Títol: **Intranet del Servei de Biblioteques i Documentació**

Volum: **1**
Estudiant: **Antonio Juan Prieto Jiménez**
Director/Ponent: **Horacio Rodríguez**
Departament: **LSI**
Data: **Tardor 2004**

Dades del Projecte:

Títol del projecte:

Nom de l'estudiant:
Titulació:
Crèdits:
Director/Ponent:
Departament:

MEMBRES DEL TRIBUNAL *(nom i signatura)*

President:

Vocal:

Secretari:

QUALIFICACIÓ

Qualificació numèrica:

Qualificació descriptiva:

Data:

Índex

Índex de taules

Raó i oportunitat

En les biblioteques es realitzen processos molt diversos: atendre serveis, gestionar el pressupost i la col·lecció, catalogar i indexar recursos electrònics, formar usuaris, mantenir equipaments informàtics, etc. Aquests processos requereixen un conjunt de documents que cal difondre a tot el personal de la biblioteca.
Qui treballa en la biblioteca genera documents que reverteixen d'una manera o altre en feines dels seus companys; a la vegada que cada persona és usuària interna d'aquests documents. Una intranet ofereix la possibilitat de gestionar de forma àgil i eficaç la documentació generada i utilitzada per una biblioteca.

Conscients de la importància d'aquesta eina, el Servei de Biblioteques i Documentació de la UPC fa temps que disposa d'una intranet. Amb el pas del temps, l'actual intranet s'ha quedat petita i obsoleta. El gran nombre de documents penjats i la falta d'una organització coherent d'aquests dificulta l'accés eficient a la informació desitjada. A més, el fet que no es mantinguin dades descriptives dels documents, fa que la informació oferta sigui pobra. Així, neix la necessitat de desenvolupar una nova intranet que aprofiti els avantatges que dóna un llenguatge web dinàmic per organitzar millor els documents i esdevingui, amb el temps, una eina eficient pel treball quotidià de les 13 biblioteques de la UPC.

Situació actual

Descripció

Organització del Servei de Biblioteques i Documentació

El Servei de Biblioteques i Documentació consta de 13 biblioteques que donen servei als diferents centres de la UPC i dels Serveis Generals que s'encarreguen de coordinar el treball comú d'aquestes biblioteques. Actualment, el Servei es regeix pel Programa Paideia que descriu 4 eixos estratègics:

- Aprenentatge: els recursos i serveis per a l'aprenentatge
- Recerca: el recursos i serveis per a la recerca i la innovació tecnològica
- Xarxa: els recursos i serveis bibliotecaris digitals
- Organització: els recursos i serveis bibliotecaris de gestió

A més, en molts casos també es considera un cinquè eix, Humanisme, que se centra en recursos i serveis d'informació orientats al desenvolupament "humanístic" dels usuaris de les biblioteques.

Aquest eixos també determinen l'organització dels components del Servei. Concretament, els Serveis Generals es divideixen 4 unitats:

- Unitat de recursos per l'aprenentatge (URA)
- Unitat de recursos de recerca (URR)
- Unitat de recursos digitals (URD)
- Unitat de gestió i desenvolupament (UGD)

La resta de biblioteques tenen una organització similar en funció de la seva mida.

Intranet actual

La intranet actual del Servei de Biblioteques està implementada amb codi Html estàtic. Té una estructura molt simple que s'ha anat modificant a mesura que han sorgit noves necessitats. Editada per personal no necessariament informàtic, utilitza carpetes compartides del servidor per penjar els documents que després s'han de referenciar editant les pàgines Html mitjançant un programa d'edició com FrontPage.

Per altra banda, les tres biblioteques més grans, la Biblioteca Rector Gabriel Ferraté, la Biblioteca de l'Escola Tècnica Superior d'Enginyeria Industrial de Barcelona i la Biblioteca del Campus de Terrassa disposen d'una intranet pròpia que es gestiona com la intranet comuna però des de les biblioteques corresponents.

Avaluació

De la següent situació podem extreure un conjunt de punts forts i punts febles:

PUNTS FEBLES

- El procés d'edició de pàgines Html fa difícil mantenir un estil similar en totes les pàgines. A més és un procés complex per al personal no informàtic.
- No es mantenen dades descriptives dels documents.

- És difícil per localitzar els documents perquè l'organització pot arribar a ser caòtica.
- Quan el número de documents és elevat, es poden generar pàgines Html amb massa informació.
- Quan interessa localitzar un document a una intranet local d'una biblioteca, és fa més difícil perquè l'estructura de les intranets no ha de ser necessàriament la mateixa.

PUNTS FORTS

- No cal analitzar el contingut d'un document que es vol penjar ja que no cal introduir dades descriptives.
- Quan es disposa d'un conjunt de documents enllaçats per un únic arxiu físic, penjar aquests documents és ràpid ja que només cal referenciar el document que enllaça la resta de documents.
- No existeix dependència d'una base de dades o d'un llenguatge de programació.

Objectius del projecte

L'objectiu del projecte és desenvolupar i implementar una nova intranet realitzada amb un llenguatge web dinàmic. Aquesta nova intranet serà l'encarregada d'emmagatzemar i donar accés als documents que s'utilitzen en els diversos processos que es duen a terme des de les biblioteques i unitats que formen el Servei de Biblioteques i Documentació de la UPC. Aquestes són les característiques que ha de complir la nova intranet:

Organització i classificació dels documents

Per facilitar l'accés als documents, es volen classificar a partir d'un **arbre temàtic** de tres nivells creat especialment pels documents de la intranet. Aquest arbre ha de ser representatiu dels àmbits més importants en els que es mouen les biblioteques de la UPC. A la vegada, a cada tema d'aquest arbre se li podrà assignar un o més eixos del programa estratègic Paideia explicat anteriorment.
Com que una intranet és una eina viva, s'ha de donar la possibilitat d'editar, afegir i treure temes de l'arbre per adaptar-se als nous àmbits que puguin sorgir en un futur.

Atributs dels documents

A més de l'organització temàtica, per cada document el sistema guardarà una fitxa descriptiva que s'utilitzarà per fer cerques i oferir més dades del document. Els atributs que es volen mantenir són:

1. Títol
2. Descripció
3. Data d'elaboració
4. Autor/s: unitats, biblioteques o organitzacions que han elaborat el document
5. Tipus de document: acta de reunió, procediment, normativa, presentació,etc.
6. Documents relacionats: altres documents als que fa referència el propi document

Sistema per omplir la fitxa dels documents automàticament

Guardar informació i classificar els documents ajuda a estalviar temps en la localització, però a la vegada implica un treball addicional en el moment de penjar els documents en la intranet. Per minimitzar aquest treball, el sistema disposarà d'una opció per penjar documents que, a través de la informació que es pot extreure de l'arxiu penjat, omplirà automàticament tots els camps del formulari d'entrada exceptuant-ne la descripció. Per realitzar aquesta feina se seguiran dues estratègies:

1. Buscar informació dins el document

Moltes vegades, els documents contenen la informació que interessa guardar i que es pot localitzar d'una manera relativament fàcil ja que acostuma a complir certs patrons. Aquesta estratègia s'aplicarà per obtenir el títol, la data de l'elaboració, els autors, el tipus de document i els documents relacionats.

2. ***Implementació d'un classificador temàtic***
   Mitjançant les paraules que conté el document que es vol penjar i els documents que ja han estat penjats prèviament es classificarà el document seguint l'arbre temàtic mencionat anteriorment.

Intranets locals

Com s'ha dit abans, el Servei de biblioteques consta de 13 biblioteques a més dels Serveis Generals que s'encarreguen de coordinar els objectius comuns d'aquestes biblioteques. Moltes d'aquestes biblioteques necessiten compartir documents que són d'àmbit intern que no necessariament interessen a la resta. Així doncs, és necessari no només implementar una intranet per gestionar els documents que interessin a totes les biblioteques, sinó a més, configurar la intranet per que pugui ser utilitzada com una intranet local per una d'aquestes biblioteques. D'aquesta manera, el sistema es comportarà com si fos no només una intranet, sinó com més d'una: la intranet comuna i les intranets de les biblioteques que ho desitgin.

Altres característiques

La intranet ha de ser, sobretot, una eina de difusió d'informació. Per això, també s'ha previst incloure una secció per la difusió d'**esdeveniments** o **novetats** que puguin ser interessants pel personal de les biblioteques. Per últim, la intranet haurà d'**enllaçar** les diferents **aplicacions web** que utilitza el personal del Servei.

Beneficis

El principal benefici que aporta la nova intranet és una **millora del trànsit de la informació** gràcies a la nova organització dels documents i la facilitat d'accés. Això, permetrà a la llarga una millor eficiència del processos propis de les biblioteques a causa de la estandarització.

Així mateix, el **manteniment de la intranet serà més simple**. Molts documents que abans no es penjaven a l'anterior intranet degut al complex procés que requeria es penjaran ara d'una forma més àgil i ràpida.

Per altra banda, **controlar l'accés al manteniment** de la intranet també serà més fàcil ja que es disposarà d'apartats per la gestió d'usuaris i permisos.

La utilització d'una intranet és una factor per mesurar la qualitat del servei ofert per una organització. El fet que el personal d'una organització disposi d'eines desenvolupades per ajudar a augmentar l'eficiència del seu treball provoca una **millora de la imatge de la pròpia organització**.

Per últim, la nova intranet permetrà **estalviar espai** en altres eines de compartició temporals de documents que abans s'utilitzaven per documents que s'haurien de trobar a la intranet i que sovint es trobaven plenes.

Arquitectura tècnica

Per hostetjar la intranet el Servei de Biblioteques ha adquirit un Servidor amb les següents característiques:

- Processador Pentium Xeon a 2,8 Ghz
- 1MB de cache
- Xipset Intel E7520 + ICH5R + PXH
- Velocitat del Bus 800 MHz
- Memòria DDR333 SDRAM-2
- Un disc de 36 GB a 10 Krpm
- Un disc addicional de 73 GB U320 SCSI a 10 Krpm

Aquest nou servidor disposarà del sistema operatiu Windows Server 2003, que incorpora el servidor web Internet Information Server 6. És tracta d'un servidor preparat per servir nativament Active Server Pages, llenguatge de programació web ja utilitzat en altres plataformes del Servei de Biblioteques com ara Bibliotècnica.

Per la base de dades, s'utilitzarà el SGBD SQL Server 2000 que s'integra fàcilment amb el servidor utilitzat ja que els dos són productes de Microsoft.

Finalment, i per raons de seguretat, el codi generat correrà sota SSL amb autentificació sobre el domini on es troba el personal de biblioteques (UPCXXI), que gestiona UPCNet.

Requeriments del sistema

Requeriments funcionals

Una intranet és sobretot una eina de difusió d'informació entre el personal d'una organització. És important que aquesta informació sigui validada per uns responsables. Cal, per tant, designar un conjunt de persones amb diversos permisos i rols dins la intranet per poder portar un control dels continguts que es pengen. Per les característiques del Servei de Biblioteques s'ha decidit definir 4 rols diferents:

1. Usuari normal. No tindrà permisos per editar continguts de la intranet, només podrà consultar i navegar per la part pública.
2. Mantenidor biblioteca. No podrà penjar continguts a la intranet directament però podrà fer sol·licituds de nous documents en nom d'una biblioteca.
3. Mantenidor. S'encarregarà de fer la gestió de documents i processar les sol·licituds que arribin des de les biblioteques.
4. Administrador. Gestionarà l'arbre temàtic, els esdeveniments del calendari i les novetats, assignarà els permisos als usuaris i mantindrà els paràmetres del classificador automàtic de documents i l'extractor d'atributs.

Com que la intranet és una eina coordinada des dels SGB, tots els mantenidors i administradors seran membres de Servei Generals. Concretament, per cada unitat dels Serveis es designarà una persona responsable de penjar els documents corresponents a la seva area i els administradors seran membres de direcció del Servei i personal informàtic encarregats del desenvolupament.

Aquest rols s'han definit per la intranet comuna però també seran aplicables a les intranets locals de les biblioteques. En aquest cas no caldrà la figura del mantenidor de biblioteca ja que l'entorn d'una biblioteca és més reduït.

Pel que fa a les funcionalitats requerides per cada un d'aquests rols, cal aclarir que una persona que tingui assignat un determinat rol dins d'una biblioteca, també podrà utilitzar les funcionalitats de la resta de rols que el precedeixen, és a dir, un administrador podrà utilitzar les funcionalitats del mantenidor, del mantenidor de biblioteca i, naturalment, de l'usuari, i així successivament.

Usuari

Llistar documents

L'usuari utilitzarà la intranet, sobretot, per accedir als documents que hi ha penjats. És per això que cal facilitar al màxim l'accés. Una de les vies per obtenir llistats dels documents és a través de l'arbre temàtic utilitzat per la seva classificació. Per tal de donar facilitats en la utilització d'aquest arbre, el sistema haurà de generar un **menú desplegable** per donar varies opcions de navegació i **obtenir diferents llistats**. Aquestes són les opcions que ha d'oferir el menú:

1) Temes clau
Els temes clau són els temes de primer nivell que es volen destacar especialment. El menú mostra els temes clau per donar accés als documents que pengen d'aquests.

2) Tots els temes
El menú tindrà un altre apartat amb tots els temes de primer nivell per poder navegar a través dels diferents nivells.

3) Eixos
Cada eix de la intranet tindrà assignat el conjunt de temes amb els que té alguna relació. El menú ha de donar l'opció d'escollir un eix i navegar pels temes que té assignat aquest eix.

4) Tipus de document
En certes ocasions l'usuari voldrà accedir a tots els documents d'un determinat tipus que pengen d'un tema concret. Per fer això, el menú disposarà d'un últim apartat on es podrà seleccionar un tipus de document i després navegar per l'arbre per veure els documents d'aquell tipus que pengen de cada tema.

Cercador

Una segona via per accedir als documents és a través del **cercador**. L'usuari tindrà l'opció de fer una cerca de documents en funció del títol, la descripció, la intranet on es troba el document, el tipus de document, l'autor i la data d'elaboració del document.

Visualització de documents

Finalment, l'usuari podrà **veure les dades de la fitxa** d'un determinat document i **visualitzar** el document al navegador amb les dades de la fitxa.

Un cas especial en la visualització dels documents és el que s'anomena Llibre clau. El **llibre clau** és un tipus document jerarquitzat que organitza un conjunt d'arxius en un arbre. El sistema també haurà de donar l'opció de visualitzar aquest tipus de document mostrant l'organització jeràrquica i donant accés als arxius que conté.

Altres

A part dels documents, un usuari de la intranet també tindrà la necessitat d'obtenir altres llistats o informacions:

1. Llistat de les intranets de les biblioteques amb accés a cada una.
2. Llistat d'aplicacions web que utilitza el personal de la intranet accessibles des de la intranet.
3. Llistat de novetats o informacions destacades.
4. Llistat d'esdeveniments introduïts, recuperables per data (dia i mes).
5. Dades d'un esdeveniment concret.
6. Dades d'una novetat concreta.

Mantenidor de biblioteca

Els mantenidors de biblioteca seran els encarregats de fer sol·licituds de nous continguts per penjar a la intranet. Aquestes són les funcionalitats que ha d'oferir el sistema:

1. Fer una sol·licitud de documents per penjar a la intranet. Aquestes sol·licituds es destinaran al responsable de l'eix que el mantenidor de biblioteca seleccioni.
2. Fer una sol·licitud de nou esdeveniment de la biblioteca per penjar a la intranet.
3. Fer una nova sol·licitud de novetat a destacar en la intranet.

Mantenidor

Els mantenidors són els encarregats de fer la gestió dels documents. Aquestes són les funcionalitats que requereixen:

Llistar documents

1. El sistema generarà el **llistat de documents per cada tema** de l'arbre temàtic amb les opcions per manipular-los.
2. També generarà llistats de documents per la **lletra inicial del títol**, **paraules clau del títol** o el llistat sencer amb les opcions per manipular-los.

Alta, edició i baixa de documents

Per cada document de la intranet caldrà mantenir un fitxa que contindrà el títol, la data d'elaboració del document, la descripció, el tipus de document, els autors del document, els temes assignats i els documents relacionats.

El mantenidor tindrà l'opció d'**afegir els documents** omplint la seva fitxa i introduint l'arxiu o l'enllaç del document al sistema. També podrà **editar** el contingut de la fitxa o l'enllaç del document en qualsevol moment i **afegir noves versions** del document introduint un nou enllaç i seleccionant una nova data d'elaboració i autors de la versió.

Per facilitar el manteniment dels documents el sistema donarà l'opció d'afegir els documents amb un **sistema d'extracció dels atributs** de la fitxa juntament amb un **classificador de documents** que introduint l'enllaç del document ompli directament la fitxa que el mantenidor validarà seguidament. Per realitzar aquesta tasca el sistema haurà de fer un **preprocés** dels enllaços introduïts on caldrà convertir el document en altres formats fins poder-lo llegir. Seguidament **aplicarà un conjunt de regles** que s'hauran emmagatzemat prèviament al sistema per extreure un conjunt d'atributs. Finalment, amb les dades calculades per la classificació temàtica de documents (la relació entre les paraules i els temes) el sistema assignarà un conjunt de temes al document processat.
Per últim existirà l'opció d'esborrar documents en dos passos. En el primer el sistema amagarà el document de la intranet deixant-lo en una secció de manteniment on, si es vol, es podrà esborrar definitivament.

Llibre clau

Com ja s'ha dit abans, dins la intranet existeix un tipus de document jerarquitzat que contindrà més d'un arxiu. Els mantenidors seran els encarregats de crear aquest tipus de document **creant nous apartats** (nodes dins l'arbre que mostra el document) o **nous arxius**.

Documents sol·licitats

Els mantenidors vinculats a un eix, s'encarregaran de gestionar les peticions que rebin dels mantenidors de biblioteques. Hauran de poder **llistar els documents sol·licitats** i **publicar els documents** que considerin oportuns.

Administrador

Gestió de l'arbre temàtic

L'arbre temàtic s'ha d'adaptar als nous àmbits que es desenvolupar a les biblioteques. És per això que el sistema ha de donar la possibilitat d'editar aquest arbre i fer-lo així més flexible als canvis.

L'administrador haurà de poder **afegir** nous **temes, editar** els temes existents**, moure** els temes de lloc i **esborrar**-los.

Així mateix, per les opcions del menú de navegació de temes clau i eixos, caldrà oferir la possibilitat de **seleccionar** els temes de primer nivell que es consideren **temes clau** i **relacionar els eixos** amb els temes.

Encara que hauria de ser menys freqüent, també s'haurà de donar la possibilitat d'**afegir, editar i esborra els eixos**.

Gestió del calendari (esdeveniments i novetats)

Tant els **esdeveniments** com les **novetats** són elements per fer difusió al personal de biblioteques. Un esdeveniment ha de tenir un títol i una data d'inici. A més, pot tenir una descripció, data de fi, hora d'inici i de fi, lloc on es realitzarà i un enllaç amb informació relacionada. Pel que fa a les novetats, tindran un títol, una descripció, una data de caducitat i un enllaç a una pàgina web amb informació relacionada.

El sistema donarà l'opció **d'afegir, editar i esborrar** tant esdeveniments com novetats i llistar els esdeveniments i novetats existents.

Com en el cas dels documents, el sistema haurà d'oferir accés als **llistats** d'esdeveniments i novetats sol·licitades pels mantenidors de la biblioteca i permetre que l'administrador les validi i les publiqui definitivament al sistema.

Gestió d'usuaris

L'administrador serà també l'encarregat de donar els **permisos** al personal de les biblioteques per gestionar els continguts de la intranet. Per fer-ho el sistema oferirà llistats de totes les persones amb permisos a la intranet que administra i les opcions per editar tant el rol assignat com el correu electrònic, i el nom dels usuaris. A més també podrà **afegir nous usuaris**.

Finalment, es podrà seleccionar per cada eix un responsable que s'encarregarà de gestionar les sol·licituds que s'enviïn al seu eix.

Sistema per l'extracció d'atributs del document i classificador temàtic

Per poder afegir documents amb extracció d'atributs de l'enllaç i classificació temàtica automàtica, caldrà que l'administrador introdueixi al sistema els paràmetres o regles que guiaran aquest procés. S'assumeix que totes aquestes funcionalitats les utilitzaran persones amb coneixements informàtics i del sistema i no caldrà, per tant, oferir una interfície web per facilitar l'execució sinó que amb un conjunt de scripts serà suficient.

Per una banda, el sistema ha de **permetre introduir noves regles** que s'utilitzaran per determinar els atributs que s'ompliran de la fitxa a partir de la lectura de l'enllaç. L'administrador generarà un arxiu amb les regles amb el format explicat més endavant i executarà el script que introdueix les regles a la base de dades.

Així mateix, pel classificador temàtic, el sistema permetrà tornar a fer la **selecció de les paraules que s'assignaran a cada tema** i de la relació de pes que hi haurà entre ambdues. En aquest cas, com en el cas anterior, això es farà mitjançant un o més scripts que seleccionarà un conjunt de documents i farà els càlculs necessaris per obtenir les paraules assignades a cada tema.

Finalment, **per provar la qualitat del sistema classificador**, l'administrador podrà utilitzar un altre script que, amb els documents que no s'han utilitzat en la selecció de les a paraules assignades a cada tema, farà una classificació dels documents i la comparará amb la classificació real mostrant els paràmetres de *precision i recall* resultants de l'execució.

Requeriments no funcionals

Seguretat

Com que l'eina que s'està desenvolupant és una intranet uns dels aspectes importants a controlar és que l'accés als continguts oferts només estigui disponible pel personal del Servei de Biblioteques. Per realitzar aquesta tasca s'utilitzaran les eines que ofereix IIS que permet configurar el servidor per donar accés només al usuaris que pertanyin a un cert domini (en el cas del Servei el domini UPCXXI gestionat per UPCNet) i, a més, es farà corre el codi de la intranet sota SSL per encriptar els documents i les pàgines generades pel servidor.
Una vegada dins la intranet, també caldrà controlar que les persones sense permisos no puguin accedir als apartats de manteniment de la intranet, i que només puguin fer servir les funcionalitats que permet el rol que tenen assignat.
Així mateix, també caldrà guardar informació dels usuaris que han penjat continguts en la intranet per poder actuar davant de comportaments no desitjables.

Interfície

En tractar-se d'una aplicació web, és important tenir en compte aspectes relacionats amb l'interfície:

1. Formularis clars i entenedors. Els administradors i mantenidors de la intranet poden no ser usuaris molt acostumats a tractar amb eines web. Per això, cal fer les seccions de manteniment el màxim de clares possibles.
2. Visualització en tots els navegadors. Existeixen una gran nombre de navegadors amb diferents versions i és important que la intranet es mostri i funcioni correctament en tots, especialment en els navegadors instal·lats per defecte en les imatges dels ordinadors del personal del Servei.
3. Navegació ràpida. Cal oferir el màxim d'opcions d'accés a altres apartats de la intranet en cada pàgina per fer més ràpida la navegació.

Qualitat

Una intranet s'ha d'adaptar als canvis que es puguin dur a terme dins l'organització a la que serveix. En poc temps poden sorgir noves necessitats. Per tant, és important afavorir el màxim la canviabilitat del sistema, tant a l'hora d'afegir noves funcionalitats com de modificar les actuals. En especial, cal que la gestió de l'arbre de temes sigui fàcil d'utilitzar per afavorir al màxim els canvis i, per tant, la flexibilitat en l'organització dels documents.

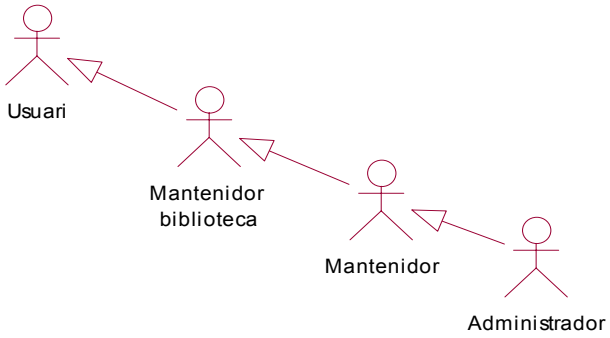
Pel que fa al sistema per a la inserció automàtica de documents, s'ha de dissenyar com un component a part de la intranet, que no afecti a la resta de funcionalitats. És important que el temps que trigui en fer tot el procés d'extracció d'atributs i classificació del document sigui mínim per estalviar el màxim de temps d'espera a l'usuari. No s'espera que el resultats siguin cent per cent fiables, però sí que obtinguin els resultats prou bons com perquè als mantenidors de la intranet els resulti més fàcil introduir els documents amb aquesta opció que no pas amb la normal, fixant especial atenció en el classificador temàtic de documents.

Model de casos d'ús

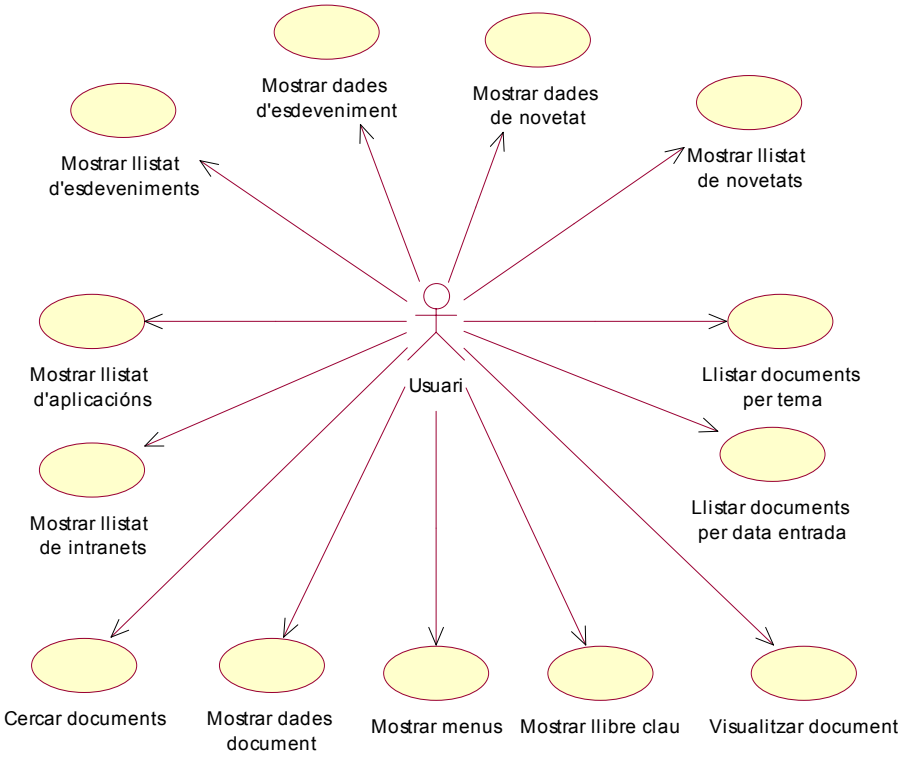
Diagrama de casos d'ús

Després de definir els requeriments funcionals es defineixen els casos d'ús que es poden extreure:

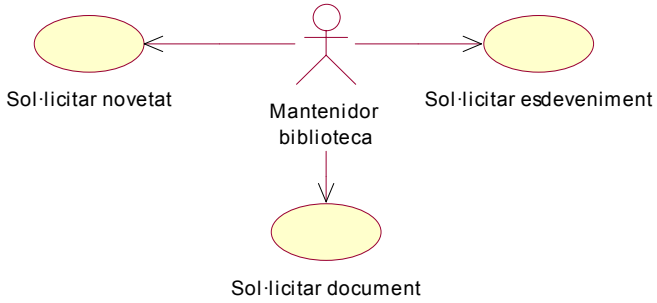
Primer de tot, es mostra un llistat dels autors del sistema amb la seva relació i els casos d'ús on participa cada actor.



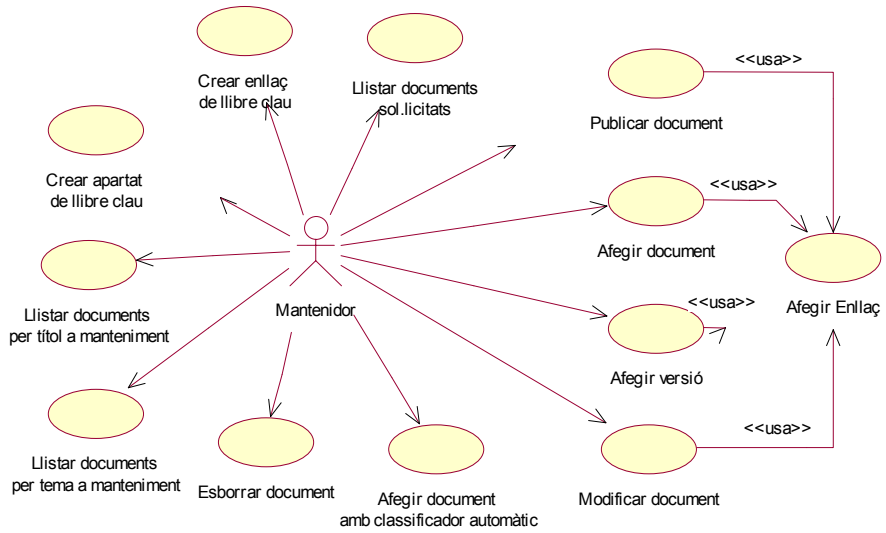
Usuari



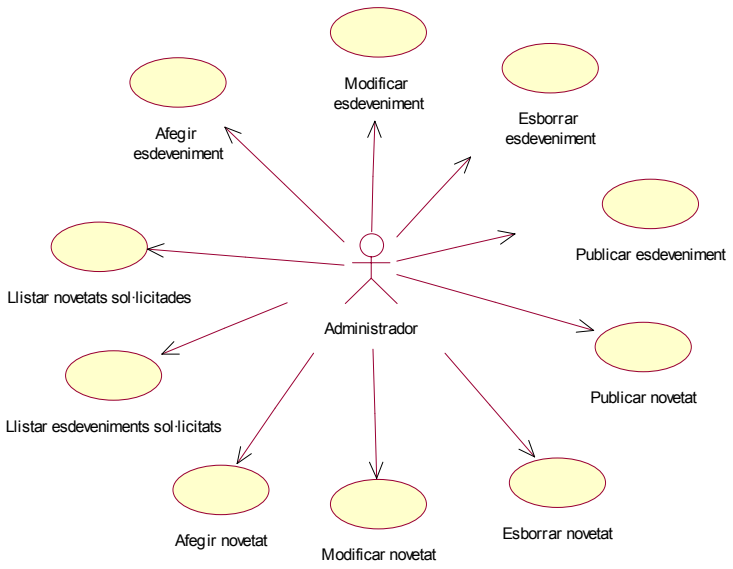
Mantenidor de biblioteca

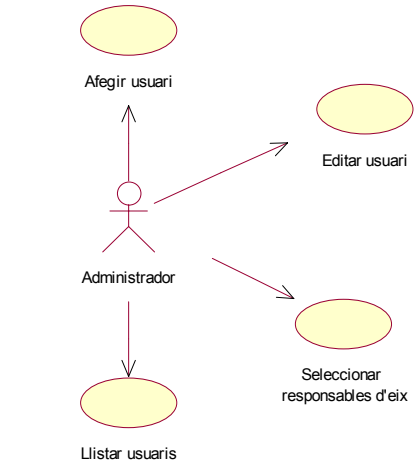


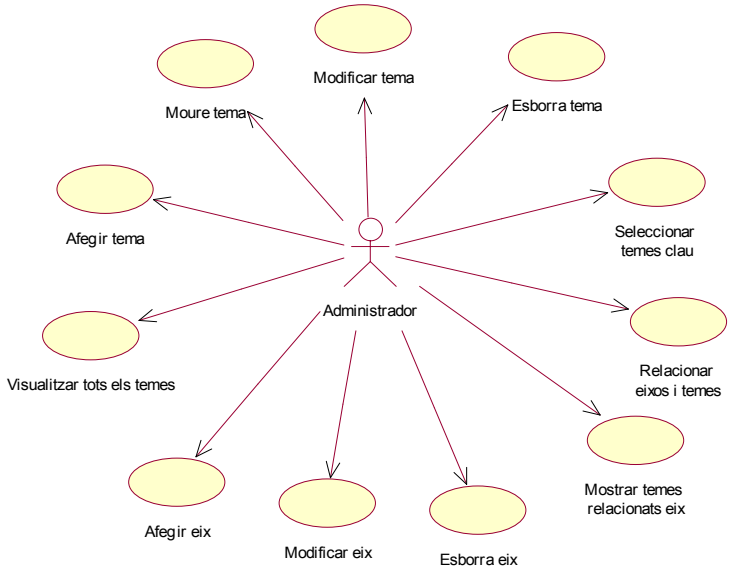
Mantenidor

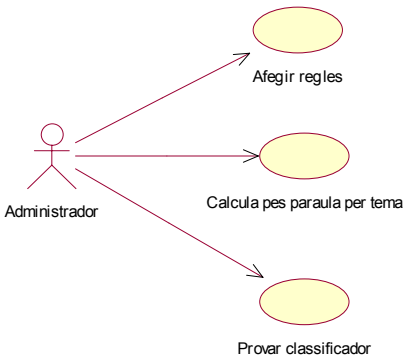


Administrador




Afegir usuari
Editar usuari
Administrador
Seleccionar
responsables d'eix
Llistar usuaris


Moure tema
Modificar tema
Esborra tema
Afegir tema
Seleccionar
temes clau
Administrador
Visualitzar tots els temes
Relacionar
eixos i temes
Mostrar temes
relacionats eix
Afegir eix
Modificar eix
Esborra eix


Afegir regles
Administrador
Calcula pes paraula per tema
Provar classificador

Especificació del casos d'ús

Usuari

Cas d'ús: Mostrar menús
Actors: Usuari(iniciador), Sistema
Propòsit: Mostrar a l'usuari els menús per la navegació dins la intranet
Resum: El sistema genera un llistat amb els menús de la intranet en funció de l'apartat on es troba, manteniment o part pública. A la part pública el sistema mostra els menús que permeten llistar els documents per tema. A la part de manteniment mostra les possibles opcions de manteniment en funció del seu rol.
Tipus: primari i essencial

Cas d'ús: Mostrar documents ordenats per data entrada
Actors: Usuari(iniciador), Sistema
Propòsit: Mostrar els últims 10 documents entrats
Resum: El sistema, quan l'usuari accedeix a la home de la intranet, mostra un llistat dels 10 últims documents entrats amb el títol, descripció, tipus de document, data d'elaboració i un enllaç per poder descarregar el document.
Tipus: primari i essencial

Cas d'ús: Llistar documents per tema
Actors: Usuari(iniciador), Sistema
Propòsit: Mostrar a l'usuari els documents que pengen d'un determinat tema
Resum: L'usuari selecciona un tema, i el sistema mostra el llistat de documents que pengen d'aquest tema amb el títol, descripció, tipus de document, data d'elaboració i un enllaç per poder descarregar el document. Si l'usuari navega per l'opció de tipus de document només es mostren el documents d'aquell tipus. En el cas de les intranets de les biblioteques, el sistema també mostra els documents de la intranet comuna que hi ha en aquest tema.
Tipus: primari i essencial

Cas d'ús: Cercar documents
Actors: Usuari(iniciador), Sistema
Propòsit: Fer una cerca de documents amb unes característiques determinades
Resum: L'usuari introdueix alguns paràmetres de la fitxa d'un document incloent la intranet on vol buscar el document i el sistema busca a la base de dades els documents que compleixin les propietats de la cerca.
Tipus: primari i essencial

Cas d'ús: Mostrar dades document
Actors: Usuari(iniciador), Sistema
Propòsit: Mostrar a l'usuari les dades de la fitxa d'un determinat document
Resum: El sistema mostra les dades de determinat document acompanyat d'un enllaç al document físic.
Tipus: primari i essencial

Cas d'ús: Mostrar llibre clau
Actors: Usuari(iniciador), Sistema
Propòsit: Mostrar a l'usuari els apartats i enllaços que conté un llibre clau
Resum: L'usuari selecciona un llibre clau per visualitzar. El sistema mostra els apartats i els enllaços estructurats tal i com es van penjar per mantenidor.
Tipus: primari i essencial

Cas d'ús: Visualitzar document
Actors: Usuari(iniciador), Sistema
Propòsit: Mostrar al navegador l'enllaç d'un document juntament amb la informació de la fitxa del document.
Resum: L'usuari selecciona un document per visualitzar. El sistema obre una nova finestra al navegador amb dos parts. L'enllaç del document seleccionat i les dades de la fitxa.
Tipus: primari i essencial

Cas d'ús: Mostrar llistat d'aplicacions
Actors: Usuari(iniciador),Sistema
Propòsit: Mostrar al usuari el conjunt d'aplicacions accessibles des de la intranet
Resum: El sistema mostra el llistat d'aplicacions accessibles des de la intranet amb un enllaç a cada una.
Tipus: primari i essencial

Cas d'ús: Mostrar llistat d'intranets
Actors: Usuari(iniciador), Sistema
Propòsit: Mostrar el conjunt d'intranets (intranet comuna i intranets de les biblioteques)
Resum: Mostra el llistat complet d'intranets amb un enllaç a cadascuna.
Tipus: primari i essencial

Cas d'ús: Mostrar llistat d'esdeveniments
Actors: Usuari(iniciador), Sistema
Propòsit: Mostrar al usuari un llistat d'esdeveniment
Resum: Si l'usuari no ha seleccionat cap data es mostra el nom dels esdeveniments no caducats. Si l'usuari ha seleccionat una data o un mes es mostra un llistat dels esdeveniments d'aquella data o mes amb el seu nom i un enllaç per mostrar més informació.
Tipus: primari i essencial

Cas d'ús: Mostrar llistat de novetats
Actors: Usuari(iniciador), Sistema
Propòsit: Mostrar al usuari un llistat de novetats actuals
Resum: El sistema mostra un llistat de les novetats no caducades amb el seu nom i un enllaç per mostrar més informació.
Tipus: primari i essencial

Cas d'ús: Mostrar dades de novetat
Actors: Usuari(iniciador),Sistema
Propòsit: Mostrar les dades assignades a una novetat
Resum: L'usuari selecciona una novetat i el sistema mostra el nom, la descripció i l'enllaç assignat a aquesta novetat.

Cas d'ús: Mostrar dades d'esdeveniment
Actors: Usuari(iniciador),Sistema
Propòsit: Mostrar les dades assignades a un determinat esdeveniment
Resum: L'usuari selecciona un esdeveniment i el sistema mostra el nom, la descripció, la data d'inici i de fi, l'horari, el lloc i l'enllaç assignat a l'esdeveniment.
Tipus: primari i essencial

Mantenidor de biblioteca

Cas d'ús: Sol·licitar nou document
Actors: Mantenidor de biblioteca(iniciador), Sistema
Propòsit: Fer una sol·licitud per penjar un determinat document a la intranet
Resum: El mantenidor de biblioteca introdueix en un formulari el títol del document, la data d'elaboració, la descripció, el tipus de document i l'autor del document que vol penjar amb l'enllaç del document, l'eix relacionat amb el document i el seu correu electrònic. El sistema guarda la nova sol·licitud i envia un missatge al mantenidor de l'eix seleccionat i al sol·licitant indicant que s'ha processat la sol·licitud. Si el document té documents relacionats el mantenidor introdueix els enllaços al sistema i el sistema ho guarda com a noves sol·licituds relacionades.
Tipus: secundari i essencial

Cas d'ús: Sol·licitar nou esdeveniment
Actors: Mantenidor de biblioteca(iniciador),Sistema
Propòsit: Fer una sol·licitud per publicar un nou esdeveniment
Resum: El mantenidor de biblioteca introdueix en un formulari el nom i la data d'inici de l'esdeveniment més la descripció, data de fi, horari, lloc i un enllaç si s'escau. El sistema guarda la sol·licitud i envia un missatge a l'administrador de la intranet i al sol·licitant indicant que s'ha processat la sol·licitud.
Tipus: secundari i essencial

Cas d'ús: Sol·licitar novetat
Actors: Mantenidor de biblioteca(iniciador),Sistema
Propòsit: Fer una sol·licitud per publicar una novetat a la intranet
Resum: El mantenidor de biblioteca introdueix en un formulari el nom de la novetat més la descripció i un enllaç si s'escau. El sistema guarda la sol·licitud i envia un missatge a l'administrador de la intranet i al sol·licitant indicant que s'ha processat la sol·licitud.
Tipus: secundari i essencial

Mantenidor

Cas d'ús: Llistar documents per tema a manteniment
Actors: Mantenidor(iniciador), Sistema
Propòsit: Mostrar al mantenidor els documents que pengen d'un determinat tema
Resum: El mantenidor selecciona un tema, i el sistema mostrar el llistat de documents que pengen d'aquest tema amb enllaços a les accions que es poden fer amb aquest tema i opcions per ordenar el llistat per diferents paràmetres.
Tipus: primari i essencial

Cas d'ús: Llistar documents per títol a manteniment
Actors: Mantenidor(iniciador), Sistema
Propòsit: Mostrar al mantenidor un llistat de documents que compleix una certa propietat al títol
Resum: El mantenidor selecciona una lletra inicial del títol o introdueix una paraula clau de títol o selecciona l'opció de veure tots els documents. El sistema mostra el llistat de documents que compleixen aquestes condicions amb enllaços a les accions que es poden fer amb aquest document i opcions per ordenar el llistat per diferents paràmetres.
Tipus: primari i essencial

Cas d'ús: Afegir document
Actors: Mantenidor(iniciador), Sistema
Propòsit: Afegir un nou document a la intranet
Resum: El mantenidor introdueix el títol del document, la data d'elaboració, el tipus de document i selecciona els autors del document, els temes relacionats i els documents relacionats. El sistema guarda aquestes dades i passa a la pantalla per seleccionar l'enllaç (cas d'ús afegir enllaç).
Tipus: primari i essencial

Cas d'ús: Afegir versió
Actors: Mantenidor(iniciador), Sistema
Propòsit: Afegir una nova versió d'un document a la intranet
Resum: El mantenidor selecciona un document i modifica les seves dades introduint una nova data d'elaboració per la nova versió. Seguidament el sistema guarda les noves dades i passa a la pantalla per seleccionar l'enllaç (cas d'ús afegir enllaç).
Tipus: primari i essencial

Cas d'ús: Modificar document
Actors: Mantenidor(iniciador), Sistema
Propòsit: Modificar les dades d'un document o l'enllaç
Resum: El mantenidor selecciona el document que vol editar. El sistema mostra en un formulari les dades de l'ultima versió del document. El mantenidor fa els canvis i el sistema guarda les noves dades. Seguidament, el sistema passa a la pantalla per seleccionar l'enllaç (cas d'ús afegir enllaç) o deixar l'enllaç actual.
Tipus: secundari i essencial

Cas d'ús: Afegir document amb classificador automàtic
Actors: Mantenidor(iniciador), Sistema
Propòsit: Afegir un document a la intranet fent servir sistema per extreure les dades de l'enllaç i el classificador temàtic
Resum: El mantenidor introdueix l'enllaç al sistema, el sistema llegeix l'enllaç, preprocesa l'arxiu penjat i amb les regles que conté i els càlculs de les paraules relacionades per cada tema extreu les dades de la fitxa i els temes relacionats i mostra un formulari amb les dades. El mantenidor valida les dades i el sistema guarda el document a la intranet.
Tipus: primari i essencial

Cas d'ús: Afegir enllaç
Actors: Mantenidor(iniciador), Sistema
Propòsit: Afegir un nou enllaç a un document
Resum: El mantenidor selecciona l'arxiu o l'enllaç del document. El sistema guarda l'enllaç del document.
Tipus: primari i essencial

Cas d'ús: Esborrar document
Actors: Mantenidor(iniciador), Administrador, Sistema
Propòsit: Esborrar un document de la intranet
Resum: El mantenidor selecciona un documents a esborrar i el sistema pregunta si vol realment esborrar el document. Si respon que sí, el sistema deixa el document a una zona de manteniment on l'administrador el podrà esborrar definitivament o recuperar-lo si cal.
Tipus: primari i essencial

Cas d'ús: Llistar documents sol·licitats
Actors: Mantenidor(iniciador), Sistema
Propòsit: Mostrar al mantenidor un llistat dels documents sol·licitats que encara no s'han publicat
Resum: El sistema el mostra els documents sol·licitats amb les opcions per esborrar i validar/publicar.
Tipus: secundari i essencial

Cas d'ús: Publicar document sol·licitat
Actors: Mantenidor(iniciador), Sistema
Propòsit: Validar les dades d'un document sol·licitat per un mantenidor de biblioteca
Resum: El mantenidor selecciona un document sol·licitat, valida les dades introduïdes pel mantenidor de biblioteques i li assigna uns temes. Un cop el mantenidor ha validat les dades el sistema publica el document per fer-lo visible per tots els usuaris de la intranet i envia un missatge al sol·licitant.
Tipus: secundari i essencial

Cas d'ús: Crear apartat de llibre clau
Actors: Mantenidor(iniciador),Sistema
Propòsit: Afegir un nou apartat a l'arbre que representa un llibre clau
Resum: Dins la pantalla d'edició de llibre clau, el mantenidor introdueix el nom del nou apartat i selecciona el lloc on es penjarà. El sistema introdueix el nou apartat al llibre clau.
Tipus: primari i essencial

Cas d'ús: Crear enllaç de llibre clau
Actors: Mantenidor(iniciador),Sistema
Propòsit: Afegir un nou arxiu a un llibre clau
Resum: Dins la pantalla d'edició de llibre clau, el mantenidor introdueix un nou arxiu i el nom del nou arxiu i selecciona el lloc on es penjarà. El sistema introdueix el nou enllaç al llibre clau.
Tipus: primari i essencial

Administrador

Gestió del calendari (esdeveniments i novetats)

Cas d'ús: Afegir esdeveniments
Actors: Administrador(iniciador), Sistema
Propòsit: Afegir al calendari un nou esdeveniment
Resum: L'administrador introdueix el nom de l'esdeveniment, la data d'inici i si cal, la data de finalització, la hora d'inici i de fi, el lloc on es produeix l'esdeveniment, la descripció i un enllaç cap una pàgina web per donar més informació. El sistema guarda el nou esdeveniment.
Tipus: primari i essencial

Cas d'ús: Editar esdeveniment
Actors: Administrador(iniciador), Sistema
Propòsit: Modificar les dades d'un esdeveniment
Resum: L'administrador selecciona un esdeveniment i el sistema mostra les dades de l'esdeveniment en un formulari. L'usuari modifica les dades del formulari i accepta els canvis.
Tipus: secundari i essencial

Cas d'ús: Esborrar esdeveniment
Actors: Administrador(iniciador), Sistema
Propòsit: Esborrar un esdeveniment.
Resum: El sistema mostra un llistat d'esdeveniments que conté. L'administrador selecciona l'esdeveniment que vol esborrar. El sistema pregunta si està segur que vol esborrar l'esdeveniment. Si l'administrador confirma, el sistema esborra de la base de dades l'esdeveniment.
Tipus: primari i essencial

Cas d'ús: Afegir novetat
Actors: Administrador(iniciador),Sistema
Propòsit: Introduir una novetat al sistema
Resum: L'administrador introdueix el nom de la novetat, la data de caducitat, la descripció i un enllaç amb informació relacionada si cal. El sistema introdueix les dades a la base de dades.
Tipus: primari i essencial

Cas d'ús: Editar novetat
Actors: Administrador(iniciador),Sistema
Propòsit: Modificar les dades d'una novetat
Resum: L'administrador selecciona una novetat, el sistema mostra les dades de la novetat en un formulari. L'administrador modifica les dades que vol i accepta els canvis. El sistema fa les modificacions en la base de dades.
Tipus: primari i essencial

Cas d'ús: Esborrar novetat
Actors: Administrador(iniciador),Sistema
Propòsit: Modificar les dades d'una novetat
Resum: El sistema mostra un llistat de novetats que conté. L'administrador selecciona la novetat que vol esborrar. El sistema pregunta si està segur que vol esborrar la novetat. Si l'administrador confirma, el sistema esborra de la base de dades la novetat.
Tipus: primari i essencial

Cas d'ús: Mostrar llistat d'esdeveniments sol:licitats
Actors: Administrador(iniciador), Sistema
Propòsit: Mostrar a l'administrador un llistat d'esdeveniment que han sol·licitats i que encara no s'han publicat
Resum: El sistema mostra el llistat d'esdeveniments sol·licitats que encara no s'han publicat amb opcions per esborrar o publicar l'esdeveniment.
Tipus: secundari i essencial

Cas d'ús: Mostrar llistat de novetats sol·licitades
Actors: Administrador(iniciador), Sistema
Propòsit: Mostrar a l'administrador un llistat de novetats sol·licitades que encara no s'han publicat
Resum: El sistema mostra un llistat de les novetats sol·licitades que encara no s'han publicat amb opcions per esborrar o publicar.
Tipus: secundari i essencial

Cas d'ús: Publicar esdeveniment
Actors: Administrador(iniciador), Sistema
Propòsit: Publicar un esdeveniment sol·licitat per un mantenidor de biblioteca
Resum: L'administrador selecciona un esdeveniment sol·licitat i valida les dades i fa públic l'esdeveniment per tots els usuaris.
Tipus: secundari i essencial

Cas d'ús: Publicar novetat
Actors: Administrador(iniciador), Sistema
Propòsit: Publicar una novetat sol·licitada per un mantenidor de biblioteca
Resum: L'administrador selecciona una novetat sol·licitada i valida les dades i fa públic l'esdeveniment per tots els usuaris.
Tipus: secundari i essencial

Gestió d'usuaris

Cas d'ús: Afegir usuari
Actors: Administrador(iniciador), Sistema
Propòsit: Afegir un nou usuari al sistema i donar-li permisos
Resum: L'administrador afegeix el nom de l'usuari (nom d'usuari UPCXXI), el correu electrònic si escau i el rol que li assigna a la intranet comuna. Si l'usuari pertany a una biblioteca concreta l'administrador introdueix també la biblioteca i el rol dins la biblioteca assignada. El sistema introdueix el nou usuari a la base de dades.
Tipus: primari i essencial

Cas d'ús: Editar usuari
Actors: Administrador(iniciador),Sistema
Propòsit: Editar el nom, el correu electrònic o la biblioteca assignada o el rol assignat d'un usuari.
Resum: L'administrador selecciona un usuari i el sistema mostra un formulari amb les dades assignades a aquest usuari. L'administrador canvia les dades que vol i accepta els canvis. El sistema guarda les noves dades.
Tipus: primari i essencial

Cas d'ús: Llistar usuaris
Actors: Administrador(iniciador),Sistema
Propòsit: Mostrar un llistat dels usuaris introduïts al sistema
Resum: A la intranet comuna mostra un llistat de tots els usuaris amb el seu correu, biblioteca assignada i el seu rol dins la intranet. A les intranets de les biblioteques mostra un llistat dels usuaris assignats a la biblioteca amb el seu rol dins la biblioteca i el seu correu electrònic.
Tipus: primari i essencial

Cas d'ús: Seleccionar responsables d'eix
Actors: Administrador(iniciador),Sistema
Propòsit: Escollir els responsables que rebran les sol·licituds fetes pels mantenidors de biblioteca per cada eix
Resum: El sistema mostra per cada eix el conjunt d'usuaris introduïts a la intranet i l'administrador selecciona els responsables de cada eix. Seguidament el sistema guarda la relació.
Tipus: primari i essencial

Gestió de l'arbre temàtic

Cas d'ús: Afegir tema
Actors: Administrador(iniciador),Sistema
Propòsit: Incloure un nou tema a l'arbre temàtic
Resum: El sistema mostra el llistat de temes on es pot afegir un nou tema (incloent l'arrel de l'arbre). L'usuari selecciona on vol penjar el nou tema i introdueix el nom, selecciona els eixos relacionats del llistat d'eixos i si és un tema de nivell 1 selecciona si vol que sigui tema clau.
Tipus: primari i essencial

Cas d'ús: Modificar tema
Actors: Administrador(iniciador),Sistema
Propòsit: Editar el nom d'un tema o els eixos amb els que està relacionat.
Resum: L'administrador selecciona el tema que vol editar del llistat complet de temes. Canvia el nom o els eixos relacionats. En cas que el tema sigui de nivell 1 també pot seleccionar si el tema és clau o no.
Tipus: primari i essencial

Cas d'ús: Moure tema
Actors: Administrador(iniciador),Sistema
Propòsit: Canviar de lloc un tema
Resum: L'administrador selecciona el tema que vol moure. El sistema mostra el llistat de temes (incloent l'arrel) on es pot moure el tema. L'administrador selecciona el tema destí del tema mogut.
Tipus: primari i essencial

Cas d'ús: Esborrar tema.
Actors: Administrador(iniciador),Sistema
Propòsit: Esborrar un tema de l'arbre temàtic
Resum: L'administrador selecciona el tema que vol esborrar. El sistema pregunta si està segur. Si respon que sí, el sistema passa a treure el tema de la base de dades. Si el tema té subtemes assignats el sistema dona la opció d'esborrar-los també o moure'ls cap altre tema. Si el tema té documents assignats esborra la relació entre el document i el tema.
Tipus: primari i essencial

Cas d'ús: Seleccionar temes clau
Actors: Administrador(iniciador),Sistema
Propòsit: Escollir els tems que es mostraran com a tema clau
Resum: El sistema mostra un llistat dels temes de nivell 1 i l'administrador marca els tems que vol marcar com tema clau. El sistema guarda la relació.
Tipus: primari i essencial

Cas d'ús: Visualitzar tots els temes
Actors: Administrador(iniciador),Sistema
Propòsit: Mostrar l'arbre temàtic completament desplegat
Resum: El sistema mostra tots els temes de l'arbre temàtic mostrant la jerarquia establerta.
Tipus: primari i essencial

Cas d'ús: Afegir eix
Actors: Administrador(iniciador),Sistema
Propòsit: Incloure un nou eix
Resum: L'administrador introdueix el nom de l'eix i el sistema l'inclou a la taula.
Tipus: primari i essencial

Cas d'ús: Modificar eix
Actors: Administrador(iniciador),Sistema
Propòsit: Modificar el nom d'un eix
Resum: L'administrador introdueix el nou nom de l'eix i el sistema el canvia a la base de dades.
Tipus: primari i essencial

Cas d'ús: Esborra eix
Actors: Administrador(iniciador),Sistema
Propòsit: Esborrar un conjunt d'eixos
Resum: L'administrador selecciona els eixos que vol esborrar i els sistema els treu de la base de dades.
Tipus: primari i essencial

Cas d'ús: Relacionar eixos i temes
Actors: Administrador(iniciador),Sistema
Propòsit: Seleccionar per un determinat eix els temes relacionats
Resum: L'administrador selecciona un eix i el sistema mostra tots els temes de l'arbre temàtic, marcant els temes que estan relacionats amb l'eix seleccionat. L'administrador marca o desmarca els temes que vol i accepta els canvis. El sistema guarda la relació entre l'eix i els temes que ha deixa't marcat l'administrador.
Tipus: primari i essencial

Cas d'ús: Mostrar temes relacionats eix
Actors: Administrador(iniciador), Sistema
Propòsit: Mostrar el conjunt de temes relacionats amb un determinat eix
Resum: L'usuari selecciona un eix i mostra els temes amb els està relacionat mostrant la jerarquia que existeix entre ells.
Tipus: primari i essencial

Sistema per l'extracció d' atributs del document i classificador temàtic

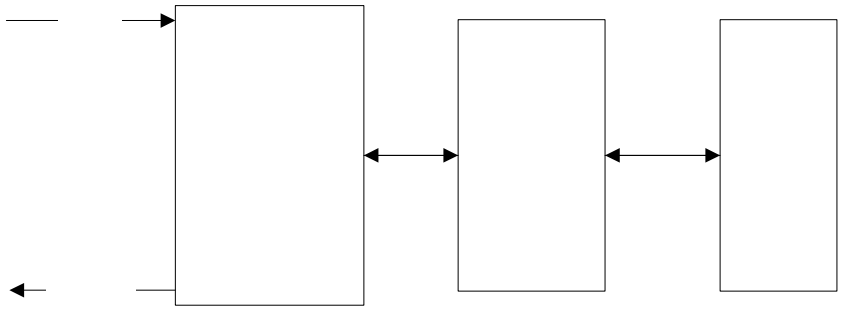
Cas d'ús: Afegir regles
Actors: Administrador(iniciador), Sistema
Propòsit: Modificar les regles que s'utilitzen per l'extracció d'atributs d'un enllaç
Resum: L'administrador prepara un arxiu de text amb un conjunt de regles amb el format establert. L'administrador introdueix l'arxiu al sistema i el sistema reemplaça les regles existents per les regles que llegeix de l'arxiu introduït.
Tipus: primari i essencial

Cas d'ús: Calcula pes paraula per tema
Actors: Administrador(iniciador), Sistema
Propòsit: Assignar un conjunt de paraules amb un pes determinat a cada tema de la intranet per utilitzar en la classificació temàtica dels documents
Resum: L'administrador demana al sistema fer un nou càlcul de la relació de paraula i tema. El sistema llegeix un conjunt dels documents que es troben a la intranet i a partir de les paraules que extreu de cada document i de la relació existent d'aquest amb els temes, fa els càlculs necessaris per guardar un llistat de paraules relacionades amb els temes de l'arbre temàtic amb un pes determinat.
Tipus: primari i essencial

Cas d'ús: Provar classificador
Actors: Administrador(iniciador), Sistema
Propòsit: Provar diferents paràmetres del classificador temàtic per afinar després i veure el comportament del classificador.
Resum: L'administrador introdueix els paràmetres que vol provar al sistema. El sistema utilitza els documents que no s'han fet servir per fer els càlculs de la relació entre paraules i temes per comprovar mitjançant dos paràmetres (*precision i recall*) l'efectivitat del classificador.
Tipus: primari i essencial

Model Arquitectònic

El disseny de l'arquitectura s'ha realitzat en **2 capes**. En la primera, les pàgines Asp o Scripts reben les peticions de l'usuari i generen les vistes a partir de les dades que demanen als components del domini. Aquests components, s'encarreguen de fer les peticions al SGBD que guarda les dades. D'aquesta manera, s'aconsegueix separar el codi Html retornat, és a dir la visualització, de la gestió de les dades.



Orientació a objectes

Encara que el llenguatge amb el que s'ha desenvolupat la intranet no està dissenyat plenament per fer servir la orientació a objectes ja que no que disposa de moltes característiques d'aquest paradigma de programació, com pot ser la herència, s'utilitzarà aquesta paradigma perquè ens permet separar les funcionalitats del sistema afavorint així la reusabilitat i cohesió del nostre sistema.

Per poder obtenir o guardar les dades de cada objecte del domini que ho requereixi, es farà que els objectes del domini disposin de dues funcions bàsiques:

- Carrega(rs): que rep les dades provinents d'una consulta a la base de dades i carrega els atributs de l'objecte corresponent.
- Desmonta(): que guarda les dades de l'objecte corresponent o l'introdueix a la base de dades assignant un identificador.

Petició

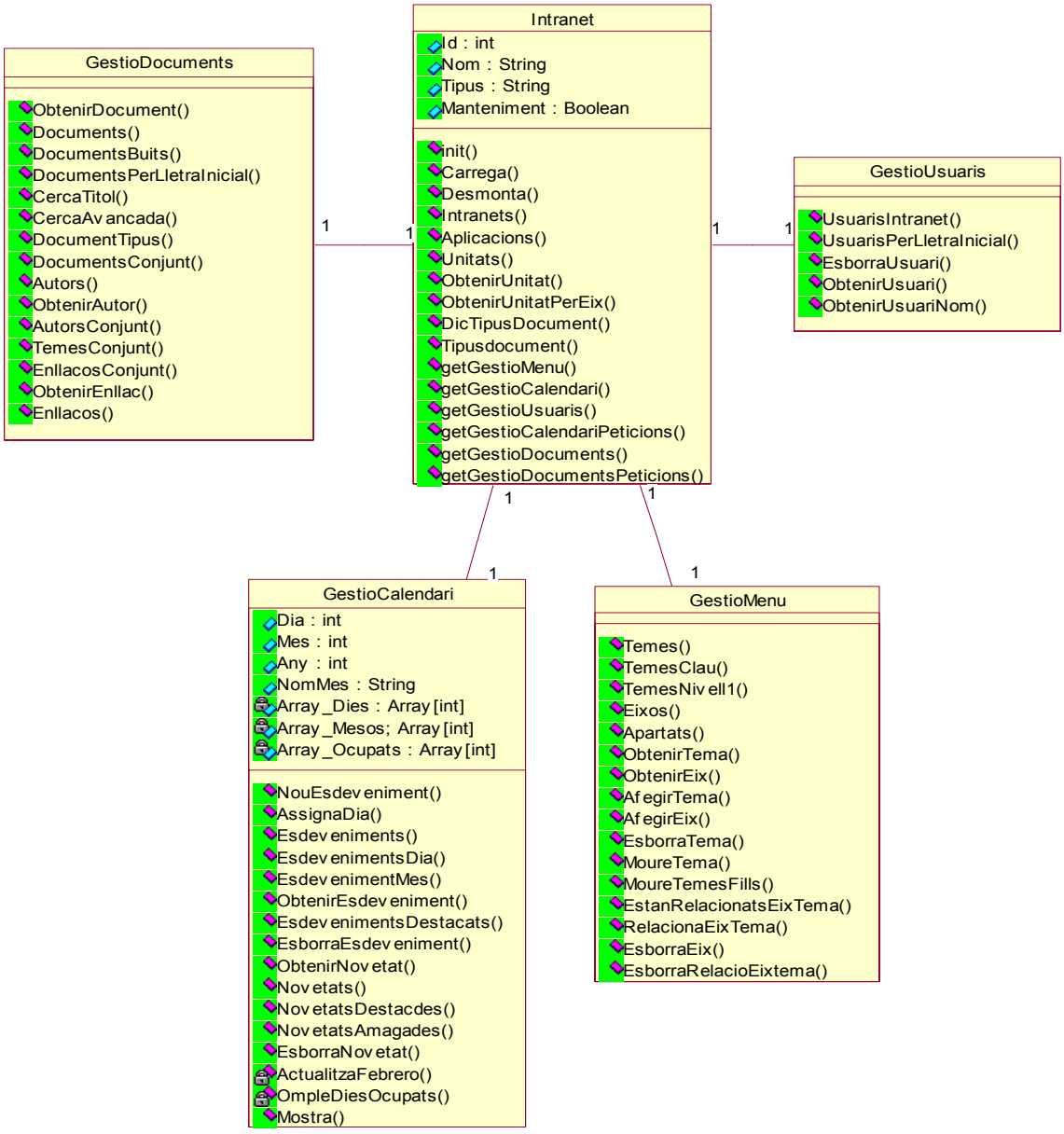
Model Conceptual

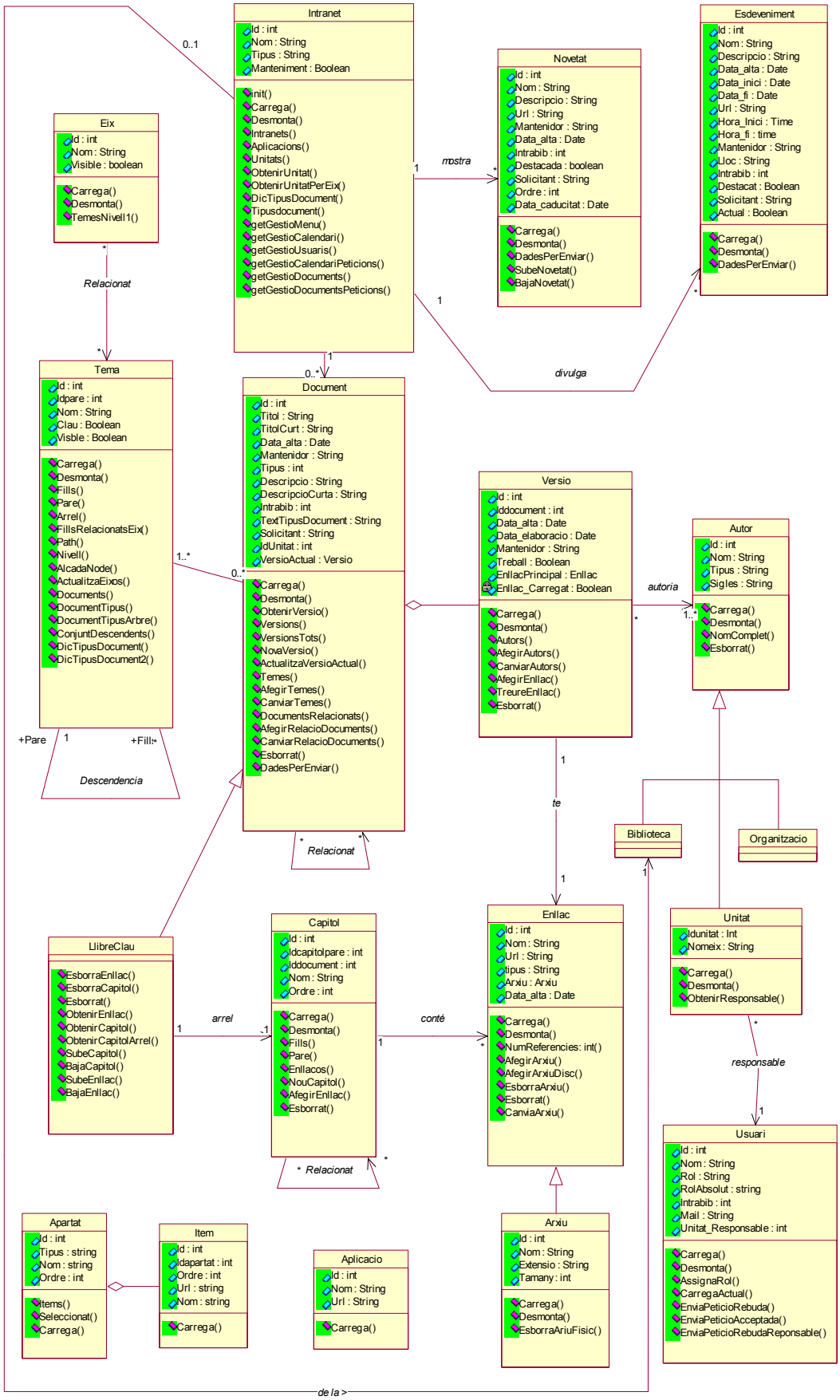
El model conceptual s'ha separat en tres parts:

1) Les classes per la gestió del quatre àmbits que ocupa la intranet més la classe intranet:
    - GestioMenu: gestiona els objectes del menú incloent els temes i eixos
    - GestioDocuments: gestiona els documents i els objectes relacionats
    - GestioUsuaris: gestiona els usuaris de la intranet i els permisos d'aquests
    - GestioCalendari: gestiona els esdeveniments i novetats de la intranet
2) La resta de classes del domini que utilitzen aquestes quatre classes.
3) Les classes especifiques del sistema d'IA

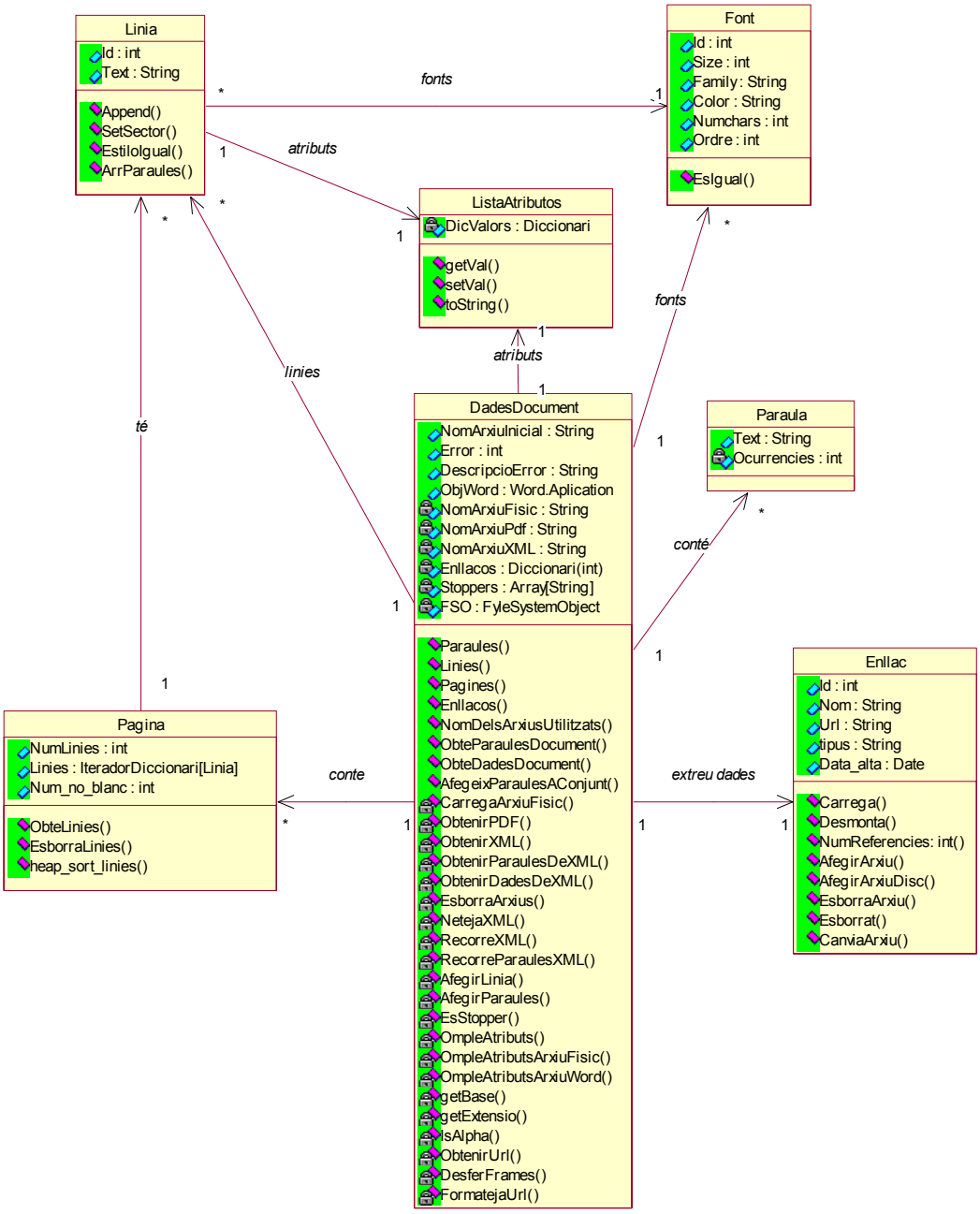
Objectes de gestió

Per separar les àrees que tracta cada intranet, s'ha decidit dividir les funcions generals de cada àmbit en objectes de gestió deixant la resta de funcions a la intranet. Aquests objectes fan el paper de diccionari dels objectes que gestionen i donen opcions per editar-los.

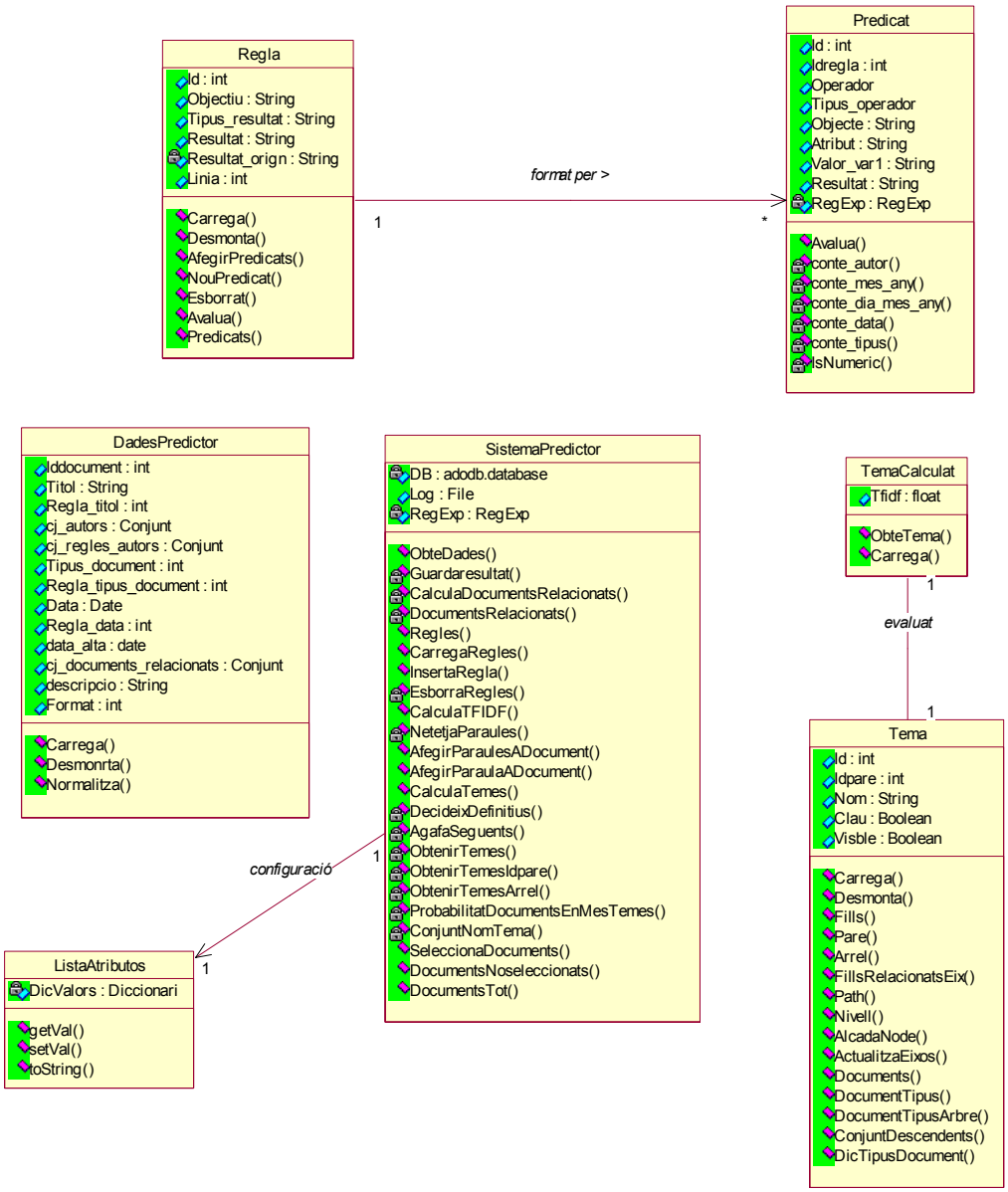



Intranet
Id : int
Nom : String
Tipus : String
Manteniment : Boolean
init()
Carrega()
Desmonta()
Intranets()
Aplicacions()
Unitats()
ObtenirUnitat()
ObtenirUnitatPerEix()
DicTipusDocument()
Tipusdocument()
getGestioMenu()
getGestioCalendari()
getGestioUsuaris()
getGestioCalendariPeticions()
getGestioDocuments()
getGestioDocumentsPeticions()
0..1
1
mostra
*
Novetat
Id : int
Nom : String
Descripcio : String
Url : String
Mantenidor : String
Data_alta : Date
Intrabib : int
Destacada : boolean
Solicitant : String
Ordre : int
Data_caducitat : Date
Carrega()
Desmonta()
DadesPerEnviar()
SubeNovetat()
BajaNovetat()
Esdeveniment
Id : int
Nom : String
Descripcio : String
Data_alta : Date
Data_inici : Date
Data_fi : Date
Url : String
Hora_Inici : Time
Hora_fi : time
Mantenidor : String
Lloc : String
Intrabib : int
Destacat : Boolean
Solicitant : String
Actual : Boolean
Carrega()
Desmonta()
DadesPerEnviar()
1
divulga
*
Eix
Id : int
Nom : String
Visible : boolean
Carrega()
Desmonta()
TemesNivell1()
*
Relacionat
*
Tema
Id : int
Idpare : int
Nom : String
Clau : Boolean
Visble : Boolean
Carrega()
Desmonta()
Fills()
Pare()
Arrel()
FillsRelacionatsEix()
Path()
Nivell()
AlcadaNode()
ActualitzaEixos()
Documents()
DocumentTipus()
DocumentTipusArbre()
ConjuntDescendents()
DicTipusDocument()
DicTipusDocument2()
1..*
0..*
+Pare
1
+Fill
*
Descendencia
1
0..*
Document
Id : int
Titol : String
TitolCurt : String
Data_alta : Date
Mantenidor : String
Tipus : int
Descripcio : String
DescripcioCurta : String
Intrabib : int
TextTipusDocument : String
Solicitant : String
idUnitat : int
VersioActual : Versio
Carrega()
Desmonta()
ObtenirVersio()
Versions()
VersionsTots()
NovaVersio()
ActualitzaVersioActual()
Temes()
AfegirTemes()
CanviarTemes()
DocumentsRelacionats()
AfegirRelacioDocuments()
CanviarRelacioDocuments()
Esborrat()
DadesPerEnviar()
*
Relacionat
*
Versio
Id : int
Iddocument : int
Data_alta : Date
Data_elaboracio : Date
Mantenidor : String
Treball : Boolean
EnllacPrincipal : Enllac
Enllac_Carregat : Boolean
Carrega()
Desmonta()
Autors()
AfegirAutors()
CanviarAutors()
AfegirEnllac()
TreureEnllac()
Esborrat()
*
autoria
1..*
Autor
Id : int
Nom : String
Tipus : String
Sigles : String
Carrega()
Desmonta()
NomComplet()
Esborrat()
1
te
1
Biblioteca
Organitzacio
1
LlibreClau
EsborraEnllac()
EsborraCapitol()
Esborrat()
ObtenirEnllac()
ObtenirCapitol()
ObtenirCapitolArrel()
SubeCapitol()
BajaCapitol()
SubeEnllac()
BajaEnllac()
1
arrel
1..1
Capitol
Id : int
Idcapitolpare : int
Iddocument : int
Nom : String
Ordre : int
Carrega()
Desmonta()
Fills()
Pare()
Enllacos()
NouCapitol()
AfegirEnllac()
Esborrat()
1
conté
*
* Relacionat
*
Enllac
Id : int
Nom : String
Url : String
tipus : String
Arxiu : Arxiu
Data_alta : Date
Carrega()
Desmonta()
NumReferencies: int()
AfegirArxiu()
AfegirArxiuDisc()
EsborraArxiu()
Esborrat()
CanviaArxiu()
Unitat
Idunitat : Int
Nomeix : String
Carrega()
Desmonta()
ObtenirResponsable()
*
responsable
1
Usuari
Id : int
Nom : String
Rol : String
RolAbsolut : string
Intrabib : int
Mail : String
Unitat_Responsable : int
Carrega()
Desmonta()
AssignaRol()
CarregaActual()
EnviaPeticioRebuda()
EnviaPeticioAcceptada()
EnviaPeticioRebudaReponsable()
Apartat
Id : int
Tipus : string
Nom : string
Ordre : int
Items()
Seleccionat()
Carrega()
Item
Id : int
Idapartat : int
Ordre : int
Url : string
Nom : string
Carrega()
Aplicacio
Id : int
Nom : String
Url : String
Carrega()
Arxiu
Id : int
Nom : String
Extensio : String
Tamany : int
Carrega()
Desmonta()
EsborraAriuFisic()
de la >

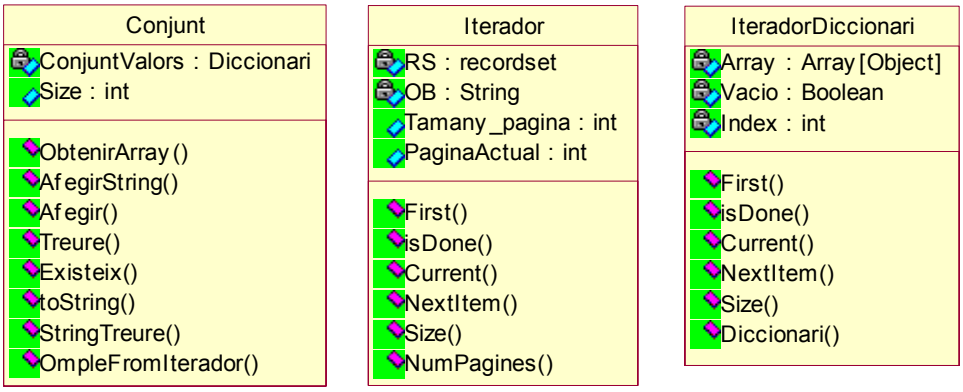
Aquests són els objectes que s'utilitzen per fer el preprocés d'un enllaç determinat. En aquest model, cal destacar l'objecte *LlistaAtributs* que conté les dades que després es faran servir per avaluar les regles tan per les línies del document com per les dades generals i que permet afegir tants atributs com es vulgui ja que està implementat amb un diccionari que guarda tuples (atribut,valor). A més també cal destacar la classe Paraula, on es guarden el número d'ocurrències de cada paraula de l'enllaç. La resta (*Font* i *Pagina*) són objectes que només s'utilitzen durant la lectura del XML de forma temporal.



Aquestes són la resta de classes que utilitza el sistema per introduir documents automàticament. La classe *SistemaPredictor* és el controlador del procés i utilitza les classes de *Regla*, *Predicat* i *TemaCalculat* per obtenir els resultats. A més fa servir la classe *DadesPredictor* com un DTO per guardar les dades que s'extreuen i les emmagatzemar a la base de dades. Per últim, es torna a utilitzar la classe *LlistaAtributs* per guarda la configuració del *SistemaPredictor*.



Finalment, es mostra tres classes que es faran servir per manipular dades. Concretament es tracta de dos Iteradors, el primer obté les dades d'un objecte RecordSet de la base de dades creant objectes del tipus assignat a la variable OB i el segon obté les dades d'un diccionari assignat prèviament. L'altre classe és un conjunt que guarda identificadors d'objectes i que està implementat fent servir un diccionari (objecte natiu d'ASP).



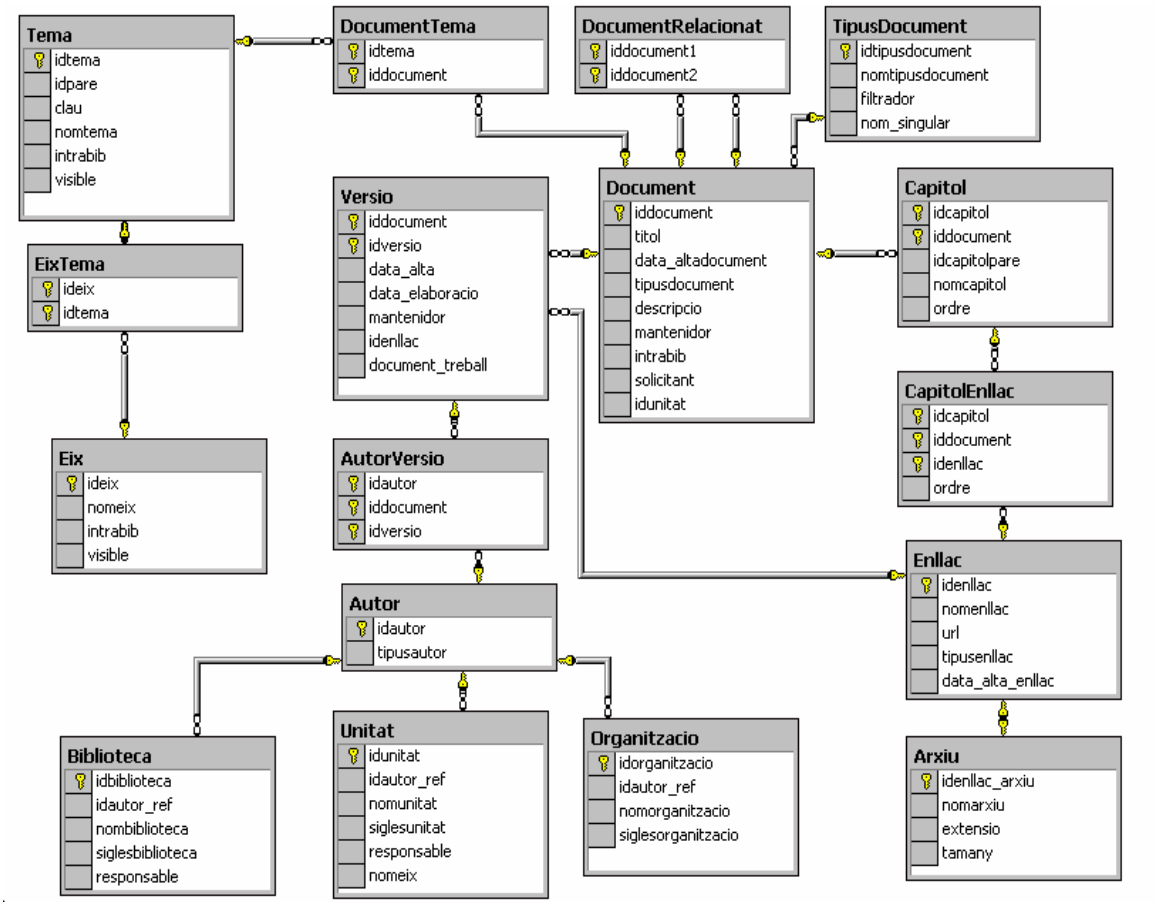
Model Relacional

A partir de les classes del model conceptual, s'obtenen el següent model relacional.
Per mostrar el resultat s'ha utilitzat tres diagrames generats per l' SQL Server:

- Taules document i relacionats
- Taules de la resta de classes de la intranet
- Taules utilitzades per la classificació automàtica de documents

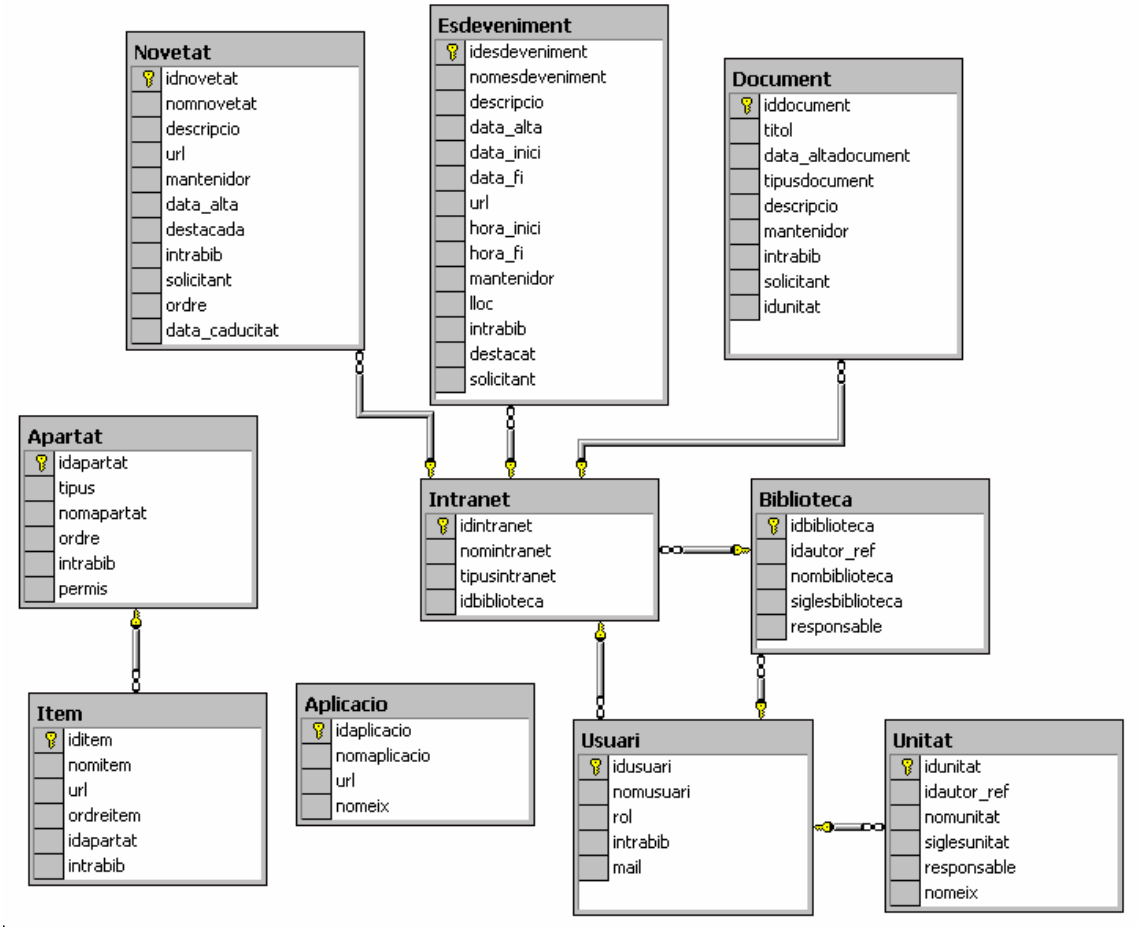
Taula document i relacionades

Com es pot veure de la traducció del model conceptual al model relacional han sorgit noves taules per traduir relacions n-àries. És el cas de les taules *EixTema*, *CapitolEnllac*, *AutorVersio*, *DocumentTema* i *DocumentRelacionat*. A més s'ha inclòs una taula *Tipusdocument* per guardar els valors que pot tenir aquest atribut del document.
La resta d'associacions s'han resolt fent servir alguna de les taules relacionades, cal destacar sobretot les relacions que guarden una jerarquia, com el cas del temes o els capítols del llibre clau que s'han resolt fent servit un nou atribut idpare que fa referència al tema o capítol ascendent.
Finalment, les dos especialitzacions del model conceptual, el llibre clau i els autors s'han resolt de dos maneres diferents. En el cas del llibre clau no s'ha materialitzat, i únicament hem inclòs les taules relacionades dels capítols i els enllaços. Pel que fa a l'autor, hem inclòs una referència a la taula autor des de les seves especialitzacions.



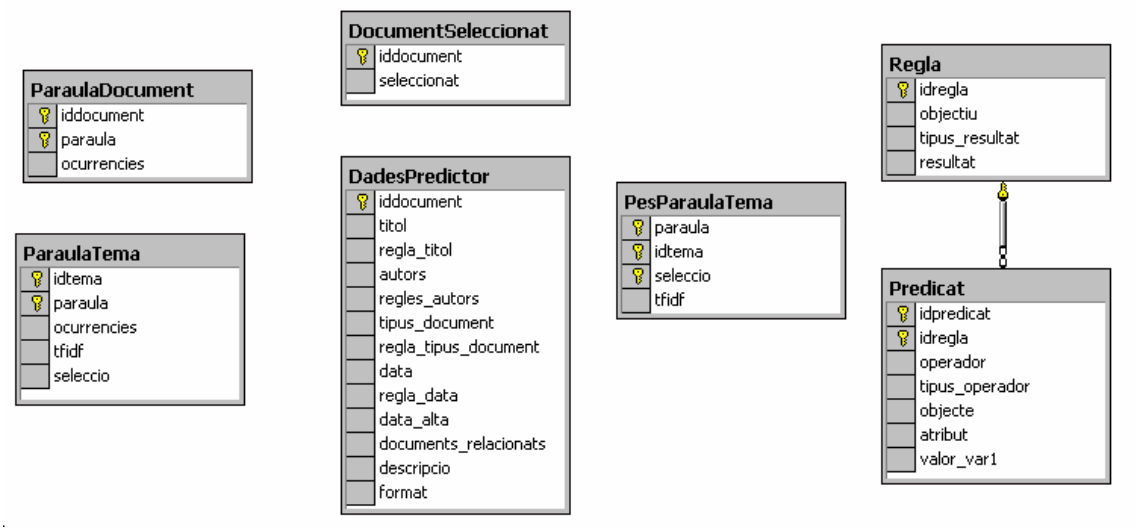
Resta del diagrama

Aquí podem veure la resta del model relacional. Hi ha tres taules repetides respecte l'anterior diagrama per mostrar la relació que la taula *Usuari* amb la *Biblioteca* a la que pertany i la *Unitat* de la que és responsable a més de la taula *Document* i la relació amb la intranet on es troba. També es pot veure dos taules no relacionades amb la taula Intranet, la taula *Aplicació* on és guarden les aplicacions accessibles des de la intranet i les taules *Apartat* i *Ítem* on es guarden els menús de la intranet.



Taules utilitzades en la classificació automàtica

Les taules *ParaulaDocument* i *ParaulaTema* són taules que s'utilitzen temporalment per calcular el pes de les paraules per cada tema. La taula *DocumentSeleccionat* s'utilitza per separar els documents que són per fer proves dels que utilitzem per seleccionar les paraules. La taula *DadesPredictor* s'utilitzarà per guardar les dades extretes d'un determinat document, per després poder extreure conclusions. Finalment, les tres últimes taules, *Regla*, *Predicat* i *PesParaulaTema* són les que es fan servir per extreure els atributs de la fitxa d'un document.

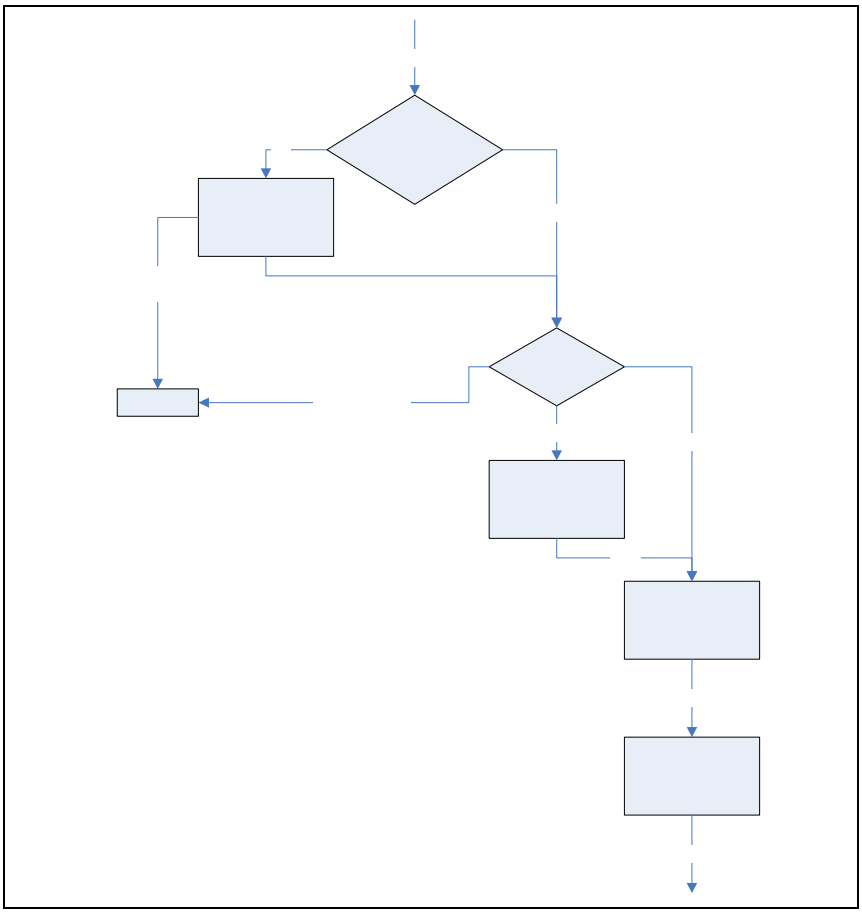


Anàlisi i disseny del sistema d'intel·ligència artificial

Com s'ha dit anteriorment, un dels objectius del projecte és construir un sistema que, donat l'enllaç d'un document, extregui del text els atributs de la fitxa i el classifiqui dins l'arbre temàtic de la intranet. En aquest apartat, volem mostrar les passes adoptades per la construcció d'aquest sistema tan pel que fa a l'extracció d'atributs, com la classificació i el preprocés previ dels enllaços.

Preprocés dels documents

A la nova intranet es podran penjar documents en diversos formats tot i que la majoria estaran en format Word, PDF o PowerPoint. A més, també es podran penjar enllaços a documents que es troben fora del servidor. És per això, que cal definir un preprocés dels enllaços per poder extreure la informació en un format homogeni a l'hora d'aplicar els algoritmes que s'han establert. Aquests són els passos adoptats en el preprocés:



Taula 1: Esquema del preprocés

Obtenir arxiu físic

Si el document es troba en un altre servidor caldrà baixar l'arxiu en una carpeta temporal. Aquesta tasca només es podran realitzar si el protocol de transferència que utilitza el servidor per servir aquest document és HTTP. Si es tracta d'un protocol diferent (com l'HTTPS) es cancel·la el preprocés.

Per altra banda, cal tenir en compte el format del document que ens baixem, ja que molts d'ells poden ser documents Html generats dinàmicament amb una extensió diferent a Html. Finalment, si es tracta d'un document Html s'haurà de vigilar que el document no contingui Frames ja que es podrien baixar arxius sense contingut. Si és aquest el cas, cal baixar l'últim frame amb contingut.

Convertir a PDF

Si el document no està en format PDF mitjançant una eina de què disposa el Servei de Biblioteques es convertirà l'arxiu. Aquesta eina, anomenada autoPDFmatic, forma part d'un projecte final de carrera i s'utilitza per transformar arxius a PDF d'una manera automàtica des d'un servidor que serveix pàgines ASP. A la taula 2 es pot veure els formats convertibles.

Format / extensió
Text .txt, .ans, .asc
Rich text format .rtf
MS Word .doc
MS WordPerfect .wpd
MS Works .wps
MS Excel .xl, .xls, .xlc
MS PowerPoint .ppt, .pps
PostScript .ps, .eps
Html .html, .htm

Taula 2: Formats covertibles

Lògicament, si el format del document no és PDF ni és cap d'aquests formats especificats, no es continuarà el preprocés, encara que això passarà poques vegades ja que la majoria de documents penjats a la intranet són convertibles.

Convertir a XML

Per finalitzar el preprocés, es farà servir un executable anomenat pdftohtml.exe que s'ha extret de la pàgina *http://sourceforge.net/projects/pdftohtml/* i que forma part d'un projecte de software lliure que es basa en el paquet xpdf. Aquesta projecte ofereix eines per extreure el text que contenen els documents PDF en format Html o en un format XML determinat. En concret, es farà servir el format XML, ja que aporta la informació que volem estructurada i fàcil de processar. El XML organitza el document en les diferents pàgines que conté, que a la vegada estan organitzades en les línies de text que podem trobar dins. A més, per cada línia de text, el document XML li assigna una conjunt d'atributs per conèixer la posició i el format de lletra en que apareixen. A la taula 3 es pot veure el DTD que compleixen aquests arxius XML i a la taula 4 un exemple d'un document processat.


```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE pdf2xml SYSTEM "pdf2xml.dtd">

<pdf2xml>
<page number="1" position="absolute" top="0" left="0" height="918" width="1188">
        <fontspec id="0" size="12" family="Times" color="#000000"/>
        ...
        <fontspec id="8" size="28" family="Times" color="#800000"/>
        <fontspec id="9" size="19" family="Times" color="#800080"/>
<text top="130" left="106" width="216" height="60" font="1">PAIDEIA</text>
<text top="194" left="106" width="128" height="34" font="2">"...llegeix,</text>
<text top="194" left="235" width="8" height="34" font="3"> </text>
<text top="194" left="243" width="105" height="34" font="4">estudia,</text>
<text top="194" left="348" width="8" height="34" font="5"> </text>
<text top="194" left="357" width="85" height="34" font="6">aprèn,</text>
<text top="194" left="442" width="8" height="34" font="5"> </text>
<text top="194" left="450" width="145" height="34" font="7">connecta't,</text>
<text top="194" left="595" width="8" height="34" font="5"> </text>
<text top="194" left="603" width="137" height="34" font="8">investiga!"</text>
<text top="416" left="106" width="705" height="60" font="1">Pla de formació del Servei de</text>
<text top="478" left="106" width="682" height="60" font="1">Biblioteques i Documentació</text>
<text top="727" left="106" width="463" height="24" font="9">SERVEI DE BIBLIOTEQUES I
DOCUMENTACIÓ</text>
<text top="752" left="106" width="452" height="24" font="9">UNIVERSITAT POLITÈCNICA DE
CATALUNYA</text>
<text top="848" left="1073" width="8" height="17" font="0">1</text>
</page>
<page number="2" position="absolute" top="0" left="0" height="918" width="1188">
...
</page>
</pdf2xml>
```

Taula 3: Exemple de document generat

```
<?xml version="1.0"?>
<!ELEMENT pdf2xml (page+)>
<!ELEMENT page (fontspec*, text*)>
<!ATTLIST page
        number CDATA #REQUIRED
        position CDATA #REQUIRED
        top CDATA #REQUIRED
        left CDATA #REQUIRED
        height CDATA #REQUIRED
        width CDATA #REQUIRED
>
<!ELEMENT fontspec EMPTY>
<!ATTLIST fontspec
        id CDATA #REQUIRED
        size CDATA #REQUIRED
        family CDATA #REQUIRED
        color CDATA #REQUIRED
>

<!ELEMENT text (#PCDATA | b | i)*>
<!ATTLIST text
        top CDATA #REQUIRED
        left CDATA #REQUIRED
        width CDATA #REQUIRED
        height CDATA #REQUIRED
        font CDATA #REQUIRED
>
<!ELEMENT b (#PCDATA)>
<!ELEMENT i (#PCDATA)>
```

Taula 4: DTD del XML generat

Obtenir dades

Una vegada s'ha obtingut l'arxiu XML es disposa a calcular i emmagatzemar les dades que es faran servir per l'extracció d'atributs del document. En aquest procés es llegirà l'arxiu XML mitjançant un objecte DOM que permet recorre l'arbre d'etiquetes i s'omplirà l'objecte *DadesDocument* utilitzat per emmagatzemar la informació que es considera interessants. Tota aquest informació, tret del conjunt de paraules per la classificació temàtica del document, només s'extraurà de la primera pàgina del XML, ja que la resta no conté, normalment, les dades que volem extreure.

Una vegada finalitzat el preprocés del document es disposa de les següents dades per fer l'extracció d'atributs i la classificació temàtica:

1. Conjunt d'atributs del document (veure taula 5).

   Aquest atributs s'han escollit observant característiques dels documents penjats que es podien relacionar amb els atributs del document que volem extreure. La majoria tenen una relació amb l'atribut tipus de document. Per exemple, l'extensió pot ajudar a saber si el document és una presentació, o el número de paraules pot indicar si es tracta d'una plantilla de document, o el *títol* i el *nom_arxiu* pot contenir el nom del tipus del document. Cal destacar també l'atribut *estandard* que indica si la primera pàgina del document segueix un patró típic on apareixen el títol, la data d'elaboració i els autors del document en llocs molt concrets.

Atribut	Descripció
Estandard	Indica si la primera pàgina té un format típic en els documents del Servei.
Es_url	Indica si l'arxiu es trobava inicialment al servidor o és un arxiu extern.
Nom_arxiu	Nom de l'arxiu.
Extensio	Extensió de l'arxiu.
Num_paraules	Nombre de paraules que conté el document.
Num_numeros	Nombre de números que conté el document.
Títol	Títol extret del document (amb les regles que s'explicaran més endavant)

Taula 5: Atributs de l'enllaç

2. Conjunt de línies de text del document amb atributs assignats a cada una. En el preprocés les línies que hem obtingut de la primera pàgina de l'arxiu XML s'hauran de concatenar en el cas que tinguin el mateix format i siguin correlatives. A més, en alguns casos caldrà ordenar-les perquè el XML resultant no sempre està ordenat. A la taula 6 podem veure el llistat d'atributs que s'extreu per cada línia.

La majoria dels atributs s'ha escollit perquè ajuden a trobar la línia de text que és el títol del document. És el cas d'atributs com T_font, Pos_in_pag o ordre_tamany. Finalment, hi ha altres atributs que fan referència a la posició de la línia de text en la pàgina. És el cas de sector, pos_in_pag i pos_in_pag_invert. Aquests atributs s'han escollit perquè els atributs que es busquen, acostumen a estar en zones determinades de la pàgina, a l'inici o al final, o a la meitat de la pàgina en el cas del títol. Altres atributs no han estat molt rellevants, com el cas de negreta, però s'han mantingut per si després es volien utilitzar.

Atribut	Descripció
Text	Contingut de la línia.
Sector	Dividint la pàgina en 4 sector per alçades. Número de sector on apareix la línia.
Pos_in_pag	Posició que ocupa la línia en relació a la resta de línies de la pàgina.
Pos_in_pag_invert	Posició que ocupa la línia en relació a la resta de línies de la pàgina.
Longitud	Longitud del text.
T_font	Mida de la font de la línia de text.
Ordre_tamany	Posició que ocupa la mida de la font de la línia ordenant de més gran a més petit.
Pagina	Número de pàgina on apareix la línia.
Negreta	Indica si la línia de text està en negreta
Left	Posició horitzontal del text.
Top	Posició vertical del text.

Taula 6: Atributs de la línia de text

3. Conjunt de paraules que apareixen en el document amb el número d'aparicions per cada una d'elles (sense contemplar les paraules que s'han considerat *StopWords*).
4. Conjunt d'enllaços extrets. Si l'arxiu penjat està en format Word es podrà extreure fàcilment el conjunt d'enllaços que conté aquest arxiu.

Extracció d'atributs

Una vegada s'han obtingut les dades de l'enllaç, el següent pas serà extreure els atributs de la fitxa del document. Aquests atributs són:

1. Títol
2. Data d'elaboració
3. Autors
4. Tipus de document
5. Documents relacionats

Per l'extracció dels quatre primers es farà servir un conjunt de regles que prèviament s'haurà inserit al sistema mitjançant un Script. Aquestes regles es podran reemplaçar per noves per millorar el procés d'extracció d'atributs. L'últim atribut, documents relacionats, s'extraurà dels enllaços que conté l'arxiu físic.

Format de les regles

Les regles que s'utilitzaran per l'extracció de dades hauran de complir un format per tal de que el sistema les processi i les pugui inserir. Per cada regla s'hauran de determinar 3 parts: l'objectiu de la regla, el conjunt de predicats que ha de satisfer la regla i el resultat que ha de retornar quan la regla sigui avaluada com a certa.

Aquesta és l'expressió regular de les regles:

O : P (and P)* -> R

On O és l'objectiu de la regla
P és un predicat
R és el resultat que ha de retornar la regla

Com es pot veure a la fórmula, el que ha de complir la regla sempre és una conjunció de predicats.

Seguidament, s'explica més detalladament cada una de les parts:

Objectiu

L'objectiu designa per quin dels atributs que ves vol extreure s'aplica la regla.
El valors possibles són: autor, titol, data, tipus_document.

Conjunt de predicats

El conjunt de predicats són un conjunt de funcions o expressions booleanes que, en alguns casos, obtenen un resultat per la regla. Hi ha tres tipus de predicats en funció del tipus d'operant i el nombre de paràmetres:

1. Expressions booleanes per comprovar si un atribut compleix una certa condició.
2. Funció booleana d'un paràmetre.
3. Funció booleana de dos paràmetres.

Aquest és el format dels tres tipus de predicats:

1	(atribut OP constant)
2	(nom_funcio(atribut))
3	(nom_funcio(atribut,constant))

Taula 7: Format dels predicats

On OP és qualsevol operador de comparació que accepti el llenguatge ASP
atribut fa referència a alguna de les dades emmagatzemades en el preprocés
constant és un valor numèric o una paraula (sense espais)
nom_funcio és el nom d'alguna funció definida a la classe Predicat

Pel que fa als atributs hi ha de dos tipus en funció de si pertanyen al document o a una línia de text del document. Si es tracta d'un atribut del document caldrà posar el nom de l'atribut i, si es tracta d'un atribut d'una línia, caldrà posar *linia.* més el nom de l'atribut de la línia. Aquest atributs estan descrits en l'apartat anterior: el preprocés.

Les funcions d'un o dos paràmetres estan implementades a la classe Predicat. Moltes d'elles estan dissenyades per trobar patrons per un atribut concret i, per tant, també guarden el valor de l'atribut trobat. Aquest és el llistat de funcions disponibles:

Funció	Descripció
conte_en(txt,par)	Comprova si la paraula par es troba dins de txt.
conte_autor(atr)	Comprova si al paràmetre atr es troba el nom complet d'algun dels autors guardats a la base de dades. Si en troba algun, es guarda els identificadors i retorna cert.
conte_mes_any(atr)	Comprova si el paràmetre atr conté el nom d'un mes seguit d'un any. Si el troba es guarda la data del primer dia del mes trobat i retorna cert.
conte_dia_mes_any(atr)	Comprova si el paràmetre atr conté el número de dia seguit del nom d'un mes seguit d'un any. Si el troba es guarda la data trobada i retorna cert.
conte_data(atr)	Comprova si el paràmetre atr conté una data en format dd/mm/aaaa o similars. Si la troba la guarda i retorna cert.
conte_tipus(atr)	Comprova si el paràmetre atr conté el nom d'algun tipus de document guardat a la base de dades. Si el troba es guarda l'identificador i retorna cert.

Taula 8: Funcions definides a la classe Predicat

Si es creu necessari és pot augmentar aquest llistat afegint funcions amb les característiques descrites a la classe Predicat.

Resultats que retorna la regla

Aquest es l'últim component de la regla. Pot ser de dos tipus:

1. Regles que retornen un valor constant. En aquest cas el format del resultat de la regla seria *directe [#valor de la constant#]*.
2. Regles que retornen el resultat generat per l'avaluació d'algun dels predicats de la regla. En aquest cas s'expressa amb la paraula predicat, seguit del número de predicat que es vol retornar, ordenant de 1 a *#predicats* d'esquerra a dreta.

```
//Regles de titol
titol: (estandard = 1) and (linia.sector > 1) and (linia.sector < 4) and (linia.longitud < 200) and (linia.pagina = 1) -> linia.text
titol: (estandard = 0) and (linia.ordre_tamany = 1) and (linia.longitud < 200) and (linia.pos_in_pagina < 4) and (linia.pagina = 1) -> linia.text
titol: (estandard = 0) and (linia.ordre_tamany = 2) and (linia.longitud < 200) and (linia.pos_in_pagina < 4) and (linia.pagina = 1) -> linia.text
titol: (estandard = 0) and (linia.longitud < 200) and (linia.pos_in_pagina < 4) and (linia.pagina = 1) -> linia.text

//Regles d'autor
autor: (estandard = 1) and (linia.sector > 2) and (conte_autor(linia.text)) -> predicat 3
autor: (estandard = 1) and (linia.sector = 1) and (conte_en(linia.text,recerca)) -> directe 3
autor: (estandard = 1) and (linia.sector = 1) and (conte_en(linia.text,aprenentatge)) -> directe 2
autor: (estandard = 1) and (linia.sector = 1) and (conte_en(linia.text,xarxa)) -> directe 4
autor: (estandard = 1) and (linia.sector = 1) and (conte_en(linia.text,organitzacio)) -> directe 5
autor: (estandard = 0) and (linia.pos_in_pagina_invert < 4) and (conte_autor(linia.text)) and (linia.longitud < 200) -> predicat 3
autor: (estandard = 0) and (linia.pos_in_pagina < 4) and (conte_autor(linia.text)) and (linia.longitud < 200) -> predicat 3
autor: (estandard = 0) and (tipus_document = 1) and (conte_en(titol,recerca)) -> directe 3
autor: (estandard = 0) and (tipus_document = 1) and (conte_en(titol,aprenentatge)) -> directe 2
autor: (estandard = 0) and (tipus_document = 1) and (conte_en(titol,xarxa)) -> directe 4
autor: (estandard = 0) and (tipus_document = 1) and (conte_en(titol,digital)) -> directe 4
autor: (estandard = 0) and (tipus_document = 1) and (conte_en(titol,servei)) and (conte_en(titol,obtenció)) and (conte_en(titol,documents)) -> directe 3
autor: (estandard = 0) and (tipus_document = 1) and (conte_en(titol,revistes)) -> directe 3
autor: (estandard = 0) and (tipus_document = 1) and (conte_en(titol,temàtic)) -> directe 3
autor: (estandard = 0) and (tipus_document = 1) and (conte_en(titol,temàtiques)) -> directe 3
autor: (estandard = 0) and (tipus_document = 1) and (conte_en(titol,catalogació)) -> directe 2
autor: (estandard = 0) and (tipus_document = 1) and (conte_en(titol,adquisicions)) -> directe 2
autor: (estandard = 0) and (tipus_document = 1) and (conte_en(titol,préstec)) -> directe 2
autor: (estandard = 0) and (tipus_document = 1) and (conte_en(titol, SOD)) -> directe 3
autor: (estandard = 0) and (tipus_document = 1) and (conte_en(titol,caps)) -> directe 0

//Regles per la data
data: (estandard = 1) and (linia.sector = 4) and (conte_mes_any(linia.text)) -> predicat 3
data: (conte_data(titol)) -> predicat 1
data: (estandard = 0) and (linia.sector = 1) and (conte_data(linia.text)) -> predicat 3
data: (estandard = 0) and (linia.pos_in_pagina_invert < 4) and  (conte_data(linia.text)) -> predicat 3
data: (estandard = 0) and (linia.sector = 1) and (conte_dia_mes_any(linia.text)) -> predicat 3
data: (estandard = 0) and (linia.pos_in_pagina_invert < 4) and  (conte_dia_mes_any(linia.text)) -> predicat 3
data: (estandard = 0) and (linia.sector = 1) and (conte_mes_any(linia.text)) -> predicat 3
data: (estandard = 0) and (linia.pos_in_pagina_invert < 4) and  (conte_mes_any(linia.text)) -> predicat 3
data: (estandard = 0) and (linia.sector = 4) and (conte_data(linia.text)) -> predicat 3

//Tipus document
tipus_document: (extensio = ppt) and (num_paraules > 10) -> directe 6
tipus_document: (num_paraules < 9) -> directe 13
tipus_document: (conte_en(titol,reunió)) and (estandard = 0) -> directe 1
tipus_document: (conte_en(nom_arxiu,acta)) and (estandard = 0)  -> directe 1
tipus_document: (conte_en(nom_arxiu,plantilla)) -> directe 13
tipus_document: (conte_en(titol,plantilla)) -> directe 13
tipus_document: (conte_en(nom_arxiu,reunio)) and (estandard = 0) -> directe 1
tipus_document: (conte_tipus(titol)) -> predicat 1
tipus_document: (conte_tipus(nom_arxiu)) -> predicat 1
```

Taula 9: Conjunt de regles carregades inicialment

Avaluació de les regles

Com es pot veure al format de les regles hi ha dos tipus de predicats. Uns fan referència a atributs generals del document, com pot ser l'extensió de l'arxiu i altres fan referència a atributs de les línies que s'han seleccionat prèviament. Per fer una avaluació de les regles més eficients, primer es mirarà si el document compleix tots els predicats del document i després es comprova si existeix alguna línia del document que compleixi la resta de predicats.

L'ordre en què s'insereixen les regles és molt important ja que per les regles del títol, data i tipus de document el sistema acceptarà el resultat que obtingui la primera regla que sigui avaluada com a certa. En canvi, per les regles que tenen per objectiu l'autor el sistema acceptarà totes les regles que s'avaluïn com a certes.

Així mateix, també apliquem un ordre a l'hora de fer l'extracció dels quatre atributs:

1. Títol
2. Data d'elaboració
3. Tipus de document
4. Autors

Això permet, utilitzar l'atribut extret a títol per fer nous predicats a la resta d'atributs que es volen extreure (com ara buscar una data dins del títol del document).

Construcció de les regles

Per construir les regles s'ha adoptat diferents estratègies en funció de l'atribut objectiu:

- **Títol**. Pel títol s'han construït regles que tenien en compte la mida de la lletra de les línies de text candidates i la posició del text dins la primera pàgina. Si la primera pàgina complia el patró estàndard es busca el títol a la part central de la pàgina, sinó a l'inici de la pàgina.
- **Tipus de document**. Per extreure el tipus de document s'han fet regles que tenien en compte atributs del document com pot ser l'extensió i altres regles que busquen noms de tipus de document en línies de text importants del document com pot ser el títol o el nom de l'arxiu.
- **Data**. Per extreure la data hem cercat a les línies de l'inici i del fi de la primera pàgina del document parts de text que puguin ser dates en diversos formats. A més, també hem cercat la data en el títol del document i en el nom de l'arxiu.
- **Autors**. Finalment, els autors els hem cercat en la part final del document buscant en les últimes línies de text autors que es troben a la base de dades. També hem assignat directament autors per l'aparició de paraules que tenen a veure amb temàtiques que tracta normalment una unitat del servei.

Per millorar les regles i comprovar el resultat de les regles construïdes hem realitzat vàries proves. Que ens han permès afegir noves regles per corregir errors o descartar algunes que no resultaven molt fiables.

Normalització

Una vegada s'ha avaluat totes les regles i extrets cada un dels atributs que es busquen és el moment per normalitzar alguns valors obtinguts. En aquesta normalització, cal destacar tres tasques:

1. Treure el Servei de Biblioteques del llistat d'autors si també està alguna unitat del Servei.
   Una de les restriccions que ha de complir la fitxa d'un document és que no pot tenir com autors una unitat del Servei i el propi Servei.
2. Corregir el títol de les actes.
   S'ha dissenyat un manual d'estil pels títols dels documents de la intranet. Una de les regles més importants d'aquest manual d'estil fa referència al format dels títols de les actes de reunions. El sistema, sempre que sigui possible, ha de normalitzar els títols de les actes que extreu.
3. No permetre assignar el tipus llibre clau.
   El llibre clau és un tipus de document més complex que no es pot introduir de forma automàtica. Així que, si el sistema troba algun document del tipus llibre clau no li assignarà cap format específic.

Documents relacionats

L'últim atribut que es vol extreure són els documents relacionats. Aquests documents, en la majoria dels casos, són documents referenciats dins el propi enllaç del document mitjançant *hyperlinks* que s'utilitzen en els documents Word. És per això que no caldrà aplicar regles per fer l'extracció, sinó que, únicament cal recórrer l'arxiu Word amb l'ajuda de l'objecte Word.Application, una eina per manipular documents Word des de llenguatges ASP, i llegir els enllaços (*hyperlinks*) que conté el document. Amb aquests enllaços, es cerca dins la base de dades si hi ha algun document que estigui associat amb l'enllaç i es crea un llistat de documents relacionats.

Classificador de documents

Dels atributs de la fitxa del document que cal omplir per penjar un document la classificació temàtica és, sens dubte, l'atribut més interessant. Amb l'aparició d'Internet i la proliferació de motors de cerca que s'encarreguen d'indexar els documents de la xarxa per temàtiques o paraules clau han evolucionat una sèrie de tècniques per resoldre problemes d'indexació. Una d'aquesta tècniques anomenada TFIDF es basa en la obtenció de paraules rellevants i la relació d'aquestes amb un pes. Aquesta és la tècnica que s'aplicarà per fer la classificació temàtica:

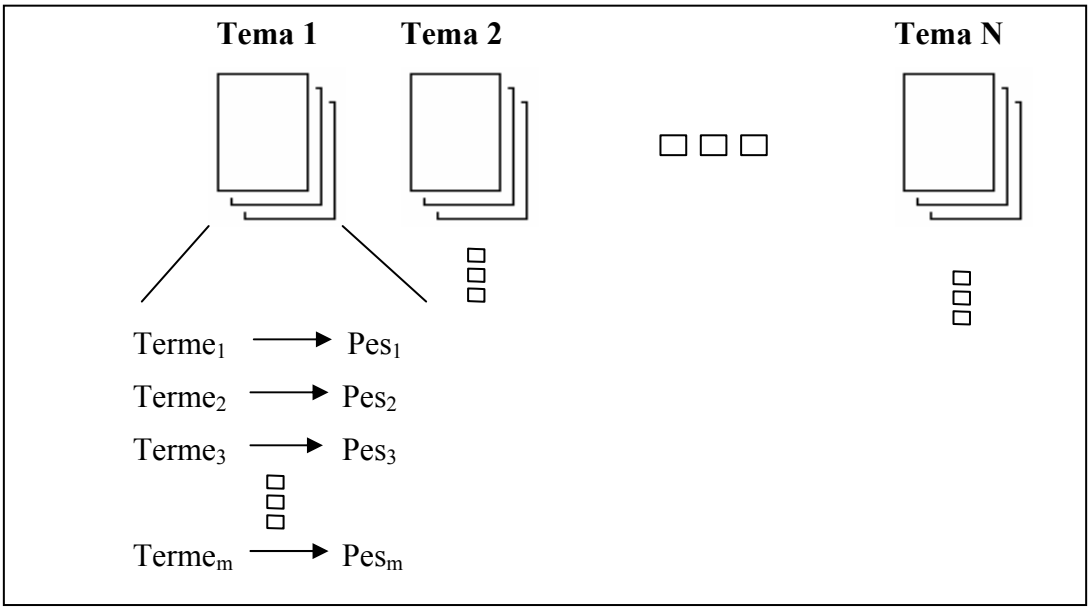
TFIDF

El TFIDF és una tècnica per conèixer la rellevància que té un terme dins un document o conjunt de documents. Utilitza la freqüència dels termes (TF- Term Frequency) i la seva freqüència inversa dins del conjunt total de documents (IDF - Inverse Document Frequency). En el nostre cas, consisteix a seleccionar el conjunt de documents de cada tema i calcular la freqüència relativa de les paraules que ocorren dins els documents i multiplicar per la freqüència inversa de la paraula concreta al conjunt de temes. Aquest càlcul es pot expressar amb la següent formula:

$$W_{pt} = FR_{pt} * \log\left(\frac{N}{N_p}\right)$$

On W_{pt} és el pes assignat a la paraula p en el tema t
FR_{pt} és la freqüència relativa de la paraula p en el tema t
N és el número de temes
Np és el número de temes on apareix la paraula p

Una vegada fet els càlculs, per cada tema es disposarà del conjunt de paraules que contenen els seus documents amb un pes assignat.



Llavors, s'haurà d'escollir unes quantes paraules de les que millor pes obtinguin per utilitzar-les en el càlcul del temes assignats.

La classificació d'un document consistirà, llavors, en calcular per cada tema la suma dels pesos que tenen les paraules del document i seleccionar el tema o temes que millor resultat obtinguin. Expressat en la formula el càlcul que cal fer per cada tema és:

$$F_{dt} = \sum_{p \in pars(d)} (W_{pt} * O_{pd})$$

On F_{dt} és el càlcul resultat pel document d en el tema t
W_{pt} és el pes assignat de la paraula p en el tema t
O_{pd} és el número d'ocurrències de la paraula p en el document d

Decisions de disseny

Per completar la tècnica utilitzada s'han pres les següents decisions:

Augmentar el valor de les paraules del títol

Es considera que les paraules que apareixen al títol, sovint, tenen una rellevància superior dins la temàtica del document que la resta de paraules. És per això, que per tenir en compte aquest fet, cal incloure en les paraules del títol una rellevància tres vegades superior a la resta de paraules del document afegint-les a la llista d'ocurrències tres vegades més per cada vegada que es troben dins del títol.

StopWords

Les paraules que apareixen freqüentment en tots els temes tindran un poder de resolució mínim, ja que la segona part de la fórmula del TFIDF assignarà valors molt baixos. És per això que s'establirà un llista de paraules buides (stopwords) que inclouen aquelles paraules amb freqüències excessivament altes i que normalment són articles, preposicions, adverbis, etc. D'aquesta manera, es reduirà l'espai de paraules del document simplificant una mica els càlculs.
Aquest llistat de paraules s'ha extret de la pàgina personal del professor Lluís Padró on ofereix un parell de llistats de StopWords pel català i pel castellà. A més s'han afegit algunes trobades després de fer algunes proves.

Selecció de les paraules per cada tema

Un pas important en l'elaboració del classificador és l'elecció de les paraules que es faran servir per cada tema. En el nostre cas s'ha escollit seleccionar les 20 primeres paraules que obtenen millor resultat de cada tema. Això permet tenir un conjunt ampli de paraules per tema tot i que moltes poden ser poc rellevants. A més, sempre es disposarà de l'opció de treure, afegir o modificar paraules de la base de dades per tal de millorar el classificador

Avaluació del classificador

Per poder avaluar el resultat del classificador, es divideix el conjunt de documents de que es disposa en dos corpus. El primer (un 80% més o menys del documents) s'utilitzarà per calcular la relació de les paraules amb els temes. La resta es farà servir per avaluar el classificador.

A més, es farà servir dues mesures clàssiques d'efectivitat:

- *precision*: percentatge de temes assignats pel classificador que s'han assignat correctament.
- *recall*: percentatge de temes assignats realment que el classificador ha trobat.

A part del *precision* i del *recall* també es calcularà un paràmetre F1 que es calcula en funció d'aquet dos, i que compleix la següent expressió:

$$F1 = \frac{2 * precision * recall}{presicion + recall}$$

Pel nostre sistema, és important assolir bons resultats tant pel *recall* com pel *precision*, ja que si el sistema prova moltes possibilitats i encerta poques o assigna temes poques vegades, els usuaris no tindran una bona impressió i acabaran per no utilitzar-ho. És per això, que ens basarem en la mesura f1, que és una mitjana de les dues mesures anteriors, per comprovar l'efectivitat del sistema tot i que es veu que el sistema s'arrisca poc s'intentarà afavorir una mica el *recall* encara que s'obtinguin pitjors resultats a f1.

Característiques del corpus

Abans de començar a avaluar el resultats del classificador cal conèixer una sèrie de dades de l'estat de la intranet en el moment de les proves:

Documents

Encara faltaven per penjar una gran part dels documents que cal traslladar de la intranet antiga. La intranet disposa de 400 documents:

- 389 passen correctament el preprocés (97,25 %)
- 11 no passen el preprocés (0,75%)
  - 9 aplicacions a que utilitzen autentificació (protocol HTTPS)
  - 2 documents en format no transformable a PDF (Zips o imatges)

Arbre temàtic

Abans de començar a penjar els documents és va construir un arbre de temes tenint en compte les necessitats del Servei. Aquest arbre temàtic disposa de 264 temes. Molts d'aquest temes no contenen documents, perquè encara no s'han penjat els documents que fan referència a aquest tema o perquè són temes que molt probablement no contindran cap i s'hauran d'eliminar.

Documents assignats	Nº de temes
0	148
1	52
2	21
3	10
4	7
> 4	26

Taula 10: Nombre de documents per tema

Documents en més d'un tema

Per últim, es mostra a la taula 11, dels 389 documents que s'han processat correctament quans estan classificats en més d'un tema:

Nº de temes on s'han classificat	Nº de documents
1 tema	363
2 temes	22
3 temes	4

Taula 11: Nombre de temes per document

Com es pot veure a la taula 11, la majoria de documents tenen només un tema assignat i el màxim de temes assignats és només de tres.

Sumant els temes assignats en tots els documents es disposa d'un total de 419 temes assignats per 389 documents. Dels 389 documents es farà servir 358 per l'aprenentatge i 31 documents com a corpus de testeig que no participaran en l'aprenentatge i que representen en aquest cas unes **34 assignacions reals a temes**.

Algoritme inicial

Una vegada calculat el TFIDF i seleccionades les paraules més representatives de cada tema amb el seu respectiu pes, es comença a treballar amb el corpus de testeig per construir l'algoritme que s'utilitzarà per fer la classificació.

En aquest apartat, es construirà un primer algoritme molt senzill que, més tard, es millorarà observant els resultats obtinguts. Aquest algoritme, primer obté la relació del document amb cada tema a partir de les paraules que hi ha a la base de dades ordenant-los de millor a pitjor. Llavors, selecciona el primer tema si sobrepassa un llindar i selecciona la resta si el càlcul obtingut té un valor pròxim al primer.

A la taula 12 es mostra el pseudocodi de l'algoritme amb dos 2 paràmetres configurables per adaptar-los a les nostres necessitats:

LIM_TFIDF: valor mínim per seleccionar el primer tema.
RELACIO_TFIDF: valor entre 0 i 1 que indica el percentatge del valor del primer tema que han de superar els temes següents per ser seleccionats.

Al pseudocodi s'utilitzarà a més objectes *Conjunt* i objectes *Iterador* amb les seves operacions típiques.

```
Funcio CalculaTemes() : Conjunt

 cj_candidats:  Conjunt
calculs : Iterador

calculs =  ObtenirTemes(document)

Si calculs.size > 0 llavors
   tfidf_ini = calculs.current.TFIDF

    Si tfidf_ini > LIM_TFIDF llavors
       Mentres no calculs.isDone

            Si tfidf_ini * RELACIO_TFIDF <  calculs.current.TFIDF llavors
               cj_candidats.Afegir(calculs.current.Tema)
            Fsi

           calculs.nextitem
       FMentres
    Fsi
Fsi

retorna  cj_candidats

FFuncio
```

Taula 12: Pseudocodi del primer algoritme per seleccionar els temes

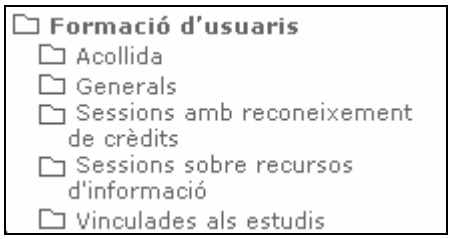
Es prova l'algoritme amb els paràmetres configurats per tal que només seleccioni un tema per document (resultats a la taula 13).

Resultats obtinguts (31 documents)	
Total temes	34
Temes encertats	12
Temes provats	31
Precision	0,387
Recall	0,353
F1	0,369

Taula 13: Prova algoritme inicial seleccionant només un tema

Es pot observar que els resultats no són gaire bons. Mirant els temes seleccionats pel sistema es pot apreciar que hi ha una tendència a seleccionar temes que s'han calculat amb pocs documents i per tant amb poques paraules. Això es deu sobretot a la utilització de la freqüència relativa a l'hora de calcular el TFIDF. Els temes que contenen poques paraules tenen freqüències relatives molt més altes per les paraules representatives del tema que els temes amb moltes paraules.

Aquesta tendència es pot apreciar a la taula 14. Pels temes amb més documents i per tant amb més paraules la suma dels pesos calculats és inferior que pels temes amb pocs documents. Per exemple, amb *Acollida*, un tema amb 36 documents, la suma dels pesos és molt inferior a la del tema *Formació d'usuaris*, amb només 8 documents.



Tema	Nº documents	Nº paraules	Nº de paraules per document	Suma dels pesos calculats
Formació d'usuaris	8	4551	568,875	0.114
Acollida	36	17664	490,7	0.084
Generals	0	0	-	0
Sessions amb reconeixement de crèdits	25	27407	1096,28	0.0862
Sessions sobre recursos d'informació	1	371	371	0.137
Vinculades als estudis	0	0	-	0

Taula 14: Exemple de relació entre nombre de paraules i pesos associats

Una de les possibles solucions és deixar d'utilitzar la freqüència relativa i utilitzar la freqüència absoluta a l'hora de calcular el TFIDF. Això tampoc aporta grans resultats ja que els temes amb més documents són ara molt més probables.

Normalitzar la probabilitat

Finalment, s'ha optat per normalitzar la probabilitat de què se seleccioni un tema, és a dir, normalitzar els pesos de les paraules seleccionades per cada tema per fer que el resultat de la suma de totes sigui el mateix per cada tema.
Això té una conseqüència negativa: es perd la referència entre temes. Una paraula més significativa en un tema, si té moltes paraules significatives, pot arribar a tenir un valor inferior que en altres temes on tingui menys paraules significatives tot i ser més rellevant. Per això, també caldrà mirar d'editar els pesos de les paraules obtingudes si veuen incoherències d'aquest estil.

Fent una prova amb els pesos normalitzats i els paràmetres configurats perquè seleccioni un tema per document (com abans) s'obtenen els resultats que mostra la taula 15. Es pot veure que el resultat és ara millor. Es decideix doncs, realitzar més proves variant el valor del paràmetre RELACIO_TFIDF (taula 16).

Resultats obtinguts	
Total temes	34
Temes encertats	18
Temes provats	31
Precision	0,581
Recall	0,529
F1	0,554

Taula 15: Prova algoritme inicial seleccionant només un tema amb el pesos normalitzats

Resultats obtinguts			
RELACIO_TFIDF	0.9	0.75	0.5
Total temes	34	34	34
Temes encertats	21	23	251
Temes provats	45	56	102
Precision	0,467	0,41	0,245
Recall	0,618	0,676	0,735
F1	0,532	0,511	0,368

Taula 16: Algoritme inicial diferents valors per RELACIO_TFIDF

2º Algoritme

Observant els temes seleccionats en cada una de les proves anteriors es poden extreure conclusions per millorar l'algoritme:

1. Moltes vegades els temes que millor valora el sistema per un determinat document estan relacionats en la jerarquia de l'arbre, fills o germans dins la jerarquia. Aquests temes acostumen a tenir una valoració similar. És per això que el sistema té una dificultat especial a l'hora de seleccionar el tema relacionat del document ja que no sempre el que millor resultat obté és el tema que s'ha d'assignar al document. Cal doncs, una nova valoració per poder escollir el tema assignat amb més fiabilitat. Una opció, l'opció finalment escollida, és tornar a fer el càlcul del TFIDF però ara només considerant els temes que es troben en un mateix subarbre. D'aquesta manera els termes que tenen en comú tots aquest temes, rebran un pes inferior, i els termes més específics, continuaran sent significatius si no apareixen a la resta de temes.

   Per donar un exemple es mostra un llistat d'un conjunt de paraules seleccionades en els càlculs pel tema que penja de formació d'usuaris, sessions amb reconeixements de crèdits (taula 17).

   A la taula es pot veure que amb el nou càlcul del TFIDF, paraules típiques d'una temàtica de formació com són exercicis, curs o classe, deixen d'aparèixer mentre que altres com temps o treball obtenen una rellevància superior.

Paraula	TFIDF general	TFIDF subarbre formació
Dia	6,86727914	35,7427056
Exercicis	6,83175659	-
Curs	6,72713059	-
Classe	6,42135192	-
Mòdul	6,3408076	10,4111406
Alumnes	6,22460772	-
Sessions	5,77538526	-
Minuts	5,68470533	7,09549072
Pràctica	5,50401594	-
Temps	-	6,49867374
Treballs	-	5,0397878

Taula 17: Pes d'algunes paraules pel tema Sessions amb reconeixement de crèdits

2. Com es pot veure en una de les taules que mostra les característiques del corpus, el número de temes que s'assignen a un document mai passa de 3 i rarament passa d'un. A més, hi ha temes on és molt probable tenir documents que es troben en més d'un tema mentre que altres no acostumen a tenir documents que estiguin en més d'un tema. És per això, que si es vol que el classificador sigui precís no hauria de seleccionar molts temes per document i, en el cas que així ho faixi, s'hauria de decantar per temes on sigui més probable tenir documents amb més d'un tema. Caldrà doncs, establir alguna mesura per conèixer la probabilitat que un tema tingui documents amb més d'un tema assignat per actuar en funció d'aquest càlcul.

3. Un altre cas que es produeix sovint, és que el sistema selecciona molts candidats, tots ells amb un valoració molt semblant i no gaire alta. Això passa amb documents que pertanyen a temes en què no s'han calculat les paraules rellevants amb prous documents i, per tant, el sistema no té prou informació per classificar bé el document. En aquest casos la millor opció, segurament, serà descartar tots els temes.

Tenint en compte aquestes observacions s'ha creat un nou algoritme que selecciona un conjunt de candidats amb l'algoritme anterior i, en funció de les característiques i del nombre dels temes candidats, fa la selecció definitiva. Dins l'algoritme s'han inserit un conjunt de paràmetres per poder estudiar la influència d'aquest nous canvis. A les taules 18,19 i 20 es mostra el pseudocodi del nou algoritme distribuït en tres funcions.

```
Funcio CalculaTemes(document): Conjunt

cj_candidats:  Conjunt
calculs : Iterador

calculs =  ObtenirTemes(document)

Si calculs.size > 0 llavors
  tfidf_ini = calculs.current.TFIDF

   Si tfidf_ini > LIM_TFIDF llavors
      Mentres no calculs.isDone

          Si tfidf_ini * RELACIO_TFIDF <  calculs.current.TFIDF llavors
             cj_candidats.Afegir(calculs.current.Tema)
          Fsi
```

```
            calculs.nextitem
        FMentres
    Fsi
Fsi

Si AGAFA_CANDIDATS llavors
   retorna cj_candidats
Fsi

Si |cj_candidats| = 1 llavors
   retorna cj_candidats
Sino Si  |cj_candidats| <= LIM_MAX_TEMES llavors
   retorna DesideixDefinitius(cj_candidats,calculs,doc)
Sino
   Si LIM_DESCARTA_MOLTS < tfidf_ini llavors
      retorna DecideixDefinitius(cj_candidats,calculs,doc)
  Fsi
Fsi
retorna ∅
FFuncio
```

Taula 18: Pseudocodi de la funció CalculaTemes del segon algoritme

```
Funció DecideixDefinitius(cj_candidats:Conjunt,calculs:Iterador,doc:Document): Conjunt

...
calculem a les variables mateix_pare i mateix_arrel si el conjunt de candidats  tenen el mateix
tema pare o el mateix tema arrel a més d'altres variables auxiliars relacionades
...

Si mateix_pare llavors
          'Seleccionem el tema que mes valor treu en relació al tfidf calculat només
          'amb els temes que penjen del tema amb  id = idpare_aux (calculat abans)
   Si ProbabilitatDocumentsEnMesTemes(cj_candidats) > LIM_PERC_DOC_REP llavors
       retorna cj_candidats
   Sino
      'Només seleccionem 1, els documents dels temes seleccionats no solen apareixer en més
      temes_idpare = ObtenirTemesIdpare(idpare_aux,doc)
      Si temes_idpare.size > 0 llavors
         Si temes_idpare.current.tfidf > 0 llavors
            retorna  {temes_idpare.current.Tema}
         Fsi
      Fsi
      calculs.first
      retorna {calculs.current.Tema}
   Fsi
Sino si mateixa_arrel llavors
   'Seleccionem el tema que mes valor treu en relació al tfidf calculat només
   'amb els temes que penjen del subarbre (si la opció està activada)
   Si UTIL_SUBARBRE llavors
      temes_arrel = ObtenirTemesArrel(idarrel_aux,doct)
      Si |temes_arrel| > 0 llavors
         Si temes_arrel.current.tfidf > 0  llavors
            retorna { temes_arrel.current.Tema} U  AgafaSeguents(temes_arrel)
```

```
        Sino
            retorna {calculs.current.Tema} U  AgafaSeguents(calculs)
        Fsi
    Fsi
  Sino
    retorna {calculs.current.Tema} U  AgafaSeguents(calculs)
  Fsi
Sino
  Si |cj_candidats| = 2 llavors
    Si ProbabilitatDocumentsEnMesTemes(cj_candidats) > LIM_PERC_DOC_REP llavors
        retorna cj_candidats
    Sino
        calculs.First
        retorna { calculs.current.Tema}
    Fsi
  Sino
    Si |cj_arrels| <= 3 llavors   ' (nomes retornem si els resultats no estan molts dispersos)
        calculs.first
        retorna { calculs.current.Tema}
      Fsi
   Fsi
Fsi
'Si al final del procés no s'ha retornat, retornem un conjunt buit
retorna ∅
FFuncio
```

Taula 19: Pseudocodi de la funció DesideixDefinitius

```
Funcio AgafaSeguents(it:Iterador):Conjunt

cj: Conjunt

'Afegim el primer
it.First
cj.Afegir(it.current.Tema)
tfidf_aux = (it.current.TFIDF)

'Afegim dos més si el valor calculat es proper
Per  i=1 a 2
   Si not it.isDone llavors
    Si it.current.tfidf > tdidf_aux * LLINDAR_AGAFA_SEG llavors
        cj.Afegir(it.current.Tema)
    Fsi
    it.nextitem
  Fsi
FPer

retorna cj

FFunction
```

Taula 20: Pseudocodi de la funció AgafaSeguents()

Com es pot veure, l'algoritme es complica perquè hi ha molts casos a resoldre en funció dels temes seleccionats. Pel altra banda, a part de les tres funcions descrites aquí i de la funció inicial per obtenir la primera valoració dels temes, també caldrà implementar tres funcions més que apareixen al pseudocodi:

- *ObtenirTemesIdpare(idp):* retorna els temes valorats en funció del TFIDF calculat amb els temes que pengen del tema amb identificador igual a idp.
- *ObtenirTemesArrel(idarrel*): retorna els temes valorats en funció del TFIDF calculat amb el subarbre que té com arrel el tema amb identificador igual a idarrel.
- *ProbabilitatDocumentsEnMesTemes(cj*): retorna un valor entre 0 i 1 amb la probabilitat de que un document assignat a cualsevol tema del conjunt aparegui en un altre tema. Per fer aquest càlcul només utilitzem els temes que hem utilitzat per seleccionar les paraules del TFIDF.

Aquestes tres funcions les calcula directament l'SGBD mitjançant consultes SQL amb funcions agregades.

Afinació de paràmetres

Per poder provar l'algoritme amb el conjunt de testeig disposem dels paràmetres configurables que es mostren a la taula 21.

Paràmetre	Explicació
LIM_TFIDF	Mínim valor assignat a un tema per tal que l'algoritme el seleccioni com a candidat.
RELACIO_TFIDF	Valor entre 0 i 1 per seleccionar els temes propers al tema que millor resultat dona en la primera selecció de candidats.
AGAFA_CANDIDATS	Booleà per indicar si seleccionem tots els candidats de la primera part de l'algoritme.
LIM_MAX_TEMES	Límit de temes seleccionats en la primera selecció per que es consideri que hi ha molts temes i es tracti d'una manera diferent.
DESCARTA_MOLTS	Booleà per indicar que es no s'assigni cap tema quan hi ha molts candidats.
LIM_PERC_DOC_REP	Paràmetre que es compara amb el percentatge de documents en més d'un tema que tenen un conjunt de temes per saber si determinem que aquest temes poden compartir el document tractat.
UTIL_ARBRE	Paràmetre per escollir si es vol utilitzar el càlculs del TFIDF respecte a un determinat subarbre o amb tot l'arbre temàtic després de rebre un conjunt de temes candidats amb la mateixa arrel.
LLINDAR_AGAFA_SEG	Valor entre 0 i 1 per seleccionar els temes propers al tema que millor resultat dona en la segona selecció.

Taula 21: Paràmetres configurables per l'algoritme calculatemes

Per tal d'afinar els paràmetres, es faran una sèrie de proves assignant diferents valors a cada paràmetre. Aquest valors es poden veure en la taula 22.

Donat el gran nombre de combinacions (1080) i el limitat nombre de documents pels testeig, trobarem moltes combinacions amb el mateix resultat. Per fer-nos una idea dels valors dels paràmetres que influeixen més en el resultat final farem una mitjana per les combinacions que millor resultat obtinguin agrupades pel valor final de F1 (taula 23). Ha de quedar clar que aquestes mitjanes només són una referència i que caldrà observar la taula de resultats per prendre les decisions oportunes.

Paràmetre	Valors assignats
LIM_TFIDF	0.01 / 0.25 / 0.5 / 0.75 / 1
RELACIO_TFIDF	0.5 / 0.75 / 0.9
AGAFA_CANDIDATS	No
LIM_MAX_TEMES	3 / 4 / 5
DESCARTA_MOLTS	Sí / No
LIM_PERC_DOC_REP	0.1 / 0.3
UTIL_ARBRE	Sí / No
LLINDAR_AGAFA_SEG	0.7 / 0.8 / 0.9

Taula 22: Diferents valors assignats als paràmtres en les proves

LIM TFIDF	RELACIO TFIDF	LIM MAX TEMES	DESCARTA MOLTS	UTIL SUBARBRE	LLINDAR AGAFA SEG	TEMES ENCERTATS	TEMES PROVATS	RECALL	PRECISION	F1
0,25333	0,5	4	1	0,6	0,82	21	26	0,61765	0,8077	0,7
0,25333	0,5	4,23	0,308	0,61	0,85	21	27	0,61765	0,7778	0,6885
0,25333	0,5	4,25	0,25	0,5	0,75	21	28	0,61765	0,75	0,6774
0,25333	0,5	4,25	0,25	0	0,7	21	29	0,61765	0,7241	0,6667
0,5	0,75	3	1	0,5	0,8	21	30	0,61765	0,7	0,6563
0,13	0,75	3	1	0,5	0,8	21	31	0,61765	0,6774	0,6462
0,5	0,75	4,2	0,4	0,5	0,8	21	32	0,61765	0,6563	0,6364
0,13	0,75	4,2	0,4	0,5	0,8	21	33	0,61765	0,6364	0,6269

Taula 23: Valors obtinguts en les diferents configuracions agrupats pel valor f1

De la taula 23 i observant el resultat de totes les combinacions poden determinar que el valor o conjunt de valors per cada paràmetre que millor resultat obtenen són els que es mostren en la següent taula (taula 24).

Paràmetre	Valors assignats
LIM_TFIDF	0.25 / 0.50
RELACIO_TFIDF	0.50 / 0.75
AGAFA_CANDIDATS	No
LIM_MAX_TEMES	4
DESCARTA_MOLTS	Sí /No
LIM_PERC_DOC_REP	0.1 / 0.3
UTIL_ARBRE	Sí
LLINDAR_AGAFA_SEG	0.75

Taula 24: Valors dels paràmetres que millor resultat obtenen en les proves anteriors

Ara cal fer les proves finals per decidir la millor combinació de paràmetres entre les 16 combinacions resultants. Degut al poc nombre de document pel testeig ens veiem obligats a utilitzar tot el conjunt de documents per afinar definitivament els paràmetres anteriors. Els resultats d'aquest prova es mostren en la taula 25.

Com que sobretot interessa assolir un bon recall, s'agafarà la configuració que més temes encerti i tingui la millor precisió possible. En aquest cas la configuració marcada a la taula 25.

LIM TFIDF	RELACIO TFIDF	DESCARTA MOLTS	LIM PERC DOC REP	N TOTAL TEMES	TEMES ENCERTATS	TEMES PROVATS	RECALL	PRECISION	F1
0,5	0,75	1	0,3	422	330	385	0,782	0,857	0,818
0,5	0,75	1	0,1	422	331	388	0,784	0,853	0,817
0,25	0,75	1	0,3	422	333	393	0,789	0,847	0,817
0,5	0,75	0	0,3	422	330	386	0,782	0,855	0,817
0,25	0,75	1	0,1	422	334	396	0,791	0,843	0,817
0,5	0,75	0	0,1	422	331	389	0,784	0,851	0,816
0,25	0,75	0	0,3	422	333	394	0,789	0,845	0,816
0,25	0,75	0	0,1	422	334	397	0,791	0,841	0,816
0,5	0,5	1	0,3	422	325	377	0,770	0,862	0,814
0,25	0,5	1	0,3	422	327	382	0,775	0,856	0,813
0,25	0,5	0	0,3	422	331	394	0,784	0,840	0,811
0,5	0,5	1	0,1	422	326	382	0,773	0,853	0,811
0,25	0,5	1	0,1	422	328	387	0,777	0,848	0,811
0,5	0,5	0	0,3	422	328	387	0,777	0,848	0,811
0,25	0,5	0	0,1	422	332	399	0,787	0,832	0,809
0,5	0,5	0	0,1	422	329	392	0,780	0,839	0,808

Taula 25: Resultat de l'execució de les 16 configuracions amb tots els documents disponibles

Així doncs, la configuració que s'utilitzarà serà la que es mostra a la taula 16:

Paràmetre	Valors assignats
LIM_TFIDF	0.25
RELACIO_TFIDF	0.75
AGAFA_CANDIDATS	No
LIM_MAX_TEMES	4
DESCARTA_MOLTS	Sí
LIM_PERC_DOC_REP	0.1
UTIL_ARBRE	Sí
LLINDAR_AGAFA_SEG	0.75

Taula 26: Configuració escollida

Aquesta configuració es variarà si després de tornar a calcular el TFIDF amb més documents, el resultats obtinguts varien.

Opcions descartades

Una opció per classificar els documents és anar classificant per nivells, és a dir, primer calcular la relació que el document té amb els temes de nivell 1 (juntament amb els seus fills), escollir el que millor resultat obtingui i continuar la classificació només considerant els fills del tema i el propi tema escollit.
Per les característiques del nostre arbre temàtic, on poden existir temes amb fills que tenen temàtiques molt diferents, els resultats amb aquest tipus d'algoritmes no eren gaire bons. Un exemple d'aquest temes amb temàtiques molts diverses dins la seva jerarquia és el tema de comunicació interna, on van a parar les actes de les reunions que

es fan en els diferents eixos. Dins dels descendents d'aquest tema es pot arribar a parlar de temes tan diversos com equipament informàtic, formació d'usuaris o col·leccions. Per altra banda, l'opció que aporta l'algoritme final de seleccionar dins un arbre si els candidats escollits en la primera selecció pertanyen a un mateix subarbre ja aporta les millores que hauria produït aquest tipus d'algoritme.

Una altra opció per millorar el resultats de l'algoritme que utilitzen el TFIDF per fer una classificació temàtica de documents és fer servir les arrels de les paraules en comptes de les paraules senceres. Aquesta tècnica, anomenada stemming, s'ha descartat perquè requeria molt treball, sobretot tenint en compte que la majoria de documents estan en català, una llengua amb moltes declinacions per una determinada arrel de paraula.

Millores addicionals

Moltes vegades sobta que el sistema no classifiqui bé un document malgrat que el títol del document faci una referència clara al tema associat. Això succeeix quan no es tenen documents classificats dins aquell tema o els documents que hi ha no són molt determinants. Una opció per resoldre aquest problema seria afegir les paraules del títol a tots els temes, però com que hi ha molts temes buits i no es fàcil determinar el valor que cal donar a un paraula rellevant hem optat per una altre solució. Aquesta solució, donat un tema ja classificat fa una cerca a la base de dades amb les paraules del títol per veure si troba algun tema que tingui el nom com un fragment del títol i que, a part, aquest nom no aparegui en cap document que no estigui classificat en aquell tema, per evitar errors amb paraules molt comunes. Fent una prova amb els documents pel testeig i tots els documents s'obté el següents resultats (taules 27 i 28):

Resultats obtinguts (31 docs)	
Total temes	34
Temes encertats	24
Temes provats	36
Precision	0,706
Recall	0,667
F1	0,686

Taula 27: Resultats amb la millora en el corpus de testeig

Resultats obtinguts (389 docs)	
Total temes	422
Temes encertats	341
Temes provats	403
Precision	0,808
Recall	0,846
F1	0,827

Taula 28: Resultats amb la millora en tot el corpus

Amb aquesta millora l'algoritme encerta un quants temes més que corresponien a temes pràcticament buits.

Conclusions

Després de veure els resultats obtinguts es pot dir que el classificador, tot i que no és ni molt menys cent per cent eficient si aporta unes bones mesures d'efectivitat. Més si tenim en compte el limitat nombre de documents dels que hem disposat per fer els càlculs.
Tot i això, s'haurà d'esperar a poder fer proves amb un conjunt superior de documents quan la intranet estigui totalment actualitzada per tornar a afinar els paràmetres i conèixer amb més fiabilitat els resultats del nostre algoritme.

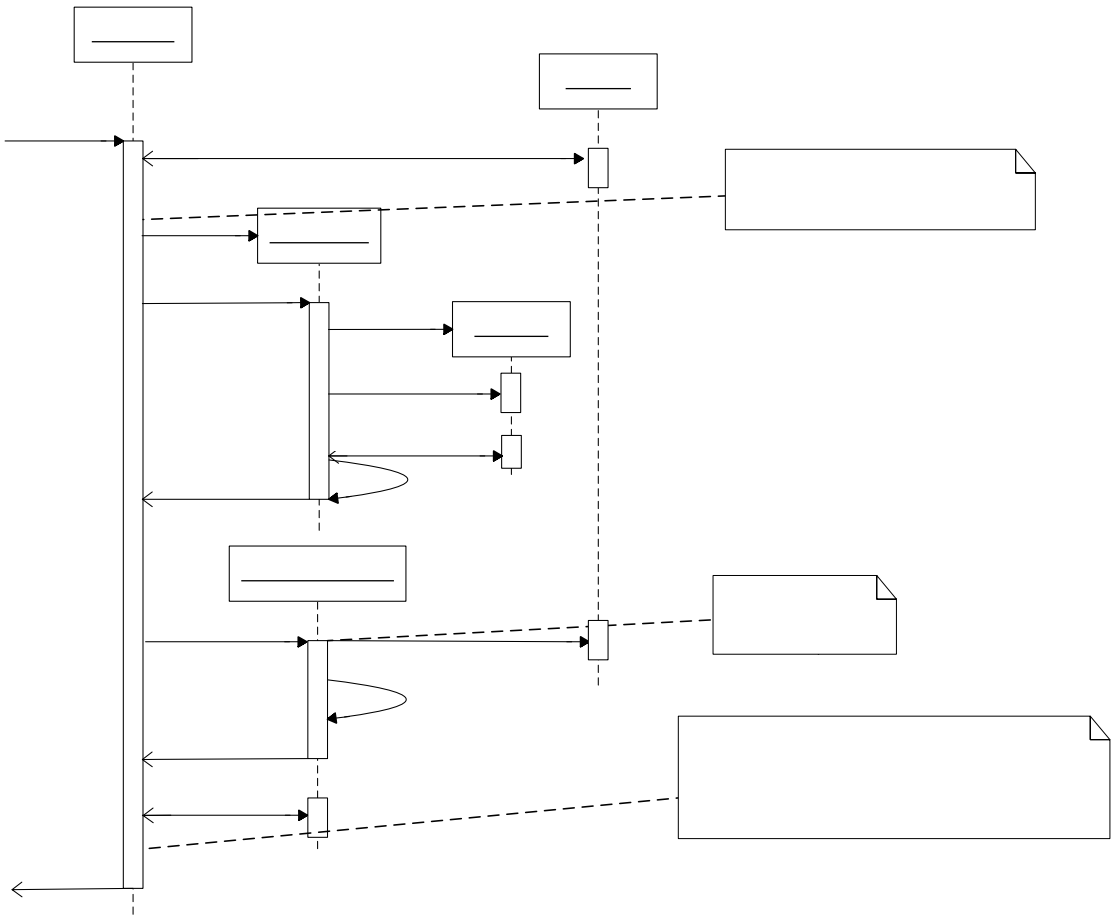
Diagrames de seqüència

Consideracions prèvies

Degut a la gran quantitat de casos d'ús i la similitud que hi ha entre molts d'ells s'ha decidit no incloure alguns diagrames molt similars. A més, a part dels diagrames generats a partir dels casos d'ús s'ha inclòs un diagrama que mostra la inicialització de cada pàgina per tal de poder utilitzar variables comunes a totes les funcions del sistema com són les variable amb els valors de la intranet actual i l'usuari actual.
Tot i que moltes funcions accedeixen al SGBD, s'ha mostrar aquesta comunicació només en alguns casos per aclarir més el comportament d'algunes funcions.
Finalment, s'han expressat les variables que es reben al fer una petició, el request per les pàgines ASP, com a *req(XXX)*.

Inicialització

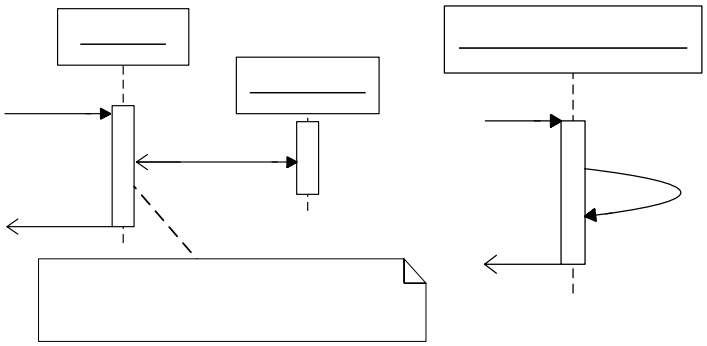
En aquest diagrama es mostra la inicialització que fan totes les pàgines per obtenir les variables Intra i UsuariActual que es faran servir en la resta de pàgines i després controlar l'accés a continguts restringits.

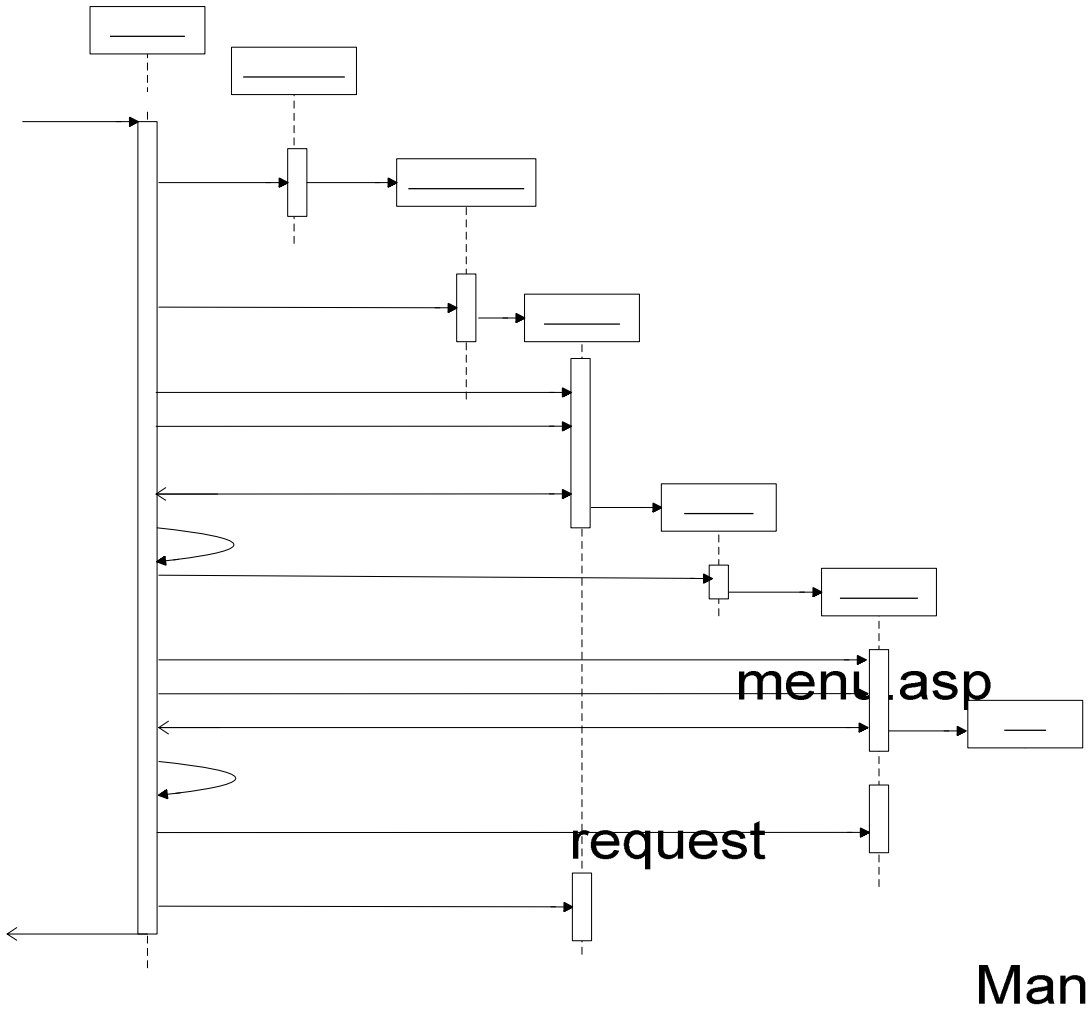


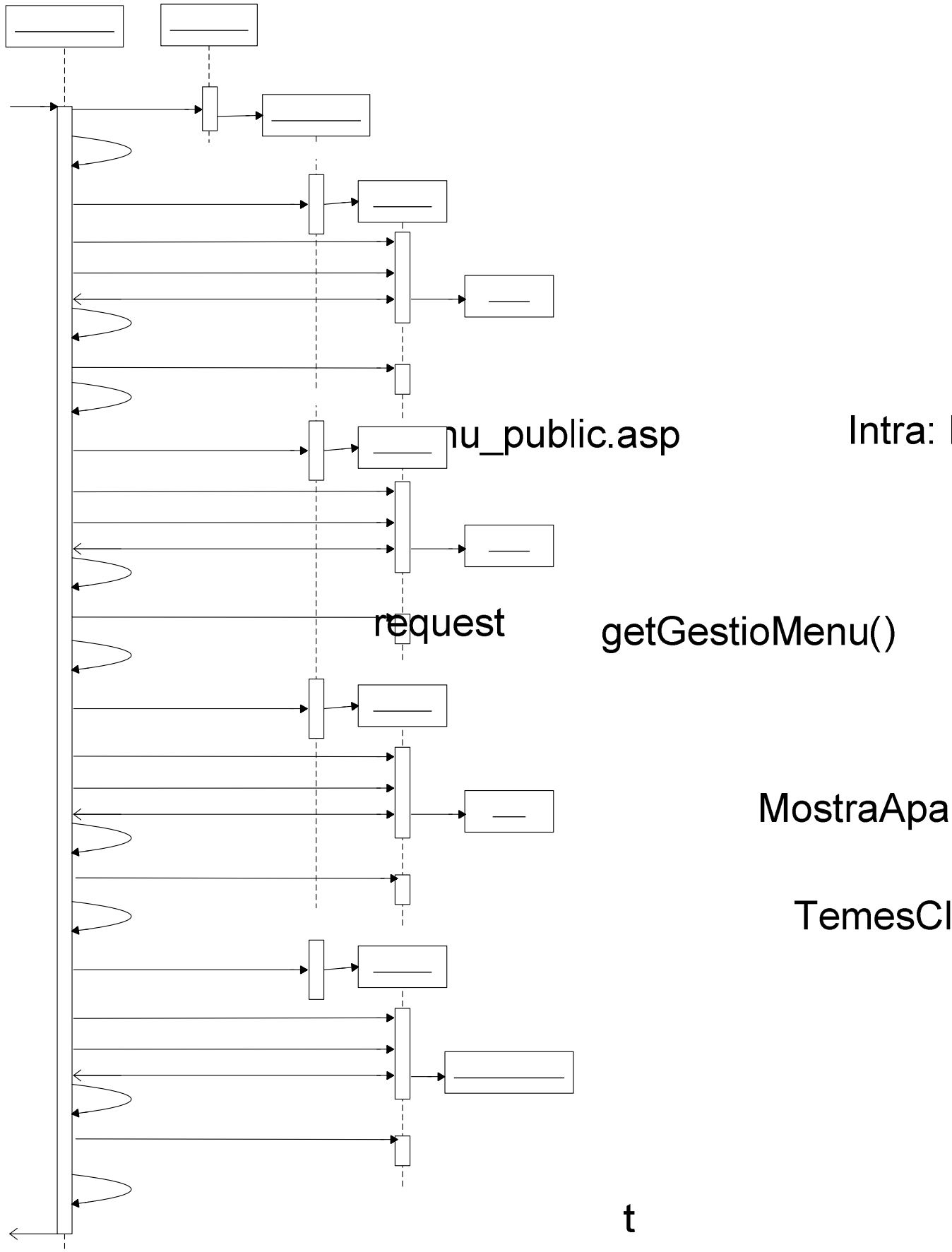
Navegació publica

Mostrar menús

Aquí es mostren com és generen els dos menús de la intranet, el públic i el de manteniment.

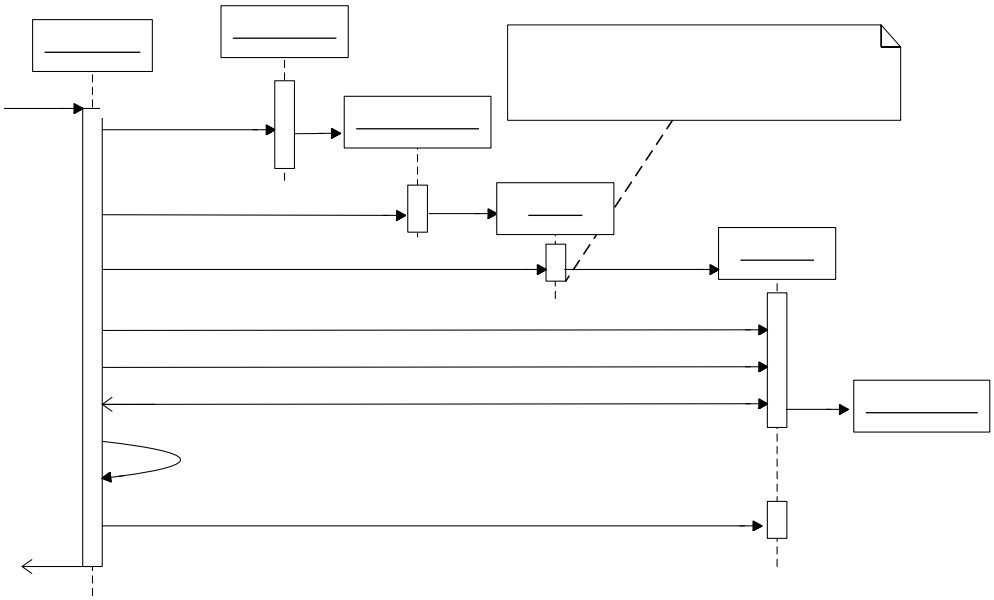





nu_public.asp
Intra:
request
getGestioMenu()
MostraApa
TemesCl
t
MostraItem

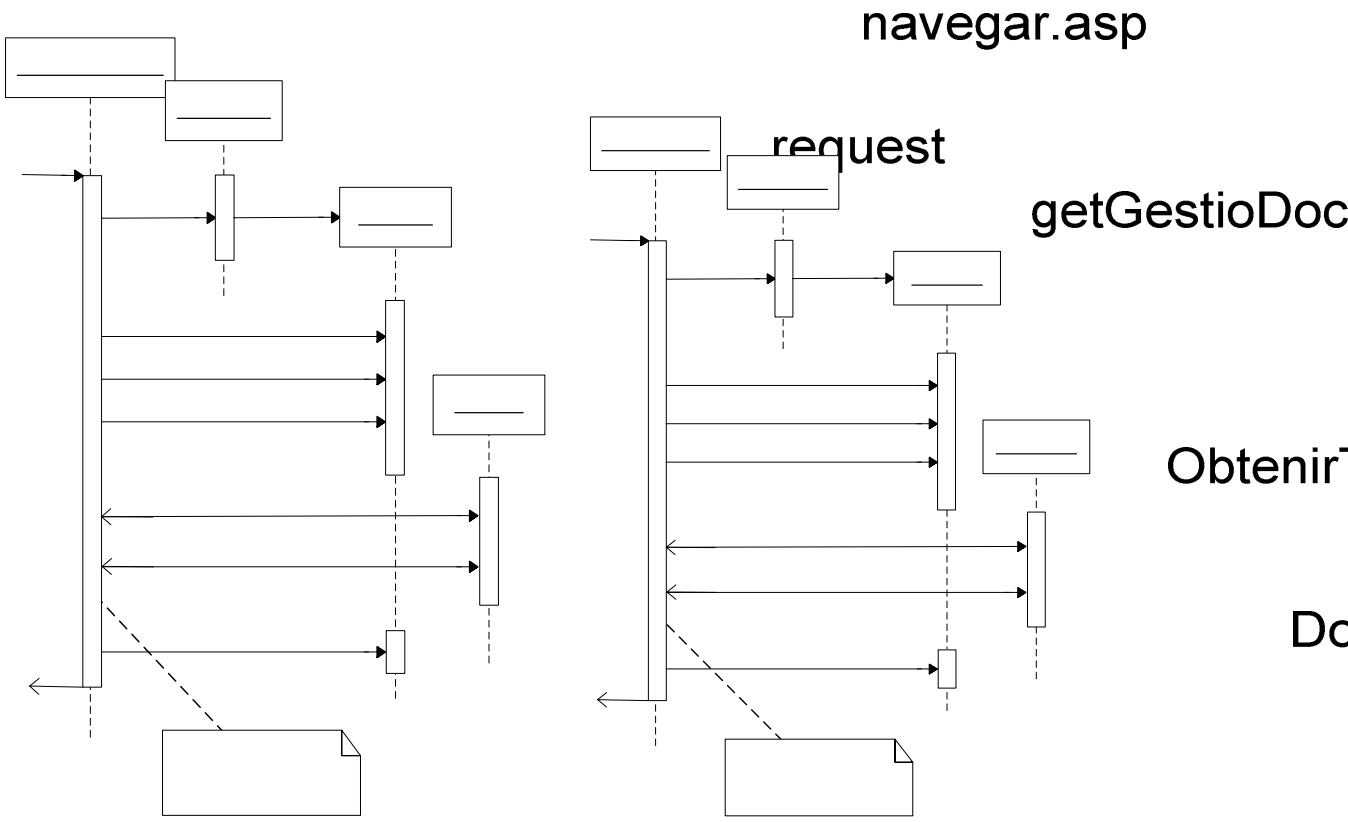
Llistar documents per tema

Aquest és un esquema bàsic de las crida per mostrar un llistat de documents d'un tema.

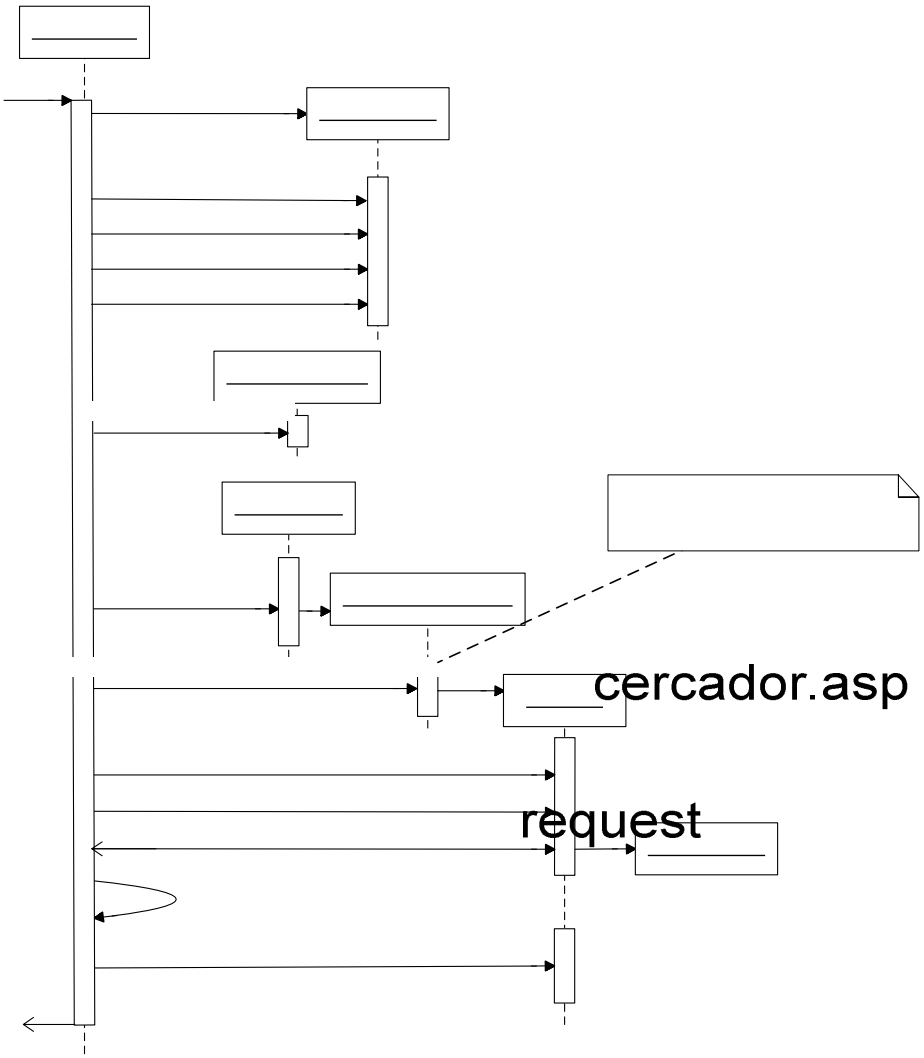


Mostrar llistat d'aplicacions/intranets

Aquest és l'esquema per llistar les aplicacions i les intranets accessibles des del sistema.

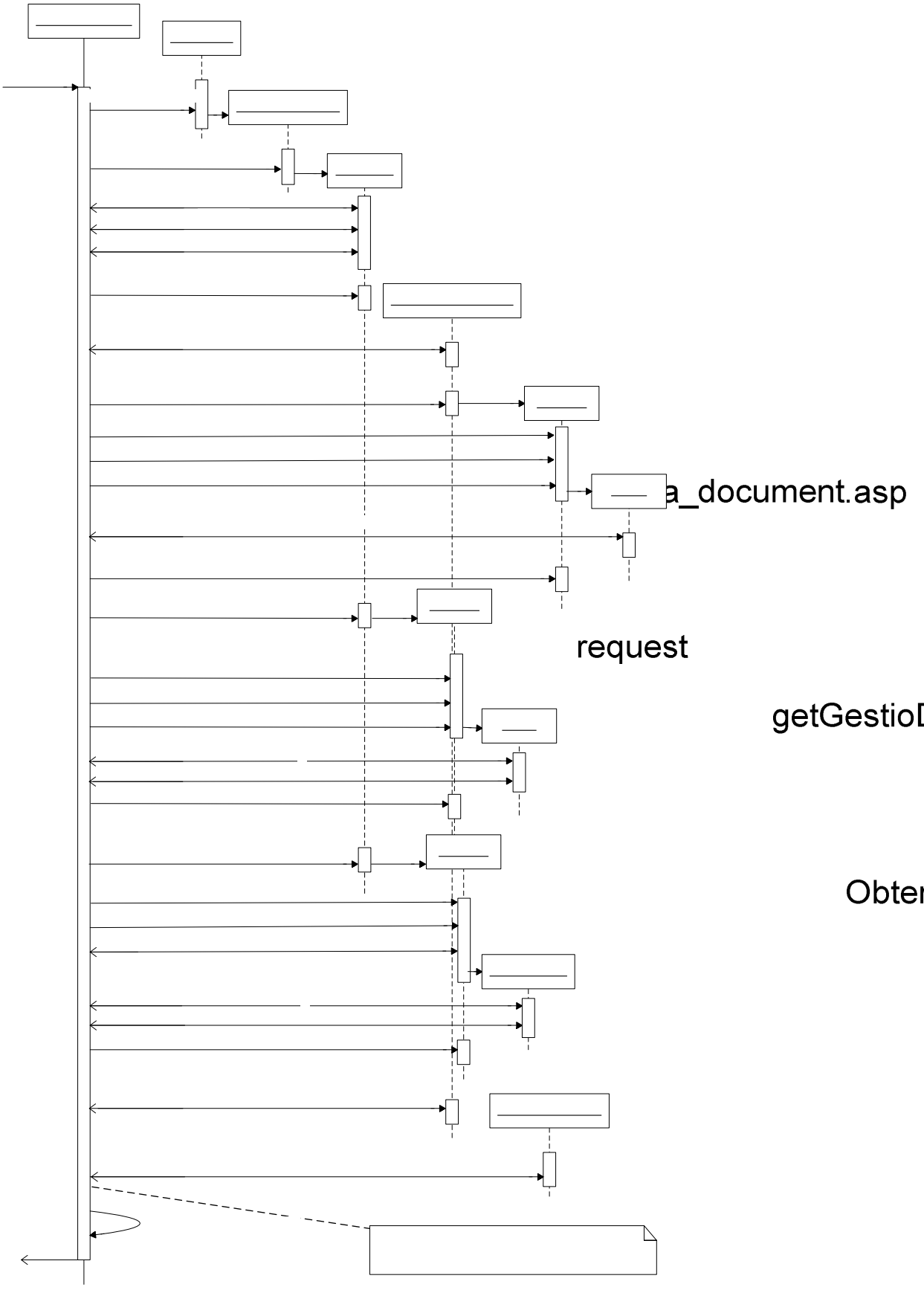


Cercar documents

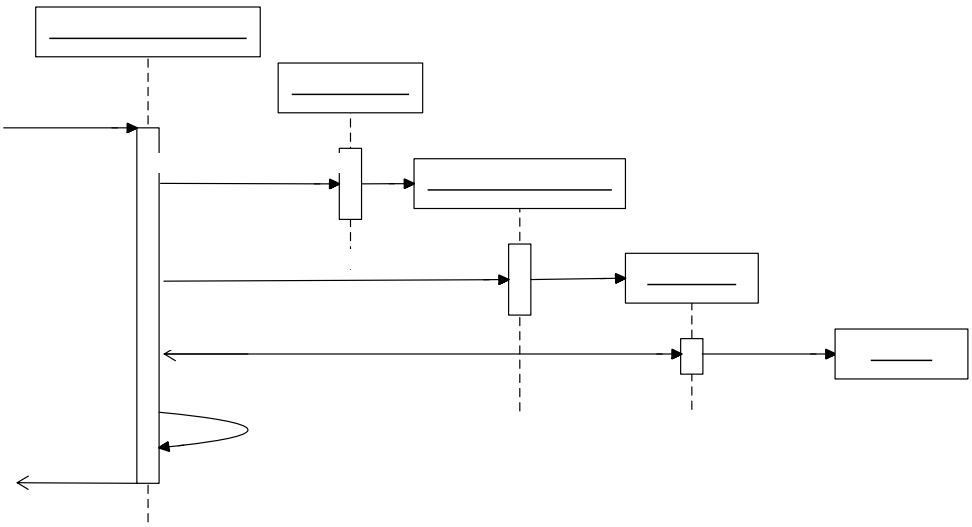


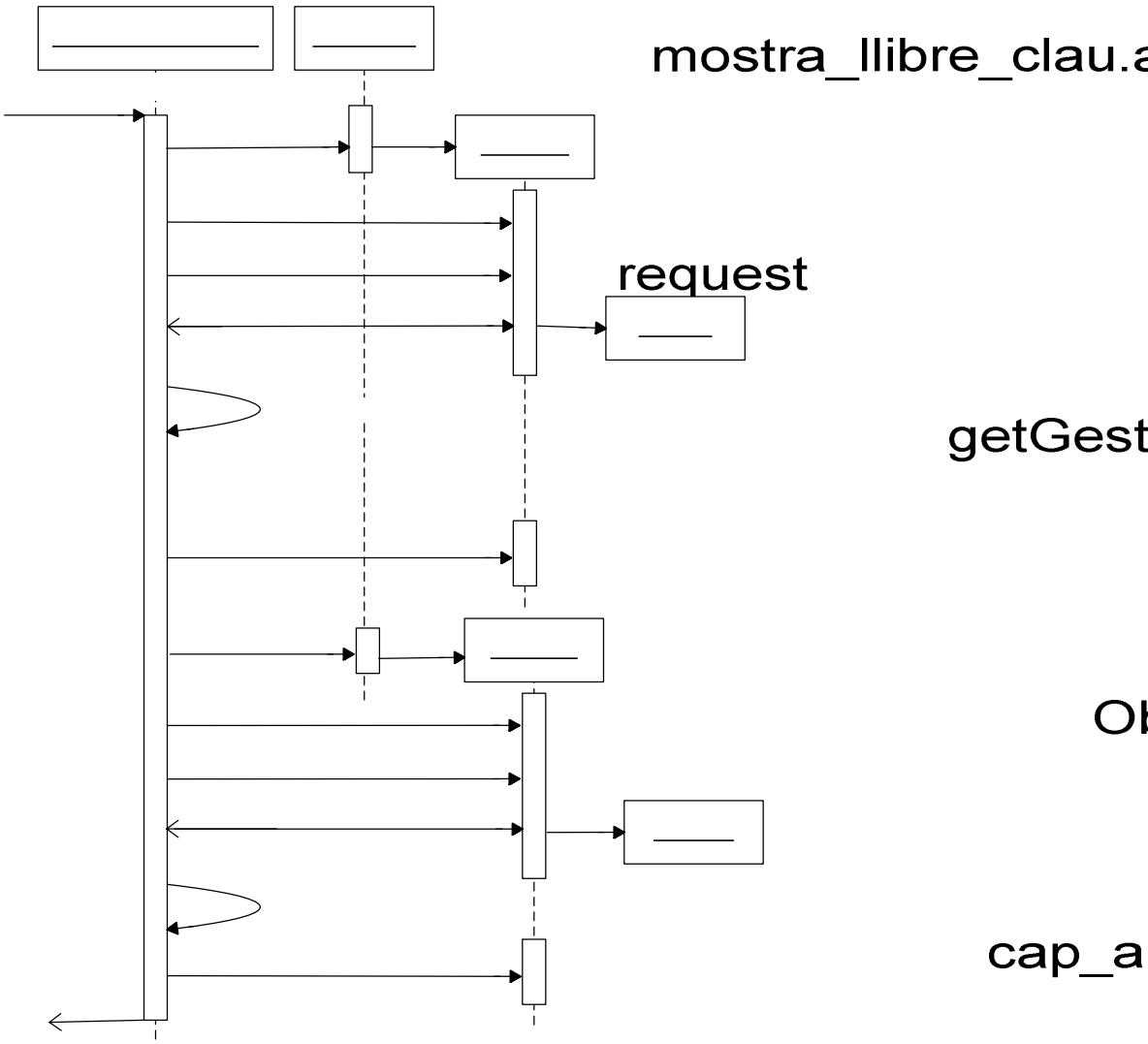

```
Titol := req("tito
tipus := req("tipu
descripcio := req("des
intrabib := req("intr


cj_auto
AfegirString(req("autors"
```

Mostrar dades document

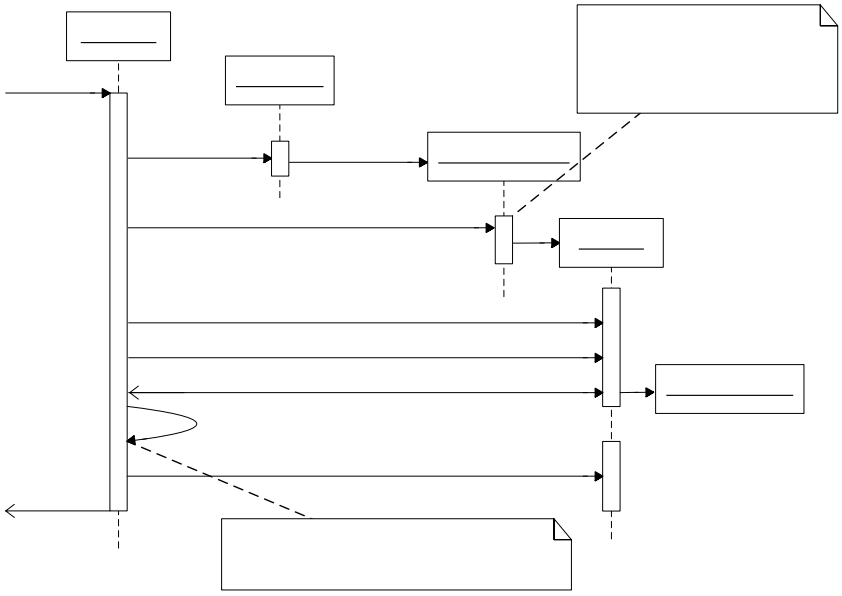


Mostrar llibre clau

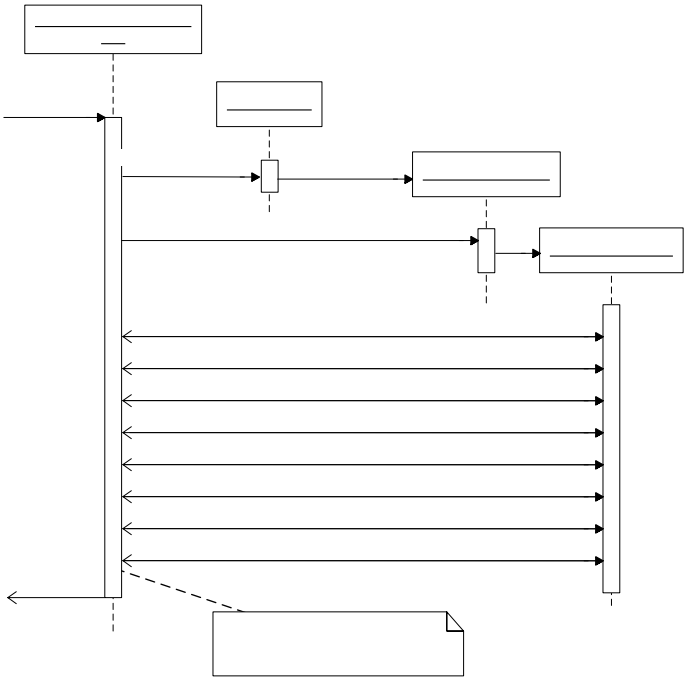




Mostrar llistat d'esdeveniments (novetats)



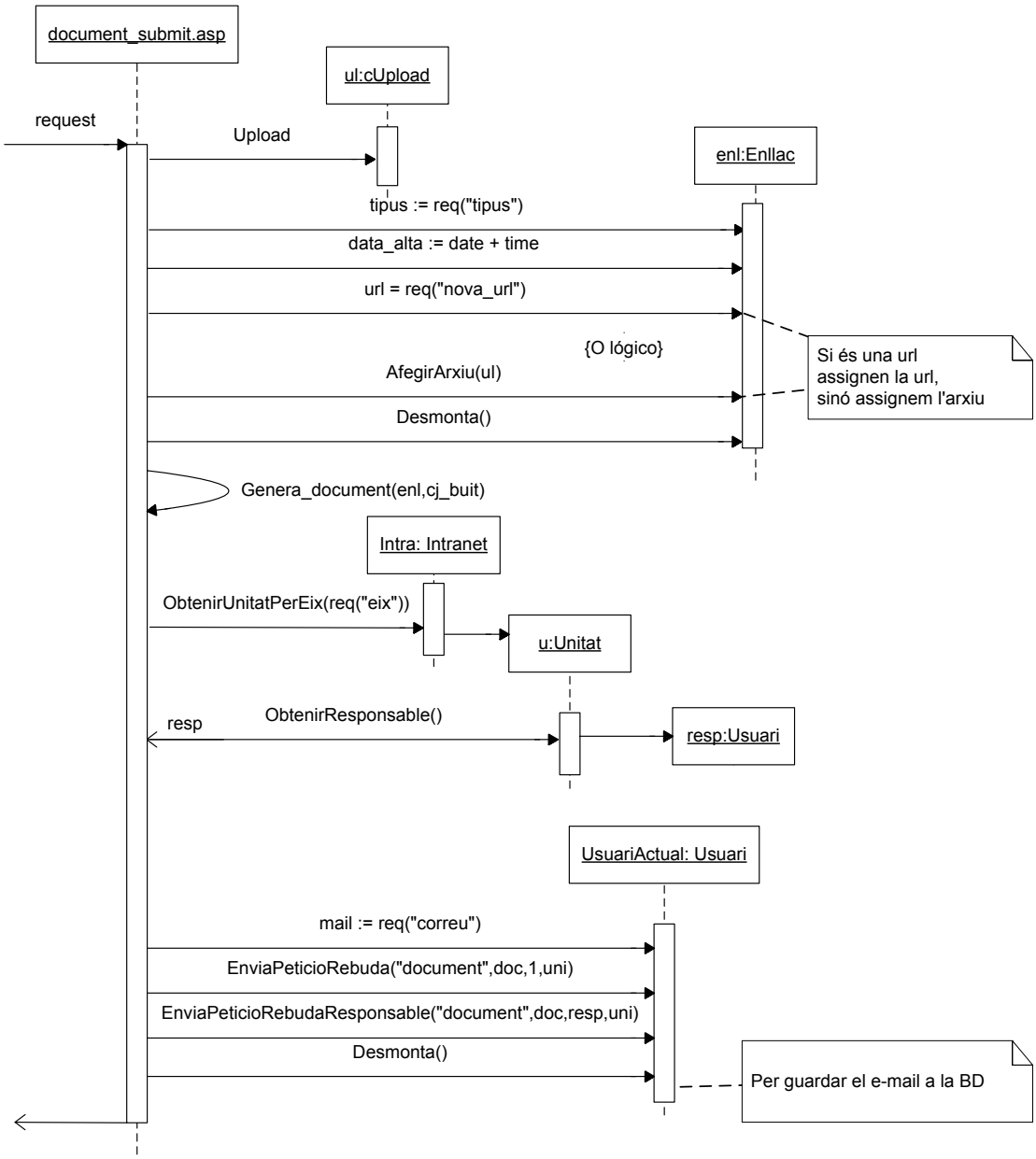
Mostrar dades d'esdeveniment (novetat)

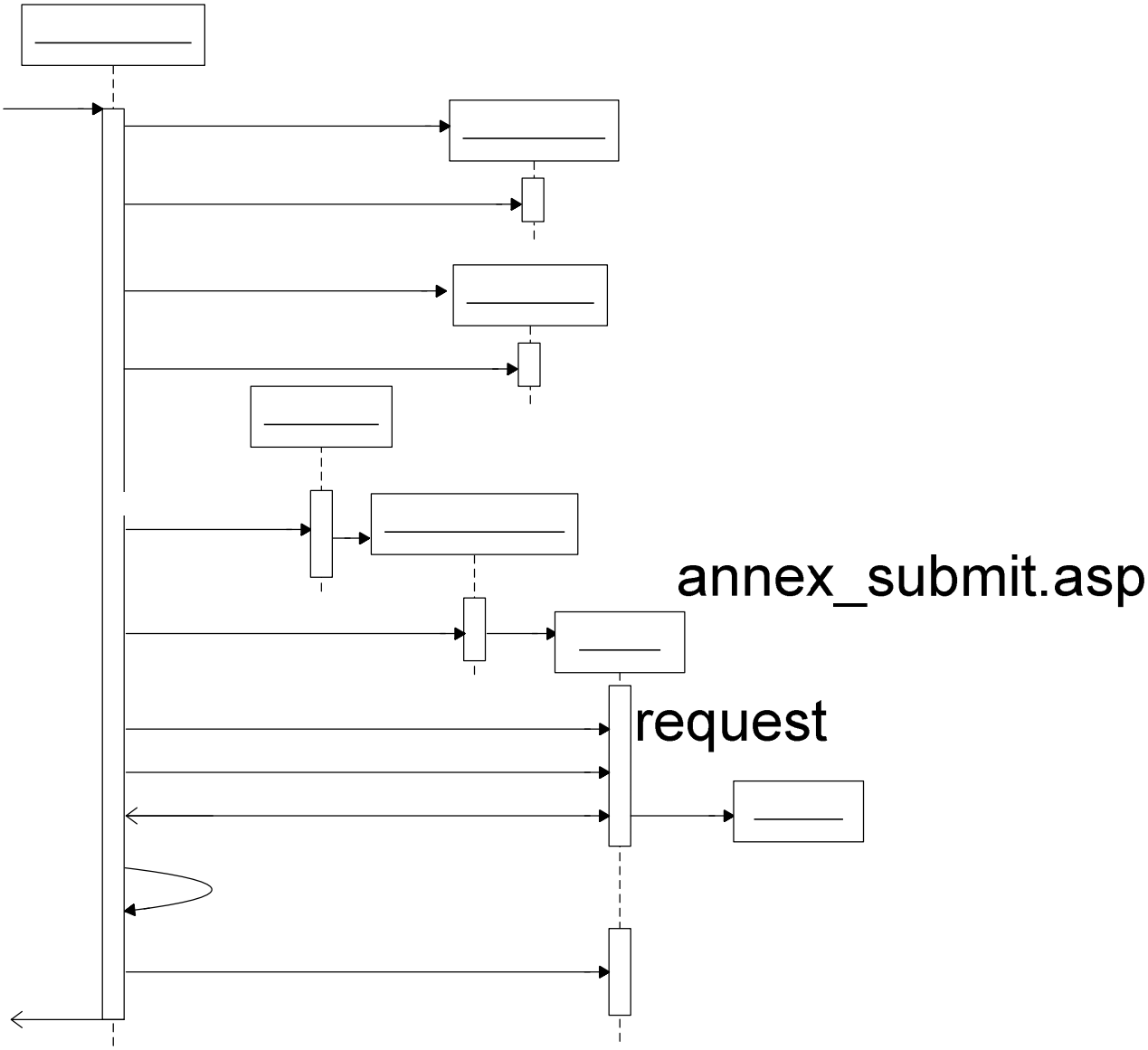


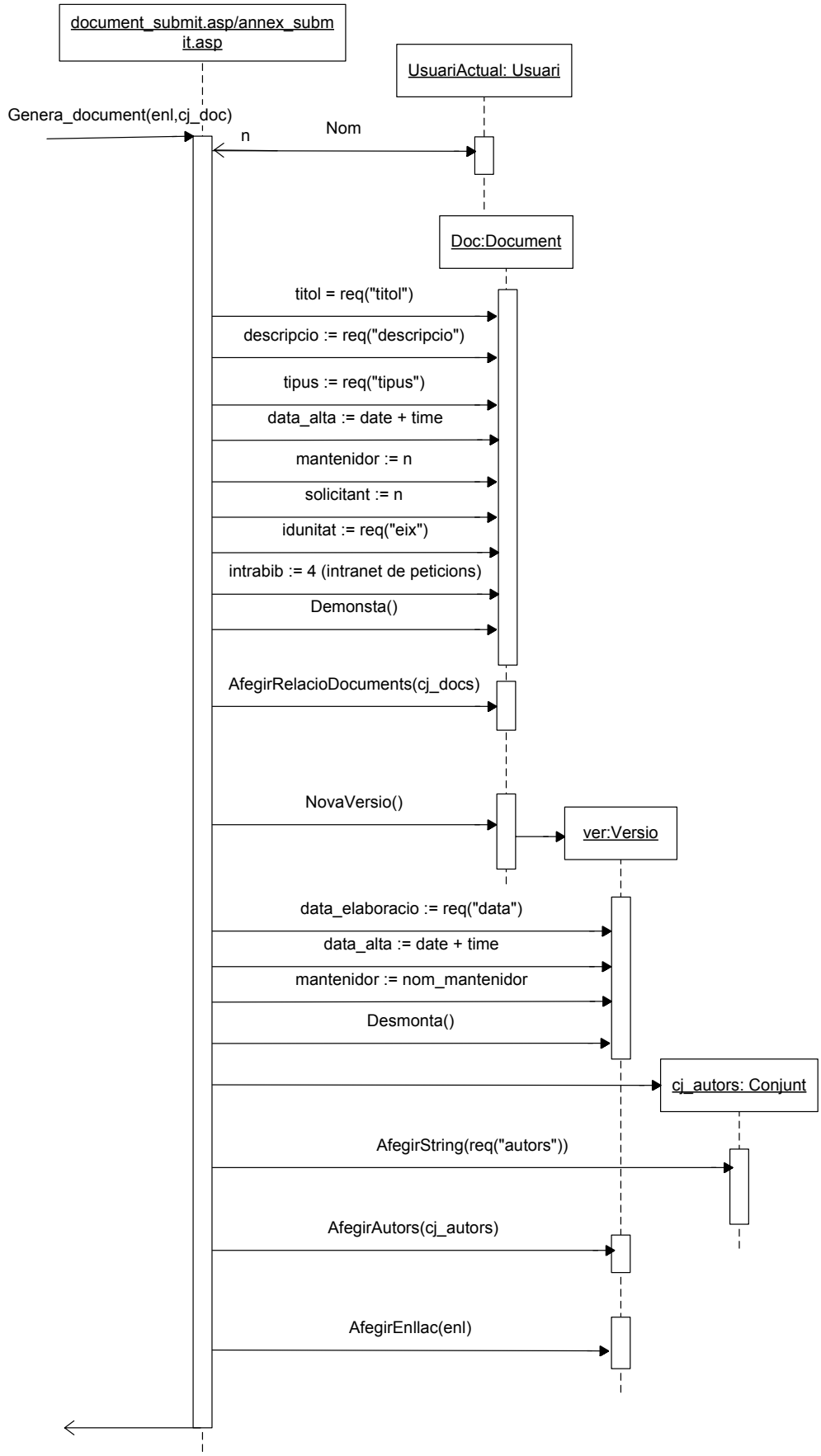
Mantenidor de biblioteca

Sol·licitar document

Aquesta funcionalitat consta de dos passos, en la primera, realitzada per document_submit.asp es crea una sol·licitud de document i s'envia a un mantenidor. En la segona s'assignen els documents relacionats amb aquesta petició.

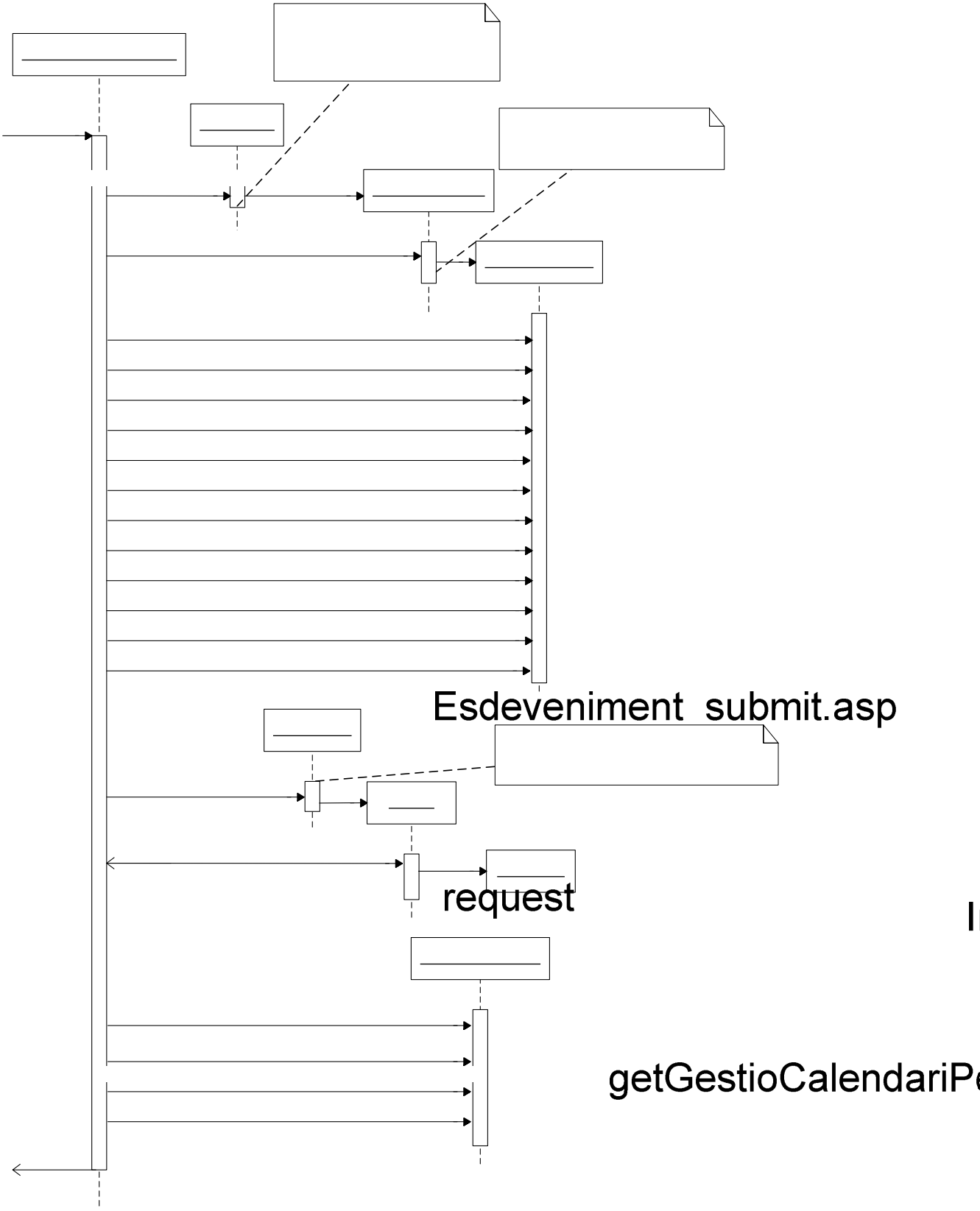



annex_submit.asp
request


document_submit.asp/annex_submit.asp
UsuariActual: Usuari
Genera_document(enl,cj_doc)
n
Nom
Doc:Document
titol = req("titol")
descripcio := req("descripcio")
tipus := req("tipus")
data_alta := date + time
mantenidor := n
solicitant := n
idunitat := req("eix")
intrabib := 4 (intranet de peticions)
Demonsta()
AfegirRelacioDocuments(cj_docs)
NovaVersio()
ver:Versio
data_elaboracio := req("data")
data_alta := date + time
mantenidor := nom_mantenidor
Desmonta()
cj_autors: Conjunt
AfegirString(req("autors"))
AfegirAutors(cj_autors)
AfegirEnllac(enl)

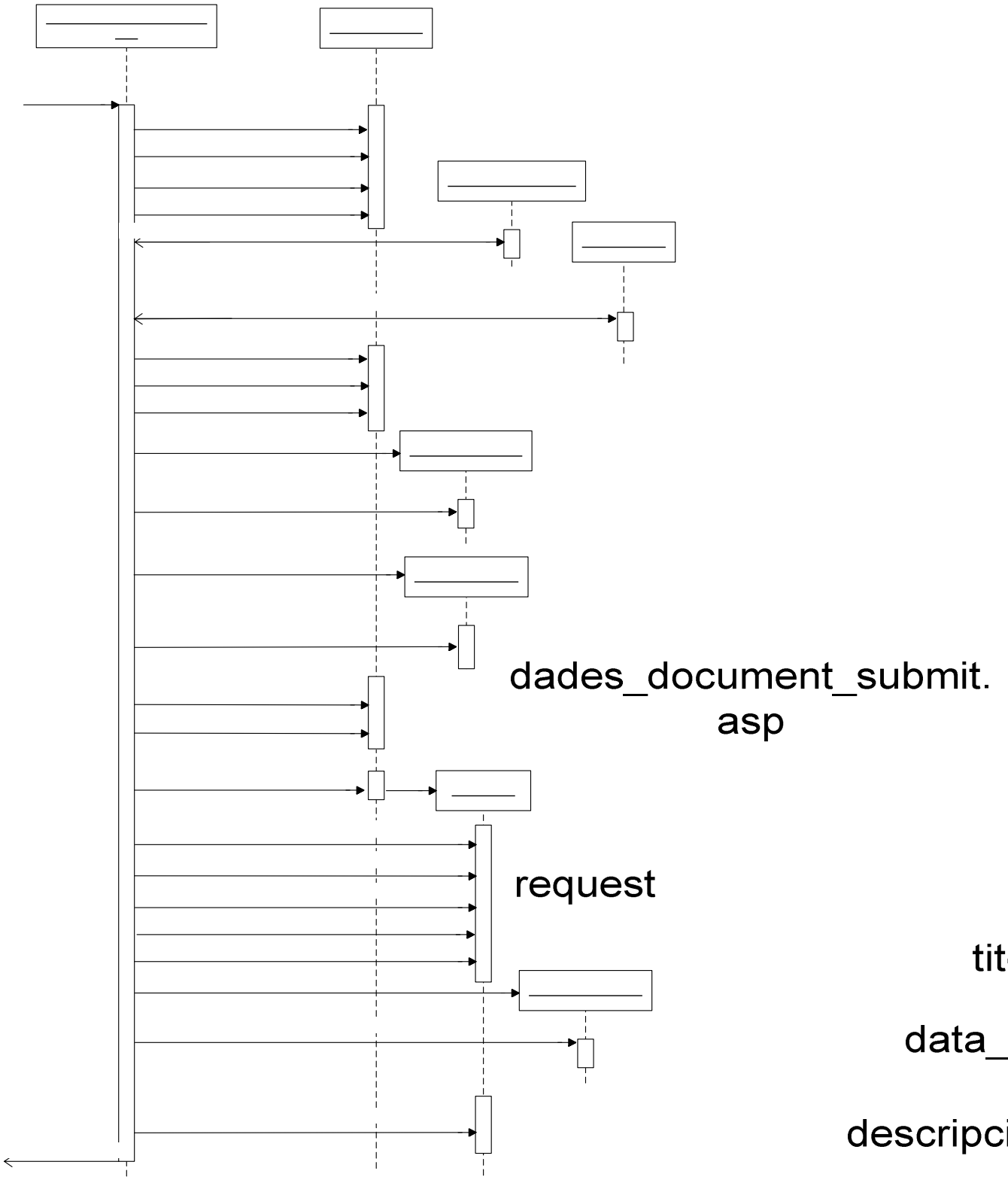
Sol·licitar esdeveniment (novetat)

Aquest es el diagrama corresponent a la sol·licitud d'un nou esdeveniment. Per les novetats, el diagrama és molt similar.

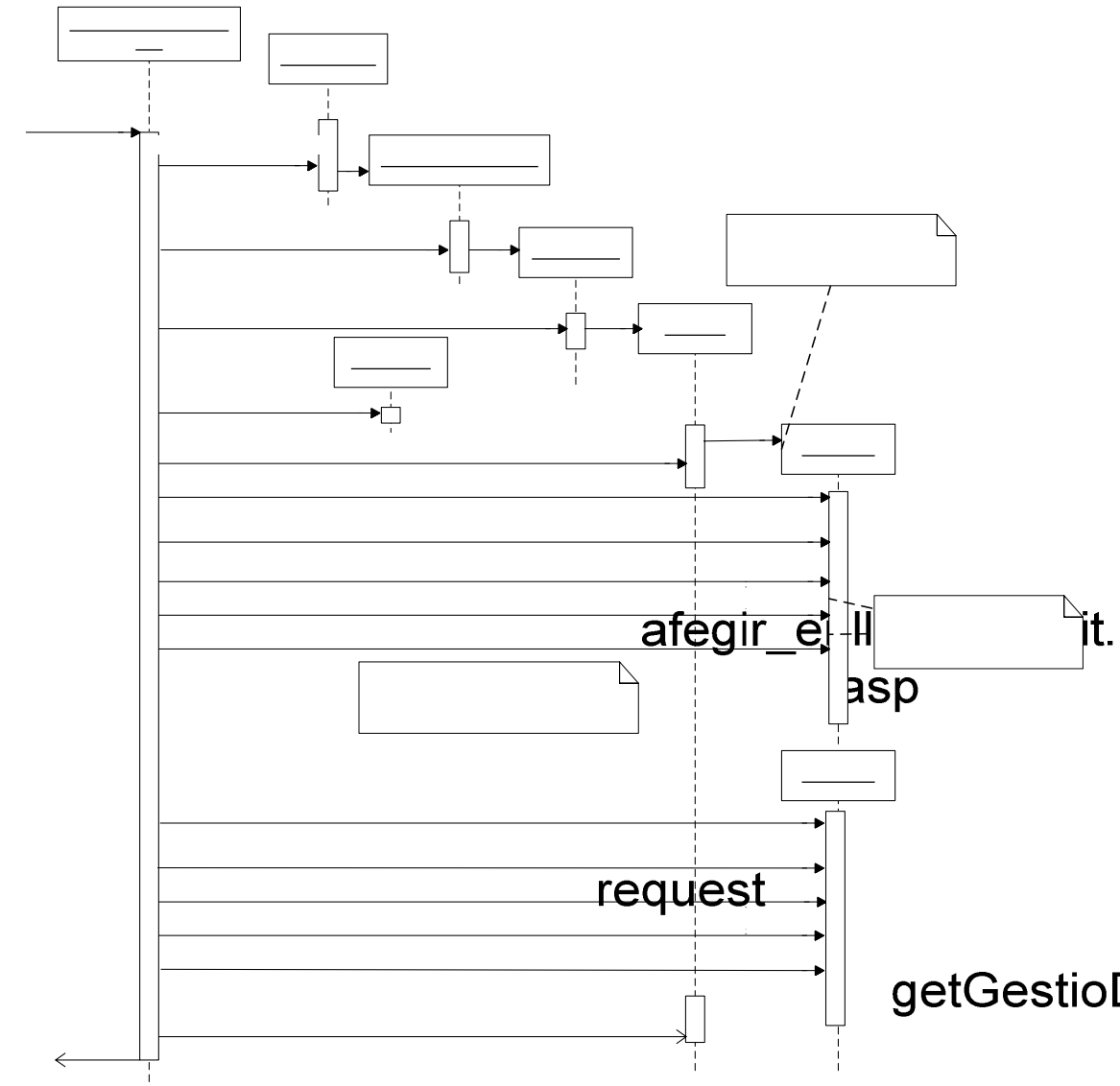


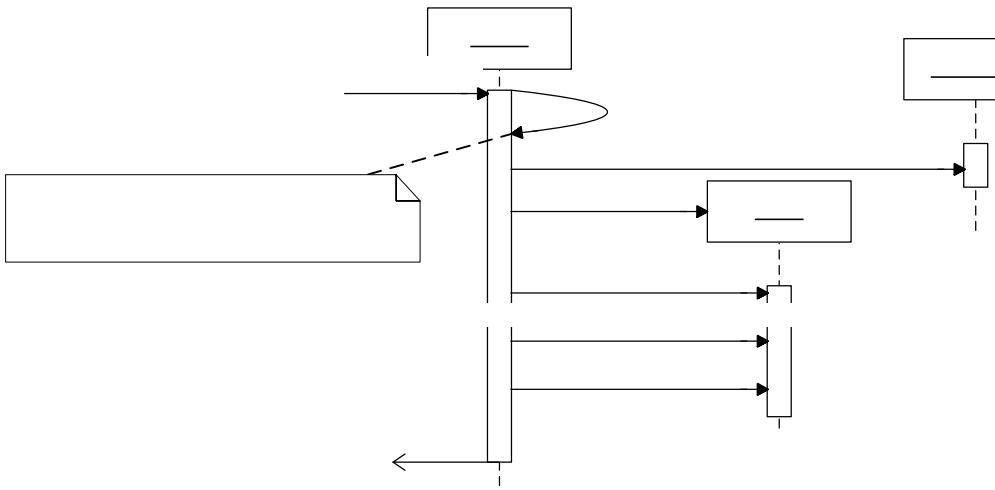
Mantenidor

Afegir document



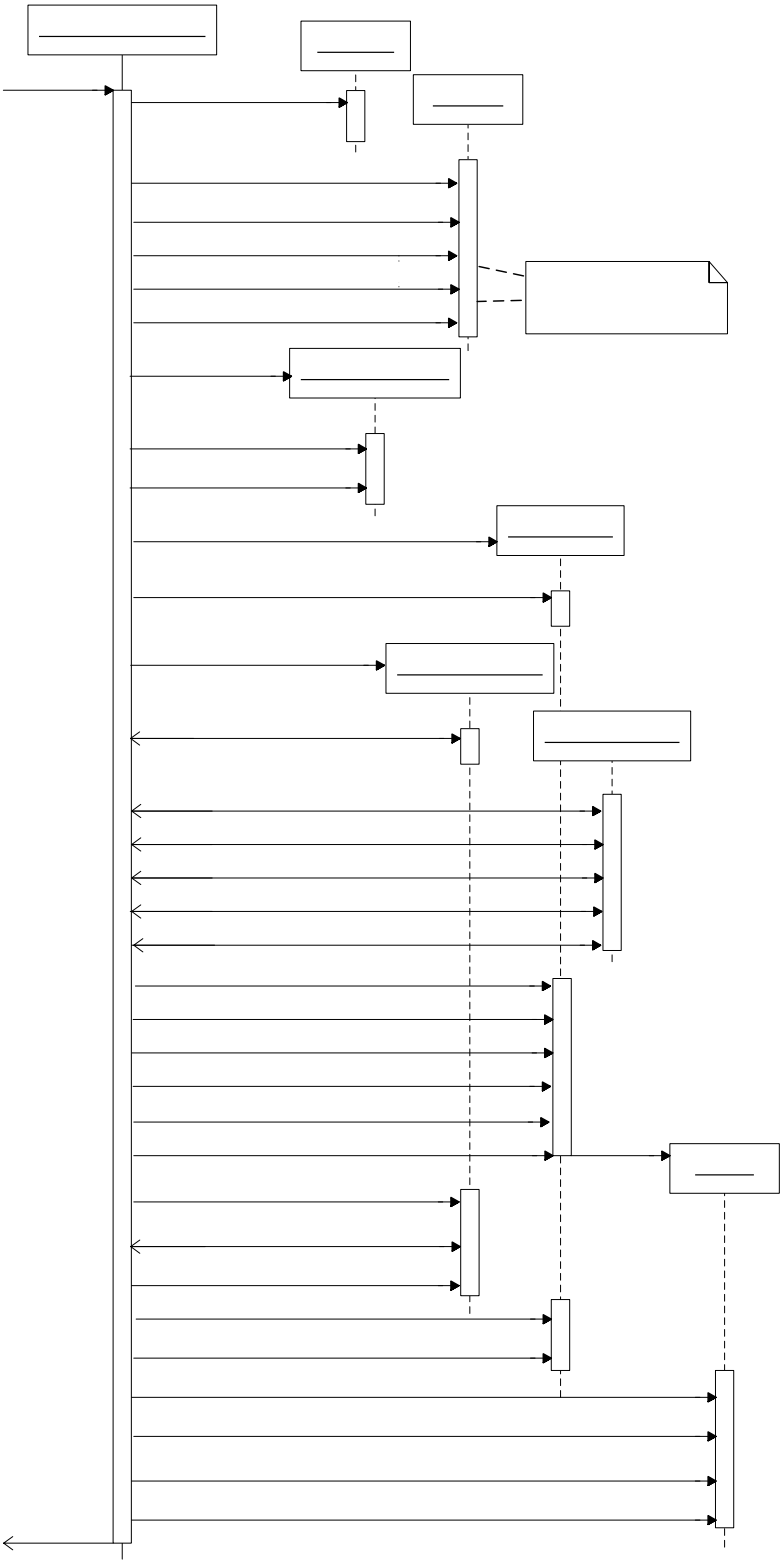
Afegir enllaç



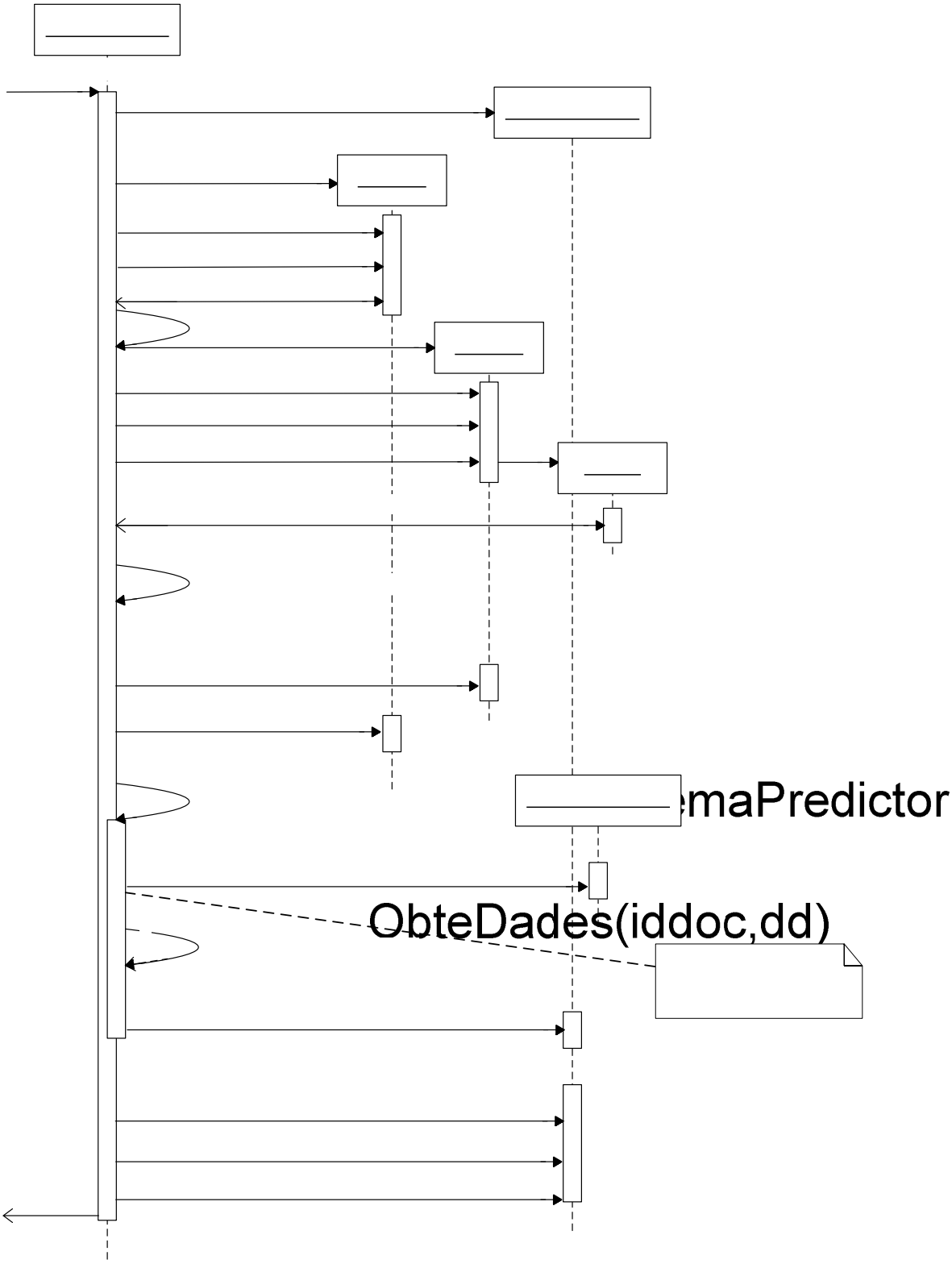


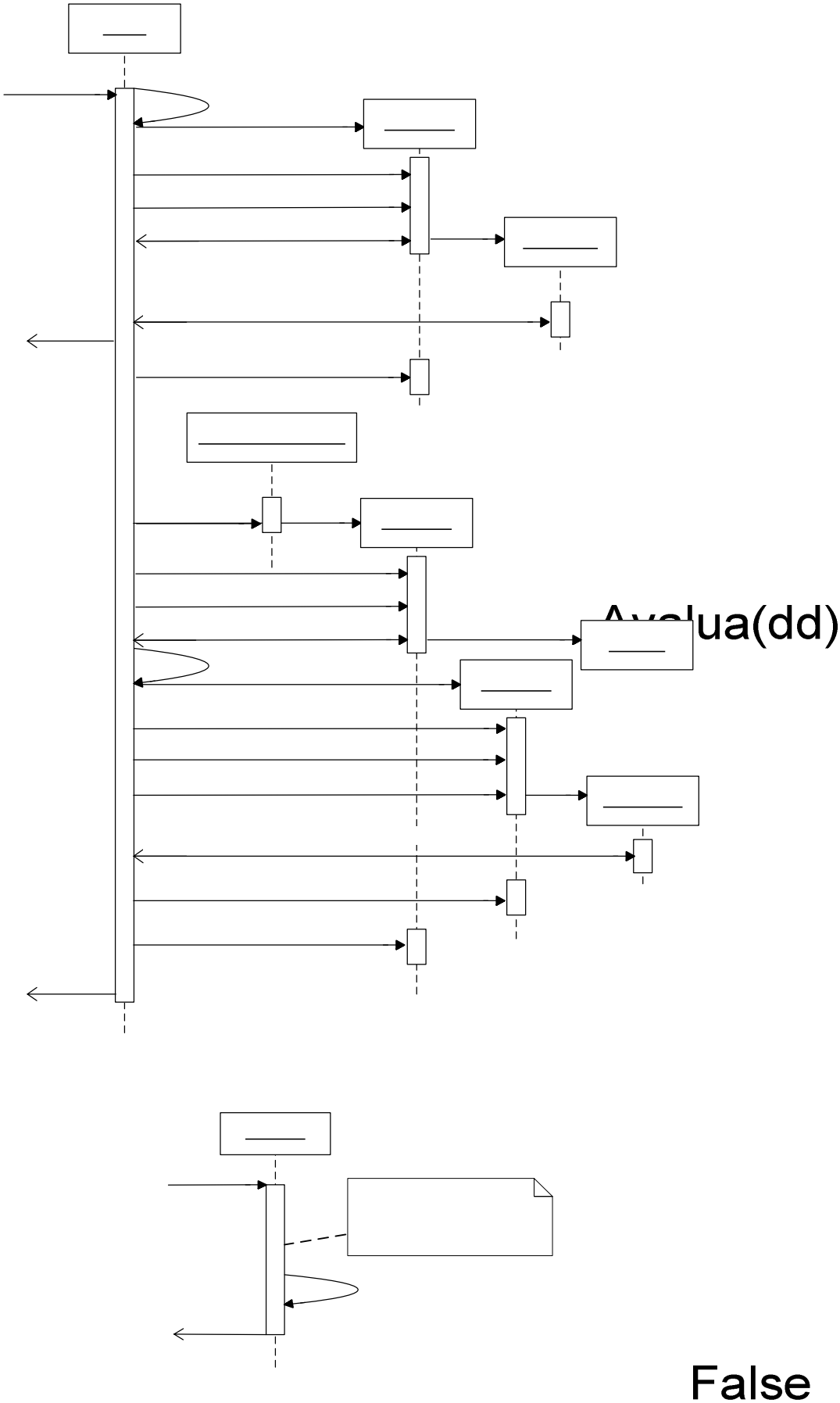
Afegir document omplint la fitxa automàticament

El disseny d'aquest cas d'ús utilitza varies funcions més complexes que es mostren més detalladament a les següents pàgines. No s'ha inclòs el diagrama de la funció *CalculaTemes(doc)* que selecciona els temes que s'assignaran al document perquè ja s'ha explicat en un apartat anterior de la memòria amb pseudocodi.

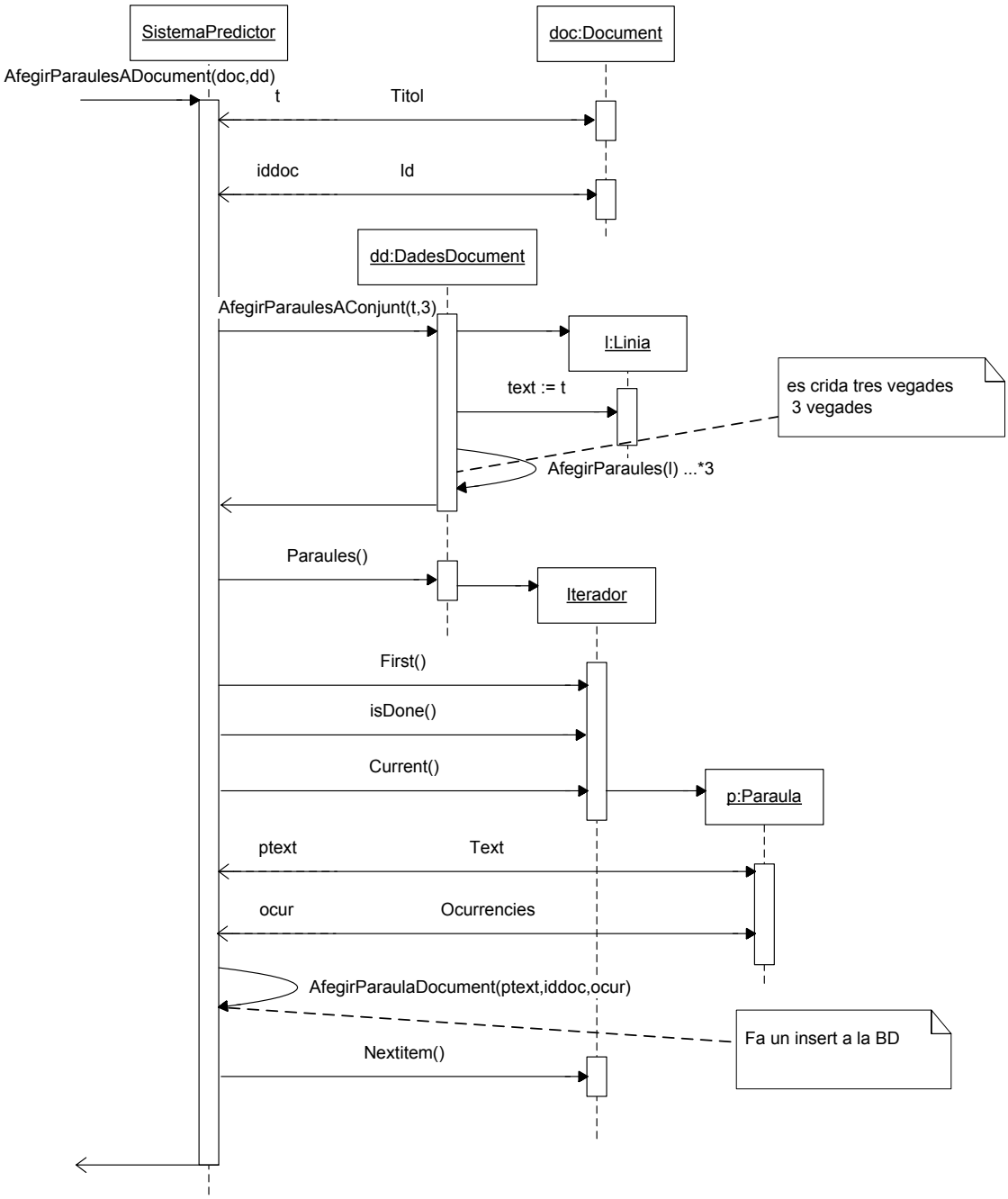


Aquesta funció, *ObteDades* de la classe *SistemaPredictor* obté les dades de títol, data, tipus i autors del document a partir de les regles que hi ha al sistema. Després busca els documents relacionats amb els enllaços que conté el document i finalment normalitza les dades. A la següent pàgina es mostra les funcions *Avalua* de la classe *Regla* i *Predicat* que s'utilitzen en aquest diagrama.

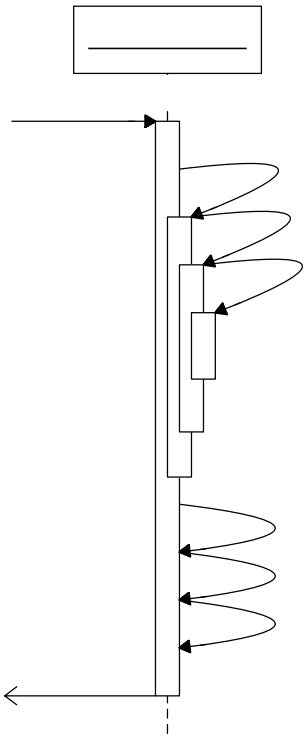



Regla
ua(dd)
False

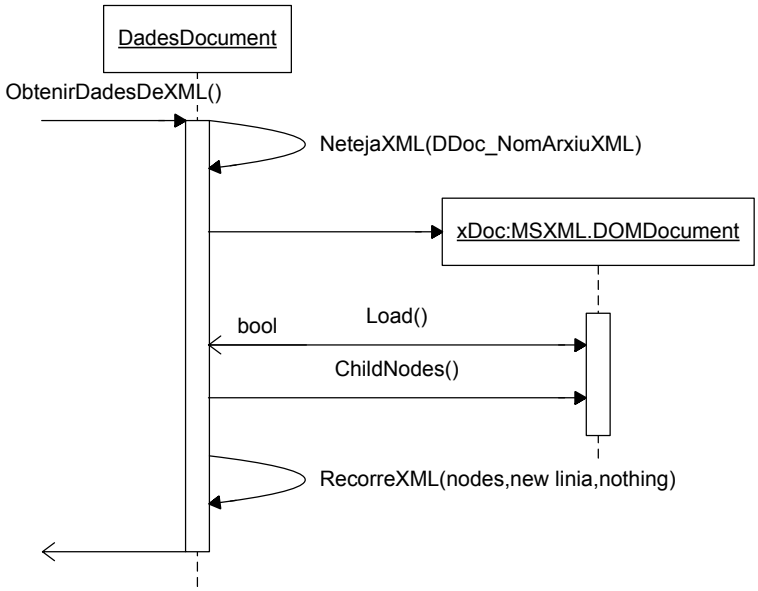
Aquesta funció, guarda a la base de dades les paraules que apareixen en un enllaç de document per poder fer la crida a les funcions que calculen la relació del document amb els temes.



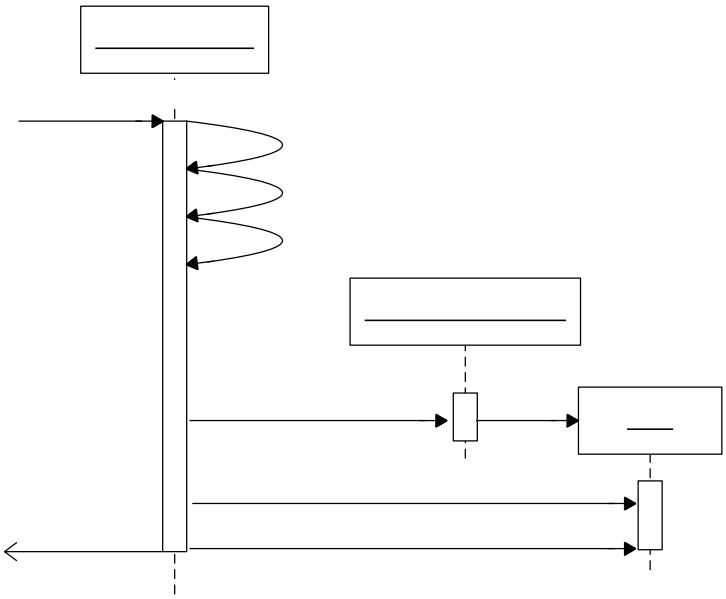
Finalment en aquest últims diagrames, es mostren alguns dels passos adoptats en el preprocés. En el primer diagrama, es mostra l'esquema dels passos necessaris per completar el preprocés. No s'ha inclòs el diagrama d'alguna d'aquestes funcions per ser molt simples o perquè els diagrames eren poc explicatius.



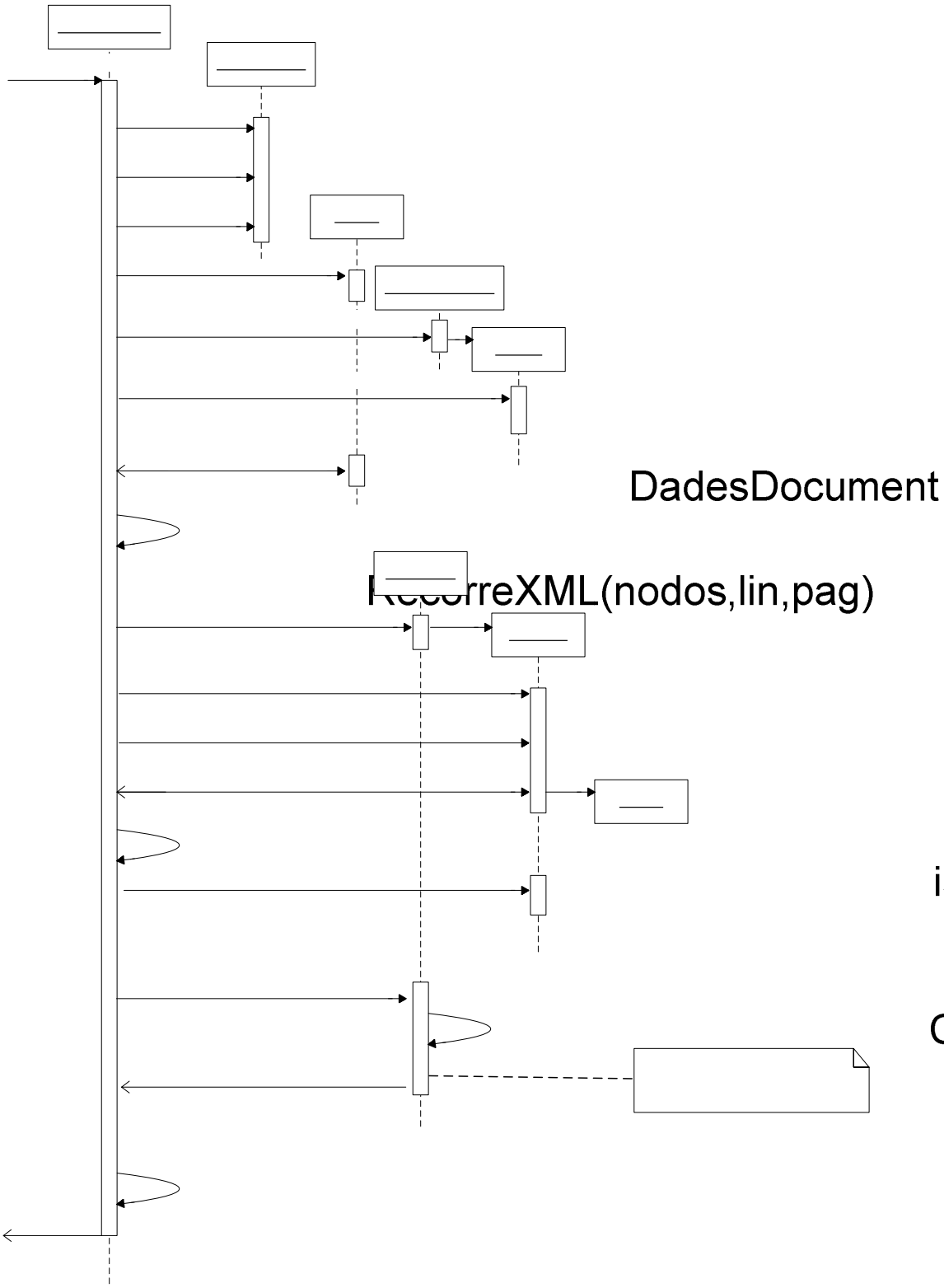
Aquest diagrama, similar al diagrama de la funció ObtenirParaulesDeXML() per obtenir les paraules de l'arxiu XML mostra les crides prèvies per fer la crida a la funció que recorre el XML i obté el conjunt de dades que es faran servir per avaluar les regles.



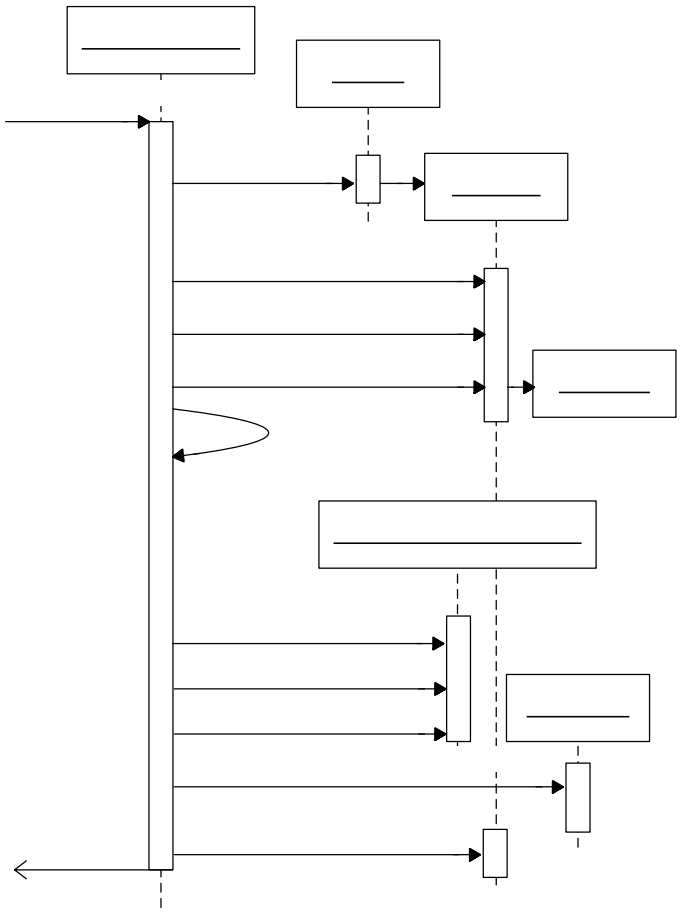
Aquest és l'esquema corresponent a la funció que descarrega un arxiu que es troba en un altre servidor. Primer fa una crida a la funció ObtenirUrl(...) que obté les dades de la url introduïda i en si es tracta d'un arxiu en format Html comprova si té frames i si els té crida a una funció recursiva que obté les dades del últim frame.



La següent funció, RecorreXML(...), es mostra un esquema de la funció recursiva que obté les línies de text i les paraules que es faran servir després. La part que es troba després de la crida a la funció recursiva es realitza en funció de si el node que estem tractant es de tipus text o de tipus pàgina i si la pàgina que tractem és la primera.


DadesDocument
RecorreXML(nodos,lin,pag)
First()
isDone(
Current(
"om

Aquesta és la funció que introdueix les paraules d'una línia de text si no són Stoppers. A més guarda el número d'ocurrències.



AfegirPa

Afegir versió

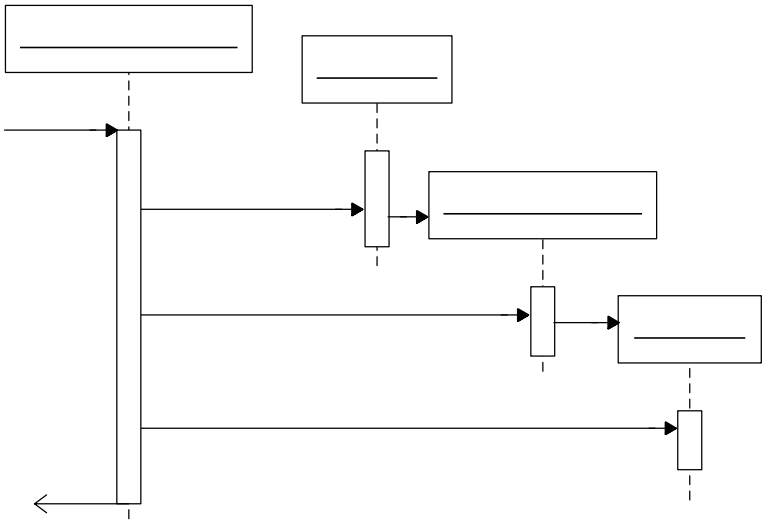
El diagrama corresponent a aquest cas d'ús és molt similar al d'afegir document. La única diferència és que no cal crear el document, sinó que s'ha de demanar a la classe *GestioDocuments* de la intranet actual amb la funció *ObtenirDocument*(id). Igualment s'han de guardar la resta de dades, menys la data d'alta del document tractat.

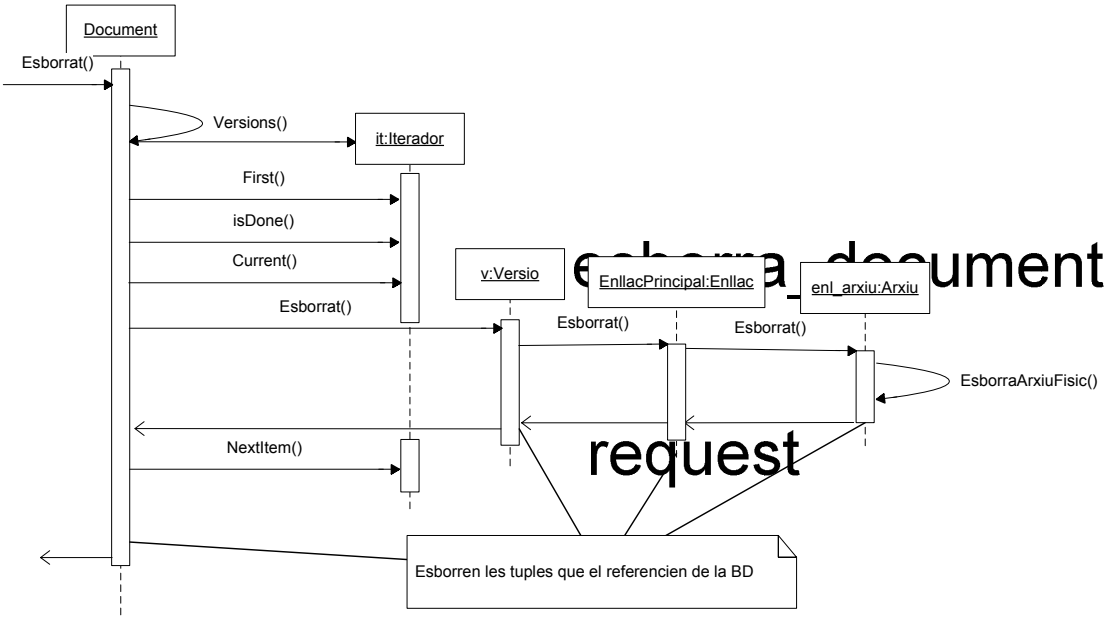
Modificar document

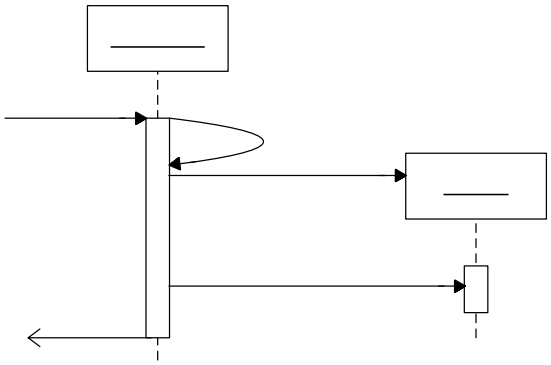
El diagrama corresponent a aquest cas d'ús és molt similar al d'afegir document. La única diferència és que no cal crear el document ni la versió, sinó que s'ha de demanar a la classe *GestioDocuments* de la intranet actual amb les funcions *ObtenirDocument(iddoc)* i *ObtenirVersio*(idver) i s'ha de substituir les funcions *AfegirXXX(cj)* per *CanviarXXX(cj)*. Igualment s'han de guardar la resta de dades, menys la data d'alta del document tractat.

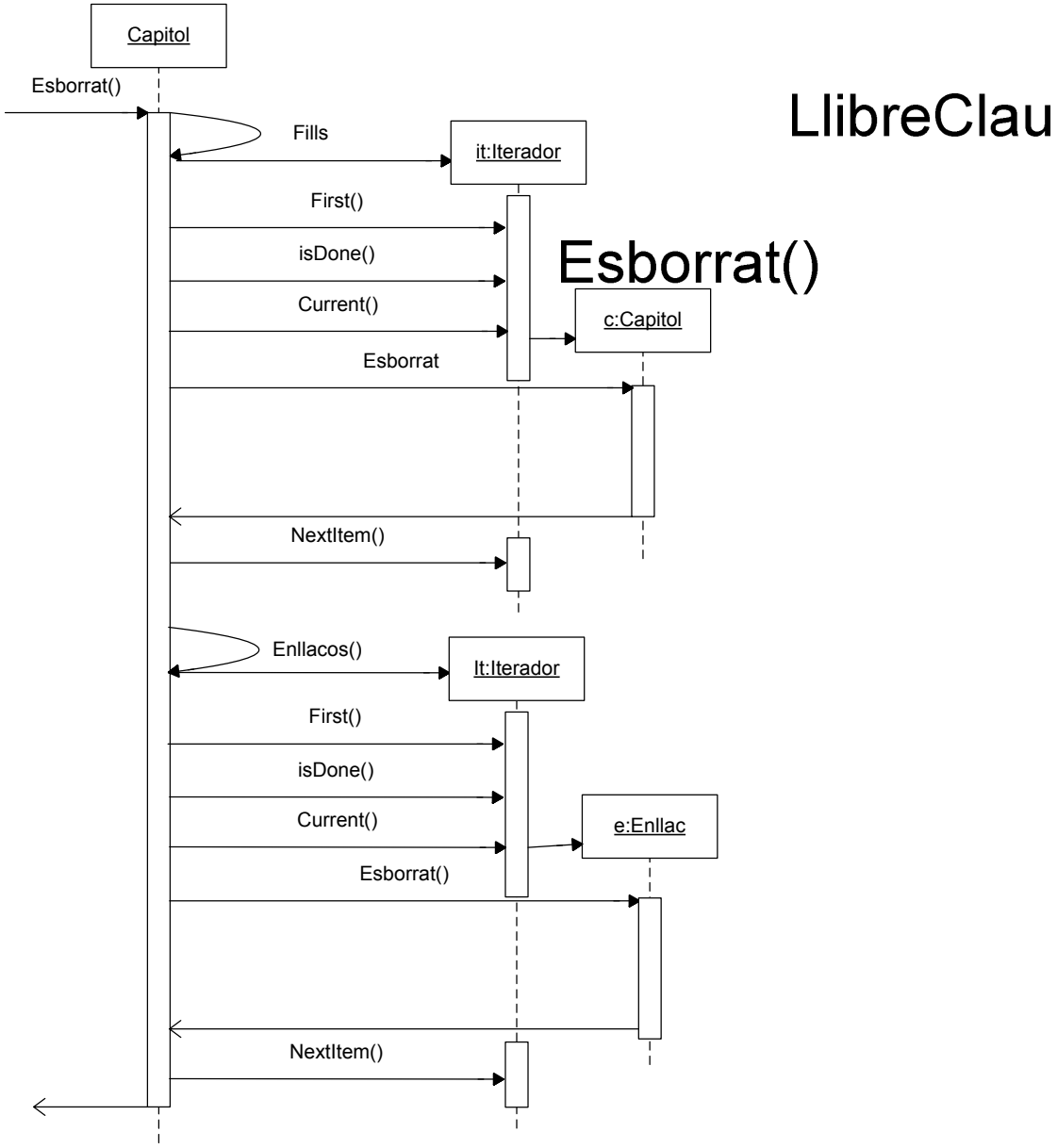
Esborrar document

En aquest diagrames es mostra com s'esborren els documents normals i els de tipus llibres clau.

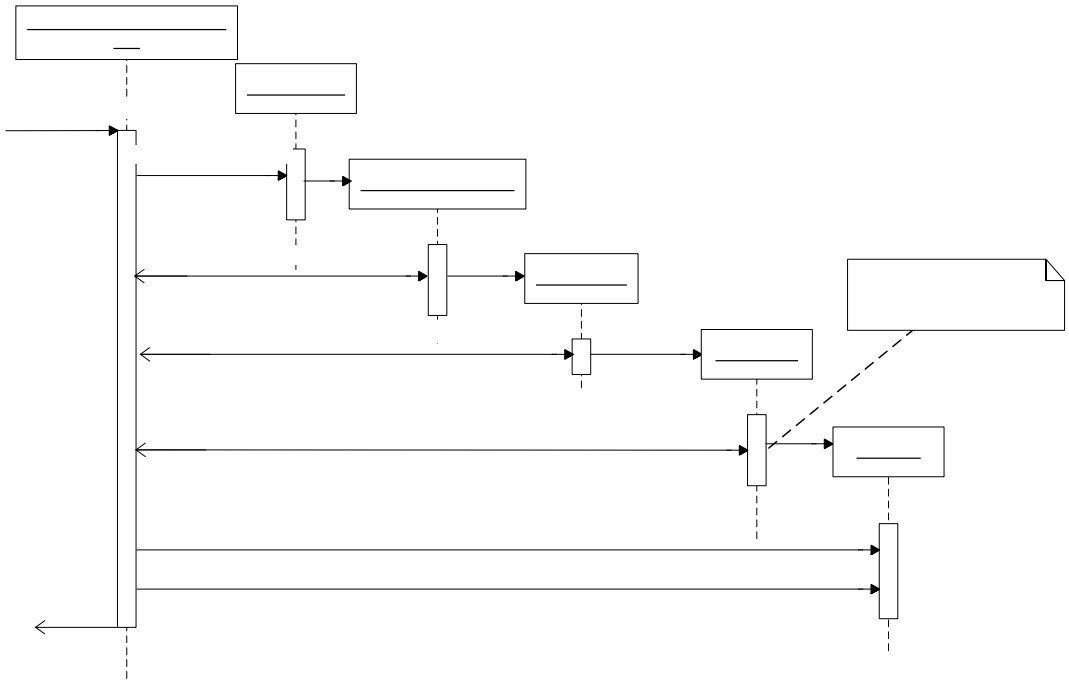




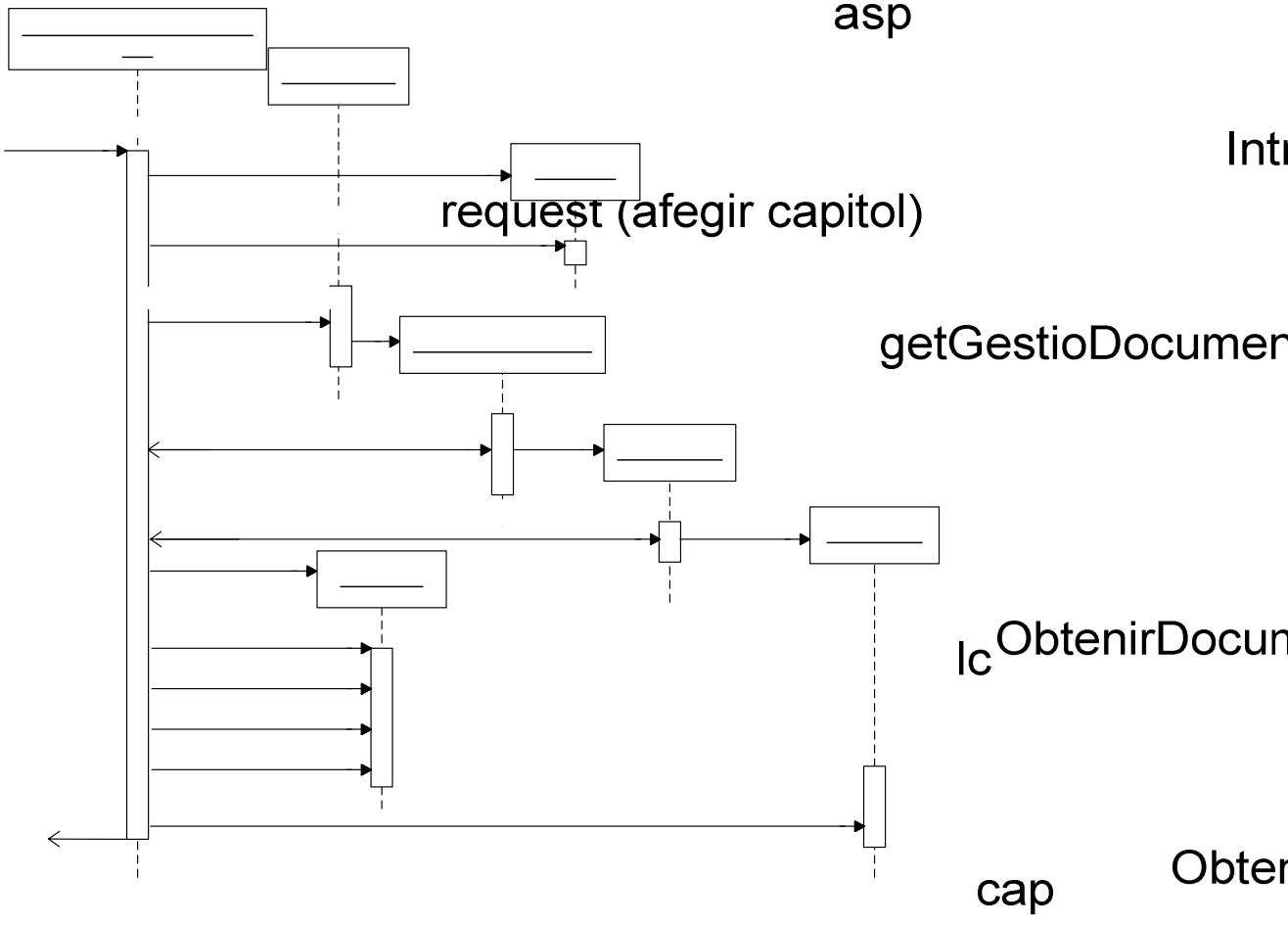



Capitol
Esborrat()
Fills
it:Iterador
First()
isDone()
Current()
c:Capitol
Esborrat
NextItem()
Enllacos()
It:Iterador
First()
isDone()
Current()
e:Enllac
Esborrat()
NextItem()
LlibreClau
Esborrat()

Crear apartat / enllaç llibre clau

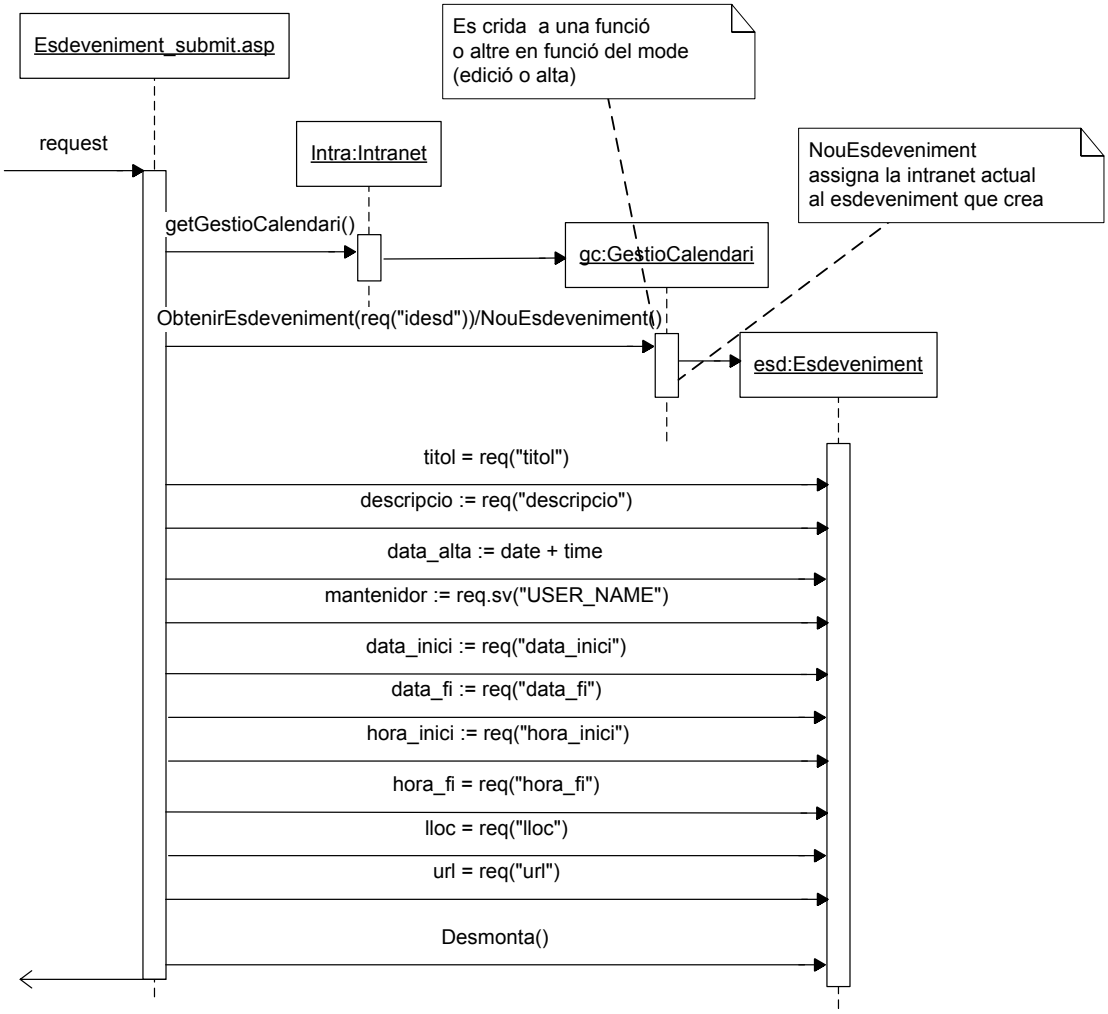


objecte_llibre_clau_submit.
asp

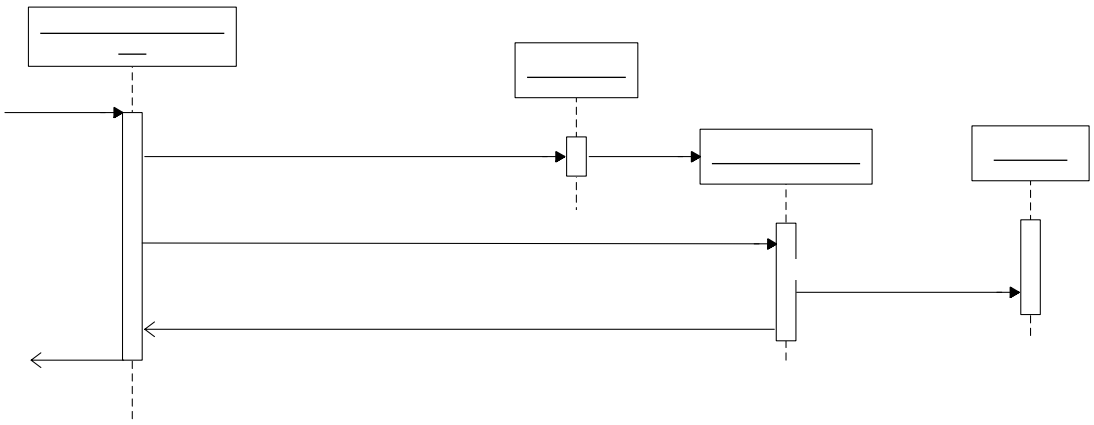


Administrador

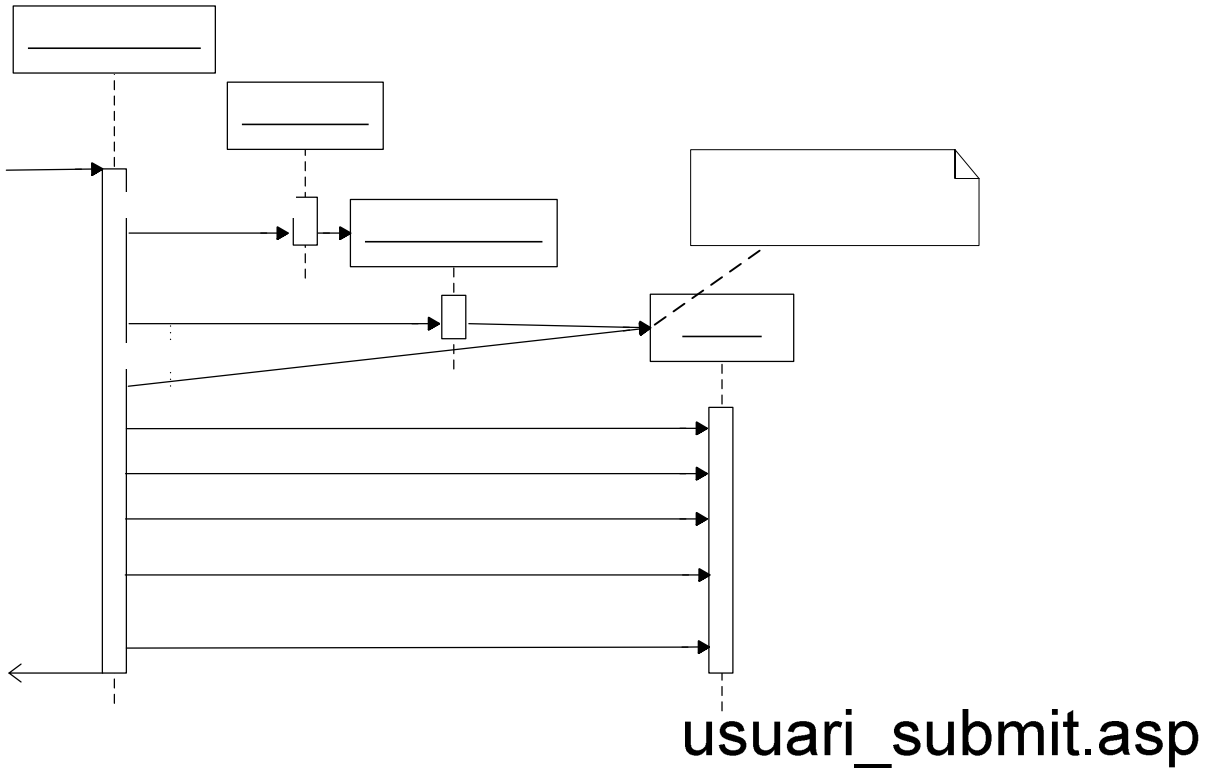
Gestió del calendari – Afegir/editar esdeveniment (novetat)



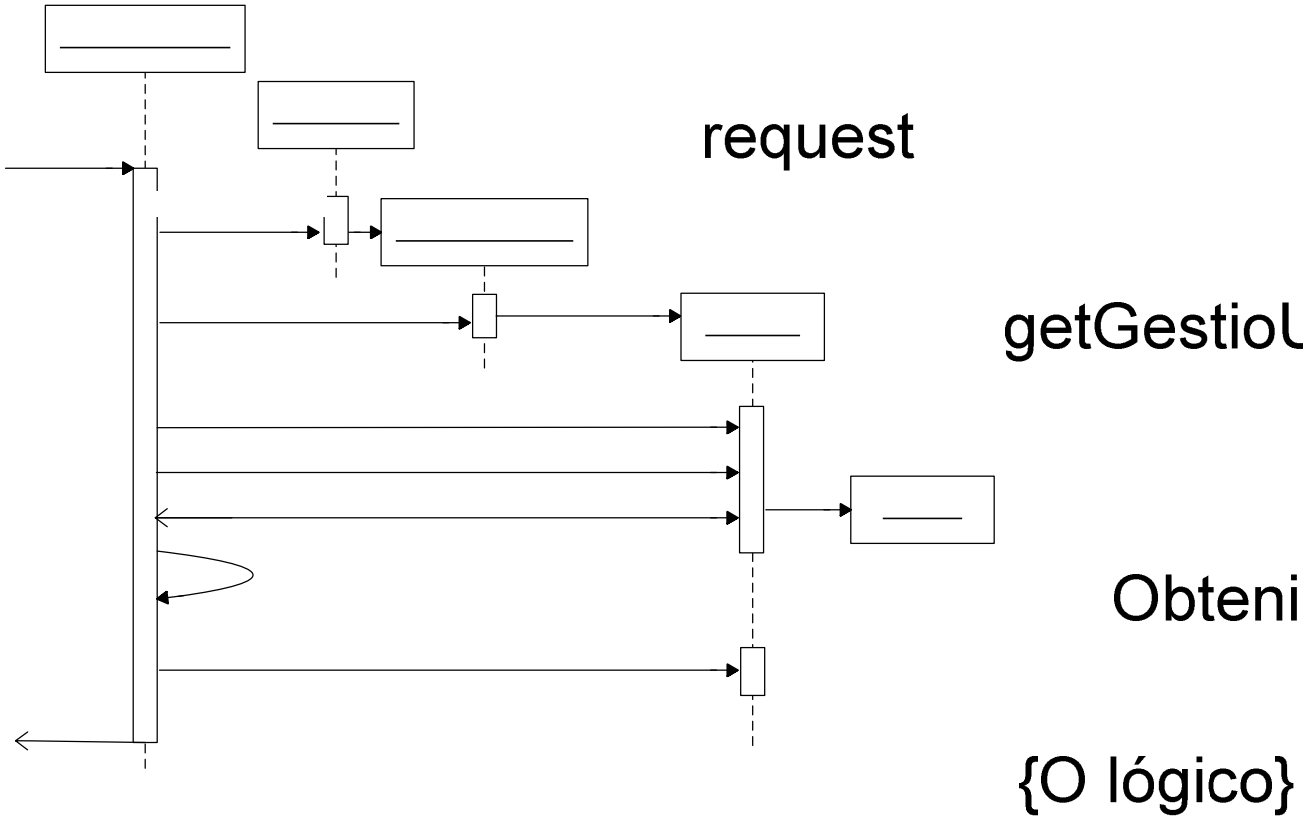
Gestió del calendari – Esborrar esdeveniment (novetat)



Gestió d'usuaris – Afegir/editar usuari

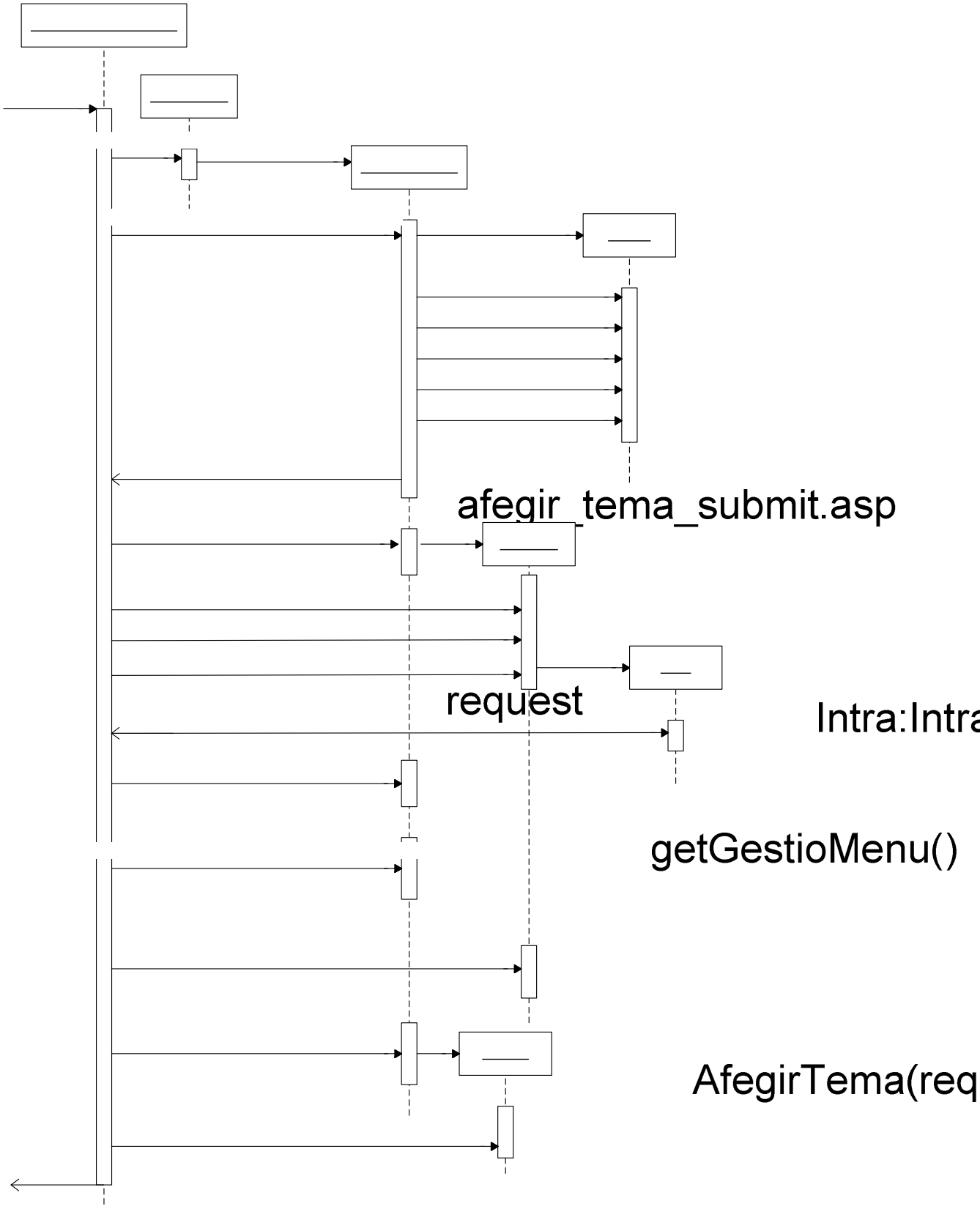


Gestió d'usuaris – Llistar usuaris

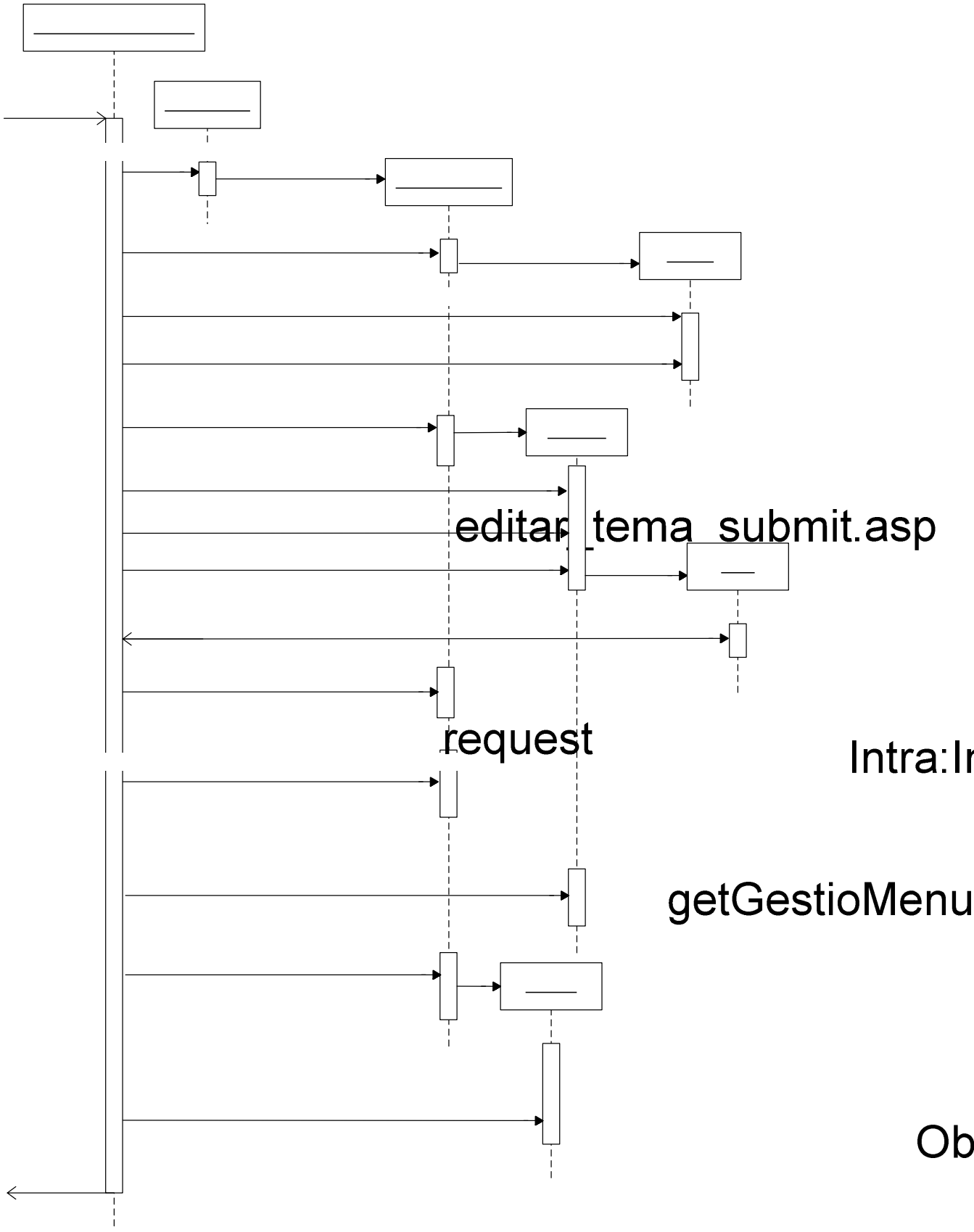


Gestió d'usuaris – Seleccionar responsables d'eix

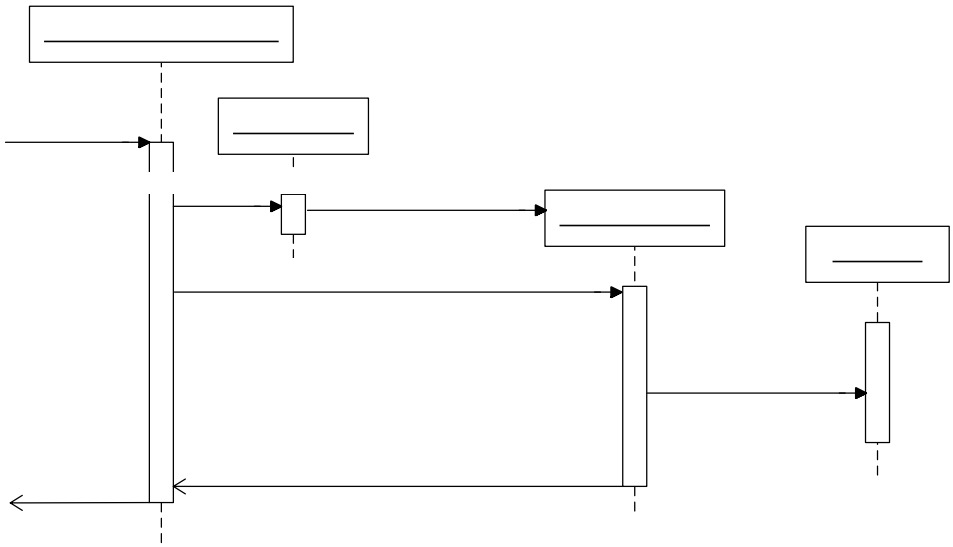
Gestió d'arbre temàtic- Afegir tema



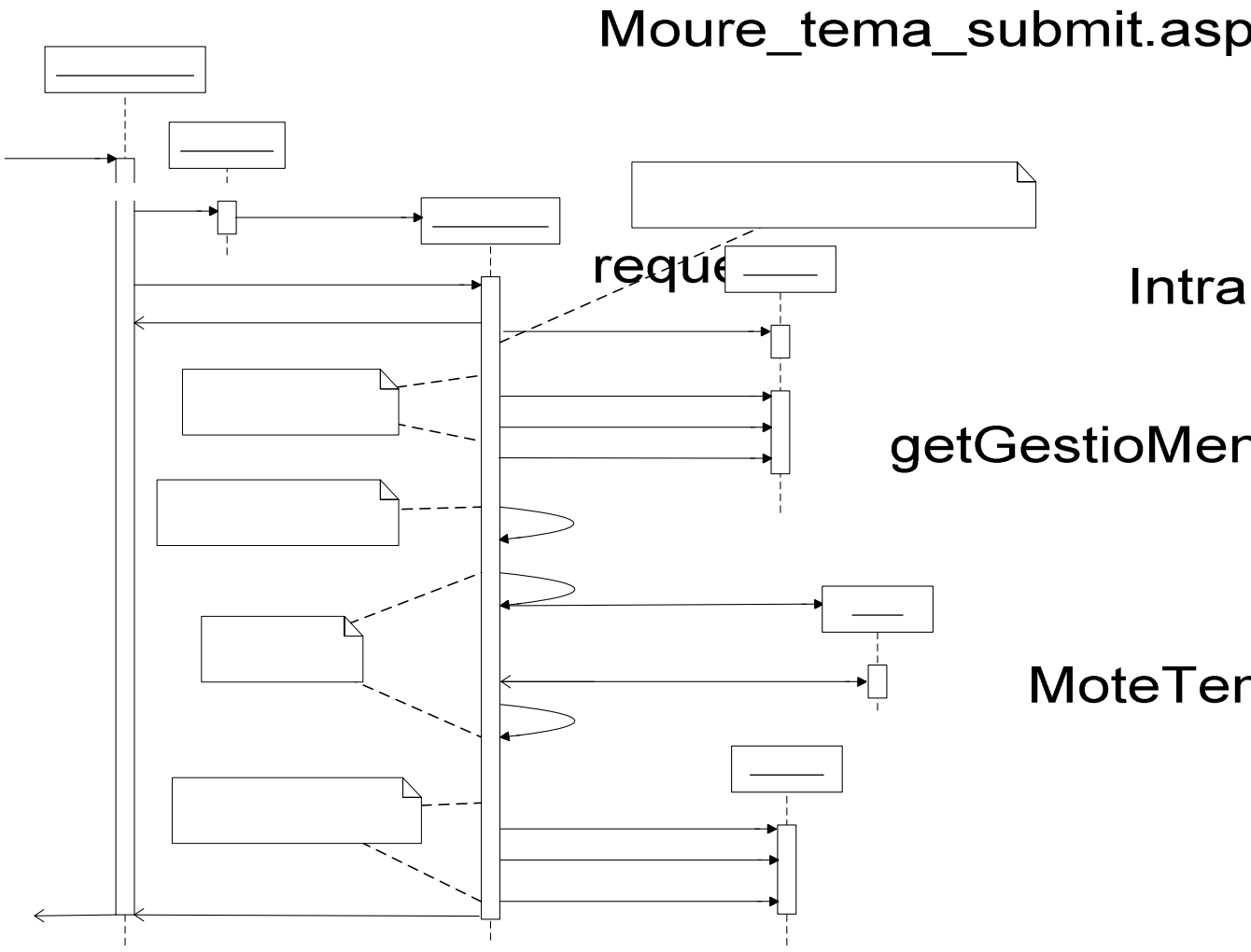
Gestió d'arbre temàtic- Editar tema



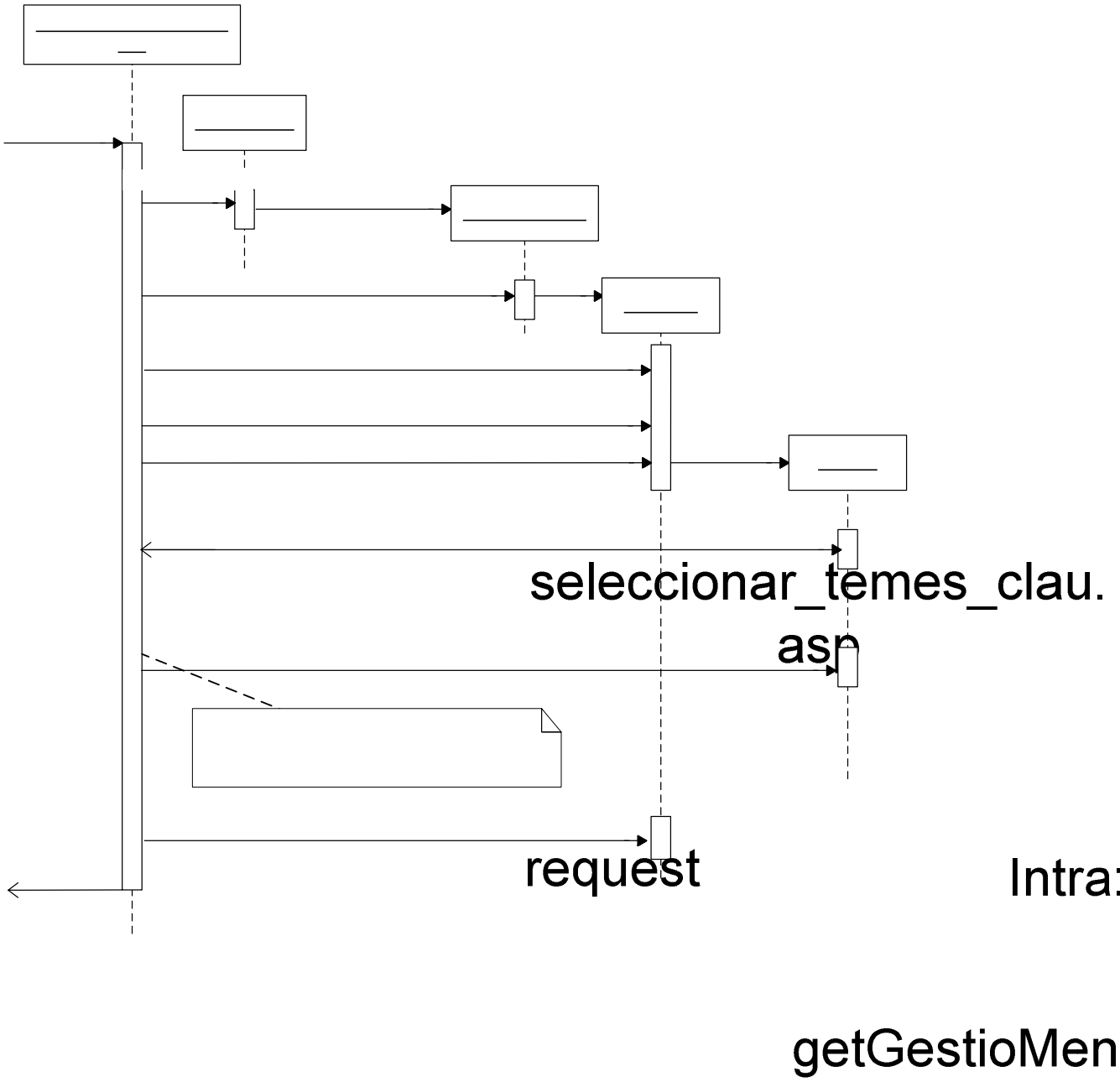
Gestió d'arbre temàtic- Moure tema



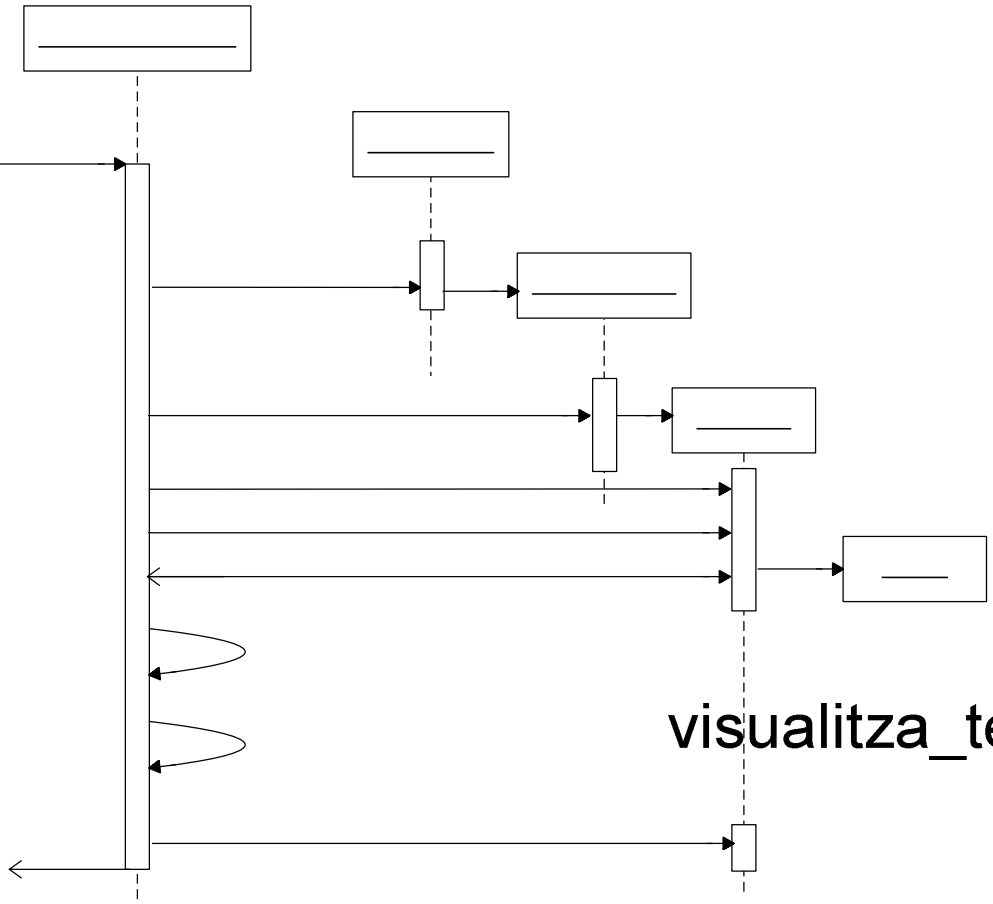
Gestió d'arbre temàtic- Esborrar tema

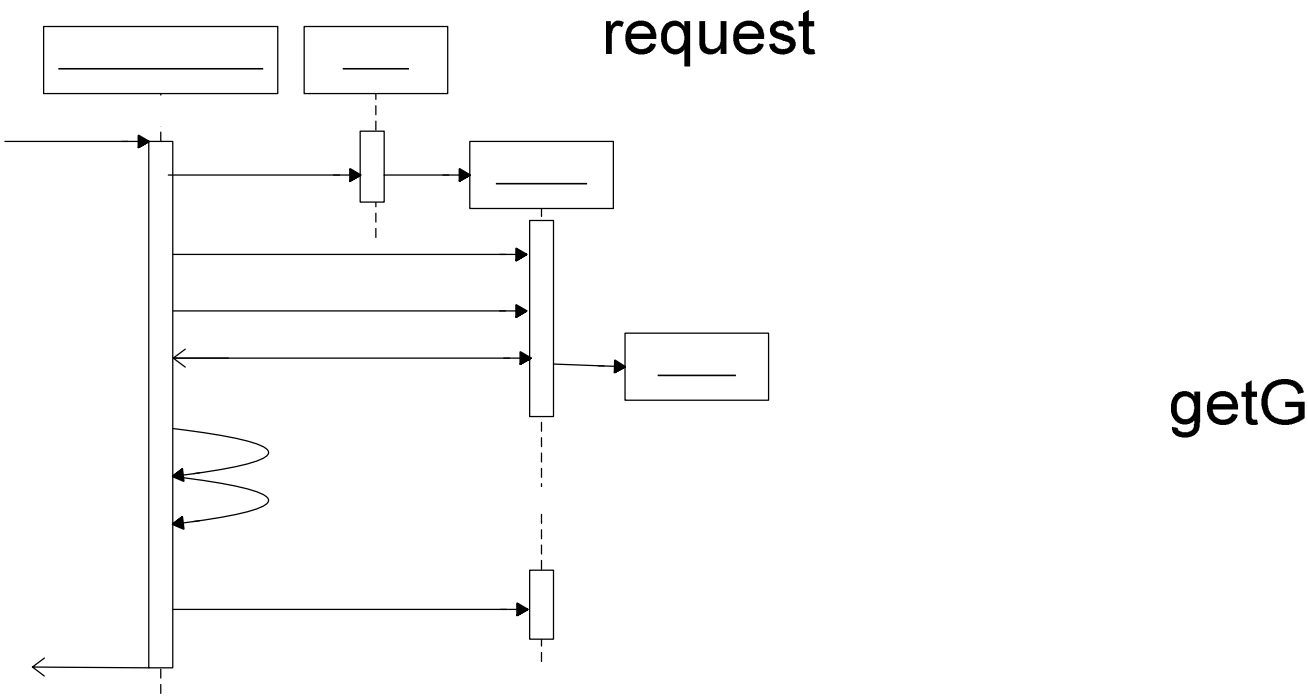


Gestió d'arbre temàtic- Seleccionar temes clau

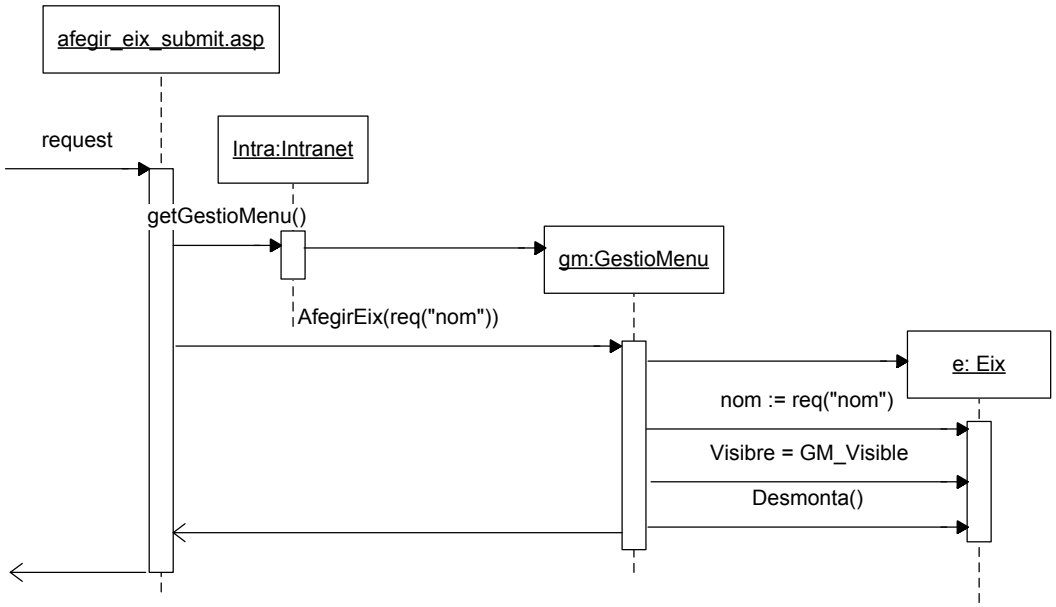


Gestió d'arbre temàtic- Visualitzar tots els temes

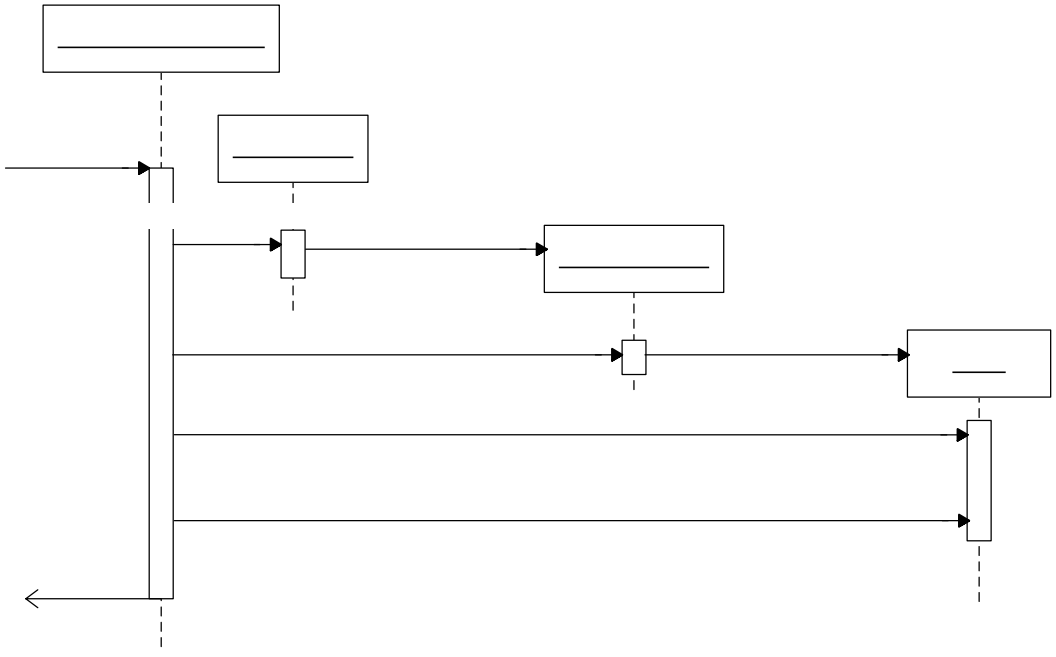




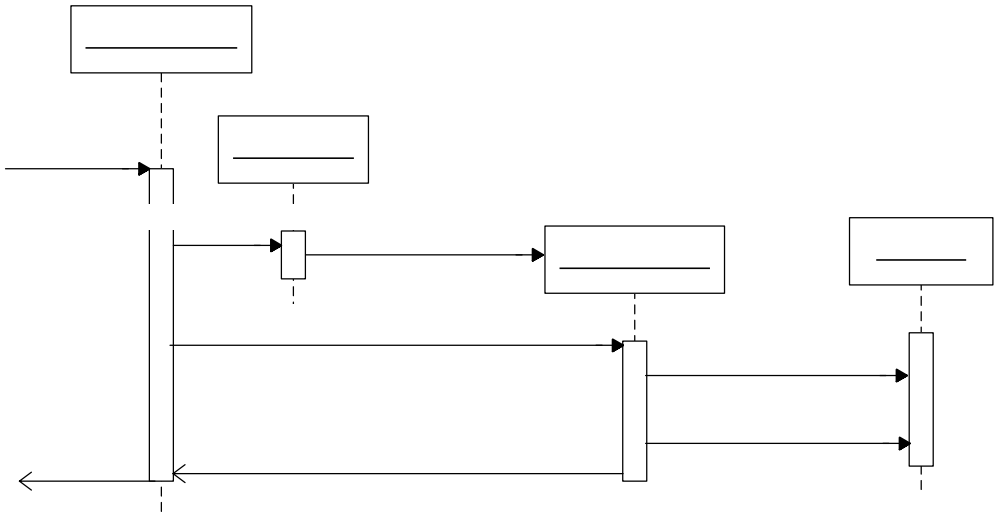
Gestió d'arbre temàtic- Afegir eix



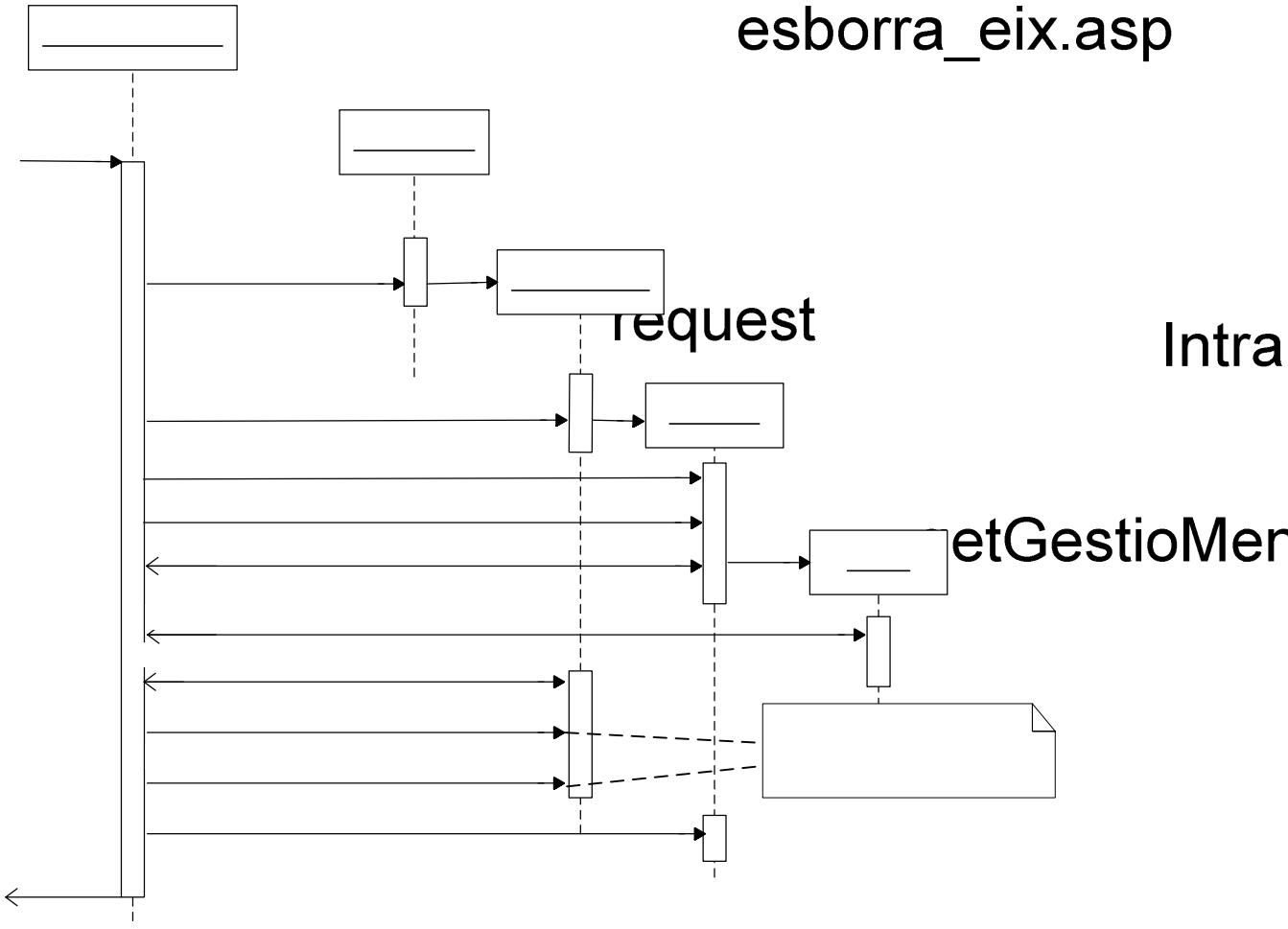
Gestió d'arbre temàtic- Editar eix



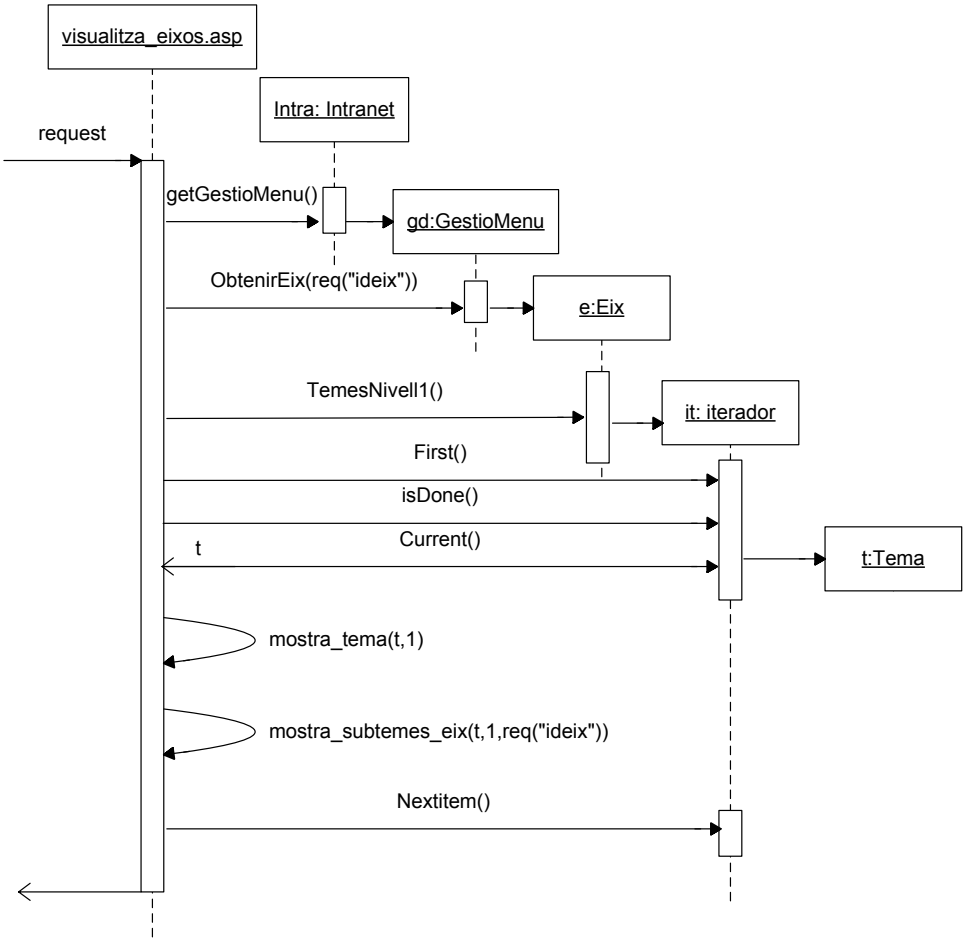
Gestió d'arbre temàtic- Esborrar eix

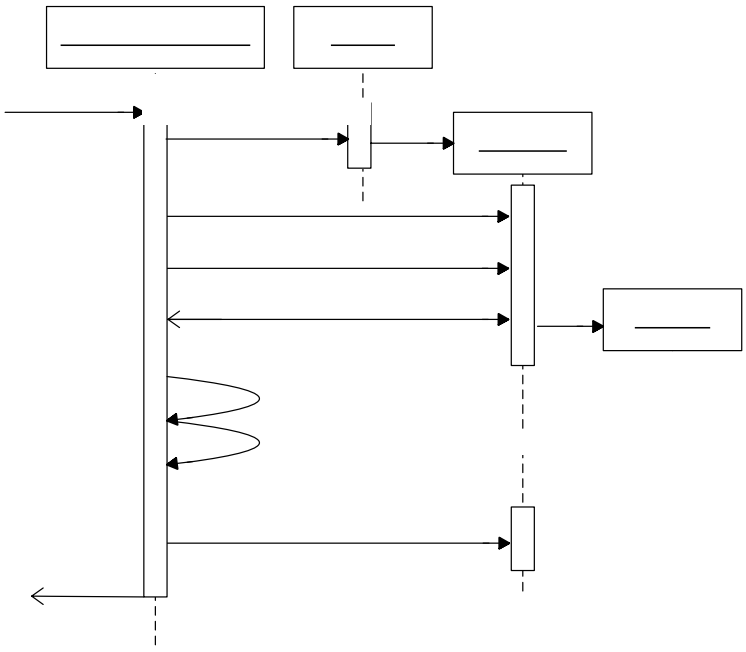


Gestió d'arbre temàtic- Relacionar eixos i temes

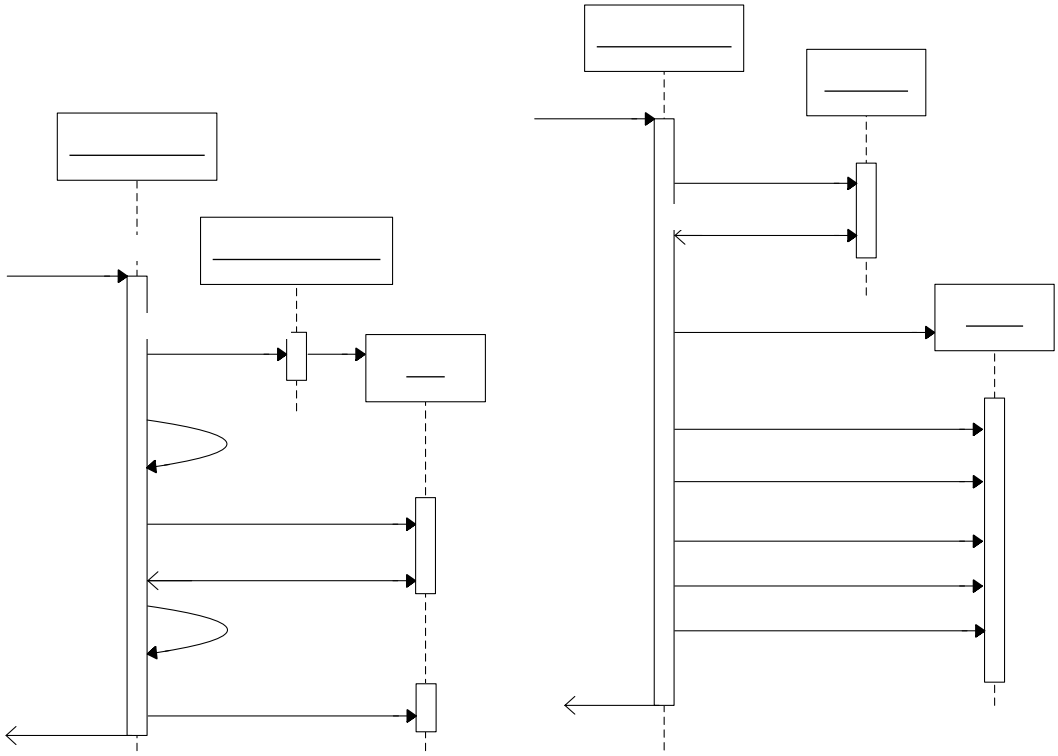


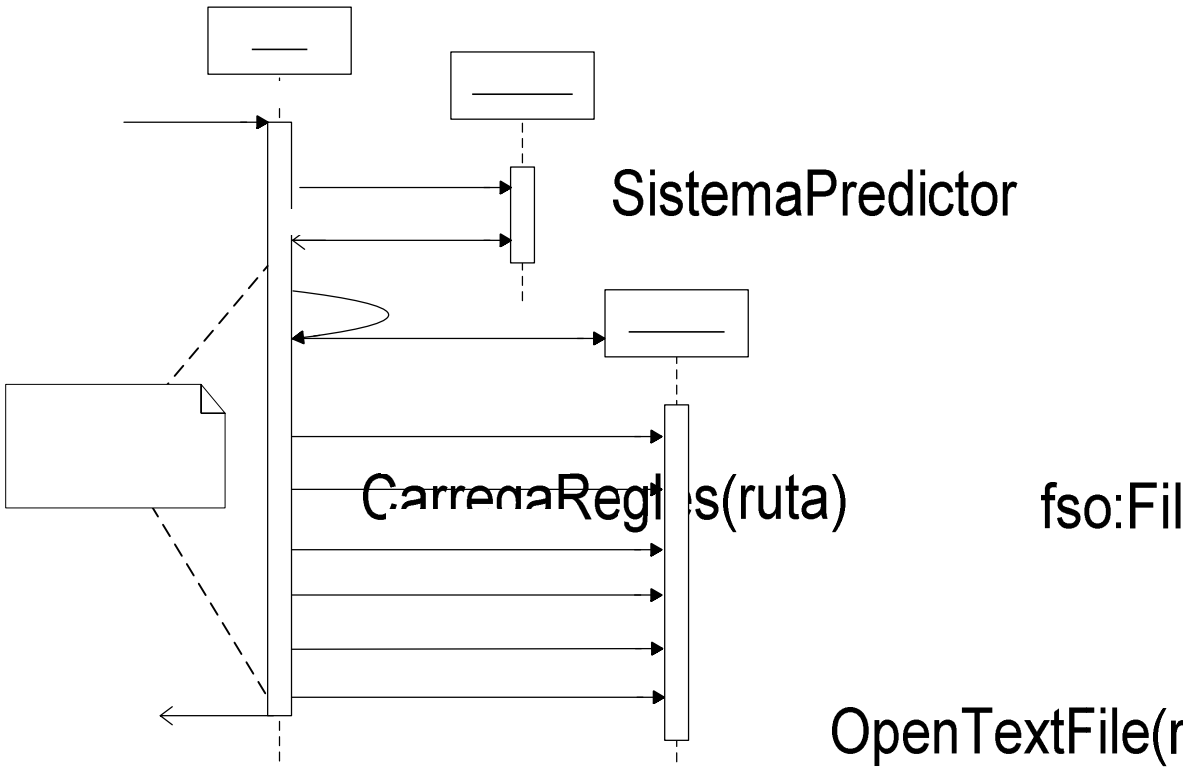
Gestió d'arbre temàtic- Mostrar temes relacionats eix





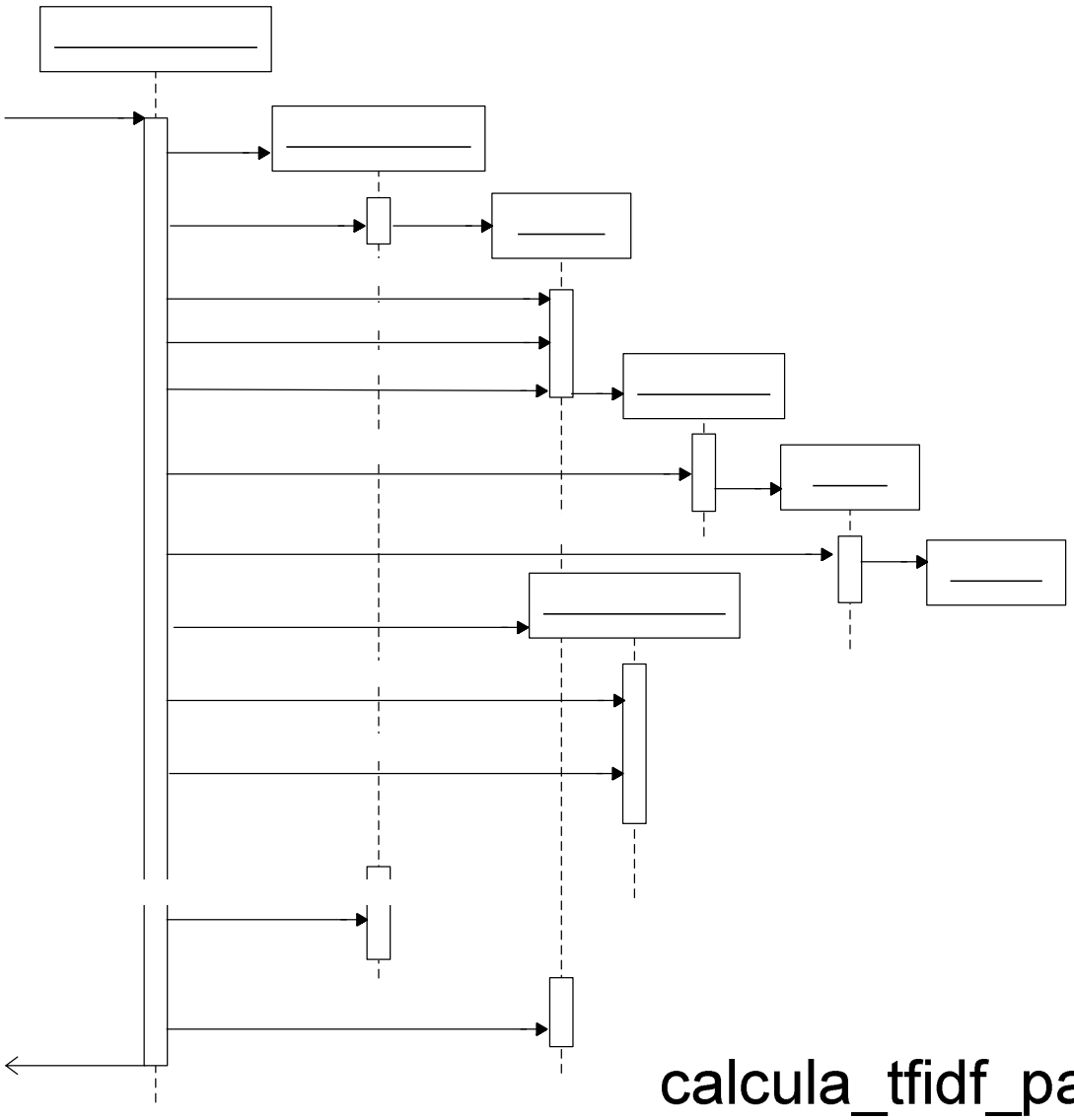
Sistema IA – Afegir regles

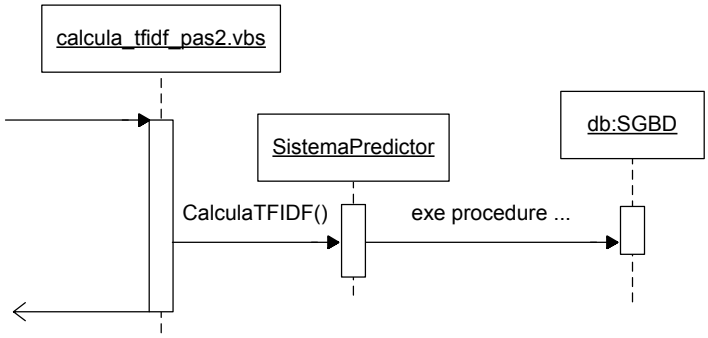


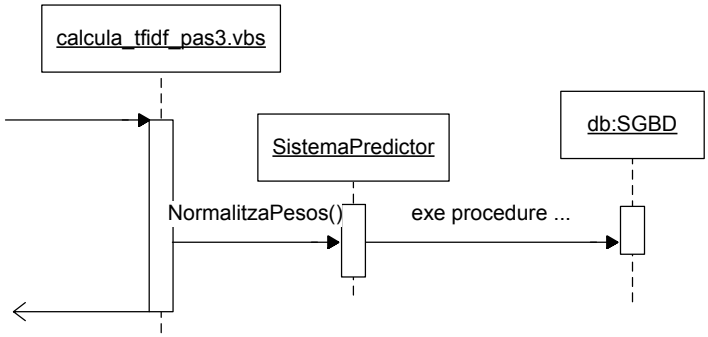


Sistema IA - Calcular TFIDF

Aquest cas d'ús es divideix en tres passos. En el primer s'insereixen les paraules del document a la base de dades. En el segon es fan els càlculs del TFIDF amb la crida a un procediment emmagatzemat. En l'últim pas es torna a cridar a un altre procediment emmagatzemat que normalitza la probabilitat de cada tema.

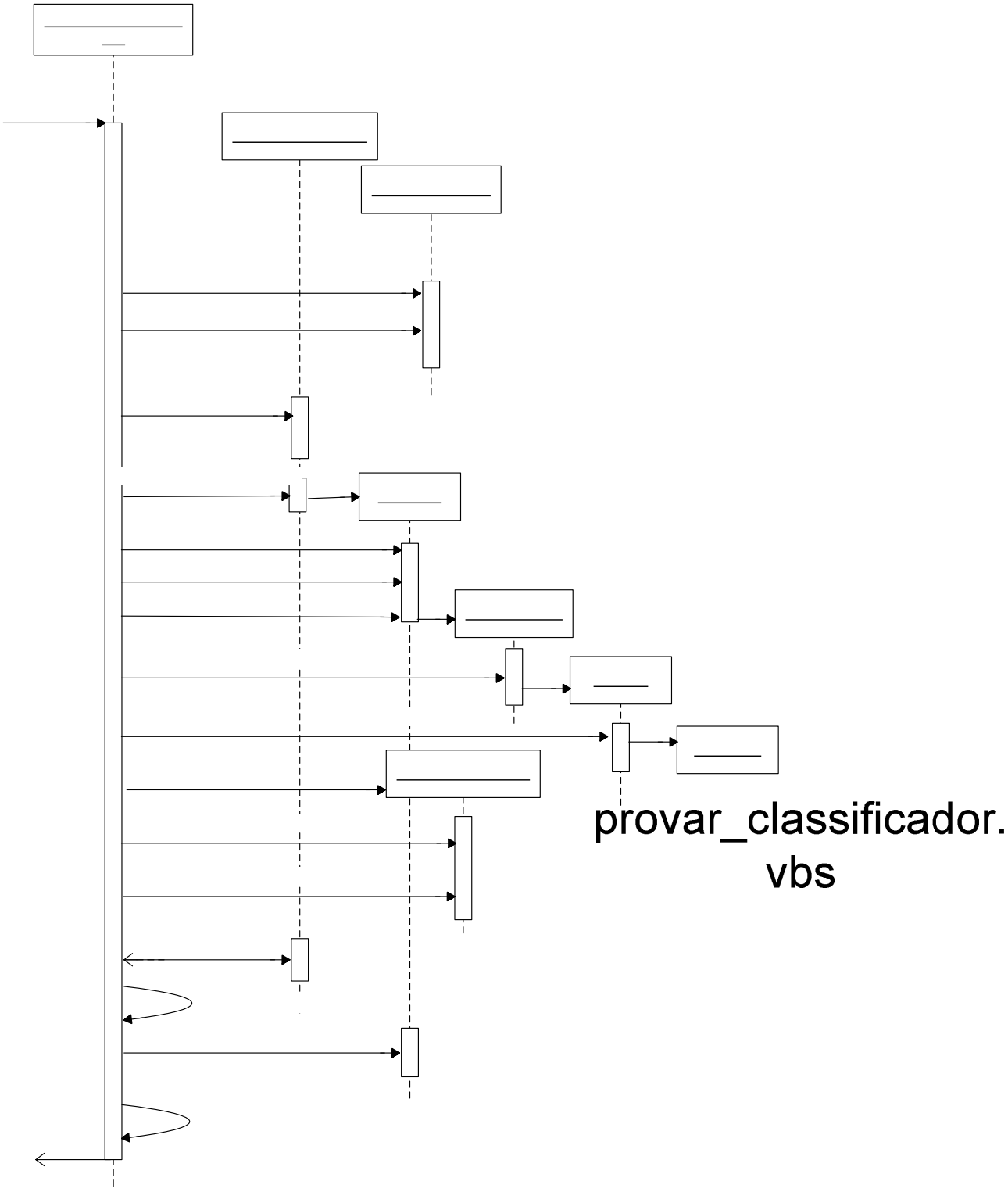



calcula_tfidf_pas2.vbs
SistemaPredictor
db:SGBD
CalculaTFIDF()
exe procedure ...


calcula_tfidf_pas3.vbs
SistemaPredictor
db:SGBD
NormalitzaPesos()
exe procedure ...

Sistema IA – Provar classificador

Aquest és el diagrama del disseny del cas d'ús provar classificador. En aquest cas només és mostra les proves amb una única configuració. En el script implementat en realitat es poden provar més d'una configuració. Algunes de les funcions que apareixen en el diagrama estan definides en el cas d'ús d'afegir document omplint la fitxa.



Decisions de codificació

Per implementar el sistema s'ha fet servir els següents llenguatges de programació:

1. ASP: Active Server Pages. És un llenguatge derivat del Visual Basic i incorpora eines pel desenvolupament ràpid d'aplicacions per un entorn web. El fet de desenvolupar les aplicacions amb aquest llenguatge és perquè és natiu al sistema Windows 2003, cosa que facilita la programació sense haver d'instal·lar res en el servidor.

   És un llenguatge de programació per aplicacions distribuïdes per internet i que permet incrustar codi en el llenguatge HTML per tal de realitzar l'accés a una o vàries bases de dades.

2. Visual Basic Script. També es un llenguatge derivat del Visual Basic i permet desenvolupar aplicacions a mode d'scripting per fer funcionar en el servidor en mode background. Aquest llenguatge s'utilitza per afegir les regles al sistema i escollir les paraules de cada tema

Ambdós llenguatges tenen molts punts de semblança, malgrat això hi ha punts que fan diferents els dos llenguatges i que no permet fer servir el codi desenvolupat per ASP dins codi VBScript de manera nativa. Per aquest motiu s'ha fet servir un petit motor de reescriptura desenvolupat per altres projectes semblants del Servei que permet fer servir el model conceptual desenvolupat per pàgines ASP dins de codi Visual Basic Script. A continuació es mostra el codi que permet aquesta conversió.

```
Sub Include(sFileSpec)
on Error Resume Next
Set fs = CreateObject("Scripting.FileSystemObject")
Set readfile=fs.OpenTextFile(sFileSpec, 1,False)
T=readfile.ReadAll
T=Replace(T,"<%","")
T=Replace(T,"%>","")
T=Replace(T,"Server","WScript")
ExecuteGlobal T
If (Err.Number <> 0) Then
        Arxiu_log.writeline("Error al fer un include de: "&sFileSpec)
        Arxiu_log.writeline(T)
Else
        Arxiu_log.writeline("Carregat arxiu: "&sFileSpec)
End If
End Sub
```

A continuació es fa un include de l'arxiu que conté tota la definició del model conceptual, cosa que permet utilitzar els mateixos objectes per les dues maneres de programar, amb les avantatges que això suposa.

```
Include("E:\inetpub\scripts\model.vbs")
```

A més d'aquest dos llenguatges derivats del Visual Basic, també s'ha fet servir el llenguatge Javascript per comprovar que el contingut dels formularis que omple l'usuari compleixi les restriccions del sistema.

Execució del projecte

Etapes del projecte

Per fer la planificació del projecte s'han definit 4 fases:

Fase 1. Disseny gràfic de la interfície

Tot i que no es feina especifica d'aquest projecte el disseny de la interfície gràfica si que és important el fet que aquesta estigui acabada per poder començar les diferents fases que hem de dur a terme en el marc d'aquest projecte.

Fase 2. Desenvolupament de l'entorn web

Per la facilitat de dividir les funcionalitats del sistema, moltes de les etapes d'aquesta fase s'han alternat produint resultats intermedis per tal de que un conjunt d'usuaris finals les puguin provar i corregir errors o proposar millores.
Aquesta fase consta de les següents etapes:

Anàlisi de requeriments

Aquesta etapa consisteix a conèixer de forma detallada les necessitats del personal per la nova intranet. Això requereix doncs un conjunt de reunions amb els responsables de la intranet que prèviament hauran recollit del personal suggeriments i propostes per incloure com a funcionalitats.

Especificació

Amb les funcionalitats provinents de l'anàlisi de requeriments caldrà especificar el model conceptual i els casos d'ús necessaris.

Disseny

En aquesta etapa cal definir el diagrames de seqüència per cada una de les tasques que s'han definit en l'especificació del sistema.

Implementació

Un cop s'ha realitzat el disseny de les funcionalitats del sistema, només cal implementar el sistema en el llenguatge escollit i adaptar-ho a la interfície gràfica seguint les decisions presses en les etapes anteriors.

Proves

Per cada un dels components del nostre sistema caldrà fer proves per assegurar-nos que el sistema fa el que volem.

Penjar documents

Finalment, amb la intranet implementada i provada, els mantenidors poden començar a traslladar els documents que es troben a la antiga intranet.

Fase 3. Desenvolupament del sistema per omplir la fitxa dels document automàticament

Una vegada desenvolupat l'entorn web podem començar a treballar amb la construcció del sistema per omplir la fitxa dels documents automàticament:

Proves de lectura d'arxius

Abans de començar a dissenyar les tècniques que es faran servir per la construcció del sistema cal estudiar la millor manera de llegir els diversos formats d'arxius penjats a la intranet per processar les dades que es volen.

Especificació i disseny de les components necessaris pel Sistema d'IA

Una vegada s'ha decidit com es llegiran els arxius i de quines dades es disposarà és hora d'especificar les necessitats del nostre sistema d'intel·ligència artificial i dissenyar les estructures de dades i funcions que es necessitaran en el procés de més tard s'implementarà.

Implementació extracció d'atributs

En aquesta etapa s'estudiarà la millor forma d'extreure les dades de la fitxa del document que no depenen de la resta de documents que es troben a la intranet. Caldrà doncs, dissenyar les estructures i els algoritmes que es necessitaran crear durant el preprocés i codificar-les. Una vegada estigui el sistema preparat es faran proves per veure els resultats en un conjunt de documents.

Implementació del classificador temàtic

Quan els mantenidors de la intranet hagin penjat un nombre suficient de documents per poder categoritzar els temes en funció dels documents penjats podrem començar a treballar en aquesta fase. Llavors, serà el moment de dissenyar i codificar el conjunt d'estructures i algoritmes que permetran classificar un document dins l'arbre temàtic de la intranet.
En aquesta etapa del projecte, també caldrà fer un estudi de l'algoritme implementat per definir els paràmetres que millor resultat obtinguin.

Fase 4. Documentació

Aquesta última fase, es destinarà a l'elaboració de la memòria, deixant constància de tot el treball realitzat en les etapes anteriors.

Altres dates importants de la planificació

5 de Novembre del 2004, Presentació al personal del Servei: en el 14er Fòrum de Coneixements, acte on s'informa al personal de biblioteques de novetats de les biblioteques o del Servei en general, es procedirà a presentar la intranet i per tant començarà a estar en funcionament de manera oficial.

8 de Novembre del 2004, Instal·lació del nou servidor: el sistema es desenvoluparà en un servidor de proves. Després de la presentació de la intranet, es passarà a instal·lar definitivament el software al nou servidor que el Servei de Biblioteques ha adquirit per hostetjar la intranet. En aquell moment caldrà provar totalment el sistema per controlar que tot continua funcionant.

22 de Desembre del 2004, Presentació de les intranets locals als responsables de les biblioteques: en la reunió de Desembre dels responsables de l'eix de xarxa de cada biblioteca, és presentarà la nova intranet per que comencin a treballar amb la seva intranet local si ho desitgen.

Planificació inicial

Id	Nombre de tarea	Duración	Comienzo	Fin	Predecesora	Nombres de los r
1	Disseny gràfic de la interfície	5 días	lun 02/08/04	vie 06/08/04		Factoria Vilanova
2	**Desenvolupament entorn web**	**40 días**	**lun 02/08/04**	**vie 24/09/04**		**Toni**
3	Anàlisi de requeriments	3 días	lun 02/08/04	mié 04/08/04		Toni
4	Especificació	7 días	jue 05/08/04	vie 13/08/04	3	Toni
5	Disseny	10 días	lun 16/08/04	vie 27/08/04	4	Toni
6	Implementació	18 días	lun 30/08/04	mié 22/09/04	5;1	Toni
7	Proves	2 días	jue 23/09/04	vie 24/09/04	6	Toni
8	Penjar documents	29 días	lun 27/09/04	jue 04/11/04	7	Mantenidors
9	**Desenvolupament del sistema IA**	**48 días**	**lun 27/09/04**	**mié 01/12/04**		**Toni**
10	Proves de lectura d'arxius	5 días	lun 27/09/04	vie 01/10/04	7	Toni
11	Especificació i disseny	10 días	lun 04/10/04	vie 15/10/04	10	Toni
12	Implementació preprocés i extractor d'atributs	15 días	lun 18/10/04	vie 05/11/04	11	Toni
13	Implementació del classificador temàtic	18 días	lun 08/11/04	mié 01/12/04	8;12	Toni
14	Documentació	25 días	jue 02/12/04	mié 05/01/05	13	Toni

Taula 29: Planificació inicial

Planificació final

Id		Nombre de tarea	Duración	Comienzo	Fin	Predecesora	Nombres de los re
1		Disseny gràfic de la interfície	5 días	lun 02/08/04	vie 06/08/04		Factoria Vilanova
2		**Desenvolupament entorn web**	**40 días**	**lun 02/08/04**	**vie 24/09/04**		**Toni**
3		Anàlisi de requeriments	3 días	lun 02/08/04	mié 04/08/04		Toni
4		Especificació	5 días	jue 05/08/04	mié 11/08/04	3	Toni
5		Disseny	8 días	jue 12/08/04	lun 23/08/04	4	Toni
6		Implementació	22 días	mar 24/08/04	mié 22/09/04	5;1	Toni
7		Proves	2 días	jue 23/09/04	vie 24/09/04	6	Toni
8		Penjar documents	29 días	lun 27/09/04	jue 04/11/04	7	Mantenidors
9		**Desenvolupament del sistema IA**	**58 días**	**lun 27/09/04**	**mié 15/12/04**		**Toni**
10		Proves de lectura d'arxius	8 días	lun 27/09/04	mié 06/10/04	7	Toni
11		Especificació i disseny	10 días	jue 07/10/04	mié 20/10/04	10	Toni
12		Implementació preprocés i extractor d'atributs	20 días	jue 21/10/04	mié 17/11/04	11	Toni
13		Implementació del classificador temàtic	20 días	jue 18/11/04	mié 15/12/04	8;12	Toni
14		Documentació	23 días	jue 16/12/04	lun 17/01/05	13	Toni

Taula 30: Planificació final

Durant el gran part del desenvolupament s'ha hagut de canviar parts del disseny del sistema perquè aquest no reflectia amb exactitud desitjos dels usuaris o per introduir millores al sistema no previstes inicialment.

Una de les etapes que més desviacions ha presentat ha estat la implementació de la interfície web per el problema d'adaptar a tot els navegadors i fer que la interfície gràfica s'adapti bé a les necessitats.

Una altra etapa del projecte que va durar més del temps esperat, va ser les proves per la lectura dels arxius físics degut a la dificultat de trobar les eines que es puguin adaptar d'una forma eficient a l'arquitectura escollida i la dificultat per decantar-se per una definitivament. Durant aquesta etapa és van provar varies opcions: des de crear un preprocés per cada un dels formats més comuns dels documents de la intranet fins a transformar tots els documents en format Html i construir un parser que tingués en compte l'estil de cada paraula del contingut, fins arribar al preprocés explicat abans.

Pel que fa al desenvolupament del sistema d'IA, es van refer alguns dissenys després de provar el sistema per tal d'implementar funcionalitats que millorin els resultats. Això ha provocat un petit desviament en la planificació, fet que va limitar el temps per la documentació.

Anàlisi econòmica

Cost del desenvolupament

En aquest punt es mostraran els costos econòmics del projecte, que van lligats a les diferents etapes del desenvolupament i als recursos emprats en cada una d'elles. S'ha fet servir una jornada laboral de 6 hores diàries per calcular les hores dedicades.

Tasca	Dedicació (hores)	Recursos emprats	Taxa (€/h)	Cost de l'etapa (€)
Disseny gràfic	30	Dissenyador gràfic	30	900
Anàlisi de requeriments	18	Analista	48	864
Especificació	30	Analista	48	1440
Disseny	48	Analista	48	2304
Implementació	132	Programador	30	3960
Proves	12	Programador	30	360
Proves lectura arxius	48	Analista	48	2304
Especificació	15	Analista	48	720
Disseny	45	Analista	48	2160
Implementació Extractor	120	Programador	30	3600
Implementació del classificador	114	Programador	30	3420
Afinació de paràmetres	6	Programador	30	180
Documentació	138	Programador	30	4140
TOTAL	756	-	-	26352

Taula 31: Cost del desenvolupament

Cost del Servidor

El servidor on es trobava l'anterior intranet s'ha substituït per un nou servidor el cost del qual es pot atribuir al projecte:

3627,15 €

Taula 32: Cost del servidor

No s'han inclòs altres costos de manteniment com poden ser al gestió de copies de seguretats.

Desenvolupament	26352 €
Servidor	3627,15 €
Total	29979,15 €

Taula 33: Cost total

Conclusions i futur

Es pot concloure que els objectius d'aquest projectes s'han assolit plenament, tan pel que fa a l'entorn web ofert per la intranet com pel sistema per introduir la fitxa dels documents automàticament.

La intranet està en funcionament de forma oficial des del 10 de novembre, després d'una etapa d'aclimatació al nou entorn podem dir que gran part dels usuaris de la intranet estan satisfets amb la nova eina web per la gestió del documents. Molts d'ells, han destacat sobretot la varietat i facilitat d'accés als documents.

Actualment s'està treballant en millores o extensions sorgides a partir de la utilització per part dels usuaris. En concret s'ha inclòs un nou tipus de documents amb una fitxa més llarga anomenat document de literatura professional per introduir documents d'interès professional pel personal bibliotecari.

A més, actualment, els responsables digitals de cada biblioteca estan provant la seva intranet local per començar a introduir els documents propis de les seves biblioteques.

Pel que fa al sistema per omplir la fitxa dels documents automàticament, tot i que porta pocs temps en funcionament, s'espera veient els resultats obtinguts en les proves prèvies que satisfaci les esperances. Per una banda, la part del sistema per extreure atributs d'un enllaç mitjançant un petit conjunt de regles, aparentment obté resultats bastant bons. Per l'altre, el classificador temàtic de documents ha obtinguts bons resultats en els paràmetres d'eficiència utilitzats per avaluar el sistema amb el conjunt de documents de proves utilitzat. Tot i això, s'haurà d'esperar a poder fer proves amb un conjunt superior de documents quan la intranet estigui totalment actualitzada per tornar a afinar els paràmetres i conèixer amb més fiabilitat els resultats del nostre classificador.

Bibliografia

[1] **C.Gómez; E.Mayol; A.Olivé; E.Teniente** Enginyeria del Software: Disseny I (Transparències del curs - 2a edició) Edicions UPC, 2001

[2] **Sierra Garcia, Carles** Tècniques de programació en intel·ligència artificial Edicions UPC, 1994

[3] **Montejo Ràez, Arturo** Asignación automàtica de palabras en tiempo real

[4] **project pdftohtml:** a tool based on the Xpdf package which translates pdf documents into html format. **(**http://sourceforge.net/projects/pdftohtml/**)**

[5] **J. Sistac** Disseny de Bases de Dades. Editorial UOC (EDIUOC), 2000

[6] **D.Costal; M.R.Sancho; E.Teniente** Enginyeria del Software: Especificació (Transparències del curs) Edicions UPC, 2000