar, però es prou agradable i usable. Tots els elements es veuen a simple vista i és molt intuïtiu.

Les tasques que ens havíem proposat per complir els objectius, s'han anat executant en els punts anteriors. Aquestes tasques són les següents:

- **Fer un estudi de les eines més interessants per implementar aquest projecte:** Aquest estudi s'ha dut a terme en el punt 7 del present projecte.

- **Trobar una metodologia que ens pugui servir per dissenyar i implementar el portal d'una manera eficient:** S'ha escollit en el punt 3, i s'ha aplicat el punt 8.
- **Dissenyar l'aplicació:** S'ha mostrat en el punt 8.
- **Implementar-la:** S'ha explicat els punts més interessants en el punt 9.
- **Fer la instal·lació tant de l'entorn de treball com de l'entorn de producció:** S'ha mostrat al punt 10.
- **Fer el testeig:** El testeig s'ha mostrat en el punt 9, a l'apartat de proves.
- **Fer una optimització:** Aquesta optimització s'ha mostrat en el punt 10.
- **Fer un estudi de les principals amenaces de seguretat que ens poden afectar, i en cas d'haver-n'hi corregir-les:** S'ha fet un exercici d'estudi de les principals amenaces que ens podien afectar, i s'ha comentat en el punt 10 quines eren les accions que preniem per minimitzar el seu efecte.

Ens podem donar per satisfets, ja que el nivell d'assoliment dels objectius del projecte ha sigut aproximadament del 100%.

11. Conclusions

11.1 Conclusions sobre els requisits

En quan als requisits de la nostra aplicació, també podem estar satisfets. Hem complert la majoria de requisits amb una solució acceptable, tot i que alguna es pot millorar.

Els requisits del sistema eren els següents:

- **Crear un sistema que ens permeti unificar les entrades d'autors per evitar les confusions en els resultats.**

El sistema de detecció d'autoritats conflictives funciona perfectament i permet a l'administrador seleccionar l'autoritat correcta en cas de ser necessari.

DUGi: Administració

Autoritats per revisar:

Nous ítems:

Estat índex:

Cache:

actiu ●
reiniciar

Netejar cache

Autoritats conflictives per l'ítem:

URL

Últimes consultes:

Data	IP	URL
02-May-2013 13:04:00	66.249.76.72	item/http://hdl.handle.net/10256/3425
02-May-2013 13:03:34	66.249.76.72	browse-titles/dc.subject:%22Bonet%20Guin%C3%B3,%20Eduard%20--%20Interviews%22/0/25
02-May-2013 13:02:45	46.161.41.24	author/RogetiArmengol,Elena
02-May-2013 13:02:34	66.249.76.72	item/http://hdl.handle.net/10256.2@1195
02-May-2013 13:01:32	66.249.76.72	browse-titles/(dc.creator:%22Roman%C3%AD,%20Joan%20M,%22%20OR%20dc.contributor:%22Roman%C3%AD,%20Joan%20M,%22%20OF
02-May-2013 13:01:26	180.76.6.227	subject/Methane
02-May-2013 13:00:32	66.249.76.72	item/http://hdl.handle.net/10256.1@356

Enviar

© 2011 Universitat de Girona | Biblioteca UdG

- **Implementar un client d'OAI-PMH completament estàndard i que ens permeti aprofitar totes les opcions del protocol.**

S'ha implementat un client d'OAI-PMH que permet recol·lectar de tots els repositoris que necessitem, que detecta els ítems eliminats i que permet recol·lecció selectiva per sets. De moment el client només treballa amb el format oai_dc.

- **Enriquir les cerques donant opcions de filtre i refinament.**

L'aplicació web permet anar refinant la cerca des del caixetí de cerca. Els núvols d'autors i matèries permeten refinar les cerques d'una manera més intuïtiva. També es permet refinar la cerca seleccionant i des-seleccionant repositoris.

La captura següent mostra com hem refinat la cerca inicial des del caixetí, afegint que el camp autor contingui *vives*, i com la hem tornat a refinar des-seleccionant tots els repositoris menys el *DUGiFonsEspecials*.

DUGi
Portal de cerca conjunta als documents digitals de la UdG

Cercar:

Llistar Títols
S'han trobat 12 ítems
Llistant ítems des de 0 a 12:

Format	Data	Títol	Autor
	19 març 1954	Carta de 1954-03-19 a José Ferrater Mora des de Barcelona (Espanya)	Vicens i Vives, Jaume
	23 abril 1954	Carta de 1954-04-23 a José Ferrater Mora des de Barcelona (Espanya)	Vicens i Vives, Jaume
	2 juny 1954	Carta de 1954-06-02 a José Ferrater Mora des de Barcelona (Espanya)	Vicens i Vives, Jaume
	21 octubre 1954	Carta de 1954-10-21 a José Ferrater Mora des de Barcelona (Espanya)	Vicens i Vives, Jaume

CARTA DE 1954-03-19 A JOSÉ FERRATER DE BARCELONA (ESPANYA)

Resum: Vicens Vives escriu a Ferrater per mantenir contacte amb de la situació que viu Catalunya i els Catalans en aque història i poder parlar dels seus llibres

Pàgines: 2 p.

Format: application/pdf

Altres identificadors: ID1_5496

Accés al document: <http://hdl.handle.net/10256.2/7149>

Llenguatge: cat

Editor: Universitat de Girona. Càtedra Ferrater Mora

Drets: Tots els drets reservats

Títol: Carta de 1954-03-19 a José Ferrater Mora des de Barcelona

Tipus: letter

Repositori: [DUGiFonsEspecials](#)

La captura següent mostra com hem refinat la cerca escollint l'autor *Thió, Alfons Maria*.

Cercar:

Llistar Títols
S'ha trobat un ítem:

Format	Data	Títol	Autor
	24 febrer 1948	Carta de 1948-02-24 d'Alfons Maria Thió a Jaume Vicens Vives des de [València]	Thió, Alfons Maria

Matèries

Autors

Thió, Alfons Maria

- **Proporcionar un protocol de cerca prou potent per anar de la cerca més senzilla a la més complexa.**

S'ha aconseguit crear un protocol de cerca que permet fer cerques molt complexes utilitzant els operadors que proporciona SOLR.

Les següents captures de pantalla pertanyen a l'ajuda de cerca i pretenen mostrar lo complexa o simple que pot ser la cerca al DUGi.

La següent captura mostra els operadors bàsics per a la cerca complexa. També es poden observar els camps per a un usuari anglès, a l'exemple que es comenta a sota.

Advanced Search: Operators

- *: wildcard for more than one character
 - ex: author:terri*
 - ex: isbn:*
If we were searching for items with isbn
- "": It's used to search an exact text (with spaces). It's neither case-sensitive nor sensitive to accents.
 - ex: "Els Materials de construcció de la bòbila romana d'Ermedàs"
- AND: "and operator"
 - ex: esquerra AND fractals
It performs the same search as: esquerra fractals
- OR: "or operator"
 - ex: esquerra OR fractals
- ! / NOT: "not operator"
 - ex: esquerra !fractals
 - ex: esquerra NOT fractals
 - ex: !isbn:*
If we were looking for items without isbn
- (): It's used to group subsearches
 - ex: (title:catleg OR title:web) AND author:esquerra
 - ex: title:(catleg OR web) AND author:esquerra

La següent captura mostra els operadors complexos que proporciona el SOLR.

Cerca Avançada: Operadors Complexos

- [x TO y]: Operador de rang incliusiu. Serveix per a buscar entre rangs de dates o de lletres incloent-hi els valors llindar.
 - ex: titol:[ava TO aval]
 - ex: data:[1800 TO 1880]
 - ex: data:[* TO 1900]
 - ex: data:[1900 TO *]
- {x TO y}: Operador de rang exclusiu. Serveix per a buscar entre rangs de dates o de lletres excloent-hi els valors llindar.
 - ex: titol:{ava TO aval}
 - ex: data:{1800 TO 1880}
 - ex: data:{* TO 1900}
 - ex: data:{1900 TO *}
- ~X: Cerca entre semblants (X lletres de diferència)
 - ex: esqueera~0.87
També retornaria "esquerra" ja que li demanem un 87% de semblança (pot fallar 1 lletra de 8)
 - ex: esqueera~
També retornaria "esquerra", "esquerra", "esquerres"...
- ~X: Cerca de proximitat a x paraules
 - ex: "matematica fractal"~10
Estaria buscant frases que continguin "matemàtica" i "fractal" separades un màxim de 10 paraules

I s'ha definit els camps que es poden utilitzar per cercar. Aquests camps són els mateixos que s'ha definit al SOLR, amb l'addició de noms de camp en l'idioma que s'estigui utilitzant en aquell moment.

La següent captura mostra els camps definits que ens permeten fer la cerca avançada.

Cerca Avançada: Camps

La cerca avançada és la mateixa que la simple, però afegint-hi camps i operadors.

Els camps que podem fer servir són els següents (**sense accents**):

1. identificador o `dc.identifier.uri`
2. issn o `dc.identifier.issn`
3. isbn o `dc.identifier.isbn`
4. `dc.contributor.other`
5. director o `dc.contributor`
6. autor o `dc.creator`
7. títol o `dc.title`
8. matèria o `dc.subject`
9. resum o `dc.description.abstract`
10. editor o `dc.publisher`
11. format o `dc.format.mimetype`
12. data o `dc.date`
13. drets o `dc.rights`
14. tipus o `dc.type`
15. llenguatge o `dc.language`
16. repositori

Utilitzem ":" per separar el camp i el text a cercar.

- **ex:** `titol:estudi`
Si busquéssim ítems que continguin "estudi" al títol

És a dir, si estic utilitzant el català, podré cercar per *dc.title* o per *titol* per fer una cerca per títol. Per contra, si estic consultant el DUGi en anglès, puc fer la cerca utilitzant el camp *title*, i el DUGi em retornarà els mateixos resultats.

Ens ha semblat interessant proporcionar aquesta possibilitat, per aconseguir un sistema de cerca més intuïtiu.

A més a més, al haver definit quins caps es copien al camp de cerca per defecte, el sistema permet cercar als camps més significatius sense haver d'especificar-ho. Aquesta seria la cerca més senzilla possible.

Cerca Simple

La cerca simple no és sensible ni a accents ni a majúscules/minúscules i afecta els següents camps:

- `dc.identifier.uri`
- `dc.title`
- `dc.description.abstract`
- `dc.publisher`
- `dc.creator`
- `dc.contributor`
- `dc.contributor.other`
- `dc.subject`

L'espai entre paraules actua com a operador "T".

- **ex:** Terricabras
- **ex:** biblioteca filosofia

- **Simplificar al màxim la interfície per aconseguir una alta usabilitat.**

S'ha intentat simplificar la interfície. Només hi ha el necessari per cercar ítems.

De totes maneres, la opció de seleccionar i des-seleccionar repositoris no ha tingut gaire acceptació. Hem rebut comentaris d'usuaris que no n'havien entès l'ús. És molt important que la interfície s'entengui sola encara que es proporcioni un manual d'ús i inclús un vídeo d'ús. La majoria d'usuaris només volen trobar el que busquen i perdre-hi el mínim de temps possible.

- **Donar la màxima visibilitat als autors, matèries i documents.**

Els llistats d'autors i matèries, així com els núvols d'aquests, permeten accedir als ítems d'un autor o d'una matèria ràpidament. A més a més, els núvols, al ser aleatoris, proporcionen als cercadors ítems diferents enllaçats al nom d'autor i a la matèria. D'aquesta manera, es dona molta visibilitat als autors i matèries.

Tot i que s'havia formulat com a requeriment, per problemes de temps, no s'ha pogut donar aquesta visibilitat als documents. Al ser aquests vídeos, àudios i documents de text, es volia proporcionar una interfície per veure'ls directament des del DUGi, però no s'ha pogut acabar.

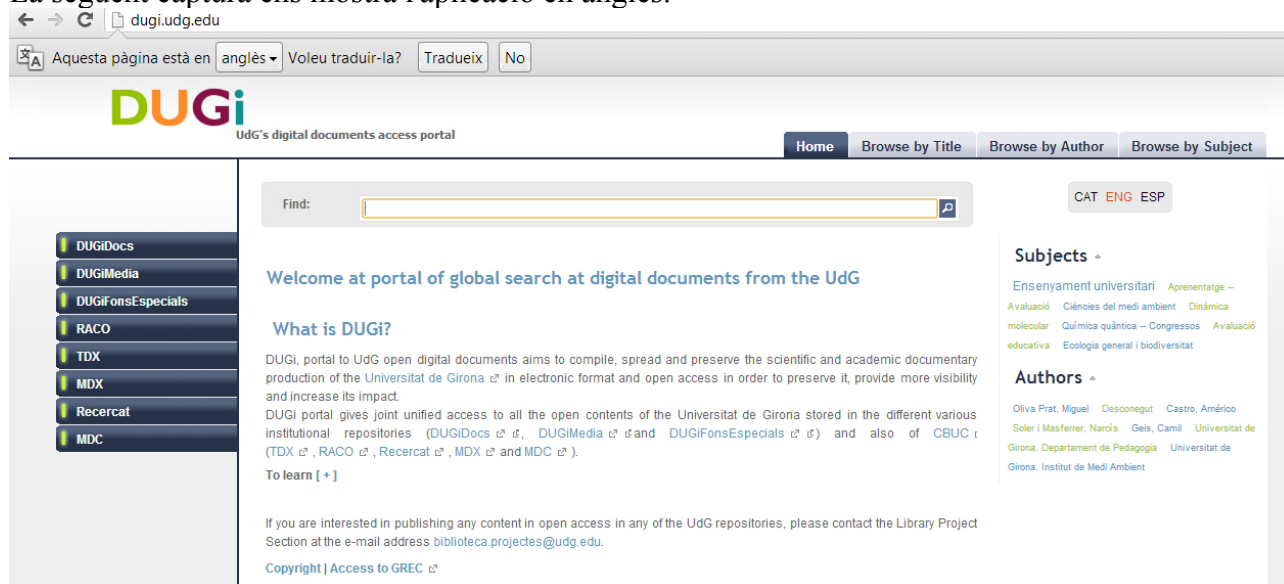
- **Implementar un sistema de memòria cau per fer les cerques més ràpides.**

El sistema de memòria cau explicat al punt 9 funciona satisfactòriament. De totes maneres, les cerques de SOLR són realment ràpides. Però la velocitat de càrrega disminueix en el cas dels llistats d'autors i matèries, ja que com també hem explicat al punt 9, ens cal recuperar la llista entera del SOLR, i ordenar-la nosaltres tenint en compte els accents.

- **Ha de tenir una interfície multilingüe. Com a mínim ha de donar suport per als idiomes català, espanyol i anglès.**

El sistema de multi-llenguatge funciona correctament. És el que es necessitava.

La següent captura ens mostra l'aplicació en anglès.



- **Ha de ser segur.**

Hem aplicat tots els nostres coneixements sobre seguretat i estem contents del resultat. Pot ser que se'ns hagi escapat alguna cosa, però creiem que és un sistema segur sempre i quan mantinguem els sistemes al dia d'actualitzacions i estem alerta als avisos de seguretat.

- **Ha de consumir pocs recursos.**

El DUGi està format per un servidor web **apache2** i un **SOLR** corrent sobre **jetty**, que és el

servidor de servlets menys pesat. Es podria haver utilitzat un servidor menys pesat que l'apache2, però aquest ja dona molts bons resultats amb el nombre de clients que tenim.

La següent captura mostra el consum de memòria del DUGi.

```
$ free -m
```

	total	used	free	shared	buffers	cached
Mem:	4054	1794	2259	0	375	977
-/+ buffers/cache:		441	3613			
Swap:	3812	0	3812			

En aquesta altra podem veure que el sistema no gasta gens de cpu en aquest moment.

```
$ top
```

top - 21:18:35 up 55 days, 5:28, 1 user, load average: 0.00, 0.00, 0.00
Tasks: 61 total, 2 running, 59 sleeping, 0 stopped, 0 zombie
Cpu(s): 1.0%us, 0.0%sy, 0.0%ni, 98.3%id, 0.0%wa, 0.0%hi, 0.7%si, 0.0%st
Mem: 4152236k total, 1838580k used, 2313656k free, 384760k buffers
Swap: 3903784k total, 0k used, 3903784k free, 1001128k cached

PID	USER	PR	NI	VIRT	RES	SHR	S	%CPU	%MEM	TIME+	COMMAND
26401	dugi	20	0	29724	8944	3660	S	0.3	0.2	0:04.50	apache2
26555	dugi	20	0	29564	8764	3640	S	0.3	0.2	0:02.80	apache2
26609	dugi	20	0	29552	8748	3636	S	0.3	0.2	0:01.65	apache2
26721	dugi	20	0	29948	8664	3608	S	0.3	0.2	0:00.10	apache2
1	root	20	0	2844	1692	548	S	0.0	0.0	0:01.48	init

- **Ha d'estar format de mòduls reutilitzables.**

Aquest requisit s'ha pogut complir amb el cas del sistema de recoll·lecció, que s'utilitza també per a recoll·lectar repositoris des del panell d'administració.

11.2 Desviacions del plantejament inicial

Com comentàvem anteriorment, ens hem hagut de desviar una mica del full de ruta descrit al punt 4. El motiu ha sigut un canvi de prioritats a les meves tasques a la Biblioteca.

Al final el temps dedicat a la implementació ha sigut només de 5 setmanes, i per complir amb els requisits del sistema de refinament de cerca amb facetes als camps d'autors, matèries, llenguatges, tipus de document i dates, i l'accés directe als documents, segurament haguéssim necessitat 10 setmanes en compte de les 7 assignades en primera instància.

11.3 Crítica dels resultats

En general estem força satisfets amb els resultats del projecte. De totes maneres, hem de ser crítics amb certs aspectes dels que no acabem d'estar del tot contents.

El primer aspecte és el de la usabilitat. Pensem que la idea de seleccionar i des-seleccionar repositoris era interessant, però no vam pensar que els usuaris aliens a la Biblioteca no coneixen els repositoris de la UdG ni les seves tipologies.

Hagués sigut molt millor idea haver muntat aquest filtre per tipus de document.

L'altre punt del que no acabem d'estar satisfets és el sistema de filtrar els resultats. El sistema

implementat embruta massa la caixa de cerca, i creiem que seria molt millor separar la cerca dels filtres.

A més a més, ens ha faltat el sistema de filtre per facetes, que és molt més intuïtiu, i no requereix estudiar totes les opcions a l'ajuda per poder filtrar l'excés de resultats.

També creiem que hem comés un error amb el camp de data auxiliar. Tot i que permet ordenar els resultats per data, com a camp de tipus string, no ens permet filtrar per rangs de dates.

Si canviem el tipus de camp a data o a numèric (sense guionets), permetrem la possibilitat de filtrar per rangs de dates.

Tampoc acabem d'estar del tot satisfets amb el sistema d'estadístiques. Necessitem proporcionar més estadístiques diferents, com per exemple el recompte per ciutats i per països.

Finalment, al llistat d'ítems no permetem el canvi del nombre de resultats per pàgina, i limitar a només 20 resultats cada pàgina no ens acaba de convèncer.

11.4 Valoracions personals

Estem força satisfets del que hem après en aquest projecte.

Sobre el llenguatge PHP, he trobat i utilitzat moltes funcions per tractament d'arrays molt interessants que desconeixia completament. Així hem pogut apreciar la facilitat que ens dona el PHP per al tractament d'arrays i hashos, que per al PHP són el mateix.

Algunes d'aquestes funcions són: **array_walk**, **array_map**, **array_splice**, **array_slice**, **asort**, **ksort**, **array_chunk**, etc.

També hem agafat experiència amb el **mod_rewrite** d'apache2, que ens permet donar un acabat més professional a les aplicacions web, a més a més d'ajudar-nos a optimitzar les urls per als cercadors.

Estudiar les opcions que tenim en bases de dades NOSQL ens ha ajudat a agafar un punt de vista diferent del de les bases de dades convencionals, i a més a més agafar moltes idees per a futures aplicacions.

Ens hem divertit bastant fent aquest projecte. Sobretot el més interessant ha sigut buscar solucions pròpies als problemes que ens proposaven els requeriments. L'altra cosa que també ens ha agradat força és el d'anar buscant les expressions regulars que necessitàvem per detectar quines són les metadades d'origen.

En general el projecte ens ha fet venir ganes de fer més projectes amb aquestes tecnologies que hem estudiat. Sobretot la base de dades REDIS, que al estar estructurada en forma de hashos, ens pot donar moltes opcions.

12. Treball futur

Creiem que aquest és un projecte viu, i sempre hi haurà millores per aplicar. Els temps passen, i sorgeixen nous estils de posicionar la informació, noves tendències d'interfícies d'usuaris, nous sistemes de bases de dades, etc, i sobretot, més temps per anar millorant els sistemes existents.

Les millores que proposem són les següents:

- **Acabar d'implementar el sistema de facetes per filtrar els resultats i fer molt més amena la interfície.**

Aquesta és la principal millora que ens cal. Els usuaris ho agrairan, i així també ens posem una mica més al dia pel que fa a les tendències d'aquests tipus d'aplicacions.

També s'ha d'aprofitar per arreglar el tema del tipus de camp de les dates que hem comentat en el punt 11.

Aquestes millores ja s'estan fent actualment, i comporten a més a més una actualització del disseny de la interfície.

En la següent captura podem veure un prototip de la nova versió del DUGi sense els botons de repositoris, i amb uns camins d'Ariadna que ens permeten treure les facetes ja seleccionades. A més a més, podem observar també la barra de paginació amb l'ordenació i la selecció del nombre de resultats per pàgina.

The screenshot displays the DUGi search portal. At the top, there's a navigation bar with links: Inici, Llistat de Títols, Llistat d'Autors, and Llistat de Matèries. A search bar is present with the placeholder 'Cercar al DUGi...'. To the right, there's a language selector with buttons for CAT, ENG, and ESP.

The main content area is titled 'Llistat d'Ítems'. It shows a list of search results, each with a title, author, date, and a thumbnail image. The results are filtered by 'letter (type)'. The list includes items like 'Carta de 1951-10-27 a José Ferrater Mora des ...', 'Carta de 1948-08-27 a José Ferrater Mora des ...', 'Carta de 1953-03-10 a José Ferrater Mora des ...', 'Carta de 1950-12-30 a José Ferrater Mora des ...', 'Carta de 1984-06-07 a José R. Echeverría des ...', 'Carta de 1948-11-11 a José Ferrater Mora des ...', 'Carta de 1949-08-15 a José Ferrater Mora des ...', and 'Carta de 1955-05-25 a José Ferrater Mora des ...'.

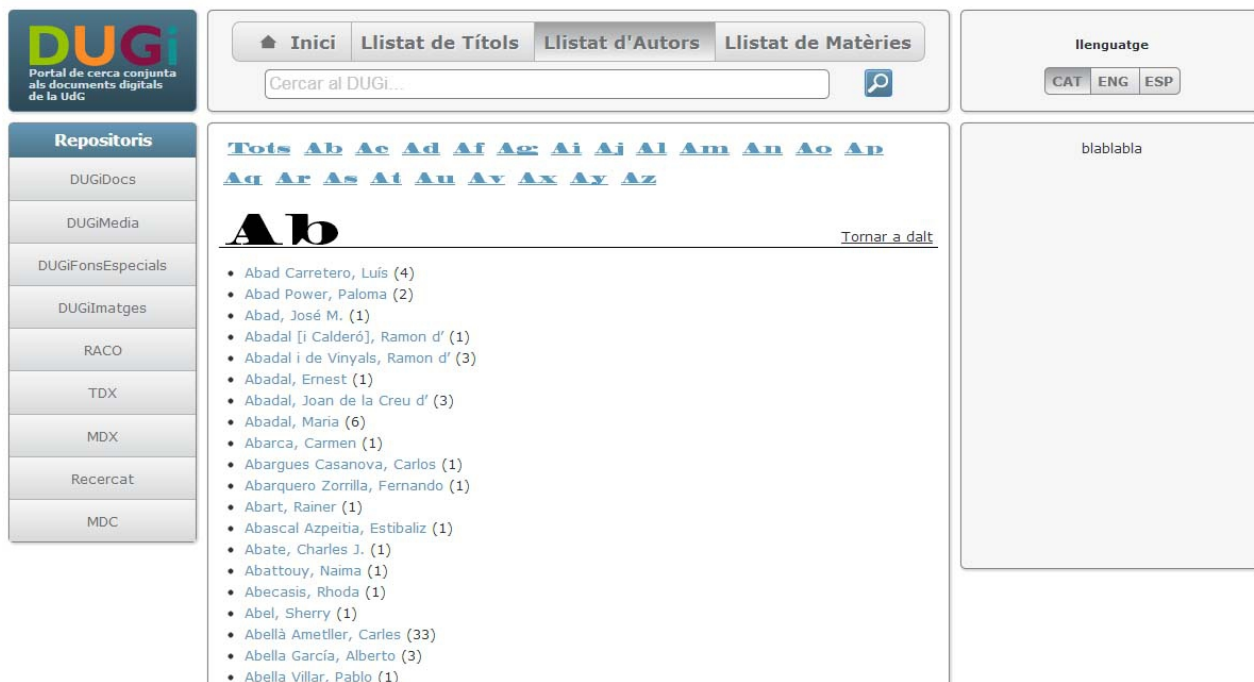
On the right side, there's a sidebar titled 'Refinar la cerca' (Refine the search). It contains several filter sections: 'type' (with 'letter' selected), 'author' (listing authors like Ferrater Mora, José, Ferrater Mora, Mercè, etc.), 'subject', 'lang' (listing languages like spa, cat, eng, etc.), and 'date-rang' (listing date ranges like 1900-01-01|1918-03-20, etc.).

- **Modificar el sistema dels llistats llargs com els d'autors i matèries.**

Actualment moltes aplicacions web permeten anar carregant resultats progressivament quan s'arriba a un cert punt de la pàgina. Creiem que pot ser interessant per temes de rendiment.

Aquestes modificacions també s'estan portant a terme actualment.

Aquesta altra captura es mostra un prototip del sistema d'scroll infinit per al llistat d'autors. S'ha cregut necessari afegir un enllaç per tornar a dalt de tot de la pàgina.



- **Nou sistema d'estadístiques**

Seria interessant canviar el sistema d'emmagatzematge del sistema d'estadístiques. La idea seria utilitzar el sistema de base de dades NOSQL Redis, que proporciona un sistema d'emmagatzematge de dades en forma de hashos, així aprofitaríem la mateixa estructura que ja tenim, però el sistema d'emmagatzematge seria més seriós.

També podríem aprofitar el Redis per guardar les dades de les autoritats conflictives pendents.

L'altra modificació de les estadístiques és una modificació de disseny i d'implementació per permetre mostrar més estadístiques interessants, com per exemple recomptes per ciutats i països.

13. Bibliografia

Apache Solr. (2012?). Solr features. Recuperat 07/01/2011, des de <http://lucene.apache.org/solr/features.html>

Arias, M., & Gavalda, R. (2012). [Ir] information retrieval. Recuperat 02/02/2012, des de <http://www-pagines.fib.upc.es/~ri/>

ASP FAQ. should I index my database table(s), and if so, how? (2003). Recuperat 15/07/2012, des de <http://databases.aspfq.com/database/should-i-index-my-database-table-s-and-if-so-how.html>

Best open source. Lucene vs solr. (2010?). Recuperat 23/04/2012, des de <http://www.findbestopensource.com/article-detail/lucene-vs-solr>

Best open source. Search engines. (2013?). Recuperat 03/05/2011, des de <http://www.findbestopensource.com/tagged/searchengine>

Cassandra wiki. (2013). Recuperat 03/05/2012, des de <http://wiki.apache.org/cassandra/>

Castillo, C. (2008). Sistemas de información II. tema 9 bases de datos de texto. http://www.tejedoresdelweb.com/wiki/images/1/14/Basesdatos_teo9_indices_texto.pdf

Cuesta Morales, P. (2004). Tomo I, apéndices técnicos. Apéndice A, indexación de información, Lucene. Motor de búsqueda para un SRI con agrupamiento. Vigo: Universidade de Vigo. <http://trevinca.ei.uvigo.es/~pcuesta/sm/practicas/Lucene.pdf>

Elasticsearch. (2013). Recuperat 12/11/2012, des de <http://www.elasticsearch.org/>

ElasticSearch : introduction and quick startup. (201-?). Recuperat 28/10/2012, des de <http://www.slideshare.net/medcl/elastic-search-quick-intro>

Esgaib, O., & Merín, M. (2012). Aplicando UWE : un proyecto de ejemplo. Recuperat 15/07/2012, des de <http://www.slideshare.net/millernegro/aplicando-uwe-un-ejemplo-didactico>

Faus, A. (2010). Escargot. Recuperat 16/01/2012, des de <https://github.com/angelf/escargot#readme>

Galiano, L. (2012). Metodología UWE aplicada a mi solución informática de mi proyecto : informe de la metodología aplicada en mi solución informática de mi proyecto. Inèdit (Instituto Universitario de Tecnología del Estado Bolívar).

<http://elproyectodeluisgaliano.blogspot.com.es/2012/11/metodologia-uwe-aplicada-mi-solucion.html>

Katta. How katta works. (2010). Recuperat 11/12/2011, des de <http://katta.sourceforge.net/documentation/how-katta-works>

Krellenstein, M. (2009). Full text search engines vs. DBMS. Missatge postejat a <http://searchhub.org/2009/09/02/full-text-search-engines-vs-dbms/>

Larman, C., & Moros Valle, B. (2002). UML y patrones: una introducción al análisis y diseño orientado a objetos y al proceso unificado (2a ed.). Madrid etc.: Prentice Hall.

Lothian, N. (2009). Solr+cassandra. Missatge postejat a <http://nicklothian.com/blog/2009/10/27/solr-cassandra-solandra/>

Lucene. (2012?). Lucene features. Recuperat 07/01/2011, des de <http://lucene.apache.org/core/features.html>

Lucene : para que empezar de cero sea más fácil. Qué es Lucene? (2008). Missatge postejat a <http://lucene-es.blogspot.com.es/2008/11/qu-es-lucene.html>

Luciani, J. (2010). Lucandra/Solandra : a Cassandra-based Lucene backend. Missatge postejat a <http://blog.sematest.com/2010/02/09/lucandra-a-cassandra-based-lucene-backend/>

Luciani, J. (2013). Solandra. Recuperat 19/05/2012, des de <https://github.com/tjake/Solandra/wiki>

Ludwig-Maximilians-Universität München. (2012). UWE-UML-based web engineering. Recuperat 02/16, 2013, des de <http://uwe.pst.ifi.lmu.de/>

Paradigma. (2009?). Apache solr. Manuscrit inèdit. <http://seminario-apache-solr.googlecode.com/files/Seminario%20Apache%20Solr.pdf>

Pérez, M. (2011). Cinco soluciones basadas en Lucene para real time search. Missatge postejat a <http://brigomp.blogspot.com.es/2011/04/cinco-soluciones-basadas-en-lucene-para.html>

Redis. (2013). Recuperat 12/11/2012, des de <http://redis.io/>

Rodgers, G. (2010). Rubberband. Missatge postejat a <https://github.com/grantr/rubberband#readme>

Rowe, S. (2012). Lucene concepts and definitions. Recuperat 23/04/2012, des de <http://wiki.apache.org/lucene-java/ConceptsAndDefinitions>

Sonnek, R. (2011). Realtime search : Solr vs Elasticsearch. Missatge postejat a <http://blog.socialcast.com/realtime-search-solr-vs-elasticsearch/>

Sphinx : about. (2013). Recuperat 11/12/2011, des de <http://sphinxsearch.com/about/sphinx/>

Tejedores del web. Bases de datos. (2008). Recuperat 23/03/2011, des de http://www.tejedoresdelweb.com/w/Bases_de_datos

UTF-8 and unicode. (2013). Recuperat 11/12/2011, des de <http://www.utf-8.com/>

UTF-8 encoding table and unicode characters. (201-?). Recuperat 11/12/2011, des de <http://www.utf8-chartable.de/>

Utilice SOLR en bases de datos relacionales MySQL. Utilice SOLR para la tabla único producto. (2013). Missatge postejat a <http://techerald.com/page/utilice-solr-en-bases-de-datos-relacionales-mysql-utilice-solr-para-la-tabla-unico-producto.html>

Viquipèdia. ASCII. (2013). Recuperat 11/12/2011, des de <http://ca.wikipedia.org/wiki/ASCII>

Viquipèdia. index (base de dades). (2013). Recuperat 23/03/2011, des de http://ca.wikipedia.org/wiki/%C3%8Dndex_%28base_de_dades%29

Viquipèdia. UTF-8. (2013). Recuperat 11/12/2011, des de <http://ca.wikipedia.org/wiki/UTF-8>

Wikipedia. abstract factory. (2013). Recuperat 31/01/2013, des de [http://es.wikipedia.org/wiki/Abstract_Factory_\(patr%C3%B3n_de_dise%C3%B1o\)](http://es.wikipedia.org/wiki/Abstract_Factory_(patr%C3%B3n_de_dise%C3%B1o))

Wikipedia. base de dades relacional. (2013). Recuperat 11/12/2011, des de http://ca.wikipedia.org/wiki/Base_de_dades_relacional

Wikipedia. distancia de levenshtein. (2013). Recuperat 31/01/2013, des de http://es.wikipedia.org/wiki/Distancia_de_Levenshtein

Wikipedia. facade. (2013). Recuperat 31/01/2013, des de [http://es.wikipedia.org/wiki/Facade_\(patr%C3%B3n_de_dise%C3%B1o\)](http://es.wikipedia.org/wiki/Facade_(patr%C3%B3n_de_dise%C3%B1o))

Wikipedia. factory method (patrón de diseño). (2013). Recuperat 31/01/2013, des de [http://es.wikipedia.org/wiki/Factory_Method_\(patr%C3%B3n_de_dise%C3%B1o\)](http://es.wikipedia.org/wiki/Factory_Method_(patr%C3%B3n_de_dise%C3%B1o))

Wikipedia. GRASP. (2013). Recuperat 31/01/2013, des de <http://es.wikipedia.org/wiki/GRASP>

Wikipedia. lucene. (2013). Recuperat 02/02/2012, des de <http://en.wikipedia.org/wiki/Lucene>

Wikipedia. patrón de diseño. (2013). Recuperat 31/01/2013, des de http://es.wikipedia.org/wiki/Patr%C3%B3n_de_dise%C3%B1o

Wikipedia. singleton. (2013). Recuperat 31/01/2013, des de http://es.wikipedia.org/wiki/Patr%C3%B3n_de_dise%C3%B1o_Singleton

Wikipedia. UTF-8. (2013). Recuperat 11/12/2011, des de <http://en.wikipedia.org/wiki/UTF-8>

Yergeau, F. (2003). UTF-8, a transformation format of ISO 10646. Manuscrit inèdit. <http://www.ietf.org/rfc/rfc3629>

Yonik, S. (2009). SearchHub. faceted search with solr. Missatge postejat a <http://searchhub.org/2009/09/02/faceted-search-with-solr/>

14. Annexos

14.1 Annex Verbs OAI-PHM

Aquest ANNEX ha estat referenciat en el punt 7.3.2 de la memòria.

Identify

Aquest verb ens permet obtenir informació del servidor. Podem saber qui n'és l'administrador, quin tractament dels documents eliminats tenim, quin format de data utilitza, etc

Vegem-ne un exemple, en aquest cas del DUGiMedia:

Fent la petició a la url: <http://diobma.udg.edu/dspace-oai/request?verb=Identify>

Ens respon:

```
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2012-11-07T16:07:37Z</responseDate>
  <request verb="Identify">http://diobma.udg.edu/dspace-oai/request</request>
  <Identify>
    <repositoryName>DUGiMedia</repositoryName>
    <baseURL>http://diobma.udg.edu/dspace-oai/request</baseURL>
    <protocolVersion>2.0</protocolVersion>
    <adminEmail>oriol.olive@udg.edu</adminEmail>
    <earliestDatestamp>2001-01-01T00:00:00Z</earliestDatestamp>
    <deletedRecord>persistent</deletedRecord>
    <granularity>YYYY-MM-DDThh:mm:ssZ</granularity>
    <compression>gzip</compression>
    <compression>deflate</compression>
    <description>
      <toolkit xmlns="http://oai.dlib.vt.edu/OAI/metadata/toolkit"
xsi:schemaLocation="http://oai.dlib.vt.edu/OAI/metadata/toolkit
http://oai.dlib.vt.edu/OAI/metadata/toolkit.xsd">
        <title>OCLC's OAICat Repository Framework</title>
        <author>
          <name>Jeffrey A. Young</name>
          <email>jyoung@oclc.org</email>
          <institution>OCLC</institution>
        </author>
        <version>1.5.48</version>
        <toolkitIcon>http://alcme.oclc.org/oaicat/oaicat_icon.gif</toolkitIcon>
        <URL>http://www.oclc.org/research/software/oai/cat.shtm</URL>
      </toolkit>
    </description>
  </Identify>
</OAI-PMH>
```

ListMetadataFormats

Ens permet preguntar, com el seu nom indica, quins són els formats de metadata suportats per el

servidor d'OAI-PMH. Ha de quedar clar que aquests formats no tenen perque ser els mateixos que suporta el repositori en si.

Fent la petició: <http://diobma.udg.edu/dspace-oai/request?verb=ListMetadataFormats>

El DUGiMedia ens respon:

```
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2012-11-07T16:42:35Z</responseDate>
  <request verb="ListMetadataFormats">http://diobma.udg.edu/dspace-
oai/request</request>
  <ListMetadataFormats>
    <metadataFormat>
      <metadataPrefix>oai_dc</metadataPrefix>
      <schema>http://www.openarchives.org/OAI/2.0/oai_dc.xsd</schema>
      <metadataNamespace>http://www.openarchives.org/OAI/2.0/oai_dc/</metadataNamespace>
    </metadataFormat>
    <metadataFormat>
      <metadataPrefix>rdf</metadataPrefix>
      <schema>http://www.openarchives.org/OAI/2.0/rdf.xsd</schema>
      <metadataNamespace>http://www.openarchives.org/OAI/2.0/rdf/</metadataNamespace>
    </metadataFormat>
    <metadataFormat>
      <metadataPrefix>ore</metadataPrefix>
      <schema>
http://tweety.lanl.gov/public/schemas/2008-06/atom-tron.sch
      </schema>
      <metadataNamespace>http://www.w3.org/2005/Atom</metadataNamespace>
    </metadataFormat>
    <metadataFormat>
      <metadataPrefix>ese</metadataPrefix>
      <schema>http://purl.org/dc/elements/1.1/</schema>
      <metadataNamespace>http://www.europeana.eu/schemas/ese/</metadataNamespace>
    </metadataFormat>
    <metadataFormat>
      <metadataPrefix>mets</metadataPrefix>
      <schema>http://www.loc.gov/standards/mets/mets.xsd</schema>
      <metadataNamespace>http://www.loc.gov/METS/</metadataNamespace>
    </metadataFormat>
  </ListMetadataFormats>
</OAI-PMH>
```

ListSets

Per tal de fer una recol·lecció més selectiva, disposem de sets o col·leccions. Així doncs, aquest verb ens permetrà recuperar-ne una llista dels disponibles al servidor.

En el nostre cas, en els repositoris consorciats (TDX, RACO i MDC), sempre recol·lectarem el set o sets que pertanyen a la Universitat de Girona.

La petició: <http://diobma.udg.edu/dspace-oai/request?verb=ListSets>

Contesta:

```
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2012-11-07T16:59:28Z</responseDate>
  <request verb="ListSets">http://diobma.udg.edu/dspace-oai/request</request>
  <ListSets>
    <set>
      <setSpec>hdl_10256.1_917</setSpec>
      <setName>2n Workshop BiblioUdG</setName>
    </set>
    <set>
      <setSpec>hdl_10256.1_1499</setSpec>
      <setName>3r Workshop BiblioUdG</setName>
    </set>
    <set>
      <setSpec>hdl_10256.1_1914</setSpec>
      <setName>4t Workshop BiblioUdG</setName>
    </set>
    <set>
      <setSpec>hdl_10256.1_610</setSpec>
      <setName>Univest</setName>
    </set>
    ...
    <set>
      <setSpec>MD1</setSpec>
      <setName>Set per mdx metodologia</setName>
    </set>
    <set>
      <setSpec>MD2</setSpec>
      <setName>Set per mdx materials docents</setName>
    </set>
    <set>
      <setSpec>RC1</setSpec>
      <setName>Set per recerca docs recerca</setName>
    </set>
    <set>
      <setSpec>RC2</setSpec>
      <setName>
        Set per mdx treballs master, final carrera i doctorats
      </setName>
    </set>
  </ListSets>
</OAI-PMH>
```

En aquest cas la llista és molt llarga ja que el DUGiMedia té moltes col·leccions que el servidor d'OAI-PMH ha convertit en sets.

ListIdentifiers

Aquest verb ens permet disposar d'una llista de tots els ítems disponibles per servir amb el protocol OAI-PMH. Ha d'estar acompanyat de l'atribut **metadataPrefix**, que fa referència al format de metadades esperat. I pot estar acompanyat dels atributs **sets**, **until** i **from**, que ens permetran seleccionar els ítems que hagin estat modificats a partir (until) i des de (from) una cerca data.

ListIndetifiers només ens retorna les capçaleres dels ítems. Així doncs, podem saber quin és el seu

identificador, quan va ser creat o modificat per última vegada, i a quin set pertany.

Amb la petició: http://diobma.udg.edu/dspace-oai/request?verb=ListIdentifiers&metadataPrefix=oai_dc&set=hdl_10256.1_3&from=2011-10-10T00:00:00Z&until=2011-10-20T00:00:00Z

Respon:

```
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2012-11-07T17:23:39Z</responseDate>
  <request set="hdl_10256.1_3" from="2011-10-10T00:00:00Z" metadataPrefix="oai_dc"
until="2011-10-20T00:00:00Z" verb="ListIdentifiers">http://diobma.udg.edu/dspace-
oai/request</request>
  <ListIdentifiers>
    <header>
      <identifier>oai:diobma.udg.edu:10256.1/6</identifier>
      <datestamp>2011-10-17T10:25:53Z</datestamp>
      <setSpec>hdl_10256.1_3</setSpec>
    </header>
    <header>
      <identifier>oai:diobma.udg.edu:10256.1/7</identifier>
      <datestamp>2011-10-17T10:29:22Z</datestamp>
      <setSpec>hdl_10256.1_3</setSpec>
    </header>
    <header>
      <identifier>oai:diobma.udg.edu:10256.1/8</identifier>
      <datestamp>2011-10-17T07:56:27Z</datestamp>
      <setSpec>hdl_10256.1_3</setSpec>
    </header>
    <header>
      <identifier>oai:diobma.udg.edu:10256.1/9</identifier>
      <datestamp>2011-10-17T11:41:18Z</datestamp>
      <setSpec>hdl_10256.1_3</setSpec>
    </header>
    <header>
      <identifier>oai:diobma.udg.edu:10256.1/13</identifier>
      <datestamp>2011-10-17T10:21:51Z</datestamp>
      <setSpec>hdl_10256.1_3</setSpec>
    </header>
    <header>
      <identifier>oai:diobma.udg.edu:10256.1/14</identifier>
      <datestamp>2011-10-17T10:20:33Z</datestamp>
      <setSpec>hdl_10256.1_3</setSpec>
    </header>
    <header>
      <identifier>oai:diobma.udg.edu:10256.1/628</identifier>
      <datestamp>2011-10-17T09:13:31Z</datestamp>
      <setSpec>hdl_10256.1_3</setSpec>
    </header>
    <header>
      <identifier>oai:diobma.udg.edu:10256.1/629</identifier>
      <datestamp>2011-10-17T11:22:07Z</datestamp>
      <setSpec>hdl_10256.1_3</setSpec>
    </header>
    <header>
      <identifier>oai:diobma.udg.edu:10256.1/1131</identifier>
```

```

<timestamp>2011-10-17T08:32:08Z</timestamp>
<setSpec>hdl_10256.1_3</setSpec>
</header>
<header>
<identifier>oai:diobma.udg.edu:10256.1/1220</identifier>
<timestamp>2011-10-17T09:50:04Z</timestamp>
<setSpec>hdl_10256.1_3</setSpec>
</header>
<header>
<identifier>oai:diobma.udg.edu:10256.1/1222</identifier>
<timestamp>2011-10-17T09:36:51Z</timestamp>
<setSpec>hdl_10256.1_3</setSpec>
</header>
<header>
<identifier>oai:diobma.udg.edu:10256.1/1290</identifier>
<timestamp>2011-10-17T10:05:21Z</timestamp>
<setSpec>hdl_10256.1_3</setSpec>
</header>
<header>
<identifier>oai:diobma.udg.edu:10256.1/1364</identifier>
<timestamp>2011-10-17T06:59:04Z</timestamp>
<setSpec>hdl_10256.1_3</setSpec>
</header>
<header>
<identifier>oai:diobma.udg.edu:10256.1/1365</identifier>
<timestamp>2011-10-17T07:14:26Z</timestamp>
<setSpec>hdl_10256.1_3</setSpec>
</header>
</ListIdentifiers>
</OAI-PMH>

```

Podem comprovar, que no hi ha cap data més gran o igual que 2011-10-20, ni anterior o igual a 2011-10-10. I evidentment, tots els ítems pertanyen al set hdl_10256.1_3.

ListRecords

Aquest verb ens permet recuperar una llista d'ítems. A diferència del verb anterior, aquest sí que conté les metadades associades a cada ítem. Al contenir les metadades, la resposta a aquesta petició pot ser bastant pesada, és per això que el servidor d'OAI-PMH en serveix la llista en grups de 100 en 100. Per aquesta raó, en aquest verb es permet l'ús del `resumptionToken`, que permetrà al servidor saber a partir de quin ítem ha de seguir, i quins eren els criteris de la selecció.

Quan s'utilitza el `resumptionToken`, aquest ja conté les dades dels altres paràmetres, i per tant aquests no calen.

El `resumptionToken` té la següent forma:

[from]/[until]/[set]/[metadataPrefix]/NúmÚltimDoc

Per exemple:

0001-01-01T00:00:00Z/9999-12-31T23:59:59Z//oai_dc/100

Aquest verb accepta els paràmetres **resumptionToken**, **set**, **from** i **until**, i és obligatori l'ús de **metadataPrefix**.

Amb pertició: http://diobma.udg.edu/dspace-oai/request?verb=ListRecords&metadataPrefix=oai_dc

Contesta:

```
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2012-11-07T17:32:39Z</responseDate>
  <request metadataPrefix="oai_dc" verb="ListRecords">http://diobma.udg.edu/dspace-
oai/request</request>
  <ListRecords>
    <record>
      <header status="deleted">
        <identifier>oai:diobma.udg.edu:10256.1/4</identifier>
        <datestamp>2011-09-09T12:39:07Z</datestamp>
        <setSpec>hdl_10256.1_3</setSpec>
      </header>
    </record>
    <record>
      <header>
        <identifier>oai:diobma.udg.edu:10256.1/5</identifier>
        <datestamp>2011-10-21T06:23:53Z</datestamp>
        <setSpec>hdl_10256.1_3</setSpec>
      </header>
      <metadata>
        <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
          <dc:title>
            Ludwig Wittgenstein: commemoració del passat, examen del present i continuïtat de
futur
          </dc:title>
          <dc:creator>Terricabras, Josep Maria</dc:creator>
          <dc:creator>Anscombe, G. E. M.</dc:creator>
          <dc:creator>Puigbert Busquets, Joan</dc:creator>
          <dc:subject>Wittgenstein, Ludwig, 1889-1951 -- Congressos</dc:subject>
          <dc:subject>Filosofia -- Congressos</dc:subject>
          <dc:description>
            Documental centrat en el filòsof Ludwig Wittgenstein en motiu del Simposi de
commemoració del seu naixement. Format, d'una banda, per un reportatge on es narra la
vida i obra del filòsof. D'altra banda, d'un resum sobre el que va significar la
celebració del simposi, com es va desenvolupar i les impressions que s'han varen
extreure
          </dc:description>
          <dc:description>
            Traspasat de la videocinta original a format digital.
          </dc:description>
          <dc:publisher>
            Universitat de Girona. Càtedra Ferrater Mora de Pensament Contemporani
          </dc:publisher>
          <dc:date>1989</dc:date>
          <dc:date>2004-09-09T11:45:43Z</dc:date>
          <dc:type>movingImage</dc:type>
          <dc:format>138060272 bytes</dc:format>
          <dc:format>200286810 bytes</dc:format>
          <dc:format>video/x-ms-wmv</dc:format>
        </oai_dc:dc>
      </metadata>
    </record>
  </ListRecords>
</OAI-PMH>
```

```

<dc:format>audio/x-pn-realaudio</dc:format>
<dc:format>video/x-flv</dc:format>
<dc:identifier>http://hdl.handle.net/10256.1/5</dc:identifier>
<dc:language>cat</dc:language>
<dc:rights>Tots els drets reservats</dc:rights>
</oai_dc:dc>
</metadata>
</record>
...
<record>
<header>
<identifier>oai:diobma.udg.edu:10256.1/137</identifier>
<datestamp>2012-06-05T11:32:54Z</datestamp>
<setSpec>hdl_10256.1_83</setSpec>
</header>
<metadata>
<oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
<dc:title>Racionalidad y totalidad</dc:title>
<dc:creator>Holz, Hans Heinz</dc:creator>
<dc:subject>Holz, Hans Heinz -- Congressos</dc:subject>
<dc:description>
La Racionalitat i Totalitat. La inevitabilitat del Pensament especulatiu i l'inici
de la inacabable confrontació amb els sistemes de Leibniz i Hegel
</dc:description>
<dc:date>2001-06</dc:date>
<dc:type>lecture</dc:type>
<dc:format>237988859 bytes</dc:format>
<dc:format>375825755 bytes</dc:format>
<dc:format>audio/x-pn-realaudio</dc:format>
<dc:format>video/x-ms-wmv</dc:format>
<dc:format>video/x-flv</dc:format>
<dc:identifier>
Holz, H.H. 'Racionalidad y totalidad' a Seminari Hans Heinz Holz. Girona:
Universitat, 2001 [consulta: 18 setembre 2008]. Necessita Windows Media. Disponible a
Internet a: http://hdl.handle.net/10256.1/137
</dc:identifier>
<dc:identifier>http://hdl.handle.net/10256.1/137</dc:identifier>
<dc:language>spa</dc:language>
<dc:relation>Seminari Hans Heinz Holz</dc:relation>
<dc:rights>Tots els drets reservats</dc:rights>
</oai_dc:dc>
</metadata>
</record>
<resumptionToken expirationDate="2012-11-07T18:32:39Z">
0001-01-01T00:00:00Z/9999-12-31T23:59:59Z//oai_dc/100
</resumptionToken>
</ListRecords>
</OAI-PMH>

```

GetRecord

Aquest verb ens permet accedir a un sol ítem, sense haver de fer un recorregut per tots ells. Ha d'estar acompanyat dels atributs **metadataPrefix** i **identifier**. No tenen sentit els atributs de **set**, **from** i **until**.

Amb la petició: http://diobma.udg.edu/dspace-oai/request?verb=GetRecord&identifier=oai:diobma.udg.edu:10256.1/5&metadataPrefix=oai_dc

Contesta:

```
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2012-11-07T17:40:35Z</responseDate>
  <request metadataPrefix="oai_dc" verb="GetRecord"
identifier="oai:diobma.udg.edu:10256.1/5">http://diobma.udg.edu/dspace-
oai/request</request>
  <GetRecord>
  <record>
  <header>
  <identifier>oai:diobma.udg.edu:10256.1/5</identifier>
  <timestamp>2011-10-21T06:23:53Z</timestamp>
  <setSpec>hdl_10256.1_3</setSpec>
  </header>
  <metadata>
  <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
  <dc:title>
    Ludwig Wittgenstein: commemoració del passat, examen del present i continuïtat de
futur
  </dc:title>
  <dc:creator>Terricabras, Josep Maria</dc:creator>
  <dc:creator>Anscombe, G. E. M.</dc:creator>
  <dc:creator>Puigbert Busquets, Joan</dc:creator>
  <dc:subject>Wittgenstein, Ludwig, 1889-1951 -- Congressos</dc:subject>
  <dc:subject>Filosofia -- Congressos</dc:subject>
  <dc:description>
    Documental centrat en el filòsof Ludwig Wittgenstein en motiu del Simposi de
commemoració del seu naixement. Format, d'una banda, per un reportatge on es narra la
vida i obra del filòsof. D'altra banda, d'un resum sobre el que va significar la
celebració del simposi, com es va desenvolupar i les impressions que s'han varen
extreure
  </dc:description>
  <dc:description>
    Traspasat de la videocinta original a format digital.
  </dc:description>
  <dc:publisher>
    Universitat de Girona. Càtedra Ferrater Mora de Pensament Contemporani
  </dc:publisher>
  <dc:date>1989</dc:date>
  <dc:date>2004-09-09T11:45:43Z</dc:date>
  <dc:type>movingImage</dc:type>
  <dc:format>138060272 bytes</dc:format>
  <dc:format>200286810 bytes</dc:format>
  <dc:format>video/x-ms-wmv</dc:format>
  <dc:format>audio/x-pn-realaudio</dc:format>
  <dc:format>video/x-flv</dc:format>
  <dc:identifier>http://hdl.handle.net/10256.1/5</dc:identifier>
  <dc:language>cat</dc:language>
```

```
<dc:rights>Tots els drets reservats</dc:rights>
</oai_dc:dc>
</metadata>
</record>
</GetRecord>
</OAI-PMH>
```

14.2 Annex APACHE SOLR

Instal·lació i proves de funcionament d'Apache SOLR

Aquest ANNEX està referenciat al punt 7.6.2 de la memòria.

14.2.1. Pre-requisits

APACHE SOLR només requereix una versió de JAVA igual o superior a la 6 per funcionar. També és necessari tenir espai de disc per a l'índex.

14.2.2. Descarrega i instal·lació

El paquet de SOLR que es pot descarregar a: <http://lucene.apache.org/solr/mirrors-solr-latest-redirect.html?>

Un cop descarregat cal descomprimir-lo.

En entorn linux:

```
$ unzip -q solr-4.2.0.zip
```

En entorns windows:

Només cal usar qualsevol descompressor com 7-zip o winzip

La carpeta *example* conté un SOLR preparat per funcionar directament. A més a més conté uns ítems d'exemple per poder-lo provar sense haver de pensar una estructura de dades.

14.2.3. Paràmetres d'indexació del camps

Dins del directori *example/solr/conf* podem editar els fitxers *schema.xml* i *solrconfig.xml* entre altres, per especificar els camps i els tipus de camp el primer, i per la configuració general del servei el segon.

Si editem el fitxer *schema.xml*, podem definir els tipus de camps i els camps. Quan especifiquem els camps, el SOLR ens permet definir una sèrie d'opcions:

- **name**: El nom del camp.
- **type**: El tipus de camp.
- **indexed**: Si el camp s'indexa o no. De tipus booleà.

- **stored**: Si el camp es guarda o no. De tipus booleà.
- **multivalued**: si el camp és multi-avaluat. De tipus booleà.
- **required**: Si el camp és requerit per l'índex. De tipus booleà.

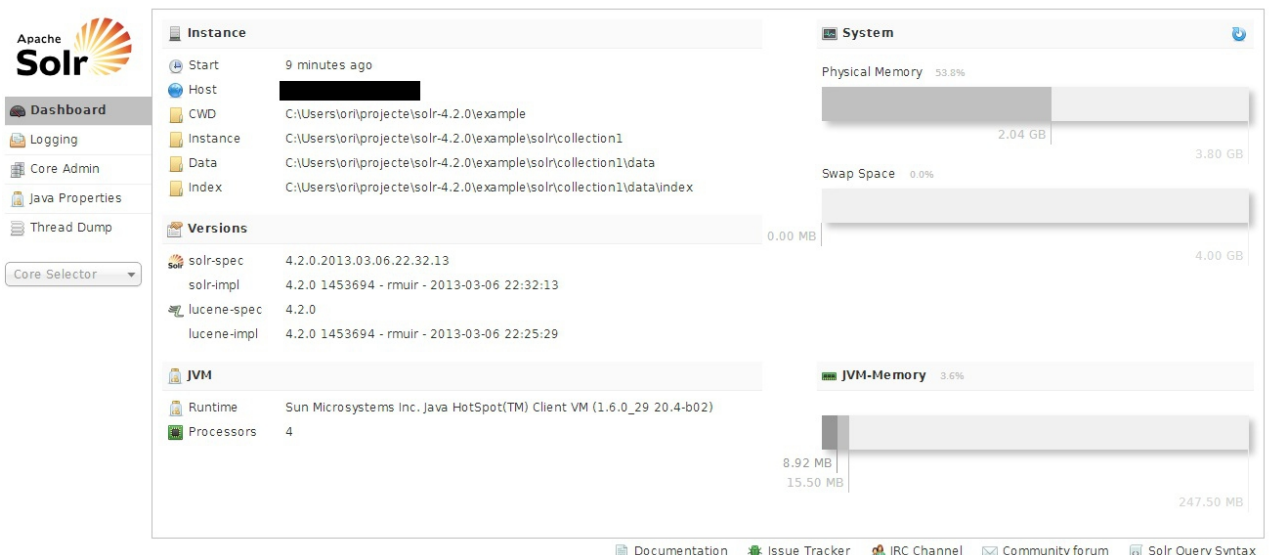
Els tipus de camps poden ser tants com la imaginació d'un. La diferència entra uns i altres radica en els filtres i els tokenitzadors que s'apliquen. Es poden fer servir els que ja aporta el SOLR, o crear-ne de nous en java.

14.2.4. Posada en funcionament

```
example$ java -jar start.jar
...
01-abr-2013 13:01:50 org.apache.solr.core.SolrCore execute
INFO: [collection1] webapp=null path=null params={event=firstSearcher&q=static+firstSearcher+warming+in+solrconfig.xml&distributed=false} hits=0 status=0 QTime=45
01-abr-2013 13:01:50 org.apache.solr.core.QuerySenderListener newSearcher
INFO: QuerySenderListener done.
01-abr-2013 13:01:50 org.apache.solr.handler.component.SpellCheckComponent$SpellCheckerListener newSearcher
INFO: Loading spell index for spellchecker: default
01-abr-2013 13:01:50 org.apache.solr.handler.component.SpellCheckComponent$SpellCheckerListener newSearcher
INFO: Loading spell index for spellchecker: wordbreak
01-abr-2013 13:01:50 org.apache.solr.core.SolrCore registerSearcher
INFO: [collection1] Registered new searcher Searcher@8ef455 main{StandardDirectoryReader(segments_1:1)}
01-abr-2013 13:01:50 org.apache.solr.core.CoreContainer registerCore
INFO: registering core: collection1
01-abr-2013 13:01:50 org.apache.solr.servlet.SolrDispatchFilter init
INFO: user.dir=C:\Users\ori\projecte\solr-4.2.0\example
01-abr-2013 13:01:50 org.apache.solr.servlet.SolrDispatchFilter init
INFO: SolrDispatchFilter.init() done
2013-04-01 13:01:50.992:INFO:oejs.AbstractConnector:Started SocketConnector@0.0.0.0:8983
```

Un cop executada l'ordre, es queda el SOLR escoltant per al port 8983. Hi podem accedir amb el navegador a l'adreça: <http://localhost:8983/solr/>.

Veiem el panell d'administració d'APACHE SOLR, que ens permet veure el rendiment del sistema i la demanada de memòria entre altres coses.



Podem fer una consulta a la nostra col·lecció de prova.

A la imatge següent podem veure que encara no tenim cap ítem al nostre índex. Per tant, no ha trobat cap resultat a la nostra consulta de `*:*`. On `*` és el comodí, i per tant estem demanant qualsevol cosa a qualsevol camp.

```

Request-Handler (qt)
/select
  - common
q
*:*
fq
sort
start, rows
0 10
fl
df
Raw Query Parameters
key1=val1&key2=val2
wt
xml
[indent]
[debugQuery]

http://localhost:8983/solr/collection1/select?q=%3A*&wt=xml&indent=true

<?xml version="1.0" encoding="UTF-8"?>
<response>
  <lst name="responseHeader">
    <int name="status">0</int>
    <int name="QTime">3</int>
    <lst name="params">
      <str name="indent">true</str>
      <str name="q">*:*</str>
      <str name="wt">xml</str>
    </lst>
  </lst>
  <result name="response" numFound="0" start="0">
  </result>
</response>
  
```

14.2.5. Indexar ítems

Per tal d'indexar ítems, el solr d'exemple inclou dues opcions senzilles: `post.jar` i `post.sh`.

En realitat per a indexar els ítems només hem d'enviar les dades en un dels formats suportats a l'adreça <http://localhost:8983/solr/update?> i fer un `commit` per tancar la transferència.

A la carpeta `exampledocs` hi trobem els fitxers que contenen els ítems.

Els XML tenen la següent estructura:

```
<add>
<doc>
  <field name="id">SOLR1000</field>
  <field name="name">Solr, the Enterprise Search Server</field>
  <field name="manu">Apache Software Foundation</field>
  <field name="cat">software</field>
  <field name="cat">search</field>
  <field name="features">Advanced Full-Text Search Capabilities using Lucene</field>
  <field name="features">Optimized for High Volume Web Traffic</field>
  <field name="features">Standards Based Open Interfaces - XML and HTTP</field>
  <field name="features">Comprehensive HTML Administration Interfaces</field>
  <field name="features">Scalability - Efficient Replication to other Solr Search
Servers</field>
  <field name="features">Flexible and Adaptable with XML configuration and
Schema</field>
  <field name="features">Good unicode support: h&#xE9;llo (hello with an accent over
the e)</field>
  <field name="price">0</field>
  <field name="popularity">10</field>
  <field name="inStock">true</field>
  <field name="incubationdate_dt">2006-01-17T00:00:00.000Z</field>
</doc>
</add>
```

Els CSV tenen la següent:

```
id,cat,name,price,inStock,author,series_t,sequence_i,genre_s
0553573403,book,A Game of Thrones,7.99,true,George R.R. Martin,"A Song of Ice and
Fire",1,fantasy
0553579908,book,A Clash of Kings,7.99,true,George R.R. Martin,"A Song of Ice and
Fire",2,fantasy
055357342X,book,A Storm of Swords,7.99,true,George R.R. Martin,"A Song of Ice and
Fire",3,fantasy
0553293354,book,Foundation,7.99,true,Isaac Asimov,Foundation Novels,1,scifi
0812521390,book,The Black Company,6.99,false,Glen Cook,The Chronicles of The Black
Company,1,fantasy
0812550706,book,Ender's Game,6.99,true,Orson Scott Card,Ender,1,scifi
0441385532,book,Jheræg,7.95,false,Steven Brust,Vlad Taltos,1,fantasy
0380014300,book,Nine Princes In Amber,6.99,true,Roger Zelazny,the Chronicles of
Amber,1,fantasy
0805080481,book,The Book of Three,5.99,true,Lloyd Alexander,The Chronicles of
Prydain,1,fantasy
080508049X,book,The Black Cauldron,5.99,true,Lloyd Alexander,The Chronicles of
Prydain,2,fantasy
```

I els JSON la següent:

```
[
{
  "id" : "978-0641723445",
  "cat" : ["book","hardcover"],
  "name" : "The Lightning Thief",
  "author" : "Rick Riordan",
  "series_t" : "Percy Jackson and the Olympians",
  "sequence_i" : 1,
  "genre_s" : "fantasy",
  "inStock" : true,
  "price" : 12.50,
```

```

    "pages_i" : 384
  }
,
{
  "id" : "978-1423103349",
  "cat" : ["book","paperback"],
  "name" : "The Sea of Monsters",
  "author" : "Rick Riordan",
  "series_t" : "Percy Jackson and the Olympians",
  "sequence_i" : 2,
  "genre_s" : "fantasy",
  "inStock" : true,
  "price" : 6.49,
  "pages_i" : 304
}
]

```

Per indexar els ítems d'exemple podem fer servir la següent comanda que ens serviria en qualsevol entorn.

En l'exemple següent hem indexat tots els fitxers de la carpeta exampledocs utilitzant la utilitat java que ens proporciona el SOLR.

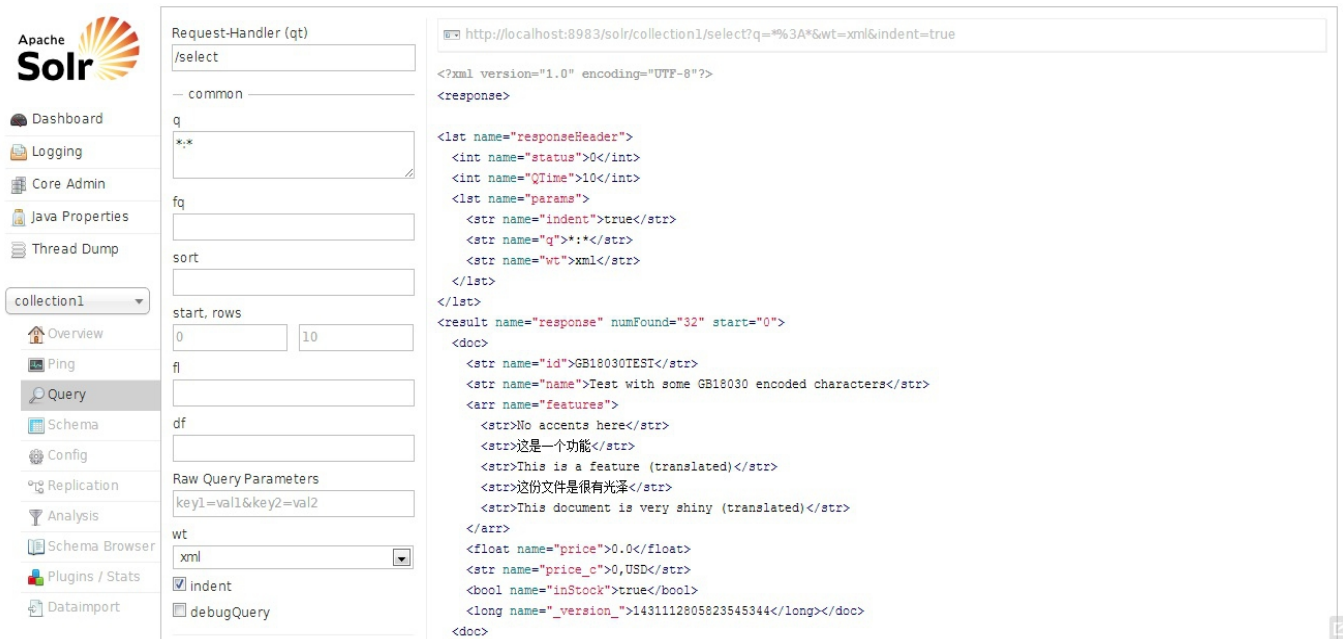
```

Example/exampledocs/$ java -jar post.jar *.xml

SimplePostTool version 1.5
Posting files to base url http://localhost:8983/solr/update using content-type application/xml..
POSTing file gb18030-example.xml
POSTing file hd.xml
POSTing file ipod_other.xml
POSTing file ipod_video.xml
POSTing file manufacturers.xml
POSTing file mem.xml
POSTing file money.xml
POSTing file monitor.xml
POSTing file monitor2.xml
POSTing file mp500.xml
POSTing file sd500.xml
POSTing file solr.xml
POSTing file utf8-example.xml
POSTing file vidcard.xml
14 files indexed.
COMMITting Solr index changes to http://localhost:8983/solr/update..
Time spent: 0:00:00.897

```


Ara ja podem tornar a consultar l'index i veurem que ja hi tenim uns quants ítems.



The screenshot shows the Apache Solr Admin UI. On the left is a sidebar with navigation links: Dashboard, Logging, Core Admin, Java Properties, Thread Dump, collection1 (selected), Overview, Ping, Query (highlighted), Schema, Config, Replication, Analysis, Schema Browser, Plugins / Stats, and Dataimport. The main area is titled 'Request-Handler (qt)' and shows a query '/select' with a 'common' facet. The query parameters are: q=* (with a dropdown arrow), fq=, sort=, start, rows=0 to 10, fl=, df=, Raw Query Parameters=key1=val1&key2=val2, wt=xml (with a dropdown arrow), and checkboxes for 'indent' (checked) and 'debugQuery'. The response is an XML document showing a list of results. The first result is a document with fields: id=GB18030TESTI, name=Test with some GB18030 encoded characters, features (an array of strings), price=0.0, price_c=0, USD, inStock=true, and _version_=1431112805823545344.

14.2.6. Actualitzar l'índex

Per actualitzar un ítem de l'índex, només cal enviar un ítem amb el mateix identificador, i SOLR ja l'actualitza tot sol.

Per eliminar ítems, podem fer-ho per identificador o per consulta.

Per ID:

```
java -Ddata=args -Dcommit=false -jar post.jar "<delete><id>SP2514N</id></delete>"
```

Per consulta:

```
java -Dcommit=false -Ddata=args -jar post.jar  
"<delete><query>name:DDR</query></delete>"
```

Tant per actualitzar registres com per eliminar-los, cal fer un *commit* al final. A la versió 1.4.2 només cal accedir a l'adreça <http://localhost:8983/solr/update?softCommit=true>. En versions anteriors calia enviar '*<commit softCommit=true/>*' a <http://localhost:8983/solr/update>.

Amb l'aplicació post.jar es fa així de simple:

```
java -jar post.jar -
```

14.2.7. Consultar l'índex

Per consultar l'índex hem de fer una petició web a l'adreça <http://localhost:8983/solr/select/>

Vegem els paràmetres més generals que ens ofereix el SOLR:

- **q**: Conté la cerca general. Pot cercar a tots els camps, a algun camp en concret o al camp per defecte, que li indiquem a la configuració.
- **fl**: Ens permet especificar quins camps volem que ens mostri.
- **fq**: És la cerca de filtre. Ens permetrà anar afegint filtres a la cerca. També pot fer directament amb el paràmetre **q** afegint *AND*. Aquest camp pot ser múltiple. Molt útil per aplicar els filtres de les facetes escollides.
- **rows**: Indica el nombre de resultats que volem.
- **start**: Indica el valor per el qual volem que comenci a mostrar els resultats. El fem servir juntament amb l'atribut **rows** per fer la paginació de resultats.
- **sort**: Ens permet ordenar els resultats. S'hi poden aplicar funcions i sempre s'ha d'indicar l'ordre, si és ascendent o descendent.
- **wt**: Ens permet especificar el format de sortida. Es poden escollir molts formats, entre els quals tenim *XML*, *JSON*, *CSV*, *PHP*, *PHPS*, *PYHTON*, *RUBY*, *PERL*...
- **indent**: Aquest paràmetre ens permet demanar que la sortida estigui indentada.

Per exemple si cerquem:

http://localhost:8983/solr/collection1/select?q=%3A*&wt=json&indent=true&rows=2&start=2&fq=text_autors:%22Oliv%C3%A9%20Comadira,%20Oriol%22

```
{
  "responseHeader":{
    "status":0,
    "QTime":10,
    "params":{
      "indent":"true",
      "start":"2",
      "q":"*.*",
      "wt":"json",
      "fq":"text_autors:\"Olivé Comadira, Oriol\"",
      "rows":"2"}},
  "response":{"numFound":1,"start":2,"docs":[]
}}
```

No ens retorna res, ja que només ha trobat un resultat, i nosaltres l'hi hem demanat que comences des del tercer.

http://localhost:8983/solr/collection1/select?q=%3A*&wt=json&indent=true&rows=2&start=0&fq=text_autors:%22Oliv%C3%A9%20Comadira,%20Oriol%22

```
{
  "responseHeader":{
    "status":0,
    "QTime":2,
    "params":{
      "indent":"true",
      "start":"0",
      "q":"*.*",
      "wt":"json",
      "fq":"text_autors:\"Olivé Comadira, Oriol\"",
```

```

    "rows":"2"}},
    "response":{"numFound":1,"start":0,"docs":[
      {
        "id":"4",
        "text_autors":"Olivé Comadira, Oriol",
        "_version_":1433837104159260672}]
    ]}
  }
}

```

Ara si que ens ha retornat el valor perquè li hem indicat que comencés per el primer.

El SOLR té molts més paràmetres, però només hem mostrat els més generals. A més a més, es poden instal·lar nombrosos plugins que ens poden permetre encara més opcions.

Per a més informació es pot visitar la wiki d'apache solr (<http://wiki.apache.org/solr/>).

14.2.8. Facetes

El SOLR ens permet fer llistats d'alguns camps, que generalment no tenen aplicats filtres ni tokenitzadors, i han d'estar indexats, però no cal que estiguin guardats.

Normalment s'utilitza la directiva **copyfield** a *schema.xml* per generar una còpia del camp en un tipus de document semblant a l'*String*.

Per fer els llistats per facetes, el SOLR ens proporciona els següents paràmetres:

- **facet**: Per activar les facetes ha de ser *true*.
- **facet.field**: Indica el camp al que s'aplica la faceta.
- **facet.sort**: Pot ser *index* o *count*. Per defecte *index* que és ordenat alfabèticament, *count* és per ordenar per nombre de resultats.
- **facet.offset**: Per fer paginacions. Indiquem a partir de quina posició ens ha de retornar els resultats.
- **facet.mincount**: Molt útil perquè no ens retorni els que no tenen resultats.
- **facet.limit**: El nombre de facetes que retorna.
- **facet.prefix**: Indica el prefix pel qual comencen les facetes que ens retornarà.

Un exemple de funcionament:

```

$ curl "http://localhost:8983/solr/select/?
q=*&rows=0&indent=true&facet=true&facet.field=text_autors&facet.sort=count&facet.limit=10"
<?xml version="1.0" encoding="UTF-8"?>
<response>

<lst name="responseHeader">
  <int name="status">0</int>
  <int name="QTime">2</int>
  <lst name="params">
    <str name="facet">true</str>
    <str name="indent">true</str>
    <str name="facet.sort">count</str>
    <str name="q">*</str>
    <str name="facet.limit">10</str>
    <str name="facet.field">text_autors</str>
    <str name="rows">0</str>
  </lst>
</response>

```

```

</lst>
</lst>
<result name="response" numFound="19099" start="0"/>
<lst name="facet_counts">
  <lst name="facet_queries"/>
  <lst name="facet_fields">
    <lst name="text_autors">
      <int name="Ferrater Mora, José">1063</int>
      <int name="Universitat de Girona. Escola Politècnica Superior">444</int>
      <int name="Universitat de Girona. Servei de Publicacions">265</int>
      <int name="Vicens Vives, Jaume">228</int>
      <int name="Ferrater Mora, Mercè">218</int>
      <int name="Santaló, Lluís">210</int>
      <int name="Brunet i Solà, Manuel">175</int>
      <int name="Marichal, Juan">169</int>
      <int name="Bertrana, Aurora">162</int>
      <int name="Muguerza, Javier">156</int>
    </lst>
  </lst>
  <lst name="facet_dates"/>
</lst>
</response>

```

Podem comprovar com els autors estan ordenats per nombre de resultats, i només mostrem els 10 primers.

```

$ curl "http://localhost:8983/solr/select/?
q=*&rows=0&indent=true&facet=true&facet.field=text_autors&facet.sort=count&facet.limi
t=10&facet.prefix=U"
<?xml version="1.0" encoding="UTF-8"?>
<response>

  <lst name="responseHeader">
    <int name="status">0</int>
    <int name="QTime">3</int>
    <lst name="params">
      <str name="facet">true</str>
      <str name="indent">true</str>
      <str name="facet.sort">count</str>
      <str name="q">*</str>
      <str name="facet.limit">10</str>
      <str name="facet.prefix">U</str>
      <str name="facet.field">text_autors</str>
      <str name="rows">0</str>
    </lst>
  </lst>
  <result name="response" numFound="19099" start="0"/>
  <lst name="facet_counts">
    <lst name="facet_queries"/>
    <lst name="facet_fields">
      <lst name="text_autors">
        <int name="Universitat de Girona. Escola Politècnica Superior">444</int>
        <int name="Universitat de Girona. Servei de Publicacions">265</int>
        <int name="Universitat de Girona. Facultat de Ciències">122</int>
        <int name="Universitat de Girona. Departament d'Informàtica i Matemàtica
Aplicada">114</int>
        <int name="Universitat de Girona. Facultat d'Educació i Psicologia">98</int>
        <int name="Universitat de Girona. Departament de Química">67</int>
        <int name="Universitat de Girona. Departament d'Enginyeria Química, Agrària
i Tecnologia Agroalimentària">57</int>

```

```

        <int name="Universitat de Girona. Institut de Medi Ambient">55</int>
        <int name="Universitat de Girona. Departament de Pedagogia">52</int>
        <int name="Universitat de Girona. Departament de Geografia, Història i
Història de l'Art">50</int>
    </lst>
</lst>
<lst name="facet_dates"/>
</lst>
</response>

```

Aquesta altra prova és amb prefix, demanant els autors que comencen per “U”.

A més a més, SOLR permet fer peticions de vàries facetes a l'hora. Llavors, als paràmetres cal indicar a quin camp tindran afectació. Per exemple, si volem que els autors comencin per “U”, en aquest cas caldrà indicar-ho així: **f.text_autors.facet.prefix=U**, ja que el nostre camp de faceta d'autors es diu text_autors.

Aquest exemple ens mostra com podem aplicar diferents paràmetres a les diferents facetes. I si no especifiquem el camp, afecta a totes les que no estiguin definides en particular.

```

$ curl "http://localhost:8983/solr/select/?
q=*&rows=0&indent=true&facet=true&facet.field=text_autors&f.text_autors.facet.sort=co
unt&f.text_autors.facet.limit=4&f.text_autors.facet.prefix=U&facet.field=text_subject&f
.text_subject.facet.limit=10&f.text_subject.facet.sort=index&facet.field=text_lang&face
t.limit=8"
<?xml version="1.0" encoding="UTF-8"?>
<response>

<lst name="responseHeader">
  <int name="status">0</int>
  <int name="QTime">3</int>
  <lst name="params">
    <str name="facet">true</str>
    <str name="f.text_autors.facet.prefix">U</str>
    <str name="indent">true</str>
    <str name="f.text_autors.facet.sort">count</str>
    <str name="q">*</str>
    <str name="facet.limit">8</str>
    <arr name="facet.field">
      <str>text_autors</str>
      <str>text_subject</str>
      <str>text_lang</str>
    </arr>
    <str name="f.text_autors.facet.limit">4</str>
    <str name="f.text_subject.facet.limit">10</str>
    <str name="f.text_subject.facet.sort">index</str>
    <str name="rows">0</str>
  </lst>
</lst>
<result name="response" numFound="19099" start="0"/>
<lst name="facet_counts">
  <lst name="facet_queries"/>
  <lst name="facet_fields">
    <lst name="text_autors">
      <int name="Universitat de Girona. Escola Politècnica Superior">444</int>
      <int name="Universitat de Girona. Servei de Publicacions">265</int>
      <int name="Universitat de Girona. Facultat de Ciències">122</int>
      <int name="Universitat de Girona. Departament d'Informàtica i Matemàtica
Aplicada">114</int>
    </lst>

```

```

    <lst name="text_subject">
      <int name="0 - Generalitats">1</int>
      <int name="0-Alquilació">1</int>
      <int name="00 - Ciència i coneixement. Investigació. Cultura.
Humanitats">6</int>
      <int name="004 - Informàtica">29</int>
      <int name="06 - Organitzacions. Associacions. Congressos. Exposicions.
Museus">1</int>
      <int name="070 - Diaris. Premsa. Periodisme. Ciències de la
informació">3</int>
      <int name="09 - Manuscrits. Llibres rars i notables">1</int>
      <int name="1 - Filosofia i psicologia">7</int>
      <int name="14th Century">2</int>
      <int name="14th and 15th Centuries">1</int>
    </lst>
    <lst name="text_lang">
      <int name="spa">6998</int>
      <int name="cat">6425</int>
      <int name="eng">2610</int>
      <int name="fra">399</int>
      <int name="ita">106</int>
      <int name="por">35</int>
      <int name="deu">31</int>
      <int name="other">14</int>
    </lst>
  </lst>
  <lst name="facet_dates"/>
</lst>
</response>

```