Construcció d'un cercador de viatges "Low Cost" utilitzant tècniques de "Web Scraping"

"VTC Seeker"

Lluís Bou Coscolla
Enginyeria en Informàtica

Jordi Ferrer Duran
Consultor

Projecte Final Carrera – Gener 2012

Índex de continguts

Agraïments i dedicatòria

La realització d'aquest projecte ha estat en bona part gràcies a l'esforç, dedicació i a les hores robades a la meva família. Em representa, per tant, una obligació ineludible dedicar-s'ho de manera sentida i, en especial, a la meva dona, ja que sense la seva ajuda i complicitat aquesta tasca hagués estat del tot impossible.

També vull agrair la col·laboració del meu consultor, en Jordi Ferrer Duran, i de tots aquells professionals docents i companys de la UOC amb els que he tingut la sort de compartir aquests darrers anys de formació.

Resum

Resum del contingut de la memòria

Els apartats inicials d'aquest memòria, punts 1, 2 i 3, es centren en l'explicació del projecte desenvolupat, dels seus objectius, del seu abast i de la seva justificació.

En el següent, punt 4, s'analitzen les eines i tecnologies que ofereix el mercat actualment per implementar el web scraping i es farà un introducció als principals conceptes tecnològics.

A continuació, en els punts 5, 6, 7 i 8, es detalla la planificació, el disseny de la solució, la tecnologia emprada, comentaris de les principals parts del codi desenvolupat i el funcionament dels servidors de cada proveïdor implementat en el cercador.

Els punts 9 i 10 contenen la conclusió a mode de resum i opinió personal sobre la feina realitzada i les possibles línies de treball per actuacions futures i de millora.

La part final conté la bibliografia emprada durant l'elaboració del projecte, un glossari de termes que faciliten la comprensió del text i les llicències de publicació del projecte.

1. Introducció

El nostre client, la companyia VTC, ens ha encarregat la construcció d'un cercador web de viatges "Low Cost" que sigui capaç de realitzar la selecció de les cinc ofertes amb el millor preu. L'objectiu de l'eina sol·licitada serà cercar, classificar i presentar els preus més competitius perquè VTC els pugui oferir als seus clients.

Actualment existeix un gran volum d'informació a Internet sobre tarifes i ofertes relacionades amb viatges, vols i hotels, que acostuma a actualitzar-se amb relativa freqüència. Per tant, la tasca de localitzar els millors preus constantment és força repetitiva i cal invertir-hi un temps valuós. El fet de disposar d'una aplicació que automatitzi aquesta cerca d'una manera senzilla i flexible pot constituir un clar avantatge competitiu per VTC i crear valor afegit al servei dels seus clients.

Per implementar la cerca i recopilació de les dades s'utilitzaran les tècniques conegudes com a "Web Scraping"[1], raspadors de webs, i les metodologies d'analitzadors semàntics per seleccionar i classificar la informació obtinguda. Pel que fa a la tecnologia, es faran servir eines i llenguatges de lliure distribució.

Un dels requeriments fonamentals en la petició del client es el de minimitzar els costos d'implementació de l'aplicatiu, fet que es tindrà en compte durant tot el procés de disseny i posterior desenvolupament del producte. En aquest sentit, l'aplicació resultant prioritzarà la seva eficiència i potència cercadora per davant d'altres aspectes com la presentació dels resultats, el disseny de les interfícies gràfiques o d'altres funcionalitats realment no indispensables.

Pel que fa a l'abast del projecte, compren el procés de recerca, disseny i implementació de l'aplicació segons els requeriments del client i la producció dels lliurables en forma de producte final, que inclou la documentació i els manuals.

No estan dins de l'abast del projecte altres funcionalitats com l'automatització de la confirmació de les reserves o la integració del producte en els sistemes informàtics del client, que podrien ser candidates a possibles millores i línies d'actuació futures.

[1] (Wikipedia, Web_scraping, 2006)

2. Problemàtica actual i justificació del projecte

L'anomenada xarxa de xarxes ha esdevingut un ampli i quasi infinit univers de possibilitats per a tots el gustos. Ofertes i demandes empresarials de tota mena, temps de lleure, oci i cultura, relacions laborals i personals, i tot allò que puguem imaginar i, potser, fins i tot, una mica més.

En definitiva, l'abast mundial i l'enorme difusió d'aquest medi ha fet que la seva evolució hagi crescut de manera exponencial en nombre i volum de dades. Aquest fet ha donat peu al naixement de tot un seguit de tècniques que cobreixen les noves necessitats que presenta aquest món virtual en continua expansió. Cada dia apareixen noves pàgines web, nous portals i serveis, alhora que en desapareixen d'altres, per manca d'èxit o per obsolescència.

En aquest sentit, es imprescindible la tasca dels cercadors, que es dediquen bàsicament a indexar continguts, classificar-los i puntuar-los. La gran majoria de cercadors utilitzen tècniques de rastreig de la xarxa. Aquesta tasca la realitzen els robots, que son programes encarregats de navegar i escorcollar de manera automatitzada per la xarxa.

De robots web existeixen diferents tipus, en funció dels seus objectius. En aquest projecte ens centrarem en l'estudi de les aranyes(spider)[2] i els rascadors(scraping), ja que la solució proposada per la implementació del motor de cerca es basarà en una combinació d'aquestes tècniques.

Davant la gran quantitat de continguts i la seva dinàmica canviant, es fa impensable i inviable que una persona es dediqui constantment a visitar els llocs web en cerca de determinades variables. Fàcilment, podem imaginar i fer-nos càrrec de la quantitat de temps, de tasques repetitives i d'esforç que hauria de dedicar una persona per estar cercant a Internet contínuament les mateixes dades per diferents variables. Ens aquest sentit, l'esforç humà seria tan esgotador com inútil ja que un programa dissenyat per aquest propòsit seria més ràpid i eficient.

En aquest moments, proliferen els cercadors especialitzats en assegurances, viatges, vols, hipoteques, dipòsits bancaris, que ofereixen als seus potencials usuaris la possibilitat de realitzar cerques de manera senzilla, exhaustiva i amb poca exigència en termes de temps. El que fan aquestes aplicacions es llençar els seus robots(programes)[3] a recórrer la xarxa i indexar els continguts que els interessin. Després els analitzen per veure si el que han trobat s'ajusta als seus objectius predefinits i en cas afirmatiu els incorporen a les seves bases de dades.

[2] (Webtaller, Robots-arania, 2005)
[3] (Webtaller, Guia_robots, 2005)

Quan un client fa una petició es fa la cerca sobre aquestes dades emmagatzemades, agilitzant els temps de resposta i acotant la cerca. Normalment, es mostren els resultats obtinguts en funció dels paràmetres d'entrada, ordenats per criteris de qualitat, preu, recomanacions o d'altres variables. També es habitual presentar la possibilitat d'enllaçar directament, mitjançant un link, amb el lloc web original que ofereix els servei o producte a efectes de tramitar la seva adquisició o per ampliar la informació sobre el mateix.

En aquest context es quan pren sentit el present projecte, ja que pretén oferir una eina útil i eficient que faciliti les tasques de cerca per a una determinada temàtica, com ho son els viatges i la necessitat d'aconseguir els millors vols i hotels al preu més competitiu per poder oferir-los als seus clients.

3. Objectius

L'objectiu principal del present projecte és investigar i obtenir el coneixement necessari sobre les diferents eines i tècniques de Web Scraping, el qual ens haurà de permetre implementar amb èxit el producte final segons els requeriments i terminis sol·licitats pel nostre client.

Altres objectius més específics serien els d'aplicar els coneixements obtinguts al llarg dels estudis, realitzar les tasques de recerca adients, generar la documentació formal del projecte, gestionar el projecte i realitzar el seguiment del grau d'assoliment de les tasques sobre els calendari previst, evitant possibles desviacions i problemàtiques derivades.

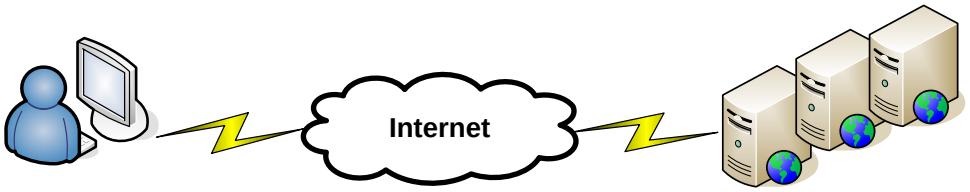
- Consolidar els coneixements adquirits al llarg dels estudis
- Treballar en diferents tècniques de cerca d'informació per Internet
- Estudiar les llibreries de Web Scraping

4. Tècniques Web Scraping

Introducció als elements tècnics

En aquest apartat es tracta de donar una visió general sobre els principals conceptes i eines que intervenen en el desenvolupament del projecte.

El navegador, browser, és l'eina que habitualment utilitza una persona per a accedir als llocs web, moure's entre les seves planes, consultar els continguts, intercanviar informació i en definitiva interactuar amb el servidor.



El servidor web rep les peticions dels client(navegador), les gestiona i retorna un flux de dades que el navegador es capaç d'entendre, processar i representar.

El protocol HTTP, Hypertext Transfer Protocol, es el que es fa servir en les transaccions web i defineix la sintaxi i la semàntica que s'utilitza en l'entorn web. Es a dir, es la manera en que es comuniquen els clients i servidors web. Es tracta d'un protocol orientat a transaccions que segueix el patró petició- resposta. El client es qui fa la petició i se'l coneix com a "user agent". La informació a transmetre se l'anomena recurs i queda identificat per un localitzador, URL.

Aquest protocol no guarda informació sobre connexions anteriors, es sense estat, fet pel qual es fan servir cookies, que son petits fitxers de text en els que es desen en el client, que contenen aquestes dades.

El codi HTML[4], HyperText Markup Language, o llenguatge de marques, es la codificació que es fa servir per la confecció de pàgines web. Fa servir etiquetes o marques per definir les diferents parts de la estructura i es pot combinar amb d'altres elements com imatges, scripts o XML.

[4] (Wikipedia, HTML, 2002)

El XML[5], eXtensible Mark Language, es un metallenguatge que serveix per definir altres llenguatges i que es pot aplicar a diferents entorns, no només a Internet. Es molt utilitzat com estàndard per a intercanviar dades entre programes i sistemes diferents d'una manera senzilla i fiable. Les seves aplicacions i possibilitats son nombroses més enllà del món web.

XSLT[6], Transformacions XSL, es un estàndard de l'organització W3C, World Wide Web Consortium, per realitzar transformacions de codi XML a HTML. S'utilitzen una sèrie de regles o normes definides en una plantilla, que servirà per donar forma al codi HTML resultant. La combinació de XSLT i SML es molt útil per generar pàgines web, separant el contingut de la presentació.

Per analitzar les dades de connexió, com s'envien a les capçaleres, la gestió de les peticions i les respostes entre el client i el servidor web, podem utilitzar eines que ens faciliten la feina i que monitoritzen els missatges que es van enviant les dues parts. Ens aquest cas utilitzarem LiveHTTPHeaders, que es un programari lliure que cobreix aquestes necessitats.

El codi finalment es tradueix a text, aquí es quan podem filtrar, ordenar i extraure els valors que ens interessen. Val dir que no totes les planes web permeten ser "raspades" ja que a vegades estan protegides contra els robots de cerca, o també perquè contenen codi o imatges de difícil interpretació.

El nostre producte simularà les peticions que faria un usuari i les enviarà als diferents servidors que tenen els proveïdors dels serveis i aquest respondran retornant els resultats de cada consulta. Cada servidor funciona d'una manera particular, es a dir, cada un disposa d'una estructura determinada que utilitza per intercanviar les dades, peticions i respostes, es la seva forma de comunicar-se amb el client.

Una de les dificultats del projecte es analitzar cada un d'aquests flux de dades i ser capaços d'automatitzar els mecanismes de cerca per aconseguir extraure les dades que ens interessen, en aquest cas les relatives als millors preus en hotels i vols per a uns determinats paràmetres de cerca.

En aquest punt, cal tenir en compte que no totes les pàgines web permeten ser "raspades", de fet en alguns països es directament il·legal. Ja sigui per temes legals o per polítiques corporatives, la realitat es que existeixen diferents mètodes anti robots, com incloure directives d'exclusió en un arxiu de text al directori arrel del servidor o, d'altres dispositius, com els "captcha", que apareixen en alguns llocs web i que simulen lletres o nombres de difícil lectura, que només poden ser interpretats correctament per una interacció humana, fet que exclou el programari automatitzat, cosa que es finalment un robot cercador.

[5] (Wikipedia, XML, 2005)
[6] (Wikipedia, Xslt, 2008)

5. Planificació

Les principals fites del projecte venen marcades per les dates d'entrega parcials de les PAC i l'entrega final del producte obtingut i de tot el material lliurable associat, com la memòria del projecte i els manuals d'usuari i d'instal·lació del producte.

Fites	Data
Lliurament PAC1 - Pla de treball	09/10/2011
PAC2 Seguiment PFC - Lliurament Parcial	13/11/2011
PAC3 Seguiment PFC - Lliurament Parcial	14/12/2011
Lliurament Final	15/01/2012
Tancament Projecte	27/01/2012

En previsió de possibles desviacions, la planificació temporal s'ha confeccionat incorporant un marge de seguretat, que contempla un pla de contingència que ens permeti una certa capacitat de maniobra per redreçar possibles imprevistos, ja siguin tasques que s'allarguen més del que seria desitjable o problemàtiques no previstes inicialment.

Tanmateix, si la planificació es desenvolupa sense desviacions, es dedicarà el temps reservat com a marge de seguretat per implementar millores en la qualitat del producte final i de tots els seus lliurables.

5.1. Calendari d'activitats

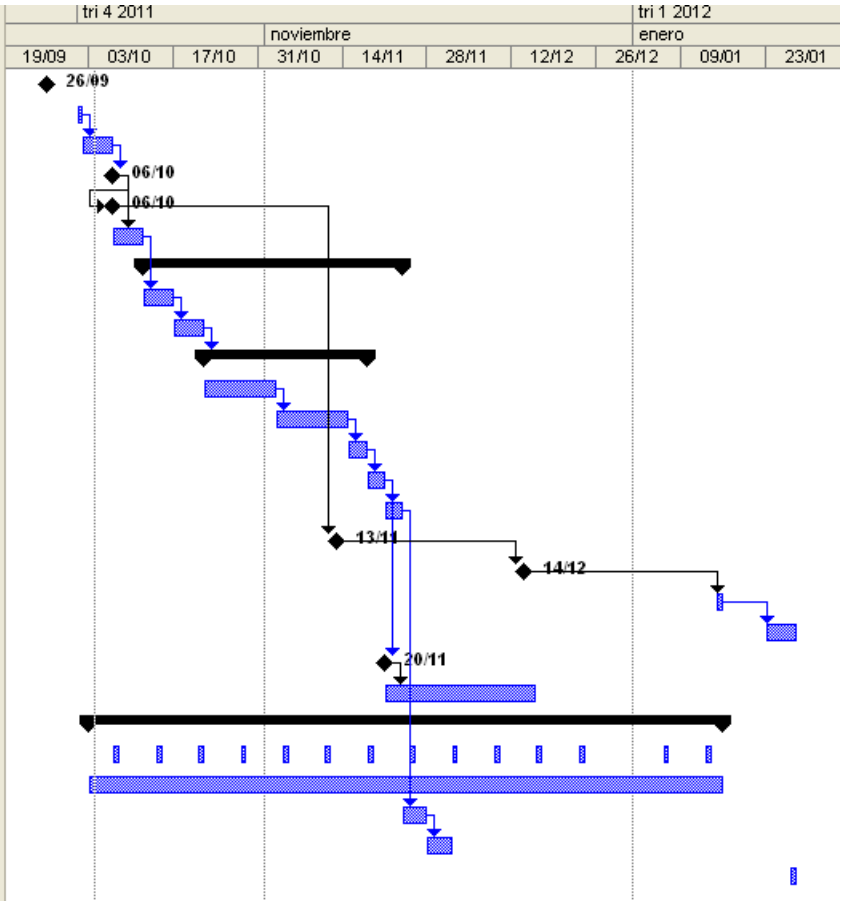
La temporalització de les tasques dissenyada en el calendari s'ha efectuat en base a jornades de treball diàries. Tanmateix, a efectes de facilitar el control i seguiment de l'evolució de les tasques a realitzar, s'ha assignat la següent valoració en hores de dedicació a cada una.

A continuació es presenta la quantificació horària de l'esforç necessari per assolir els objectius definits en el pla de treball.

Tasca	Hores	Inici	Final
T1 - Redacció Pla de treball - PAC1	10	02/10/2011	09/10/2011
T2 - Recerca Eines - Tecnologia	10	07/10/2011	11/10/2011
T3 - Definició Eines - Tecnologia	10	12/10/2011	16/10/2011
T4 - Disseny Tècnic Producte	10	17/10/2011	21/10/2011
Desenvolupament			
T5 - Motor de cerca	30	22/10/2011	02/11/2011
T6 - Analitzador semàntic	20	03/11/2011	14/11/2011
T7 - Gestió resultats	5	18/11/2011	20/11/2011
T8 - Revisió Producte – Test	5	21/11/2011	23/11/2011
T9 - Gestió Projecte - Seguiment	10	03/10/2011	15/01/2012
T10 - Documentació Memòria	20	03/10/2011	15/01/2012
T11 - Manual Usuari	5	24/11/2011	27/11/2011
T12 - Manual Instal·lació	5	28/11/2011	01/12/2011
T13 - Preguntes Tribunal - Debat Virtual	5	23/01/2012	27/01/2012

Diagrama de Gantt setmanal. (*)

	Nombre de tarea	Duración	Comienzo	Fin	Prede
1	Publicació Enunciat PFC - Kick-off Projecte	0 días	lun 26/09/11	lun 26/09/11	
2	Trobada Presencial	1 día	sáb 01/10/11	sáb 01/10/11	
3	Redacció Pla de treball - PAC1	5 días	dom 02/10/11	jue 06/10/11	2
4	Lliurament PAC1 - Pla de treball	0 días	jue 06/10/11	jue 06/10/11	3
5	Aprovació Pla del Projecte	0 días	jue 06/10/11	jue 06/10/11	4
6	Recerca Eines - Tecnologia	5 días	vie 07/10/11	mar 11/10/11	4
7	**Implementació PFC**	**43 días**	**mié 12/10/11**	**mié 23/11/11**	
8	Definició Eines - Tecnologia	5 días	mié 12/10/11	dom 16/10/11	6
9	Disseny Tècnic Producte	5 días	lun 17/10/11	vie 21/10/11	8
10	**Desenvolupament**	**27 días**	**sáb 22/10/11**	**jue 17/11/11**	**9**
11	Motor de cerca	12 días	sáb 22/10/11	mié 02/11/11	
12	Analitzador semàntic	12 días	jue 03/11/11	lun 14/11/11	11
13	Gestió Resultats	3 días	mar 15/11/11	jue 17/11/11	12
14	Proves - Test Producte	3 días	vie 18/11/11	dom 20/11/11	13
15	Revisió Producte	3 días	lun 21/11/11	mié 23/11/11	14
16	PAC2 Seguiment PFC - Lliurament Parcial	0 días	dom 13/11/11	dom 13/11/11	5
17	PAC3 Seguiment PFC - Lliurament Parcial	0 días	mié 14/12/11	mié 14/12/11	16
18	Lliurament Final	1 día	dom 15/01/12	dom 15/01/12	17
19	Preguntes Tribunal - Debat Virtual	5 días	lun 23/01/12	vie 27/01/12	18
20	Pla de contingències	0 días	dom 20/11/11	dom 20/11/11	14
21	Marge de maniobra per possibles contingències	25 días	lun 21/11/11	jue 15/12/11	20
22	**Gestió Projecte**	**100 días**	**lun 03/10/11**	**dom 15/01/12**	
23	**Seguiment Projecte**	**94 días**	**vie 07/10/11**	**vie 13/01/12**	
38	Documentació Memoria	100 días	lun 03/10/11	dom 15/01/12	
39	Manual Usuari	4 días	jue 24/11/11	dom 27/11/11	15
40	Manual Instal·lació	4 días	lun 28/11/11	jue 01/12/11	39
41	Tancament Projecte	1 día	vie 27/01/12	vie 27/01/12	



(*) La planificació de la jornada laboral esta feta a nivell de escala temporal diària. En principi, la dedicació prevista seria de 2 hores per dia. Per tant, cada dia equivaldria a 2 hores de feina efectiva.

5.2. Descripció general de les principals tasques

En aquesta fase del projecte resulta prematur concretar amb excessiva definició totes les tasques i els seus apartats, especialment pel que fa a les activitats de desenvolupament, ja que encara resten per prendre decisions importants sobre la forma en que es realitzarà aquesta implementació.

Malgrat tot, si que es poden preveure o intuir tota una sèrie de tasques i subtasques que seran necessàries per la consecució dels objectius i que es relacionen a continuació a mode de referència general que caldrà tenir en compte.

T1 - Redacció Pla de treball. Es tracta de definir els objectius i la planificació temporal del projecte. Queda subjecte a revisió i aprovació amb el client(consultor).

T2 - Recerca Eines – Tecnologia. Es una tasca vital en el desenvolupament del projecte, ja que serà la base en la que es construirà el producte final. S'utilitzaran els recursos disponibles a Internet, les bases de dades Gartner i buscadors de la biblioteca de la UOC.

T3 - Definició Eines – Tecnologia. En funció de la recerca de la tasca anterior, es generarà un informe amb les conclusions i la concreció de les eines escollides per implementar el producte. Com a subtasca també podem incloure en aquest apartat la preparació de l'entorn de treball, que consistiria en la instal·lació i configuració del programari i les eines necessàries per desenvolupar el projecte.

T4 - Disseny Tècnic Producte. Disseny de les funcionalitats i processos de l'aplicació a desenvolupar en base als requeriments de la petició del client.

Desenvolupament. Codificació de la solució dissenyada. Ens aquest punt podem agrupar les següents tasques:

T5 - Motor de cerca. Serà el procés encarregat de la cerca en funció d'un paràmetres d'entrada. Aquest paràmetres seran modificables. Com a subtasca podem incloure la recerca sobre el funcionament dels diferents proveïdors dels serveis que ofereixen les dades dels viatges, ja que cada servidor les presentarà d'una manera particular. Caldrà estudiar cada cas i veure com es pot aprofitar i gestionar aquesta informació de manera automatitzada.

T6 - Analitzador semàntic. Analitzarà i classificarà les dades obtingudes i decidirà si compleixen els requeriments de la cerca, generant les cinc millors combinacions.

T7 - Gestió resultats. A partir dels resultats obtinguts es mostraran a l'usuari en un entorn web format HTML i també format XML. S'implementarà la possibilitat de emmagatzemar aquestes sortides.

T8 - Test de proves i revisió del producte. Inclou un joc de proves complet sobre el funcionament de l'eina construïda i la correcció de les possibles anomalies detectades.

T9 - Gestió Projecte. Aquesta tasca es transversal a tot el projecte i engloba entre d'altres activitats el seguiment de l'evolució de la resta de tasques segons la planificació, el control dels riscos, la confecció de tota la documentació formal lliurable, com la memòria i els manuals d'usuari i d'instal·lació.

T10 - Redacció de la memòria del projecte. La memòria ha de reflectir i contenir tota la feina que s'ha dut a terme durant tot el procés de construcció del projecte. Per tant, la seva confecció s'ha de realitzar de manera regular i continuada per recollir les evolucions de les tasques, els avanços, desviacions i problemàtiques que es puguin produir. Forma part del material lliurable final, juntament amb el producte desenvolupat i els manuals d'usuari i d'instal·lació.

T11 - Redacció del manual d'usuari. Ha de contenir les explicacions suficients per que l'usuari pugui utilitzar el producte final de manera correcta i satisfactòria.

T12 - Redacció del manual d'instal·lació. Contindrà els requeriments de maquinari i programari necessaris, així com les instruccions d'instal·lació del producte desenvolupat.

T13 - Preguntes Tribunal - Debat Virtual. Aquesta tasca consisteix en respondre i participar de manera activa en les diferents qüestions que es plantegin des del Tribunal o durant el debat virtual.

5.3. Anàlisi de riscos

Pel que fa a l'anàlisi de riscos, la principal incertesa rau en la desconeixença de les eines i metodologies que caldrà utilitzar per implementar la cerca a Internet amb les tècniques de "Web Scraping". Per tant, es en aquest punt on es podrien produir les desviacions més importants sobre el calendari previst.

Risc per desconeixença de la complexitat tecnològica

Com acció preventiva, es proposa fer un seguiment estricte del grau d'evolució de la tasca sobre la seva planificació. També caldria simplificar i prioritzar la recerca de les solucions tecnològiques per assolir els objectius descrits al pla de treball.

Si es produeix aquest risc caldrà demanar ajuda al consultor per aclarir les dubtes i rebre orientació sobre les possibles solucions.

Risc de pèrdua dades per desastre informàtic

Com acció preventiva es realitzaran copies de tot el projecte, de manera periòdica i regular, utilitzant sistemes d'emmagatzematge auxiliars de recolzament, com unitats de discos externs o servidors remots de backup on-line.

Si es produeix aquest risc caldrà recuperar la còpia de seguretat més actualitzada per minimitzar l'impacte del problema i evitar la pèrdua de temps i els esforços realitzats, que podrien comportar desviacions importants en el desenvolupament de la correcta planificació del projecte.

Risc de desviacions i incompliment de dates

Per evitar aquest risc caldrà fer un seguiment acurat i rigorós del calendari del pla de treball. Malgrat tot, sempre hi poden succeir imprevistos, aliens al projecte, que alterin la planificació i la correcta evolució de les feines. A tal efecte, s'ha tingut en compte un marge temporal de seguretat com a pla de contingència.

Si es produeix aquest risc, s'hauran d'analitzar les causes i augmentar l'assignació dels recursos en funció de la tipologia del problema, ja sigui de manera qualitativa o quantitativa. En funció de la gravetat i l'impacte del problema caldria informar al consultor i demanar consell.

5.4. Informe Seguiment PAC2

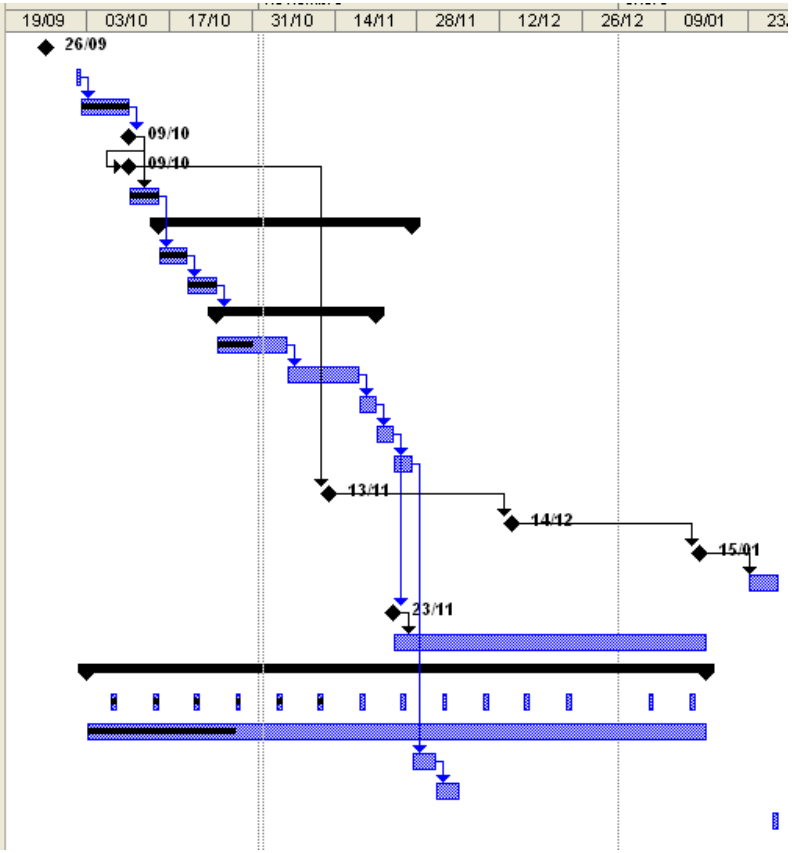
Les tasques inicials (T1, T2, T3, i T4) s'han desenvolupat segons el calendari previst en el pla de treball.

Pel que fa a la fase d'implementació del motor de cerca i de l'analitzador (T5 i T6) estic veient que es pot allargar més del planificat inicialment. Especialment, pel que fa a la captació de les dades dels servidors dels proveïdors del serveis, ja que segons el nombre, diversitat i complexitat d'aquestos la feina es multiplica degut a que cada un funciona de manera diferent i cal adaptar els procés d'extracció de les dades a cada cas en particular.

Les tasques T9 i T10, que són transversals a tot el projecte i s'executaran durant el cicle de vida del mateix, segueixen la seva planificació sense anomalies destacables.

Tasca	Hores	Inici	Final	% Progrés
T1 - Redacció Pla de treball - PAC1	10	02/10/2011	09/10/2011	100
T2 - Recerca Eines – Tecnologia	10	07/10/2011	11/10/2011	100
T3 - Definició Eines – Tecnologia	10	12/10/2011	16/10/2011	98
T4 - Disseny Tècnic Producte	10	17/10/2011	21/10/2011	98
Desenvolupament				
T5 - Motor de cerca	30	22/10/2011	02/11/2011	25
T6 - Analitzador semàntic	20	03/11/2011	14/11/2011	25
T7 - Gestió resultats	5	18/11/2011	20/11/2011	
T8 - Revisió Producte – Test	5	21/11/2011	23/11/2011	
T9 - Gestió Projecte - Seguiment	10	03/10/2011	15/01/2012	25
T10 - Documentació Memòria	20	03/10/2011	15/01/2012	25
T11 - Manual Usuari	5	24/11/2011	27/11/2011	
T12 - Manual Instal·lació	5	28/11/2011	01/12/2011	
T13 - Preguntes Tribunal - Debat Virtual	5	23/01/2012	27/01/2012	

1		Publicació Enunciat PFC - Kick-off Projecte	0 días	lun 26/09/11	lun 26/09/11	
2		Trobada Presencial	1 día	sáb 01/10/11	sáb 01/10/11	
3	✓	Redacció Pla de treball - PAC1	8 días	dom 02/10/11	dom 09/10/11	2
4		Lliurament PAC1 - Pla de treball	0 días	dom 09/10/11	dom 09/10/11	3
5		Aprovació Pla del Projecte	0 días	dom 09/10/11	dom 09/10/11	4
6	✓	Recerca Eines - Tecnologia	5 días	lun 10/10/11	vie 14/10/11	4
7		**Implementació PFC**	**43 días**	**sáb 15/10/11**	**sáb 26/11/11**	
8	✓	Definició Eines - Tecnologia	5 días	sáb 15/10/11	mié 19/10/11	6
9	✓	Disseny Tècnic Producte	5 días	jue 20/10/11	lun 24/10/11	8
10		**Desenvolupament**	**27 días**	**mar 25/10/11**	**dom 20/11/11**	**9**
11		Motor de cerca	12 días	mar 25/10/11	sáb 05/11/11	
12		Analitzador semàntic	12 días	dom 06/11/11	jue 17/11/11	11
13		Gestió Resultats	3 días	vie 18/11/11	dom 20/11/11	12
14		Proves - Test Producte	3 días	lun 21/11/11	mié 23/11/11	13
15		Revisió Producte	3 días	jue 24/11/11	sáb 26/11/11	14
16		PAC2 Seguiment PFC - Lliurament Parcial	0 días	dom 13/11/11	dom 13/11/11	5
17		PAC3 Seguiment PFC - Lliurament Parcial	0 días	mié 14/12/11	mié 14/12/11	16
18		Lliurament Final	0 días	dom 15/01/12	dom 15/01/12	17
19		Preguntes Tribunal - Debat Virtual	5 días	lun 23/01/12	vie 27/01/12	18
20		Pla de contingències	0 días	mié 23/11/11	mié 23/11/11	14
21		Marge de maniobra per possibles contingències	48 días	jue 24/11/11	dom 15/01/12	20
22		**Gestió Projecte**	**100 días**	**lun 03/10/11**	**dom 15/01/12**	
23		**Seguiment Projecte**	**94 días**	**vie 07/10/11**	**vie 13/01/12**	
38		Documentació Memoria	100 días	lun 03/10/11	dom 15/01/12	
39		Manual Usuari	4 días	dom 27/11/11	mié 30/11/11	15
40		Manual Instal·lació	4 días	jue 01/12/11	dom 04/12/11	39
41		Tancament Projecte	1 día	vie 27/01/12	vie 27/01/12	



5.5. Informe Seguiment PAC3

Les tasques T1, T2, T3 i T4 s'han desenvolupat segons el calendari previst en el pla de treball i estan totalment finalitzades.

Tal i com es va comentar en l'anterior informe de seguiment, la fase d'implementació del motor de cerca i de l'analitzador (T5 i T6) s'està allargant més del planificat inicialment a causa de les particularitats específiques de cada proveïdor.

Per contra, la tasca T7, a on es gestionen i mostren els resultats obtinguts, no sembla que representi excessives complicacions afegides i està pràcticament finalitzada la seva implementació.

Aquests fets confirmen que les estimacions prèvies, sobre la carrega de treball i la planificació temporal de les tasques realitzades en el pla de treball, no varen ser prou realistes o acurades i ens demostra la complexitat que comporta aquesta activitat de previsió en la fase inicial d'un projecte.

En aquests moments s'ha construït un prototipus estable de l'aplicació que implementa al voltant del 90% dels requeriments inclosos en l'enunciat del projecte. Funciona amb un parell de proveïdors de vols i hotels, i és capaç de realitzar cerques i mostrar els resultats. La previsió es afegir més operadors de cara a l'entrega final i s'està treballant en aquest sentit per aconseguir un producte acabat de major qualitat.

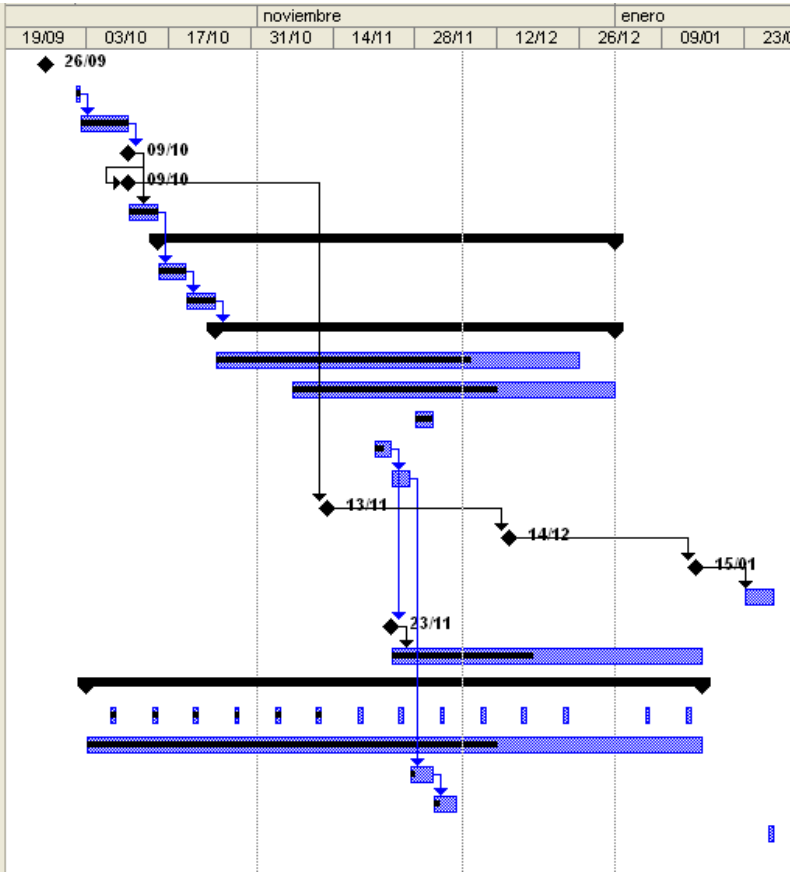
Les tasques T9 i T10, que són transversals a tot el projecte i s'executaran durant el cicle de vida del mateix, segueixen la seva planificació sense anomalies destacables. S'han iniciat les tasques T11 i T12, documentació de manuals i presentació, per avançar feina i aprofitar el temps disponible.

En general, s'ha optat per prioritzar les tasques de redacció i fer progressar en paral·lel aquelles activitats de desenvolupament que així ho permetessin, amb l'objectiu de guanyar temps i alleugerir la càrrega de treball en el tram final.

Malgrat les desviacions parcials experimentades en les tasques d'implementació T5 i T6, es confia plenament en assolir la consecució de la totalitat dels objectius del present projecte.

Tasca	Hores	Inici	Final	% Progrés
T1 - Redacció Pla de treball - PAC1	10	02/10/2011	09/10/2011	100
T2 - Recerca Eines – Tecnologia	10	07/10/2011	11/10/2011	100
T3 - Definició Eines – Tecnologia	10	12/10/2011	16/10/2011	100
T4 - Disseny Tècnic Producte	10	17/10/2011	21/10/2011	100
Desenvolupament				
T5 - Motor de cerca	30	22/10/2011	01/01/2012	75
T6 - Analitzador semàntic	20	03/11/2011	01/01/2012	75
T7 - Gestió resultats	5	18/11/2011	01/01/2012	90
T8 - Revisió Producte – Test	5	21/11/2011	01/01/2012	50
T9 - Gestió Projecte - Seguiment	10	03/10/2011	15/01/2012	70
T10 - Documentació Memòria	20	03/10/2011	15/01/2012	70
T11 - Manual Usuari	5	24/11/2011	15/01/2011	25
T12 - Manual Instal·lació	5	28/11/2011	15/01/2011	25
T13 - Preguntes Tribunal - Debat Virtual	5	23/01/2012	27/01/2012	

		Nombre de tarea	Duración	Comienzo	Fin	Predecesoras
1	✓	Publicació Enunciat PFC - Kick-off Projecte	0 días	lun 26/09/11	lun 26/09/11	
2	✓	Trobada Presencial	1 día	sáb 01/10/11	sáb 01/10/11	
3	✓	Redacció Pla de treball - PAC1	8 días	dom 02/10/11	dom 09/10/11	2
4	✓	Lliurament PAC1 - Pla de treball	0 días	dom 09/10/11	dom 09/10/11	3
5	✓	Aprovació Pla del Projecte	0 días	dom 09/10/11	dom 09/10/11	4
6	✓	Recerca Eines - Tecnologia	5 días	lun 10/10/11	vie 14/10/11	4
7		**Implementació PFC**	**73 días**	**sáb 15/10/11**	**sáb 31/12/11**	
8	✓	Definició Eines - Tecnologia	5 días	sáb 15/10/11	mié 19/10/11	6
9	✓	Disseny Tècnic Producte	5 días	jue 20/10/11	lun 24/10/11	8
10		**Desenvolupament**	**63 días**	**mar 25/10/11**	**sáb 31/12/11**	**9**
11		Motor de cerca	62 días	mar 25/10/11	dom 25/12/11	
12		Analitzador semàntic	50 días	lun 07/11/11	sáb 31/12/11	
13		Gestió Resultats	3 días	lun 28/11/11	mié 30/11/11	
14		Proves - Test Producte	3 días	lun 21/11/11	mié 23/11/11	
15		Revisió Producte	3 días	jue 24/11/11	sáb 26/11/11	14
16	✓	PAC2 Seguiment PFC - Lliurament Parcial	0 días	dom 13/11/11	dom 13/11/11	5
17	✓	PAC3 Seguiment PFC - Lliurament Parcial	0 días	mié 14/12/11	mié 14/12/11	16
18		Lliurament Final	0 días	dom 15/01/12	dom 15/01/12	17
19		Preguntes Tribunal - Debat Virtual	5 días	lun 23/01/12	vie 27/01/12	18
20		Pla de contingències	0 días	mié 23/11/11	mié 23/11/11	14
21		Marge de maniobra per possibles contingències	48 días	jue 24/11/11	dom 15/01/12	20
22		**Gestió Projecte**	**100 días**	**lun 03/10/11**	**dom 15/01/12**	
23		**Seguiment Projecte**	**94 días**	**vie 07/10/11**	**vie 13/01/12**	
38		Documentació Memoria	100 días	lun 03/10/11	dom 15/01/12	
39		Manual Usuari	4 días	dom 27/11/11	mié 30/11/11	15
40		Manual Instal·lació	4 días	jue 01/12/11	dom 04/12/11	39
41		Tancament Projecte	1 día	vie 27/01/12	vie 27/01/12	



5.6. Informe Seguiment Entrega final

Les tasques T1, T2, T3 i T4 s'han desenvolupat segons el calendari previst en el pla de treball i estan totalment finalitzades.

Tal i com es va comentar en els anteriors informes de seguiment, les fases d'implementació del motor de cerca i de l'analitzador (T5 i T6) es van allargar més del planificat inicialment a causa de les particularitats específiques de cada proveïdor.

Per contra, la tasca T7, a on es gestionen i mostren els resultats obtinguts, no ha representat excessives complicacions afegides.

Aquests fets confirmen que les estimacions prèvies, sobre la carrega de treball i la planificació temporal de les tasques realitzades en el pla de treball, no varen ser prou realistes o acurades i ens demostra la complexitat que comporta aquesta activitat de previsió en la fase inicial d'un projecte.

En aquests moments s'ha construït un prototipus estable de l'aplicació que implementa la totalitat dels requeriments inclosos en l'enunciat del projecte.

Funciona amb un parell de proveïdors de vols i hotels, i és capaç de realitzar cerques i mostrar els resultats.

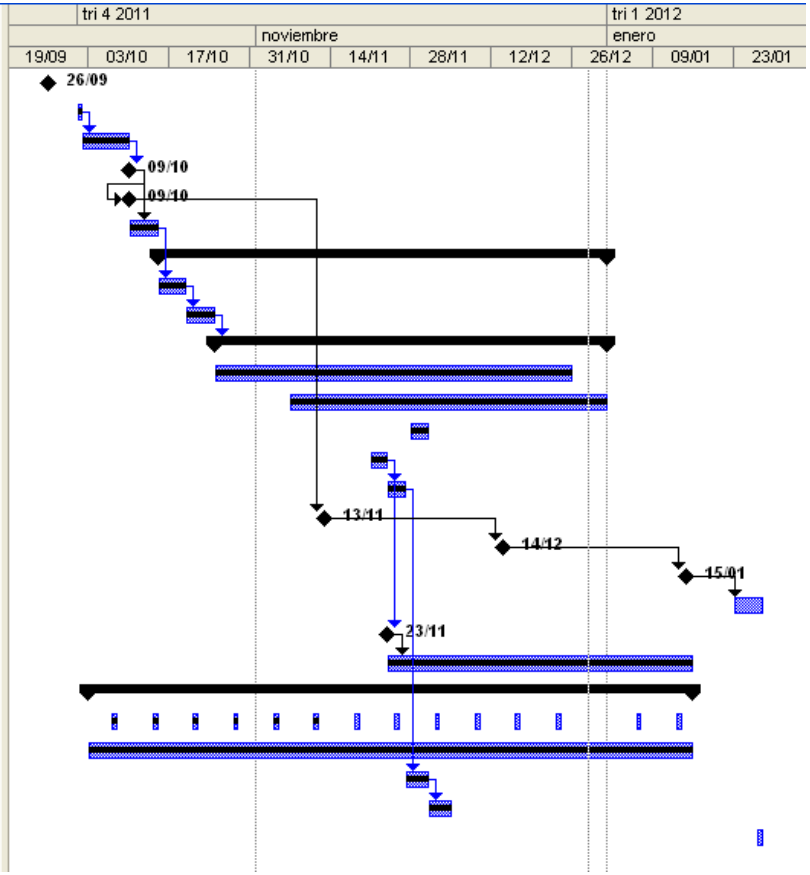
Les tasques T9 i T10, que són transversals a tot el projecte i s'executaran durant el cicle de vida del mateix, segueixen la seva planificació sense anomalies destacables.

S'han finalitzat les tasques T11 i T12, documentació de manuals i presentació, segons el calendari previst.

Malgrat les desviacions parcials experimentades en les tasques d'implementació s'han assolit la totalitat dels objectius del present projecte.

Tasca	Hores	Inici	Final	% Progrés
T1 - Redacció Pla de treball - PAC1	10	02/10/2011	09/10/2011	100
T2 - Recerca Eines – Tecnologia	10	07/10/2011	11/10/2011	100
T3 - Definició Eines – Tecnologia	10	12/10/2011	16/10/2011	100
T4 - Disseny Tècnic Producte	10	17/10/2011	21/10/2011	100
Desenvolupament				
T5 - Motor de cerca	30	22/10/2011	01/01/2012	100
T6 - Analitzador semàntic	20	03/11/2011	01/01/2012	100
T7 - Gestió resultats	5	18/11/2011	01/01/2012	100
T8 - Revisió Producte – Test	5	21/11/2011	01/01/2012	100
T9 - Gestió Projecte - Seguiment	10	03/10/2011	15/01/2012	100
T10 - Documentació Memòria	20	03/10/2011	15/01/2012	100
T11 - Manual Usuari	5	24/11/2011	15/01/2011	100
T12 - Manual Instal·lació	5	28/11/2011	15/01/2011	100
T13 - Preguntes Tribunal - Debat Virtual	5	23/01/2012	27/01/2012	

	Nombre de tarea	Duración	Comienzo	Fin	Prede
1	Publicació Enunciat PFC - Kick-off Projecte	0 días	lun 26/09/11	lun 26/09/11	
2	Trobada Presencial	1 día	sáb 01/10/11	sáb 01/10/11	
3	Redacció Pla de treball - PAC1	8 días	dom 02/10/11	dom 09/10/11	2
4	Lliurament PAC1 - Pla de treball	0 días	dom 09/10/11	dom 09/10/11	3
5	Aprovació Pla del Projecte	0 días	dom 09/10/11	dom 09/10/11	4
6	Recerca Eines - Tecnologia	5 días	lun 10/10/11	vie 14/10/11	4
7	**Implementació PFC**	**73 días**	**sáb 15/10/11**	**sáb 31/12/11**	
8	Definició Eines - Tecnologia	5 días	sáb 15/10/11	mié 19/10/11	6
9	Disseny Tècnic Producte	5 días	jue 20/10/11	lun 24/10/11	8
10	**Desenvolupament**	**63 días**	**mar 25/10/11**	**sáb 31/12/11**	**9**
11	Motor de cerca	62 días	mar 25/10/11	dom 25/12/11	
12	Analitzador semàntic	50 días	lun 07/11/11	sáb 31/12/11	
13	Gestió Resultats	3 días	lun 28/11/11	mié 30/11/11	
14	Proves - Test Producte	3 días	lun 21/11/11	mié 23/11/11	
15	Revisió Producte	3 días	jue 24/11/11	sáb 26/11/11	14
16	PAC2 Seguiment PFC - Lliurament Parcial	0 días	dom 13/11/11	dom 13/11/11	5
17	PAC3 Seguiment PFC - Lliurament Parcial	0 días	mié 14/12/11	mié 14/12/11	16
18	Lliurament Final	0 días	dom 15/01/12	dom 15/01/12	17
19	Preguntes Tribunal - Debat Virtual	5 días	lun 23/01/12	vie 27/01/12	18
20	Pla de contingències	0 días	mié 23/11/11	mié 23/11/11	14
21	Marge de maniobra per possibles contingències	48 días	jue 24/11/11	dom 15/01/12	20
22	**Gestió Projecte**	**100 días**	**lun 03/10/11**	**dom 15/01/12**	
23	**Seguiment Projecte**	**94 días**	**vie 07/10/11**	**vie 13/01/12**	
38	Documentació Memoria	100 días	lun 03/10/11	dom 15/01/12	
39	Manual Usuari	4 días	dom 27/11/11	mié 30/11/11	15
40	Manual Instal·lació	4 días	jue 01/12/11	dom 04/12/11	39
41	Tancament Projecte	1 día	vie 27/01/12	vie 27/01/12	



6. Tecnologia

Actualment existeixen diverses opcions per implementar un projecte d'aquestes característiques, des de llenguatges orientats a la web com PHP o com plataformes com Java o .net., a eines d'extracció de dades web com Harvest[7], desenvolupada en Java i de codi obert, o projectes com Nutch[8] i Lucene[9], que estan orientats a crear motors de cerca.

També trobem productes de mercat, com Web Content Extractor o Visual Web Spider del fabricant Newprosoft, que cobreixen aquestes funcionalitats i que mitjançant un assistent(wizard) es poden configurar les web i les cerques a realitzar.

Per afinitat i coneixença amb el programari utilitzat en la majoria d'assignatures dels d'estudis, s'ha decidit que la plataforma de desenvolupament escollida per la implementació del producte sigui Java i Eclipse com a entorn de treball.

Pel que fa al tractament de connexions, l'enviament de peticions i la recuperació de les respostes dels servidors, Java incorpora classes bàsiques tipus "socket" (java.net.Socket)[10] que permeten fer-ho, però de cara a simplificar la codificació existeixen llibreries més específiques que faciliten aquesta tasca.

Aquestes llibreries auxiliars faciliten les tasques d'implementació, ja que ofereixen els principals mètodes a utilitzar, les operacions més habituals, degudament encapsulades.

Una d'aquestes llibreries es la que incorpora el projecte Apache HttpComponents, que esta orientat al desenvolupament de programari que ajudi a construir aplicacions que utilitzin el protocol HTTP, com aplicacions de client, servidors, navegadors o aranyes web.

Pel que fa al tractament de dades, manipulacions i transformacions de XML es farà servir la tecnologia SAX/DOM i llibreries de Java com JAXP i JAXB.

[7] (Harvest, 2008)
[8] (Wikipedia, Nutch, 2008)
[9] (Apache, Lucene, 2008)
[10] (Wikipedia, Socket_programming, 2004)

L'estructura XML haurà de contenir les dades de la petició, els parametres d'entrada, i els resultats de la cerca, dades de sortida, ordenat per criteris de millor preu.

Dades(in)	Dades Retorn(out)	Vols	Hotels
Origen		x	
Destí		x	x
Data ini		x	x
Data fi			x
Ocupants		x	x
Nens		x	x
	Nom servei (Oferta id)	x	x
	Descripcio	x	x
	Proveidor	x	x
	Preu	x	x

6.1. Descripció de la tecnologia

A continuació es realitza una enumeració de les eines utilitzades en el procés d'implementació del present projecte, així com una introducció sobre els conceptes més destacats.

Java – llenguatge de programació i plataforma tecnològica.
Eclipse – entorn integrat de desenvolupament.
Jigloo – interface d'usuari
HTTP – protocol de comunicacions
Apache HttpComponents – llibreries de gestió HTTP

Editplus – editor de codi
Cooktop – utilitats per XML, XSLT

Firefox Mozilla – navegador per defecte
Internet Explorer - navegador per fer proves
HTTPLive Headers – analitzador de capçaleres HTTP

Microsoft Word – editor de text.
PDFCreator – conversor a format PDF.
Microsoft Project – planificació i diagrames de Gantt.
Microsoft Visio - diagrames
Microsoft Powerpoint – presentació base

SMRecorder – gravador d'escriptori en vídeo (.AVI)
SMConverter – conversió a format .FLV

6.2. HTTP

El protocol HTTP[11], Hypertext Transfer Protocol, és el que es fa servir en les transaccions web. Es tracta d'un protocol orientat a transaccions que segueix el patró petició-resposta.

Una transacció HTTP està formada per un encapçalament seguit, opcionalment, per una línia en blanc i alguna dada. L'encapçalament especificarà coses com l'acció requerida del servidor, o el tipus de dada retornada, o el codi d'estat.

L'ús de camps d'encapçalaments enviats en les transaccions HTTP li donen gran flexibilitat al protocol. Aquests camps permeten que s'enviï informació descriptiva en la transacció, permetent així l'autenticació, xifrat i identificació d'usuari.

Un encapçalament és un bloc de dades que precedeix la informació pròpiament dita, per la qual cosa moltes vegades es fa referència a ell com a metadada, ja que té dades sobre les dades,

El servidor envia el client :

- Un codi d'estat que indica si la petició va ser correcta o no. Els codis d'error típics indiquen que l'arxiu|arxivament sol•licitat no es va trobar, que la petició no es va realitzar de forma correcta o que es requereix autenticació per accedir a l'arxiu.

- La informació pròpiament dita. HTTP permet enviar documents de tot tipus i format, és ideal per transmetre dades multimèdia, com gràfics, àudio i vídeo.

Per obtenir un recurs amb l'URL http://www.example.com/index.html

1. S'obre una connexió a l'ordinador central www.example.com, port 80 que és el port per defecte per a HTTP.

2. S'envia un missatge en l'estil següent:

```
GET /index.html HTTP/1.1

 Host: www.example.com
 User-Agent: nomm-client
```

[11] (Wikipedia, HTTP, 2004)

La resposta del servidor seria:

```
HTTP/1.1 200 OK
Date: Fri, 13 Nov 2011 23:59:59 GMT
Content-Type: text/html
Content-Length: 1221

<html>
<body>
<h1>Página principal</h1>
  .
  .
  .
</body>
</html>
```

Mètodes de petició

- HTTP. Defineix uns mètodes que indiquen l'acció que es desitja que s'efectuï sobre el recurs identificat. El que aquest recurs representa, si les dades preexistents o dades que es generen de forma dinàmica, depèn de l'aplicació del servidor. Sovint, el recurs correspon a un arxiu o la sortida d'un executable que resideixen en el servidor.

- HEAD. Demana una resposta idèntica a què correspondria a una petició GET, però sense el cos de la resposta. Això es fa servir per a la recuperació de meta informació escrita en els encapçalaments de resposta, sense haver de transportar tot el contingut.

- GET. Demana una representació del recurs especificat. Per seguretat no hauria de ser usat per aplicacions que causin efectes ja que transmet informació a través de l'URI agregant paràmetres a l'URL.

- POST. Sotmet les dades que siguin processats per al recurs identificat. Les dades s'inclouran al cos de la petició. Això pot resultar útil en la creació d'un nou recurs o de les actualitzacions dels recursos existents.

- PUT. Puja, carrega o realitza un upload d'un recurs especificat, és el camí més eficient per pujar arxius a un servidor perquè en POST utilitza un missatge multipart i el missatge és descodificat pel servidor. En contrast, el mètode PUT permet escriure un arxiu en una connexió socket establerta amb el servidor. El desavantatge del mètode PUT és que els servidors d'hosting compartit habitualment no el tenen habilitat.

- TRACE. Aquest mètode sol·licita al servidor que enviï de tornada en un missatge de resposta, a la secció del cos d'entitat, tota les dades que rebi del missatge de sol·licitud. S'utilitza amb finalitats de comprovació i diagnòstic.

- OPTIONS. Torna els mètodes HTTP que el servidor suporta per a un URL específic. Això es pot ser utilitzat per comprovar la funcionalitat d'un servidor web.

6.3. XML

XML[12], de l'anglès eXtensible Markup Language, llenguatge de marques extensible, és un metallenguatge extensible, d'etiquetes, desenvolupat pel World Wide Web Consortium, W3C. És una simplificació i adaptació de l'experimentat SGML, i permet definir la gramàtica de llenguatges específics, de la mateixa manera que HTML és, alhora, un llenguatge definit per SGML. Per tant, XML no és realment un llenguatge en particular, sinó una manera de definir llenguatges per a diferents necessitats. Alguns dels llenguatges que empren XML per a la seva definició són XHTML, SVG, MathML.

XML no ha nascut només per a la seva aplicació a Internet, sinó que es proposa com a un estàndard per a l'intercanvi d'informació estructurada entre diferents plataformes. Es pot utilitzar per a bases de dades, editors de text, fulls de càlcul i per moltes altres aplicacions diverses. XML és una tecnologia relativament senzilla que té al seu voltant altres que la complementen i la fan notablement més extensa, a més de proporcionar-li unes possibilitats molt més grans. A l'actualitat té un paper molt important, ja que permet la compatibilitat entre sistemes, permetent de compartir informació d'una manera segura, fiable i fàcil.

XML prové d'un llenguatge inventat per IBM als anys setanta, anomenat GML, General Markup Language, i que va ser creat per la necessitat que tenia l'empresa d'emmagatzemar grans quantitats d'informació. Aquest llenguatge va agradar a la ISO, i al 1986 van començar a treballar per tal de normalitzar-lo, creant el llenguatge SGML, Standard General Markup Language, que era capaç d'adaptar-se a un ampli ventall de problemes. A partir d'aquest SGML s'han creat altres sistemes d'emmagatzematge d'informació.

L'any 1989 Tim Berners Lee va crear la web i, juntament amb ella, el llenguatge HTML. Aquest llenguatge es va definir en el marc del SGML, i va ser l'aplicació més coneguda d'aquest estàndard.

La tecnologia XML busca donar solució al problema d'expressar informació estructurada de la manera més abstracta i reutilitzable possible. Per informació estructurada entenem que es compon de parts ben definides, i que aquelles parts es componen d'altres parts.

Una etiqueta consisteix en una marca feta al document, que senyala una porció d'aquest com un element, un tros d'informació amb un sentit clar i definit. Les etiquetes tenen la forma <nom>, on nom és el nom de l'element senyalat.

[12] (Wikipedia, XML, 2005)

Els documents anomenats ben conformats, de l'anglès well formed, són els que acompleixen totes les definicions bàsiques de format i poden, en conseqüència, ésser analitzats correctament per qualsevol analitzador sintàctic, parser, que segueixi la norma. Distingirem aquest concepte del de validesa, que s'explica més endavant.

Els documents han de seguir una estructura estrictament jeràrquica pel que fa a les etiquetes que delimiten els seus elements. Una etiqueta ha d'estar correctament inclosa dins d'una altra. Els elements amb contingut han d'estar correctament tancats.

Els documents XML només permeten un element arrel del qual la resta en formin part, és a dir, només poden tenir un element inicial.
Els valors atributs en XML sempre han d'estar tancats entre cometes simples o dobles.

L'XML és sensible a majúscules i minúscules.

Les construccions tals com etiquetes, referències d'entitat i declaracions s'anomenen marques; són parts del document que el processador XML espera entendre. La resta del document entre marques són les dades comprensibles per les persones.

Els elements XML poden tenir contingut, més elements, caràcters o ambdós, o bé ésser elements buits. Els elements poden tenir atributs, que són una manera d'incorporar característiques o propietats als elements d'un document.

Que un document sigui ben conformat únicament parla de la seva estructura sintàctica bàsica, és a dir que es compongui d'elements, atributs i comentaris com XML mana que s'escriguin. Ara bé, cada aplicació d' XML, és a dir, cada llenguatge definit amb aquesta tecnologia, necessitarà especificar quina és exactament la relació que s'ha de verificar entre els diferents elements presents en el document. Aquesta relació entre elements s'especifica en un document extern o definició expressada com a DTD o com a Xschema.

La DTD, Document type definition, defineix els tipus d'elements, atributs i entitats permeses, i pot expressar algunes limitacions per combinar-los. Els documents XML que s'ajusten a la seva DTD s'anomenen vàlids.

A continuació, un exemple per a entendre l'estructura d'un document XML[13].

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<!DOCTYPE Edit_Missatge SYSTEM "Llista_dades_missatge.dtd"
[<!ELEMENT Edit_Missatge (Missatge)*>]>
<Edit_Missatge>
  <Missatge>
    <Remitent>
      <Nom>Nom del remitent</Nom>
      <Mail> Correu del remitent </Mail>
    </Remitent>
    <Destinatari>
      <Nom>Nom del destinatari</Nom>
      <Mail> Correu del destinatari</Mail>
    </Destinatari>
    <Text>
      <Paràgraf>
      Aquest és el meu document, amb una estructura molt senzilla. No
conté atributs ni entitats...
      </Paràgraf>
    </Text>
  </Missatge>
</Edit_Missatge>
```

Aquí hi ha l'exemple de codi de la DTD del document "Edit_missatge":

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
 <!-- Aquesta és la DTD de Edit_Missatge -->
<!ELEMENT Missatge (Remitent, Destinatari, Assumpte, Text)*>
          <!ELEMENT Remitent (Nom, Mail)>
                    <!ELEMENT Nom    (#PCDATA)>
                    <!ELEMENT Mail   (#PCDATA)>
           <!ELEMENT Destinatari (Nom, Mail)>
                    <!ELEMENT Nom    (#PCDATA)>
                    <!ELEMENT Mail   (#PCDATA)>
           <!ELEMENT Assumpte (#PCDATA)>
          <!ELEMENT Text  (Parraf)>
                    <!ELEMENT Parraf (#PCDATA)>
```

[13] (Wikipedia, Web_scraping, 2006)

6.4. XSLT

XSL[14] és l'acrònim d'Extensible Stylesheet Language, expressió anglesa traduïble com llenguatge extensible de fulles d'estil". És una família de llenguatges basats en l'estàndard XML que permet descriure com la informació continguda en un document XML qualsevol ha de ser transformada per a la seva presentació en un mitjà específic.

Aquesta família està formada per tres llenguatges:

- XSLT sigles d'Extensible Stylesheet Language Transformations, llenguatge de fulles extensibles de transformació, que permet convertir documents XML d'una sintaxi a altra (per exemple, d'un XML a un altre o a un document HTML).

- XSL-FO llenguatge de fulles extensibles de format d'objectes, que permet especificar el format visual amb el qual es vol presentar un document XML, és usat principalment per a generar documents PDF.

- XPath, o XML Path Language, és una sintaxi (no basada en XML) per a accedir o referir-se a porcions d'un document XML.

Aquestes tres especificacions són recomanacions oficials del W3C.

En el 2005 ja són suportades per alguns navegadors, per exemple Mozilla o Internet Explorer, encara que, en el seu lloc, es poden usar les CSS que són 100% compatibles encara que amb una codificació diferent.

XSLT o XSL Transformacions és un estàndard de l'organització W3C que presenta una forma de transformar documents XML en uns altres i fins i tot a formats que no són XML. Les fulles d'estil XSLT realitzen la transformació del document utilitzant una o diverses regles de plantilla: unides al document font a transformar, aquestes regles de plantilla alimenten a un processador de XSLT, el qual realitza les transformacions desitjades col·locant el resultat en un arxiu de sortida o, com en el cas d'una pàgina web, directament en un dispositiu de presentació, com el monitor d'un usuari.

Actualment, XSLT és molt usat en l'edició web, generant pàgines HTML o XHTML. La unió de XML i XSLT permet separar contingut i presentació, augmentant així la productivitat.

[14] (Wikipedia, Xslt, 2008)

6.5. HTML

HTML[15], acrònim d'Hyper Text Markup Language, que es podria traduir com llenguatge de marcat d'hipertext, és un llenguatge de marcat que deriva de l'SGML dissenyat per estructurar textos i relacionar-los en forma d'hipertext. Gràcies a Internet i als navegadors web, s'ha convertit en un dels formats més populars que existeixen per a la construcció de documents per a la web.

En el seu origen era un llenguatge dissenyat per compartir informació científica entre científics de tot el món. Era purament un llenguatge estructural, en què no hi havia forma de descriure l'aparença de les pàgines (ni tan sols la possibilitat de posar un text en negreta o cursiva). Més endavant s'hi van afegir nombroses opcions per formatar i presentar text i gràfics.

A mitjans de la dècada de 1990 van començar les ampliacions de l'HTML per aconseguir la presentació desitjada, però sempre des de diferents perspectives de diferents desenvolupadors, que van acabar amb diverses solucions no estàndards per a diferents navegadors. Això va provocar l'aparició d'un consorci que controla l'evolució de l'HTML: el W3C (World Wide Web Consortium).

Aquesta evolució tenia un punt clau que fou la separació del contingut i l'aparença. Amb la versió 4 de l'HTML es recomanava un altre mecanisme per controlar la visualització del nostre contingut HTML: els fulls d'estil, CSS, Cascading Style Sheets.

Les etiquetes bàsiques d'HTML, d'obligada presència en tot document són:

- <!DOCTYPE>: És l'etiqueta que permet definir el tipus de document HTML que s'empra. Existeixen tres tipus bàsics: l'estricte (Strict), el transicional (Transitional) i el de marcs (Frameset).

- <html>: És l'etiqueta arrel de qualsevol document HTML o XHTML.

- <head>: Defineix la capçalera del document HTML. Permet declarar meta del document que no es mostra directament en el navegador. Aquesta informació és d'especial rellevància pels indexadors i cercadors automàtics.

- <body>: Defineix el cos del document. Aquesta és la part del document HTML que es mostra en el navegador.

[15] (Wikipedia, HTML, 2002)

Dintre de la capçalera <HEAD> hi podem trobar:

- <title>: Permet definir el títol de la pàgina. En navegadors gràfics el contingut del title apareix a la barra del títol a sobre de la finestra.

- <meta>: Permet definir metainformacions del document tals com l'autor, la data de realització, la codificació del document (UTF, ISO, etc.), les paraules clau i la descripció del mateix

- <LINK>: Permet definir metadades complementàries a les del meta tals com el document anterior, el següent, el capítol al qual pertany el document, la pàgina glossari, etc.

Dintre del cos <BODY> hi podem trobar:

- <a>: Etiqueta àncora. Crea un enllaç a un altre document o a una altra zona del mateix, segons els atributs.

- <h1>, <h2>,… <h6>: capçaleres o títols del document, acostumen a distingir-se per mida.

- : Divisió estructural de la pàgina.

- <p>: Paràgraf.

-
: Salt de línia.

- <table>: Indica el començament d'una taula, després s'haurà de definir les files amb <tr> i les cel•les dintre de les files amb <td>.

- : Llista desordenada (sense numerar). Els ítems es defineixen amb <li>.

- : Llista ordenada (numerat). Els ítems es defineixen amb <li>.

- <dl>: Llista de definició. Hi ha dos tipus d'ítem; el dt i el dd.

- <dt>: Terme a definir.

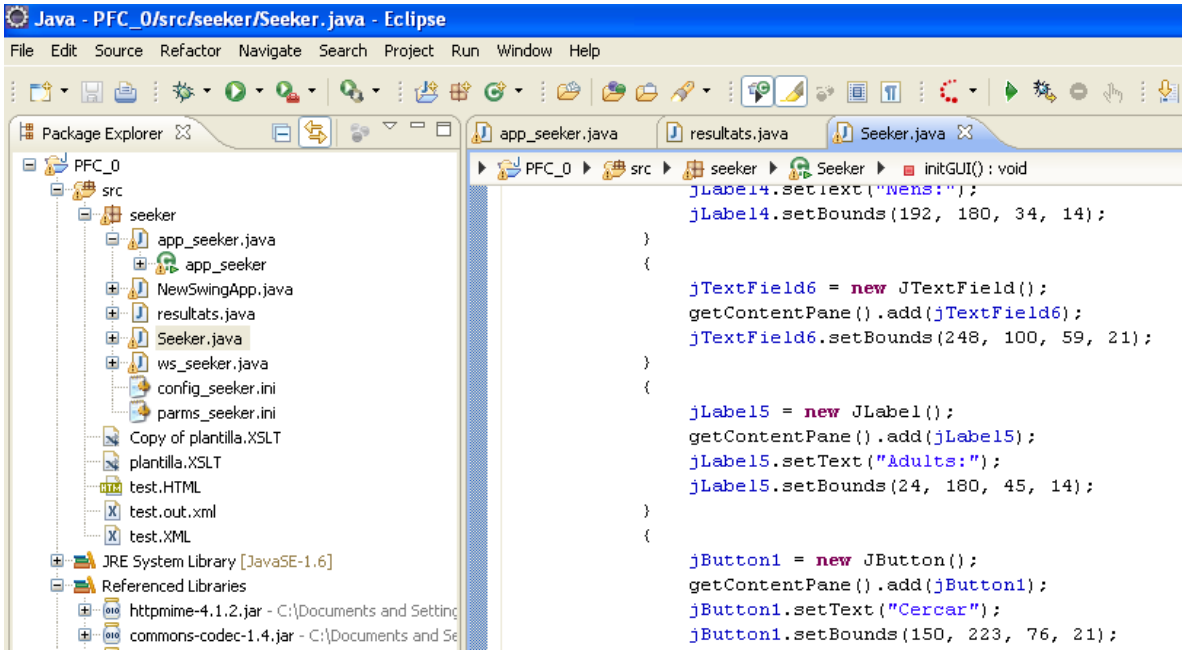
- <dd>: Definició del terme.

Excepte unes poques etiquetes, la majoria requereixen ser tancades escrivint la mateixa etiqueta precedida d'una barra "/". Exemple: <html>…</html>

6.6. Eclipse

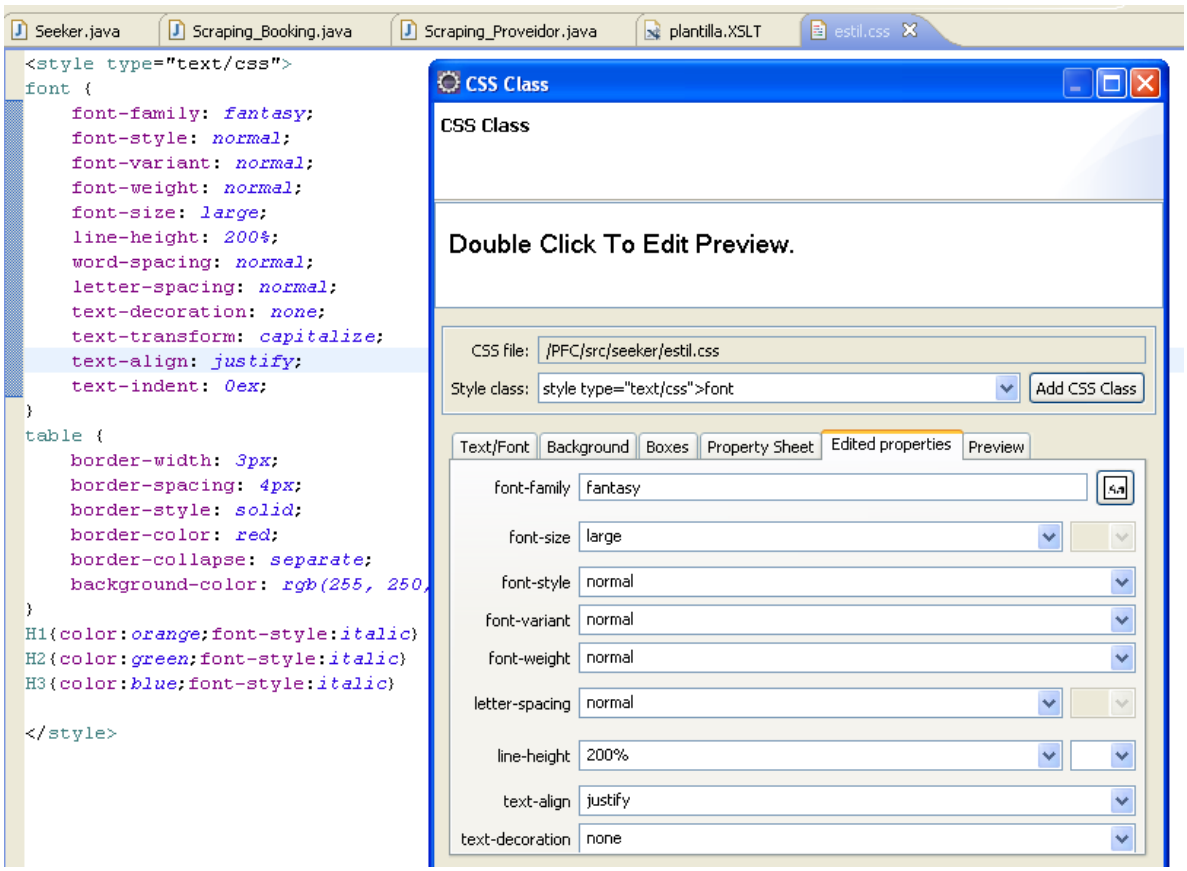
Eclipse[16] és un complet IDE, entorn integrat de desenvolupament, de codi obert programat principalment en Java, per tant, multi plataforma, que integra diferents eines i llibreries de fabricants diversos, amb l'objectiu d'ajudar i facilitar la codificació de projectes en Java, entre d'altres llenguatges, ja que també suporta per a desenvolupar C, C++, COBOL, Python, Perl, PHP, i molts altres, sempre i quan s'instal·lin els connectors corresponents per a cada llenguatge de programació.

Eclipse fou desenvolupat originalment per IBM com el successor de la seva família d'eines per a VisualAge. Tot i això, actualment Eclipse és desenvolupat per la Fundació Eclipse, una organització independent sense ànim de lucre que fomenta una comunitat de programari lliure i un conjunt de productes complementaris, capacitats i serveis.



16

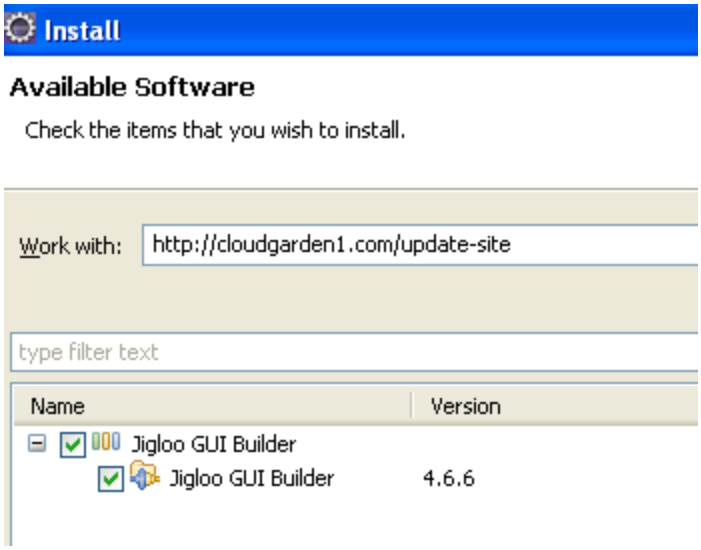
També ens pemet editar i modificar la forma i les característiques d'un full d'estil, css, per formatejar la sortida del codi HTML.



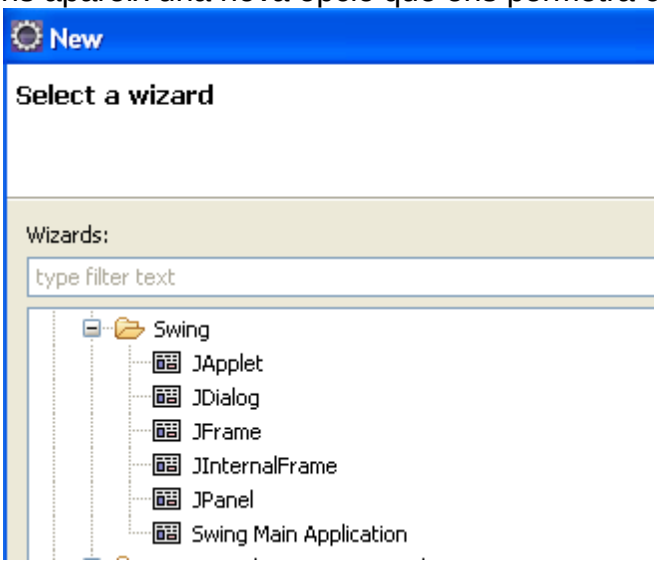
6.7. Jigloo

Jigloo[17] es un plug-in per a Eclipse que ens permet dissenyar la interfície gràfica en un mode WYSIYG, What You See Is You Get, que es podria traduir com el que veus es el que obtens. Es bassa en SWING i soporta AWT. Genera codi Java i les classes dels objectes que es visualitzen facilitant les tasques de desenvolupament, ja que s'encarrega de codificar, traduir a codi Java, tot el que es dissenya en mode visual.

La seva instal·lació és molt senzilla, ja que només cal afegir l'adreça del fabricant del programari des de l'apartat de "Programari Disponible" i el mateix Eclipse gestiona la descarrega de les llibreries necessàries.



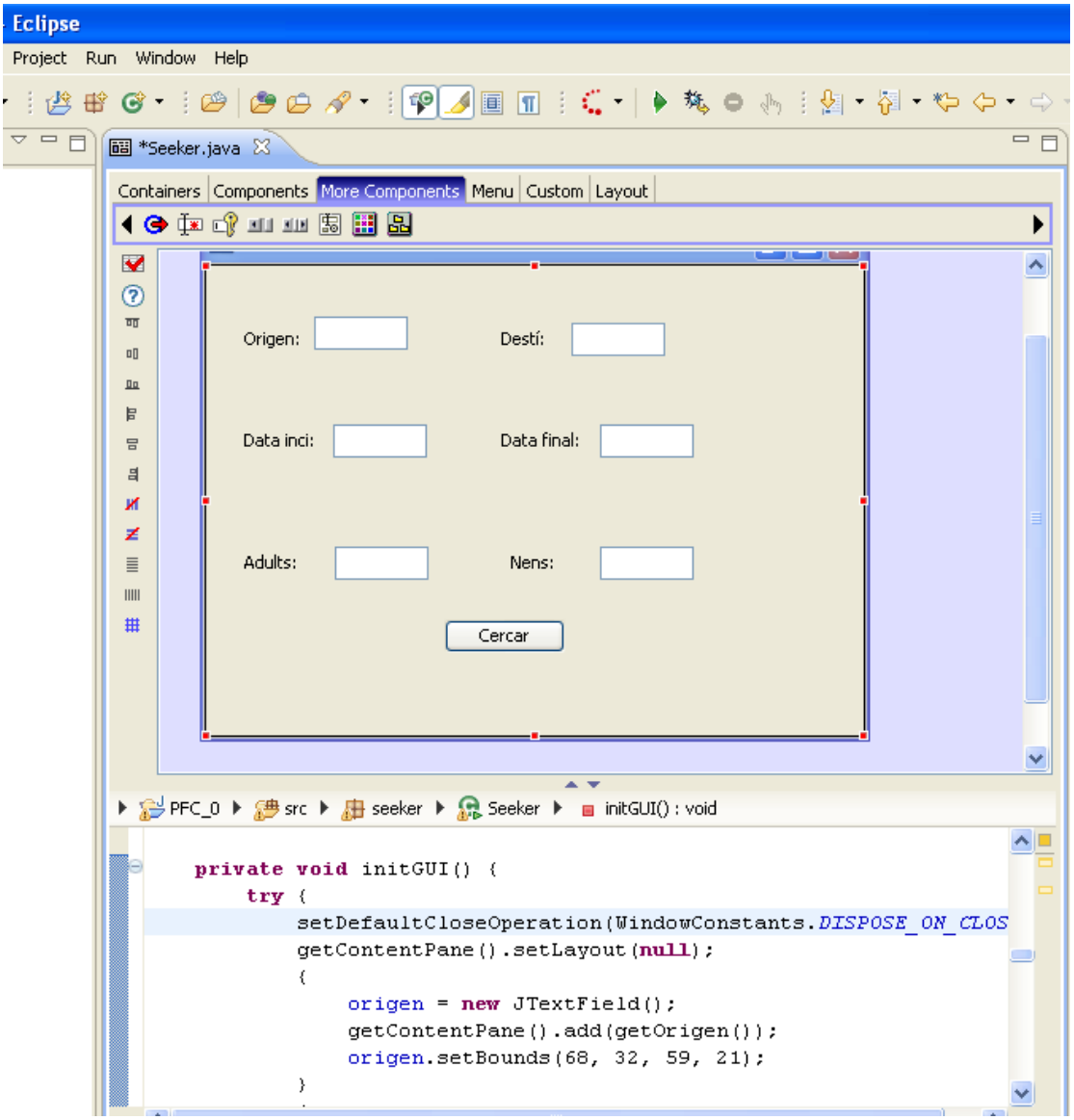
Un cop instal·lat, ens apareix una nova opció que ens permetrà crear el formulari.



[17] (Cloudgarden, 2008)

L'entorn de disseny gràfic es senzill i intuïtiu, en la línia de la majoria d'aquest tipus d'eines estil Windows, com Visual Basic, a on només cal arrossegar els objectes que volem que hi apareguin en la nostra aplicació.

Permet modificar la forma i les característiques de visualització, així com la resposta i programació dels diferents esdeveniments que es produiran en la implementació de la lògica de negoci del producte.



6.8. JAVA

Java[18] és un llenguatge de programació dissenyat el 1990 per James Gosling amb altres companys de Sun Microsystems a partir de C++. Des del seu naixement fou pensat com un llenguatge orientat a objectes. Entre el 13 de novembre de 2006 i el maig del 2007 Sun va alliberar parts de Java com a programari lliure de codi obert amb llicència GPL. És un dels llenguatges de programació més utilitzats, i s'utilitza tant per aplicacions web com per aplicacions d'escriptori.

Es un llenguatge interpretat i, per tant, pot semblar lent en comparació amb altres llenguatges, però ofereix un índex de reutilització de codi molt elevat, sent possible trobar moltes llibreries lliures de Java. És flexible i potent tot i la facilitat amb la què es programa i dels resultats que ofereix. Un dels trets que el caracteritza i que el fa una eina molt valorada a l'hora de desenvolupar aplicacions distribuïdes, és el fet que és un llenguatge multi-plataforma.

Generalment els programes de Java es compilen en un bytecode, fitxer .class, que pot córrer en una Màquina Virtual Java. Sun Microsystems disposa de tres implementacions diferents de Java: J2SE per a aplicacions d'escriptori; J2EE per a aplicacions distribuïdes i J2ME per a plataformes amb recursos més reduïts com ara mòbils o PDAs. Per a cada una de les tres implementacions és possible descarregar el JRE (entorn d'execució Java) per a executar aplicacions o el SDK (Eines per al desenvolupament d'aplicacions) per a programar aplicacions en Java, aquest últim també inclou el JRE.

Un programa desenvolupat amb Java no necessita compilar-se de nou per a poder executar-se en qualsevol de les plataformes que disposi d'una versió instal·lada de JRE prou actualitzada per al programa.

Característiques de Java[19]

- Senzill: Java s'ha creat per a que sigui un llenguatge senzill amb una sintaxi elegant. Únicament consta de tres tipus de dades primàries, eliminant els punters i l'herència múltiple

- Orientat a objectes: Java segueix els paradigmes de la programació orientada a objectes, ja que la programació amb Java es centralitza en la manipulació, creació i construcció d'objectes.

- Distribuït:Java permet la construcció d'aplicacions distribuïdes per mitjà d'una col·lecció especifica de classes.

- Interpretat: Es necessita un intèrpret per executar els programes de Java, això alenteix als programes però els hi dóna flexibilitat.

[18] (Javahispano, 2009)
[19] (Wikipedia, Java, 2005)

- Robust: Java és un llenguatge robust i fiable, s'ha escrit pensant en poder verificar errors i està molt tipificat.

- Segur: Java té pocs problemes de seguretat, característica molt important en les aplicacions distribuïdes d'Internet.

- Arquitectura neutral: Java és independent de la plataforma final on s'executarà el programa.

- Portable: Java és un llenguatge d'alt nivell i de plataforma independent, això li dóna portabilitat.

- Alt rendiment: Els compiladors Java han anat millorant les seves prestacions. Els nous compiladors coneguts com JIT permeten un rendiment molt semblant als llenguatges convencionals compilats.

- Concurrent: Java permet l'execució de múltiples fils d'execució, o diverses tasques de forma simultània.

- Dinàmic: En temps d'execució, l'entorn Java es pot ampliar mitjançant enllaços a classes que poden estar localitzades en servidors remots o en xarxa.

Altres característiques remarcables del llenguatge Java són la herència, classes abstractes, interfícies, encapsulació i polimorfisme, interfícies gràfiques d'usuari, gestió d'esdeveniments, programació concurrent i excepcions.

La sintaxi del Java deriva en gran part del C++. Però a diferència d'aquest, que combina la sintaxi per a programació genèrica, estructurada i orientada a objectes, el Java va ser dissenyat gairebé exclusivament com a llenguatge orientat a objectes. Tot el codi es troba dins d'una classe i tot és un objecte.

6.9. Expressions regulars

En informàtica, una expressió regular[20] o també anomenades regexp, acrònim de l'anglès regular expression, és una representació segons unes regles sintàctiques d'un llenguatge formal d'una porció de text genèric a buscar dins d'un altre text, com per exemple uns caràcters, paraules o patrons de text concrets.

El text genèric de l'expressió regular pot representar patrons,en anglès patterns, amb determinats caràcters que tenen un significat especial. (per exemple, en el cas del shell d'unix, el caràcter comodí "?" per representar un caràcter qualsevol, el caràcter comodí "*" per representar un nombre qualsevol de caràcters, o classes com "[abc]" per representar qualsevol dels caràcters 'a', 'b' o 'c').

Una expressió regular està escrita seguint les regles d'un llenguatge formal, que poden ser interpretades per un programa processador d'expressions regulars, capaç d'examinar un text i reconèixer-hi les parts que es corresponen (en anglès match) amb l'expressió regular especificada.

Molts processadors de textos i llenguatges de programació fan ús de les seves pròpies expressions regulars per a procediments de cerca o bé de cerca i substitució de textos.

Les expressions regulars[21] s'utilitzen des de fa anys en altres llenguatges de programació com Perl. En la versió 1.4 del JDK, Java Devloper Kit, de Sun s'inclou el paquet java.util.regex, que proporciona una sèrie de classes per poder fer ús de la potència d'aquest tipus d'expressions a Java. Abans de res no necessitem saber què és una expressió regular i per a que ens pot servir: Doncs bé, una expressió regular és un patró que descriu a una cadena de caràcters. Tots hem utilitzat alguna vegada l'expressió *.doc per buscar tots els documents en algun lloc del nostre disc dur, doncs bé, *.doc és un exemple d'una expressió regular que representa a tots els arxius|arxivaments amb extensió doc, l'asterisc significa qualsevol seqüència de caràcters (val, els que ja coneguin això diran que no és correcte, i diran bé, és mes precís parlar de *.doc però l'exemple és molt gràfic).

Les expressions regulars es regeixen per una sèrie de normes i hi ha una construcció per a qualsevol patró de caràcters. Una expressió regular només pot contenir (a part de lletres i números|nombres) els següents caràcters:

< $, ^, ., *, +, ?, [,], . >

[20] (Luauf.com, 2008)
[21] (Programacion.com, 2005)

Una expressió regular[22], ens servirà per buscar patrons en una cadena de text, per exemple trobar quantes vegades es repeteix una paraula en un text, per comprovar que una cadena té una determinada estructura, per exemple que el nom d'arxiu que ens proposen té una determinada extensió, o comprovar que un email aquesta ben escrit

El paquet java.util.regex esta format per dues classes, la classe Matcher i la classe Pattern i per una excepció, PatternSyntaxException.

La classe Pattern, patró, és la representació compilada d'una expressió regular, o el que és el mateix, representa l'expressió regular, que al paquet java.util.regex necessita estar compilat.

La classe Matcher, "encaixador", és un tipus d'objecte que es crea a partir d'un patró mitjançant la invocació del mètode Pattern.matcher. Aquest objecte és el que ens permet realitzar operacions sobre la seqüència de caràcters que volem validar o buscar.

Lògics
x|y: x o i.
xy: x seguit de i

Intervals de caràcters:
[abc]: Qualsevol dels caràcters entre claudàtors. Poden especificar-se rangs, per exemple [a-d] que equival a [abcd]).
[^abc]: Qualsevol caràcter que no hi hagi els que estan entre claudàtors.

Intervals de caràcters predefinits
. Qualsevol caràcter individual, llevat del de salt de línia.
d: Qualsevol caràcter de dígit, equivalent a [0-9].
D: Qualsevol caràcter que no sigui de dígit, equival a [^0-9].
s: Qualsevol caràcter individual d'espai en blanc (espais, tabulacions, salts de pàgina o salts de línia).
S: Qualsevol caràcter individual que no sigui un espai en blanc.
w: Qualsevol caràcter alfanumèric, equivalent a [A-Za-z0-9_].
W: Qualsevol caràcter que no sigui alfanumèric, equivalent a [^A-Za-z0-9_].

Caràcters
. : Salt de pàgina.
. : Salt de línia.
. : Retorn de carro.
. : Tabulació.

[22] (Wikipedia, Regular_expression, 2005)

Limits
^: Principi d'entrada o línia.
$: Final d'entrada o línia.
B: Final de paraula.

Quantificadors
{n}: Exactament n aparicions del caràcter anterior.
{n,m}: Com a mínim n i com a màxim m aparicions del caràcter anterior.
*: El caràcter anterior 0 o més vegades.
+: El caràcter anterior 1 o més vegades.
?: El caràcter anterior una vegada com a màxim (el caràcter és opcional).

Codi exemple.

```
import java.util.regex.Matcher;
import java.util.regex.Pattern;

public class TestMail {

    public static void main(String[] args) {
        String input = "www.test@mail.com";

        // verifica que no comenci per punt o @

        Pattern p = Pattern.compile("^\\.|^\\@");
        Matcher m = p.matcher(input);
        if ( m.find() )
            System.err.println("Adreça incorrecta");
```

6.10. Apache HttpComponents

El projecte de HttpComponents[23] d'Apatxe és el responsable de crear i mantenir un conjunt de components de Java de baix nivell centrats en HTTP i els protocols associats. Aquest projecte funciona sota la direcció de la Fundació de Programari d'Apatxe (http://www.apache.org), i és part d'una comunitat més gran de desenvolupadors i usuaris.

Visió de conjunta de HttpComponents

El Protocol de Transferència de Hyper-Text (HTTP) és potser el protocol més significatiu utilitzat en la Internet avui. Els serveis web, els electrodomèstics permesos de xarxa i el creixement de computació de xarxa continuen expandint el paper del protocol de HTTP més enllà de navegadors web orientats a l'usuari, mentre augmenten el número d'aplicacions que requereixen del suport de HTTP.

Dissenyat per extensió proporciona un suport robust pel protocol HTTP de base, els HttpComponents poden ser d'interès per a qualsevol que construeix aplicacions de client i/o servidor basades en HTTP, com navegadors web, aranyes de webs, biblioteques de transport de servei web, o sistemes que es utilitzin el protocol HTTP per a la comunicació distribuïda.

Nucli de HttpComponents

HttpCore és un conjunt de components de transport de HTTP de baix nivell que es poden utilitzar per construir un client de consum o de serveis de HTTP de la part del servidor amb un impacte mínim. HttpCore sosté dos models d'I/O: blocatge model d'I/O basat en el Java I/O clàssic i no-blocatge, esdeveniment conduït model d'I/O basat en Java NIO.

El model d'I/O de blocatge pot ser més apropiat per a dades amb temps d'espera intensius, baixos, mentre que el model non-blocking pot ser més apropiat per a temps d'espera alts on el rendiment de dades és menys important que l'habilitat per manejar milers de connexions de HTTP simultànies d'una manera eficaç de cara als recursos disponibles.

[23] (Apache, HttpComponents, 2011)

Client de HttpComponents

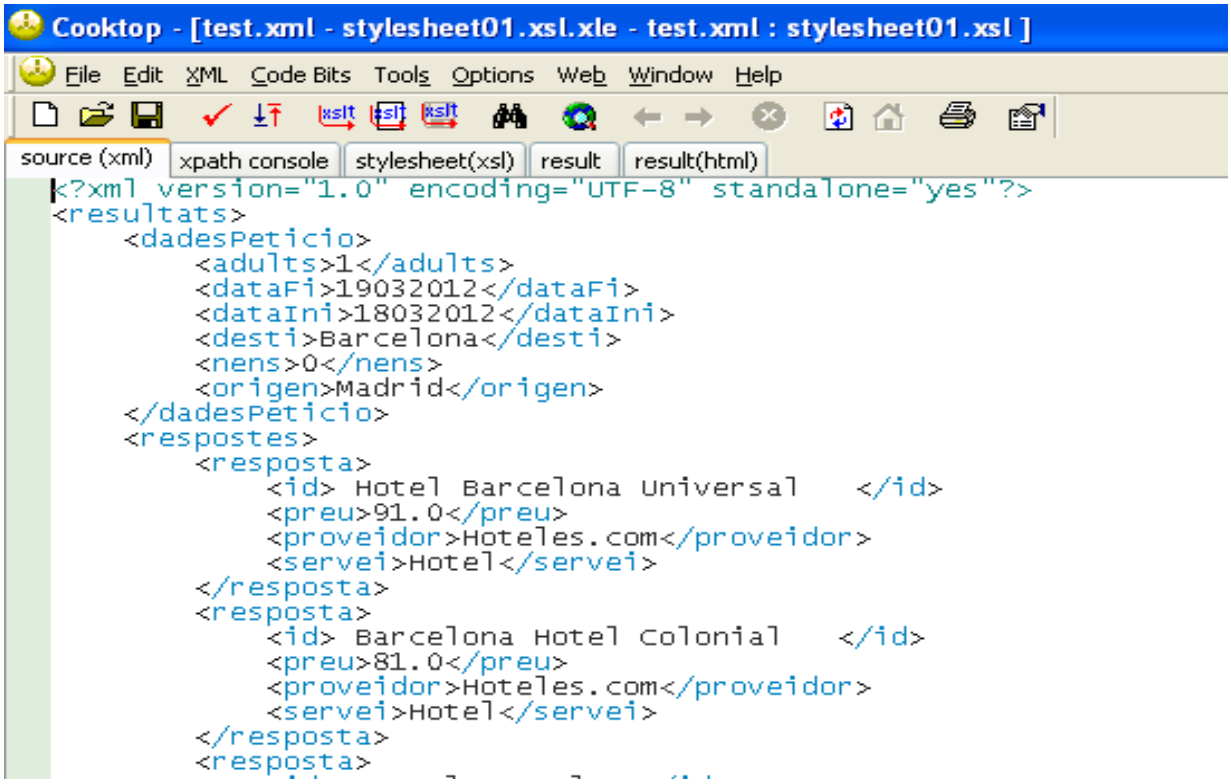
HttpClient és una aplicació agent de HTTP basada en HttpCore. També proporciona components reutilitzables per a autenticació de cara de client, gestió d'estat de HTTP, i gestió de connexió de HTTP. El Client de HttpComponents és un successor de substitució per a Commons HttpClient 3.x.

Principals característiques

- Normes basades en Java, aplicació de versions de HTTP 1.0 i 1.1
- Plena aplicació de tots els mètodes de HTTP en una marc d'OO extensible.
- Encriptació de suports amb HTTPS (HTTP sobre SSL) protocol.
- Connexions transparents a través de controladors sobre HTTP.
- Connexions de HTTPS
- Suport de gestió de connexió per a l'ús en aplicacions
- Suports per a màximes connexions per amfitrió.
- Detecta i tanca connexions caducades.
- Maneig de galetes(cookies) automàtic
- Connexions persistents que utilitzen KeepAlive.
- Accés directe al codi de resposta i encapçalaments enviats pel servidor.
- Facilitat per posar temps d'espera,timeouts, de connexió.
- Caching de resposta de HTTP/1.1.
- El codi font està lliurement disponible sota la Llicència d'Apatxe.

6.11. Cooktop

Es una eina que facilita el tractament de fitxers XML, així com les transformacions que es pugin fer utilitzant XSLT, permeten fer proves i veure els resultats de les transformacions.



Cooktop - [xmldoc05.xml - stylesheet01.xsl.xle - xmldoc05.xml : stylesheet01.xsl]

File Edit XML Code Bits Tools Options Web Window Help

source (xml) | xpath console | stylesheet(xsl) | result | result(html)

```
<xsl:stylesheet version="1.0"
    xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
    <xsl:output method="html" encoding="UTF-8"
        omit-xml-declaration="yes" indent="yes" />

    <xsl:template match="/">
        <html>
            <head>
                <title>VTC Seeker - Resultats de la cerca</title>
            </head>

            <body>
                <font>Dades de la petició</font>

                <xsl:for-each select="resultats/dadesPeticio">
                    <br>
                    </br>

                    Origen:
                    <xsl:value-of select="origen" />
```

Resultats de la transformació.

Cooktop - [test.xml - stylesheet01.xsl.xle - test.xml : stylesheet01.xsl]

File Edit XML Code Bits Tools Options Web Window Help

source (xml) | xpath console | stylesheet(xsl) | result | result(html)

Dades de la petició
Origen: Madrid
Destí: Barcelona
Data inici: 18032012
Data fi: 19032012
Adults: 1
Nens: 0
Resultats de la cerca (ordenats per preu)
1 - Hotel Millor preu: 54.0 € Hoteles.com Rialto
2 - Hotel Millor preu: 63.0 € Hoteles.com Hotel Espana
3 - Hotel Millor preu: 65.0 € Booking href="/hotel/es/royal.es.html
4 - Hotel Millor preu: 66.0 € Hoteles.com Astoria
5 - Hotel Millor preu: 67.0 € Hoteles.com Internacional Cool Local Hotel
1 - Vol Millor preu: 29.99 € Vueling code='-PVYMABC-'
2 - Vol Millor preu: 29.99 € Vueling code='-PVYMABC-'
3 - Vol Millor preu: 39.98 € Vueling code='-OVYMABC-'
4 - Vol Millor preu: 39.99 € Vueling code='-OVYBCMA-'
5 - Vol Millor preu: 39.99 € Vueling code='-OVYBCMA-'

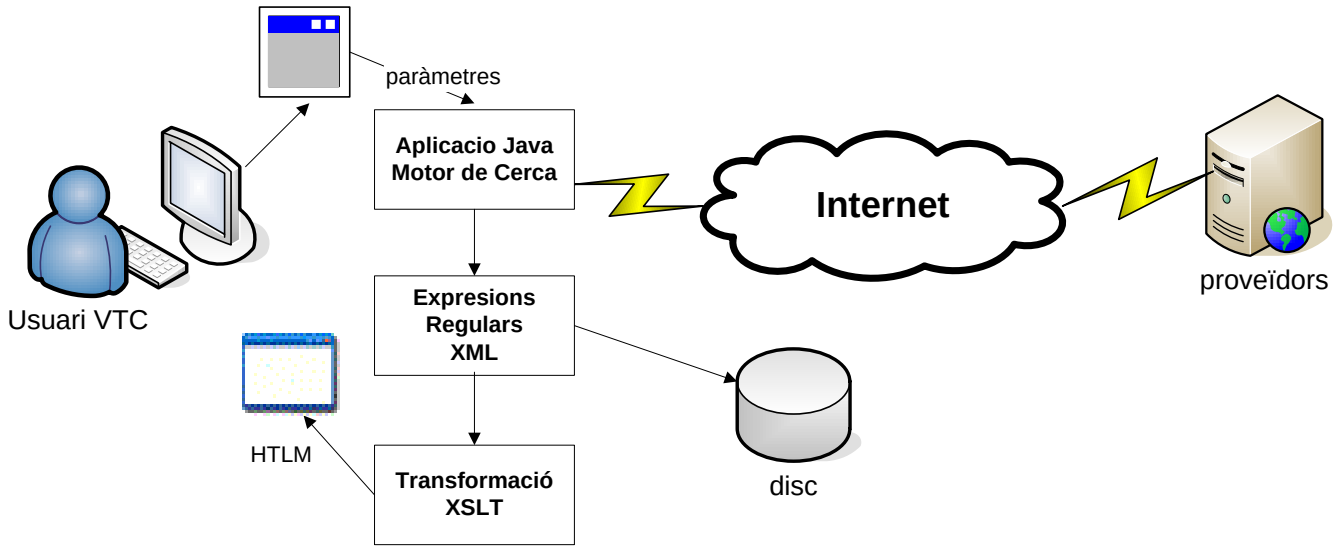
7. Disseny

La idea es implementar el producte com una aplicació d'escriptori. La interfície gràfica serà senzilla, ja que no es el principal objectiu del projecte, però alhora ha de permetre cobrir els requeriments definits en l'enunciat, que són els següents:

- Realitzar cerca
- Mostrar resultats resultats(format HTLM)
- Desar resultats cerca (format XML)
- Recuperar resultats cerca (format HTLM)

L'usuari tindrà una interfície gràfica que li permetrà introduir els paràmetres de la cerca, veure els resultats en format HTML, desar-los en format XML, així com la possibilitat de recuperar-los posteriorment.

Un cop obtenim la resposta del servidor cal analitzar-la per poder classificar-la i saber si ens interessa o no. El tractament d'aquestes dades retornades pel servidor consistirà un anàlisi sintàctic i semàntic d'aquesta estructura utilitzant expressions regulars, fet que ens permetrà la flexibilitat d'adaptació necessari per cobrir les diferents tipologies de resposta.



Casos d'us.

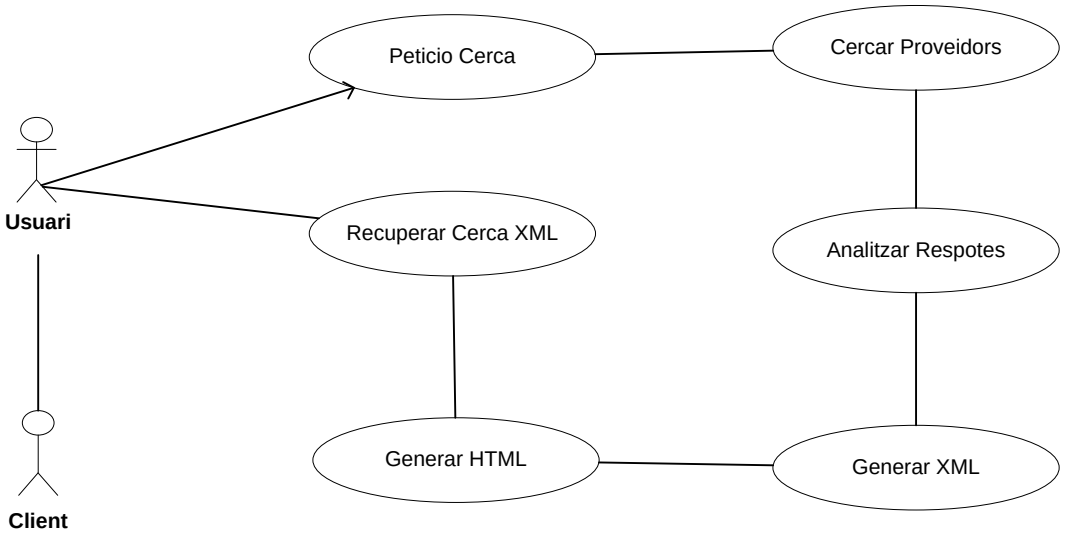
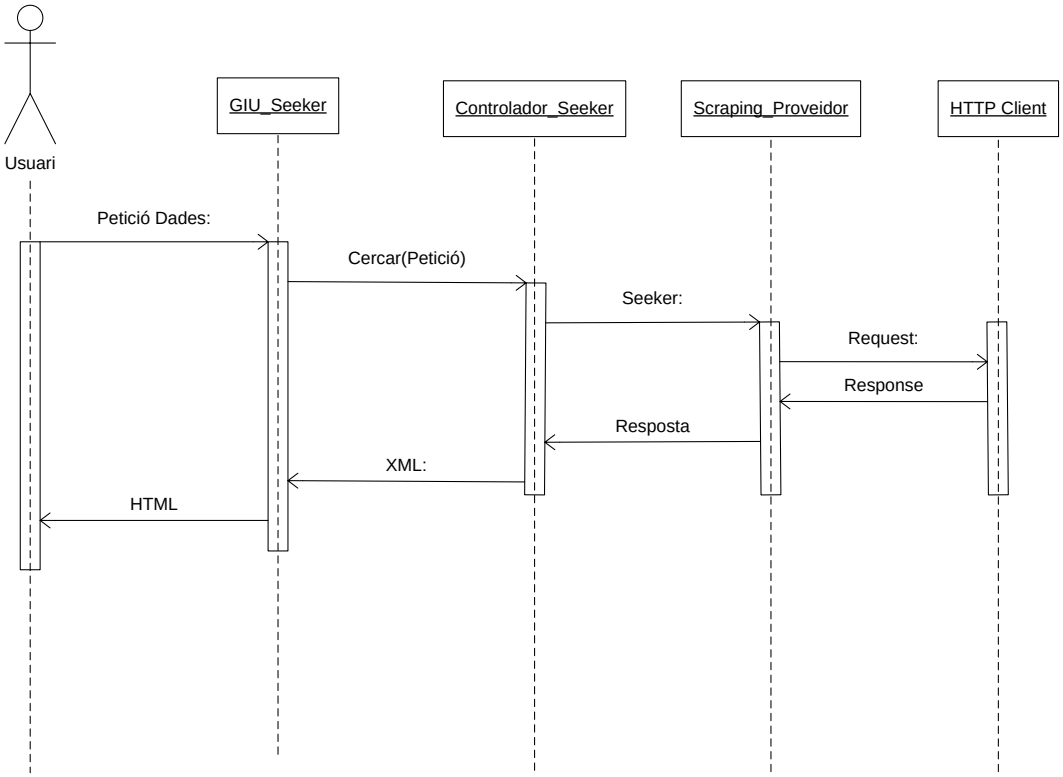
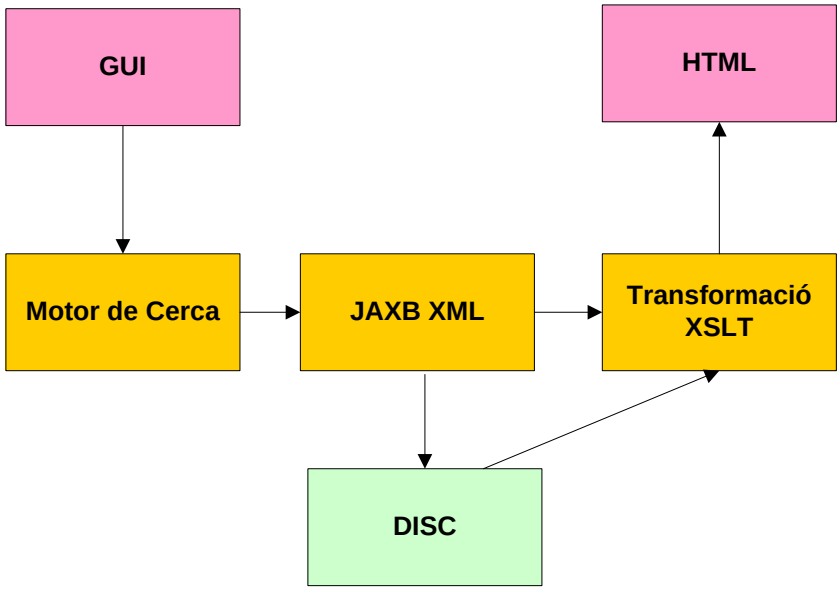


Diagrama de seqüència.



En aquesta implementació s'aplicarà un patró de disseny Model Vista Controlador (MVC) que garanteixi la independència entre les diferents capes. [24]

S'aplicarà un disseny en tres capes, presentació, lògica de negoci i persistència de dades.

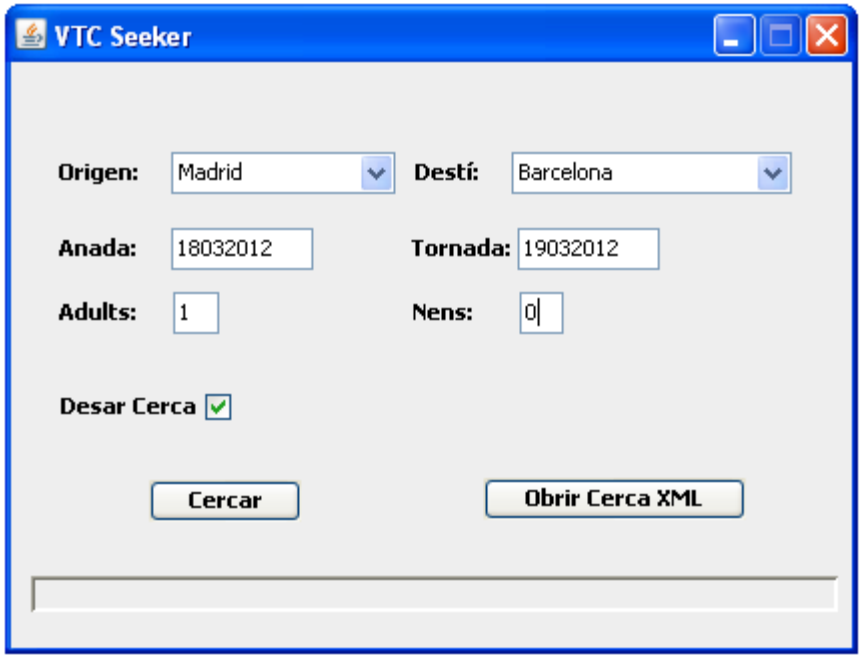


[24] (Larman, 2006)

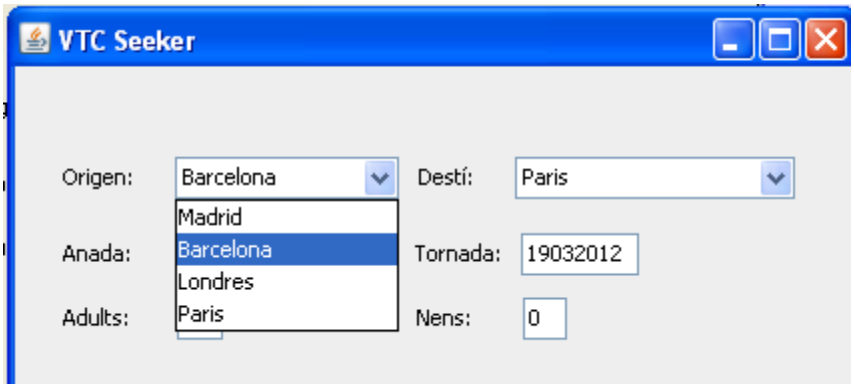
Interfície gràfica d'usuari (GUI)

Pantalla principal de l'aplicació que conté els elements gràfics que permetran a l'usuari interactuar amb el programa, de manera que podrà introduir els paràmetres de la cerca a realitzar, executar-la i veure els resultats.

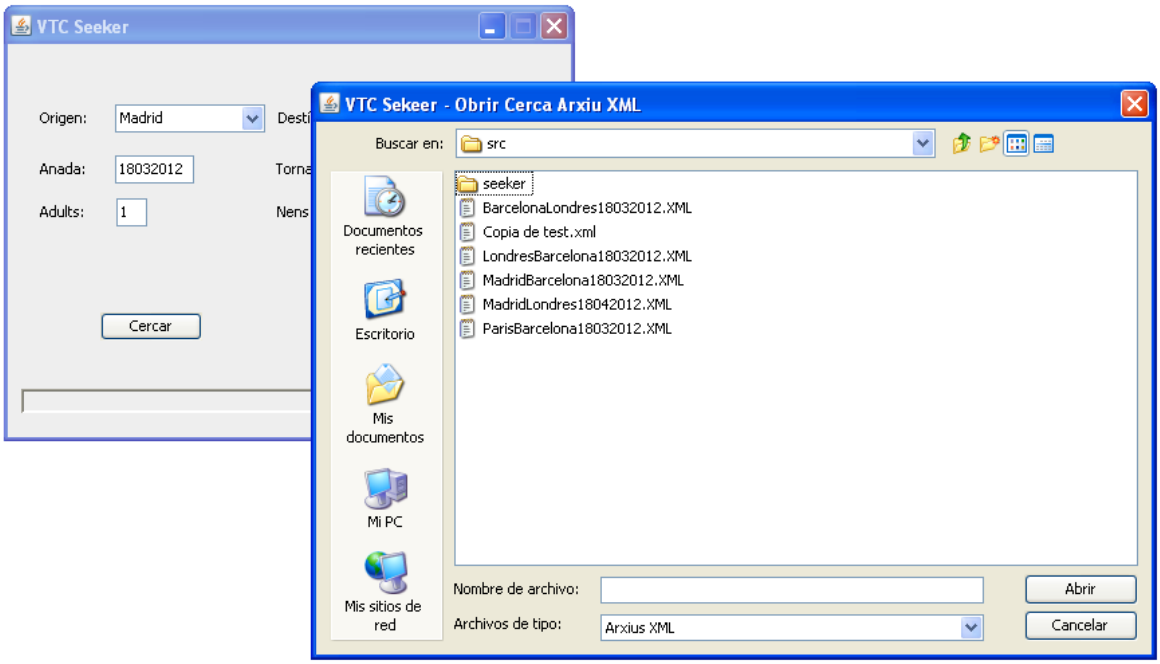
- Origen(llista ciutats configurada)
- Destí(llista ciutats configurada)
- Data inici(format data DDMMAAAA)
- Data fi(calendari,format data DDMMAAAA)
- Nombre d'ocupants adults
- Nombre d'ocupants nens
- Desar o no, el fitxer XML generat per a posterior consultes
- Botó(acció de cercar al web)
- Recuperar cerca desada en format XML(obre navegador html)



Pantalla de selecció dels paràmetres de la cerca. Les ciutats origen i destí es seleccionen d'una llista amb les ciutats configura a l'aplicació.



S'obre finestra per seleccionar el fitxer XML a mostrar en format HTML.



8. Comentaris de les parts principals del codi

En un fitxer de configuració (config_seeker.ini) tenim els noms dels proveïdors implementats en l'aplicació. Quan s'inicia l'execució el programa recupera el contingut d'aquest fitxer i invoca el nom de cada classe. En el cas d'afegir un nou operador, només caldria implementar la seva classe particular, perquè realitzi les peticions i gestioni les respostes, i afegir el nom d'aquesta nova classe en el fitxer de configuració.

8.1. Peticions

Cada proveïdor implementat presenta la seva manera particular de gestionar la recepció dels paràmetres de cerca, fet que ens obliga a desenvolupar un sistema individualitzat per generar les peticions de cada operador.

Un clar exemple d'aquesta construcció a mida de les peticions es la codificació que cada operador realitza de les poblacions dels hotels, dels aeroports origen i destí dels vols, el format de les dates d'entrada o sortida, els requisits d'accés o les restriccions que imposa cada sistema en particular.

En aquest exemple, veiem com a partir del paràmetres de la cerca, construïm la cadena de text que serà enviada com a petició a un servidor.

```
String parm_get = "http://vueling.com/booking/booking/selecciona-tu-vuelo?event=s
                + wOrigen
                + "&to1="
                + wDesti
                + "&from2="
                + wDesti
                + "&to2="
                + wOrigen
                + "&departDay1="

HttpGet httpget = new HttpGet(parm_get);
HttpResponse response = httpclient.execute(httpget);
```

8.2. Recol·lecció

Cada proveïdor implementat retorna les dades des del seu servidor d'una manera particular, fet que ens obliga a desenvolupar un sistema individualitzat per recepcionar les respostes de cada operador per poder analitzar, interpretar i recol·lectar les dades que realment ens interessen.

La resposta incorpora una gran quantitat d'informació i no tota ens resultarà útil, de fet, la gran part d'aquest flux de dades en format HTML la depreciarem i només ens quedarem amb un sèrie de dades clau, les quals són l'objectiu de la cerca.

Per accedir a aquesta informació haurem d'estudiar la manera en que cada operador retorna les dades, identificar uns determinats patrons que ens faciliten la seva localització i extraure-les amb l'ajuda de les expressions regulars.

En aquest exemple, veiem com a partir de la resposta d'un servidor, la convertim a text i la manipulem amb la classe d'utilitats, que conté les expressions regulars, per extraure les dades que ens interessen.

```
HttpEntity entity = response.getEntity();
if (entity != null) {
  line = EntityUtils.toString(entity);

  // recuperar dades que ens interessen ...
  v = Utilitats.recuperaTexts("price='-.*?-'", line, 0);

  // identificador Flight ----> code='-TVYBCAM-' price='-69.99-'
  id = Utilitats.recuperaTexts("code='-.*?-'", line, 0);
```

8.3. Conversió a XML

JAXB lliga signatures de mètode Java i missatges de WSDL, un format XML que s'utilitza per descriure serveis web, i li permet personalitzar l'aplicació mentre maneja automàticament la conversió en temps d'execució. Això fa fàcil que incorpori dades de l'XML i el processament funciona en aplicacions basades en tecnologia de Java sense haver de tenir grans coneixements de l'XML.

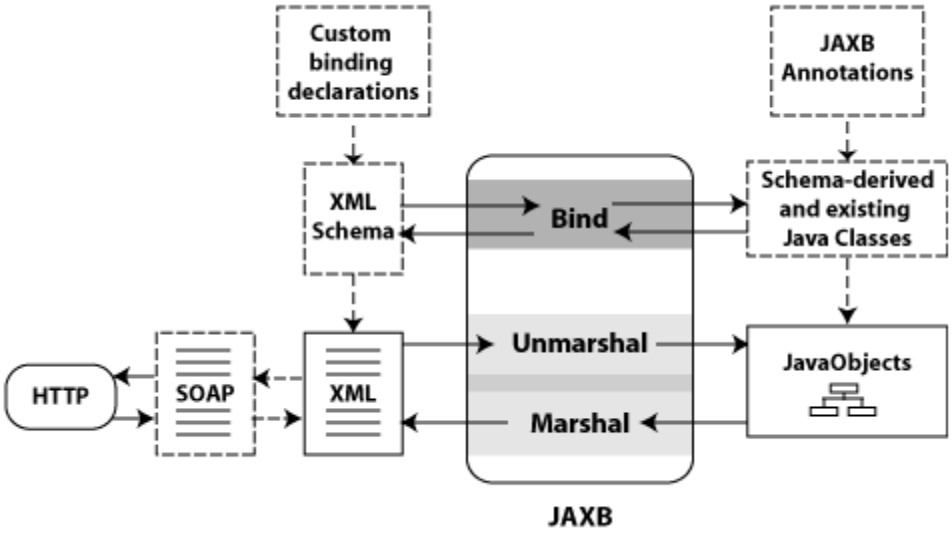
WSDL són les inicials de Web Services Description Language, un format XML que s'utilitza per descriure Serveis web, i que descriu la interfície pública dels serveis web. Està basat en XML i descriu la forma de comunicació, és a dir, els requeriments del protocol i els formats dels missatges necessaris per interactuar amb els serveis que es troben a la llista del seu catàleg.

WSDL acostuma a utilitzar-se en combinació amb SOAP i XML Schema. Un programa client que es connecta a un servei web pot llegir el seu descriptor WSDL per determinar quines funcions es troben disponibles al servidor. Els tipus de dades especials s'inclouen al fitxer WSDL en forma de XML Schema. El client pot utilitzar SOAP per fer la petició a una de les funcions de la llista del WSDL.

- Bind. Lliga esquema de l'XML a classes de Java de JAXB obtingudes d'esquema, o classes de valor. Cada classe proporciona accés al contingut mitjançant un conjunt|joc de mètodes d'accés d'estil de JavaBean

- Unmarshal. Converteix el document XML per crear un arbre d'elements de programa de Java o objectes.

- Marshal. Converteix els objectes de Java a contingut XML.[25]



[25] (Oracle, JAXB, 2008)

En aquest exemple, serialitzem la classe Resultats convertint aquest objecte i les dades que conté en el format XML, utilitzant les classes Factory i Transformer.

```
r.setContenidorPeticio(dadesPeticio);
r.setvOfertes(vResultats);

JAXBContext contextObj = JAXBContext
                .newInstance(seeker.Resultats.class);
Marshaller marshallerObj = contextObj.createMarshaller();
marshallerObj.setProperty(Marshaller.JAXB_FORMATTED_OUTPUT, true);
StringWriter oWriter = new StringWriter();
marshallerObj.marshal(r, oWriter);
String xml = oWriter.toString();
```

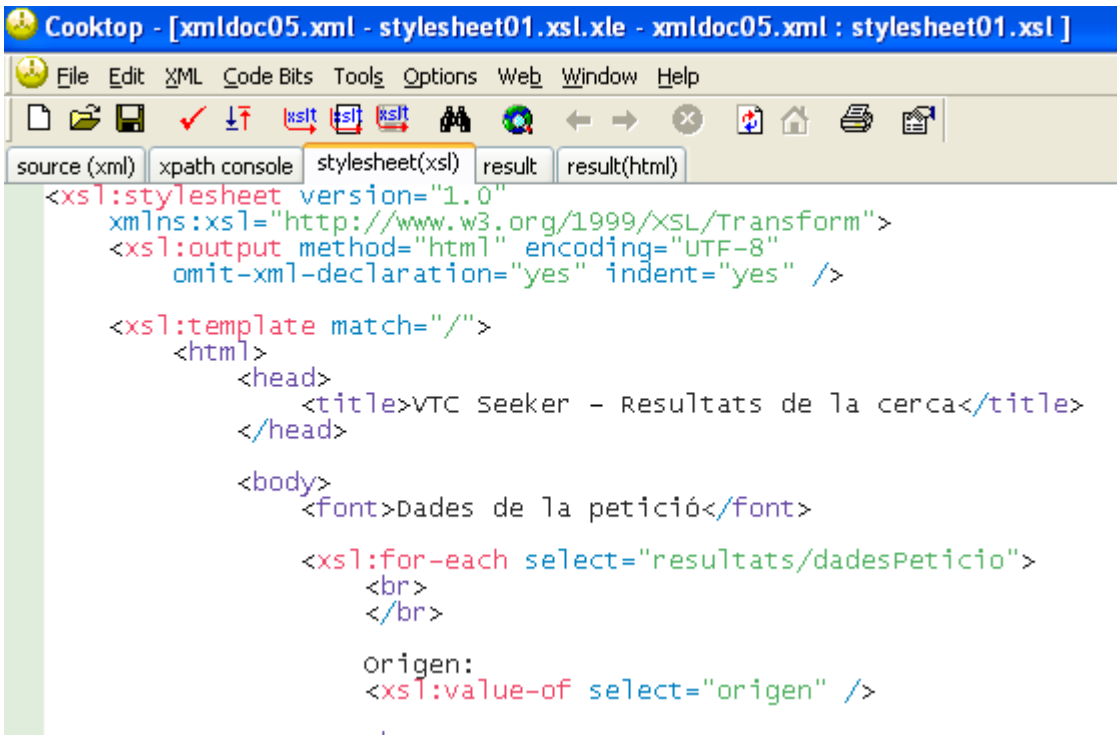
8.4. Conversió a HTML

L'API de Java per a processament de XML, o JAXP, proporciona la capacitat de validar i analitzar documents XML. A més a més a les interfícies d'anàlisi, l'API proporciona una interfície de XSLT per proporcionar transformacions de dades i estructurals en un document XML.

En aquest exemple, convertim el format XML a HTML aplicant una plantilla XSLT i utilitzant les classes Factory i Transformer.

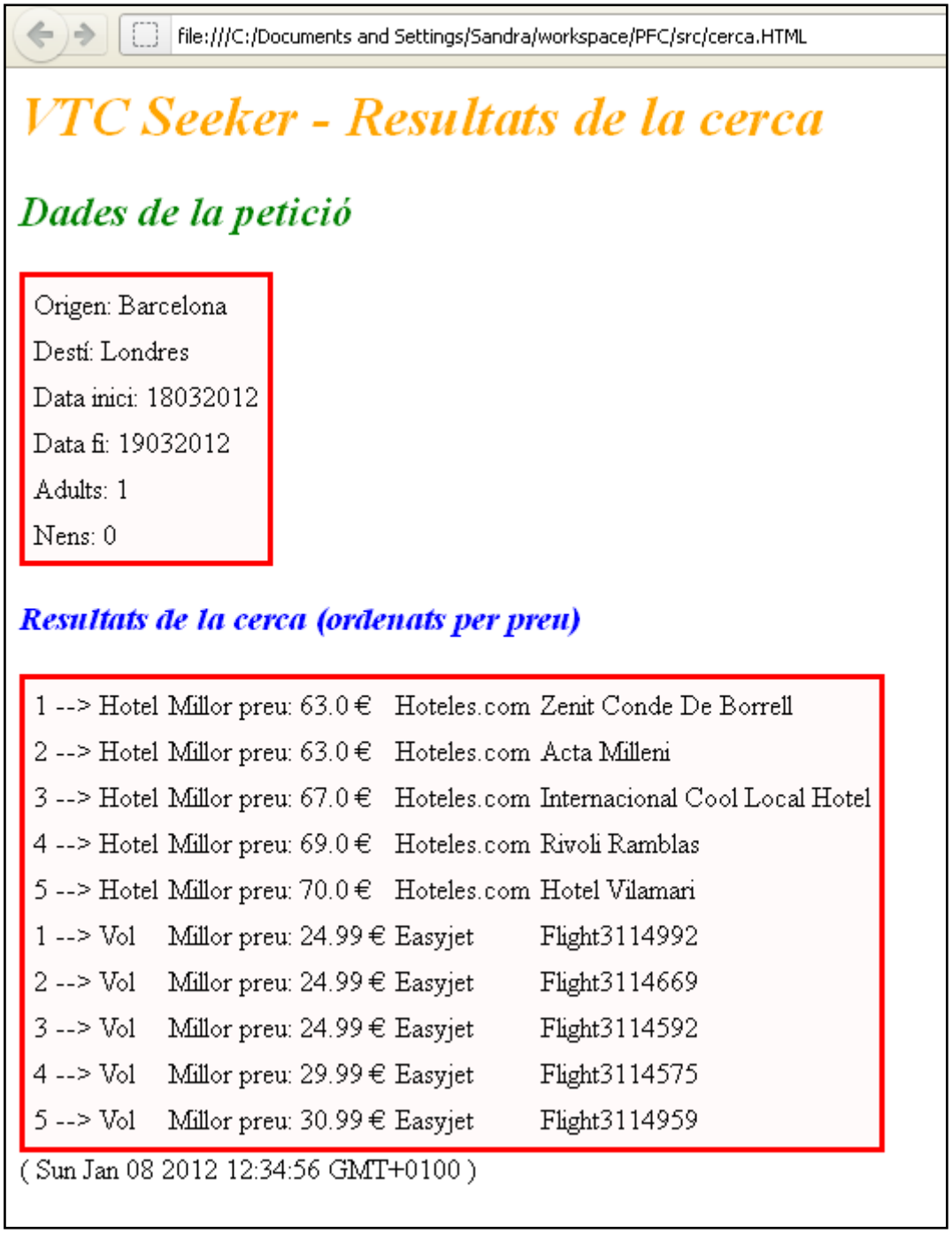
```
String path = "plantilla.XSLT";
InputStream inputStreamXSLT = Thread.currentThread()
                .getContextClassLoader().getResourceAsStream(path);
templatesXSLT = TransformerFactory.newInstance().newTemplates(
                new StreamSource(inputStreamXSLT));
Transformer transformer = templatesXSLT.newTransformer();
transformer.transform(streamSource, streamResult);
```

El contingut del fitxer “plantilla.XSLT” conté les instruccions XSL per construir el fitxer de sortida en format HTML.
i



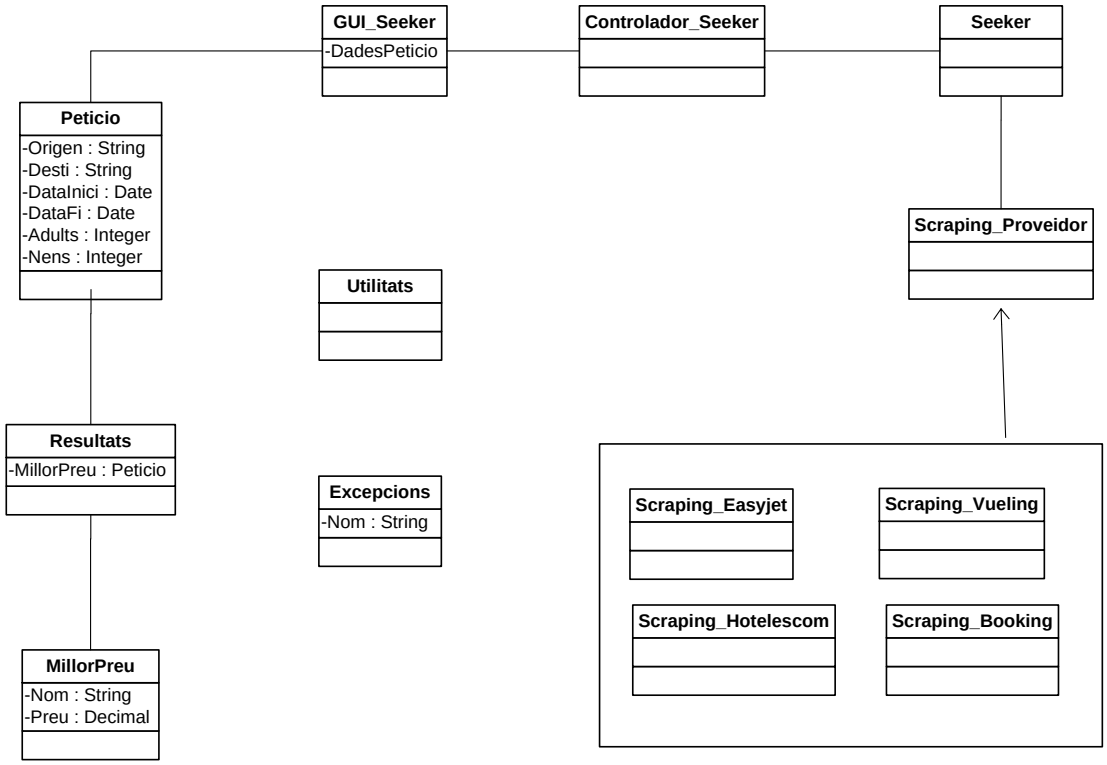
8.5. Presentació HTML

Per obri el fitxer HTML es fa servir l'objecte Desktop que és una API de JAVA.



8.6. Classes.

Diagrama de les principals classes.



Controlador_Seeker

Aquest classe s'encarrega d'interactuar entre la interfície gràfica d'usuari i la lògica de l'aplicació. Equivaldria al controlador en un patró MVC.

Seeker

Es la classe que conté la lògica de l'aplicació. Realitza les crides als cercadors implementats per a cada operador i un genera els fitxers de resultats en els formats XML i HTML.

GUI_Seeker

Aquest classe conté tots els objectes que composen la interfície gràfica d'usuari. Equivaldria a la vista en un patró MVC.

```
package seeker;
import java.awt.event.ActionEvent;
import java.awt.event.ActionListener;
import javax.swing.DebugGraphics;
import javax.swing.DefaultComboBoxModel;
import javax.swing.JButton;
import javax.swing.JComboBox;
import javax.swing.JLabel;
import javax.swing.JMenu;
import javax.swing.JMenuBar;
import javax.swing.JTextField;
import javax.swing.WindowConstants;
import javax.swing.SwingUtilities;
import java.util.Vector;
import javax.swing.ComboBoxModel;

/**
* This code was edited or generated using CloudGarden's Jigloo
* SWT/Swing GUI Builder, which is free for non-commercial use.
* A COMMERCIAL LICENSE HAS NOT BEEN PURCHASED FOR
* THIS MACHINE, SO JIGLOO OR THIS CODE CANNOT BE USED
* LEGALLY FOR ANY CORPORATE OR COMMERCIAL PURPOSE.
*/
public class GUI_Seeker extends javax.swing.JFrame {

        private JLabel labelOrigen;
        private JTextField nens;
        private JLabel jLabel2;
        private JLabel jLabel3;
        private JLabel jLabel4;
        private JLabel jLabel5;
        private JLabel jLabel6;
        private JComboBox jComboBoxOrigen;
        private JComboBox jComboBoxDesti;
        private JButton jButton2;
        private JButton jButton1;
        private JTextField dataFi;
        private JTextField dataIni;
        private JTextField adults;
        private JLabel jLabel1;

        public static void main(String[] args) {
                SwingUtilities.invokeLater(new Runnable() {
                        public void run() {
                                GUI_Seeker inst = new GUI_Seeker();
```

Scraping_Proveïdor
Aquesta classe abstracta que conté la lògica de la petició i extracció de les dades d'un proveïdor.

Scraping _Vueling
Aquesta classe que conté la lògica de la petició i extracció de les dades d'un proveïdor.

Scraping _Easyjet
Aquesta classe que conté la lògica de la petició i extracció de les dades d'un proveïdor.

Scraping _Booking
Aquesta classe que conté la lògica de la petició i extracció de les dades d'un proveïdor.

Scraping _Hotelescom
Aquesta classe que conté la lògica de la petició i extracció de les dades d'un proveïdor.

Resultats
Aquesta classe conté l'estructura del objecte amb les dades que volem serialitzar, transformar a XML.

Utilitats.java
Aquesta classe conté les funcionalitats auxiliars de l'aplicació, com son els mètodes de cerca i extracció en cadenes de text, accessos a disc, definició de constants o conversions, entre d'altres.

Fitxer auxiliars i de configuració

config_seeker.ini
Fitxer de configuració de l'aplicació, a on es parametritzen els noms de les classes responsables de la cerca de cada proveïdor implementat.

plantilla.XSLT
Es el fitxer que conté la transformació XSL per convertir el XML a format HTML.

9. Desenvolupament

Pel que fa al desenvolupament del codi, al control i a la gestió de les versions, s'ha emprat un sistema de versionat incremental que incorpora les parts implementades un cop ja provades i consolidades en el programari de l'aplicació base.

Qualsevol millora o modificació es fa sobre una nova versió que es prova independentment de la resta i, un cop superades, es realitza un test integral sobre les principals funcionalitats de l'aplicació per confirmar que la versió es estable i que es pot consolidar el codi afegit.

L'aplicació final desenvolupada, així com la resta de lliurables, s'ha elaborat mitjançant un procés de refinament iteratiu progressiu per objectius.

En una fase inicial, es va construir un prototip del motor de cerca amb l'objectiu de provar i estudiar com funcionaven les llibreries a utilitzar i de quina es la millor manera de realitzar les connexions, enviar les peticions i gestionar les respostes dels servidors.

Després es van fer proves amb diferents proveïdors de serveis de vols i hotels per analitzar com enviaven les dades i de quina es podien recuperar les que més ens interessaven.

Tot seguit l'estratègia va ser aconseguir un correcte funcionament per a un parell d'operadors de vols i d'hotels, incorporant les entrades i sortides de dades requerides. A partir d'aquí es va poder desenvolupar la resta del projecte i generar un prototip amb un percentatge alt d'objectius complerts.

La part final del desenvolupament, ha consistit en estabilitzar, fer proves, millorar la versió aconseguida, afegir més proveïdors, funcionalitats i completant els requeriments de l'enunciat del projecte.

9.1. Proveïdors implementats

Cada proveïdor implementat presenta la seva manera particular de gestionar la recepció dels paràmetres de cerca, fet que ens obliga a desenvolupar un sistema individualitzat per generar les peticions de cada operador.

Un clar exemple d'aquesta construcció a mida de les peticions es la codificació que cada operador realitza de les poblacions dels hotels, dels aeroports origen i destí dels vols, el format de les dates d'entrada o sortida, els requisits d'accés o les restriccions que imposa cada sistema en particular.

Malgrat tot, les peticions tenen un seguit de característiques que son comuns en tots els operadors.

Els proveïdors dels serveis d'hotels acostumen a presentar les següents característiques:

- població destinació
- data d'entrada
- data sortida
- nombre ocupants adults
- nombre ocupants menors

Els proveïdors dels serveis de vols acostumen a presentar les següents característiques:

- aeroport origen
- aeroport destí
- data d'anada
- data tornada
- nombre ocupants adults
- nombre ocupants menors

Cada proveïdor implementat retorna les dades des del seu servidor d'una manera particular, fet que ens obliga a desenvolupar un sistema individualitzat per recepcionar les respostes de cada operador per poder analitzar, interpretar i recol·lectar les dades que realment ens interessen.

La resposta incorpora una gran quantitat d'informació i no tota ens resultarà útil, de fet, la gran part d'aquest flux de dades en format HTML la depreciarem i només ens quedarem amb un sèrie de dades clau, les quals són l'objectiu de la cerca.

En el cas dels proveïdors dels serveis d'hotels acostumen a presentar les següents característiques:

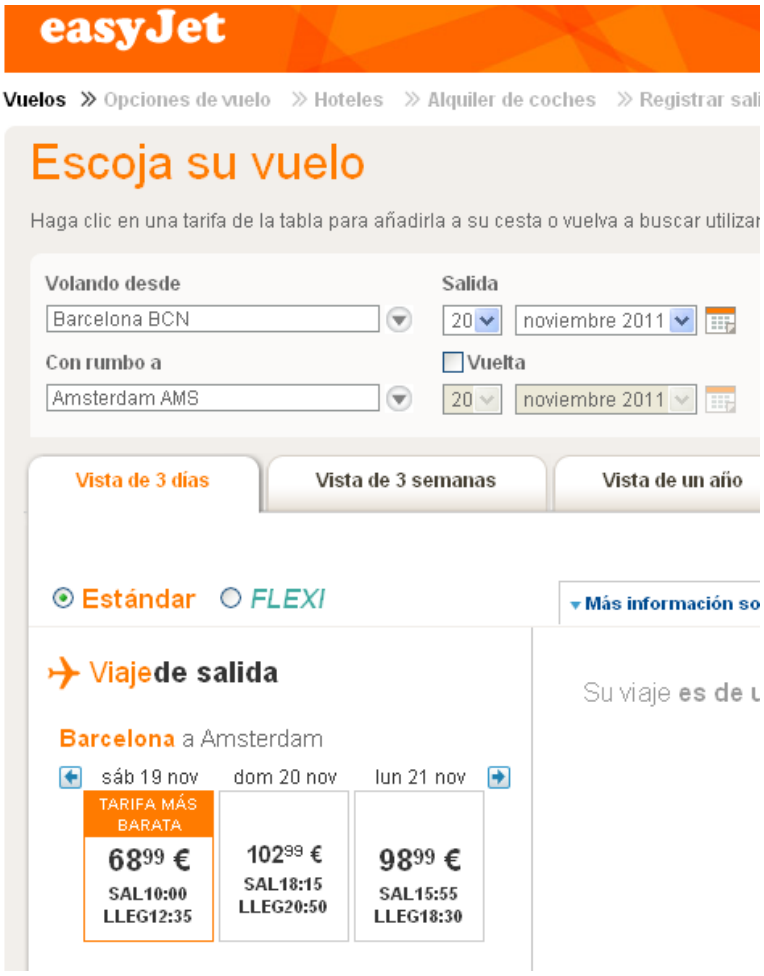
- identificador hotel
- preu

En el cas dels proveïdors dels serveis de vols acostumen a presentar les següents característiques:

- aeroport origen
- aeroport destí
- identificador vol
- preu
- data d'anada
- data tornada

Per accedir a aquesta informació haurem d'estudiar la manera en que cada operador retorna les dades, identificar uns determinats patrons que ens faciliten la seva localització i extraure-les amb l'ajuda de les expressions regulars.

9.1.1. Easyjet



En la petició es passen els següents paràmetres de cerca:

- aeroport origen
- aeroport destí
- data d'anada
- data tornada
- nombre ocupants adults
- nombre ocupants menors

```
try {
  String parm_get = "http://www.easyjet.com/es/Booking.mvc/SearchForFl
      + wOrigen
      + "&destAirportCode="
      + wDesti
      + "&departureDay="
      + wDiaIni
      + "&departureMonthYear="
      + wMesAnyIni
      + "&returnDay=00&returnMonthYear=00&numberOfAdults=" + WAdults
  HttpGet httpget = new HttpGet(parm_get);
```

Quan es retorna la resposta recuperem dades com

- identificador vol
- preu
- data d'anada
- data tornada

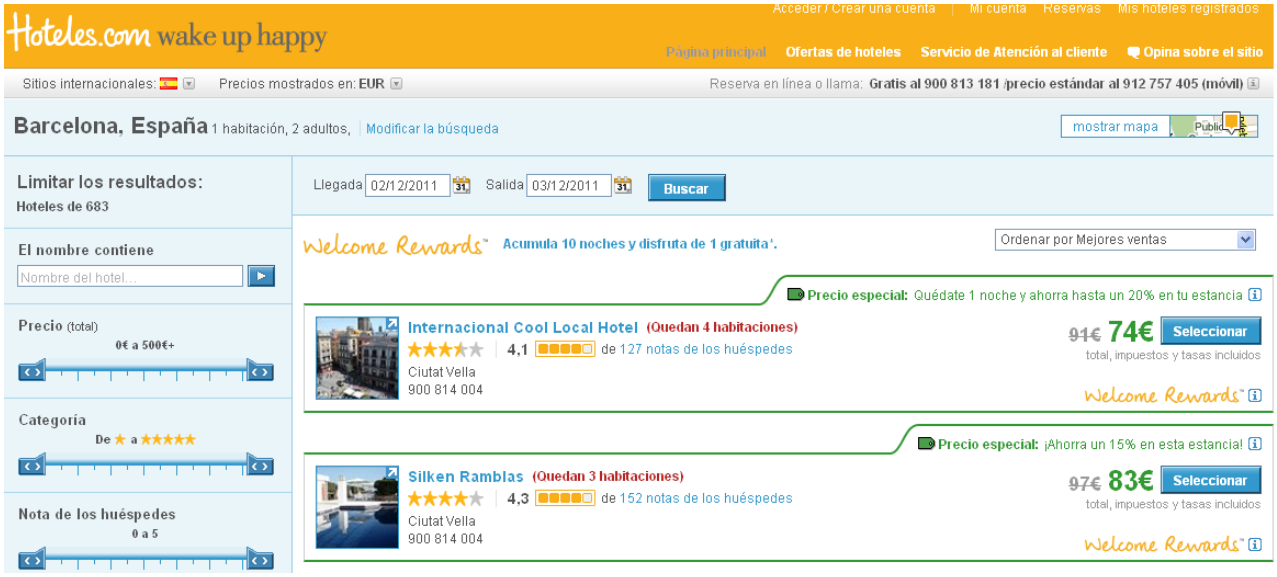
```
HttpEntity entity = response.getEntity();

if (entity != null) {
  line = EntityUtils.toString(entity);

  // class="priceSmaller"
  v = Utilitats.recuperaTexts("priceSmaller.*?€", line, 0);

  // <a id='Flight
  id = Utilitats.recuperaTexts("<a id='Flight.*?'", line, 0);
```

9.1.2. Hoteles.com



En la petició es passen els següents paràmetres de cerca:

- població destinació
- data d'entrada
- data sortida
- nombre ocupants adults
- nombre ocupants menors

```
String parm_get = "http://www.hoteles.com/search.do?destinatio
+ "%2C+Espa%C3%B1a&searchParams.arrivalDate=" + wDiaIni + "%2F"
 + wAnyIni + "&searchParams.departureDate=" + wDiaFi + "%2F" +
HttpGet httpget = new HttpGet(parm_get);
HttpResponse response = httpclient.execute(httpget);
```

Quan es retorna la resposta recuperem dades com

- identificador vol
- preu
- data d'anada
- data tornada

```
HttpEntity entity = response.getEntity();
if (entity != null) {
    InputStream instream = entity.getContent();
    line =EntityUtils.toString(entity);

    //preu
    v = Utilitats.recuperaTexts( "<ins>.*?€</ins>", line, 0);
    //id
    id = Utilitats.recuperaTexts( "<div class=\"photo_thumbnail\".*?</div>", line, 0);
```

9.1.3. Vueling

En la petició es passen els següents paràmetres de cerca:

- aeroport origen
- aeroport destí
- data d'anada
- data tornada
- nombre ocupants adults
- nombre ocupants menors

```
String parm_get = "http://vueling.com/booking/booking/selecciona-tu-vuelo?event=s
                + wOrigen
                + "&to1="
                + wDesti
                + "&from2="
                + wDesti
                + "&to2="
                + wOrigen
                + "&departDay1="

HttpGet httpget = new HttpGet(parm_get);
HttpResponse response = httpclient.execute(httpget);
```

Quan es retorna la resposta recuperem dades com

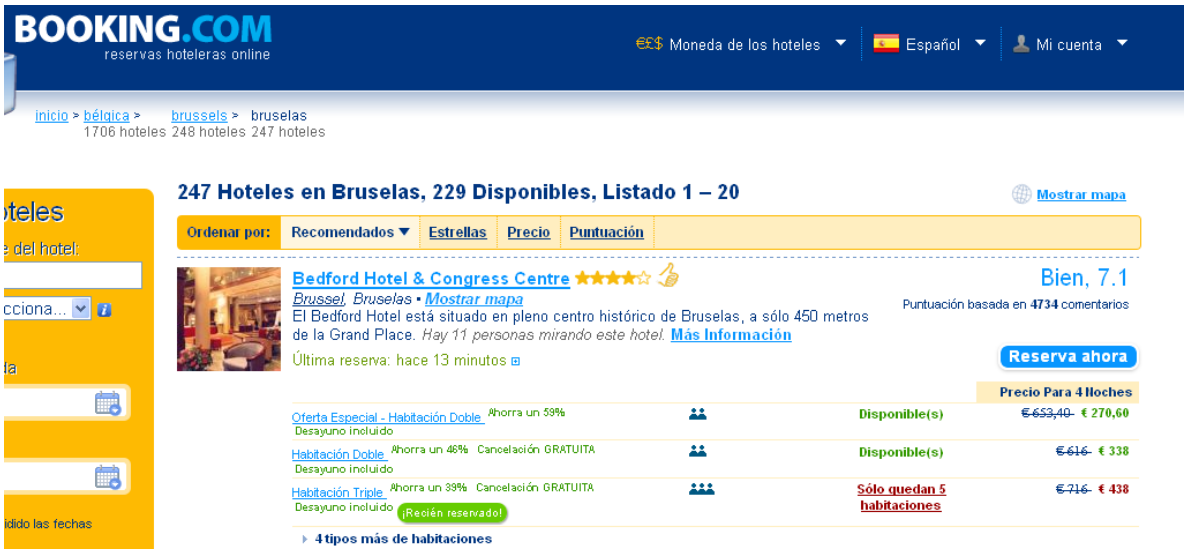
- identificador vol
- preu
- data d'anada
- data tornada

```
HttpEntity entity = response.getEntity();
if (entity != null) {
  line = EntityUtils.toString(entity);

  // recuperar dades que ens interessen ...
  v = Utilitats.recuperaTexts("price='-.*?-'", line, 0);

  // identificador Flight ----> code='-TVYBCAM-' price='-69.99-'
  id = Utilitats.recuperaTexts("code='-.*?-'", line, 0);
```

9.1.4. Booking.com



En la petició es passen els següents paràmetres de cerca:

- població destinació
- data d'entrada
- data sortida
- nombre ocupants adults
- nombre ocupants menors

```
try {
  String parm_get = "http://www.booking.com/searchresults.es.html
      + wDiaIni
      + ";checkin_year_month="
      + wMesAnyIni
      + ";checkout_monthday="
      + wDiaFi
      + ";checkout_year_month="
      + wMesAnyFi
      + ";class_interval=1;dest_id=-"
      + wDesti
      + ";dest_type=city;group_adults="
      + wAdults
      + ";group_children=" + wNens;
  HttpGet httpget = new HttpGet(parm_get);
  HttpResponse response = httpclient.execute(httpget);
```

Quan es retorna la resposta recuperem dades com

- identificador vol
- preu
- data d'anada
- data tornada

```
if (entity != null) {
  line = EntityUtils.toString(entity);

  // recuperar dades que ens interessen ... preu id ...
  v = Utilitats.recuperaTexts(
      "title=\"El precio de tu habitaci.*?. Est", line, 0);

  // href="/hotel/
  id = Utilitats.recuperaTexts("<td class=\"roomName\">.*?.html",line, 0);
```

9.2. Ciutats i aeroports configurats

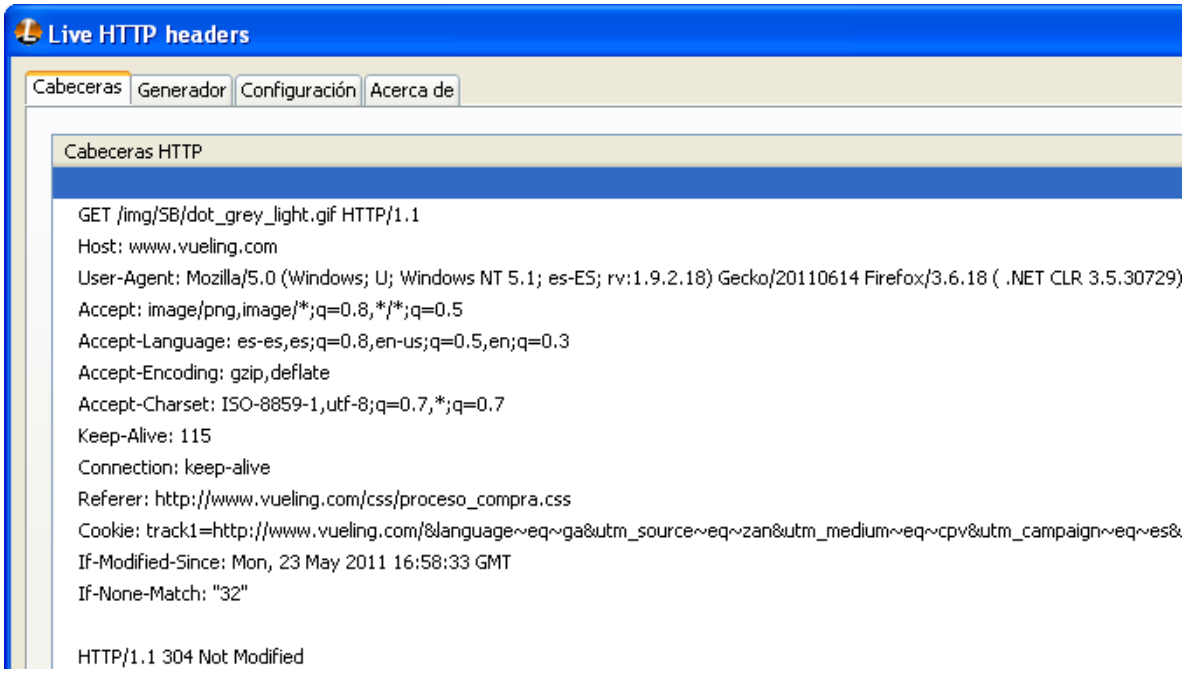
Relació de ciutats i aeroports configurats en l'aplicació.

Aeroport	Ciutat	País
ALC	Alicante	España
LEI	Almería	España
OVD	Oviedo	España
BCN	Barcelona	España
BIO	Bilbao	España
FUE	Fuerteventura	España
IBZ	Ibiza	España
ACE	Lanzarote	España
MAD	Madrid	España
AGP	Málaga	España
PMI	Mallorca	España
MAH	Menorca	España
MJV	Murcia	España
SCQ	Santiago	España
SVQ	Sevilla	España
TFS	Tenerife	España
VLC	Valencia	España
CDG	Paris	França
LGW	Londres	Regne Unit

9.3. Exemple de petició automatitzada

El nostre programa simula que es connecta a una determinada pàgina web i realitza una petició com si es tractés d'un usuari humà. En aquest cas serà la de la companyia Vueling a la que enviarem una petició per un vol a Amsterdam.

Una eina important per conèixer el funcionament intern de cada lloc web i la manera de com es poden realitzar peticions amb els paràmetres de cada cerca es el programa Live HTTP Headers, amb el qual podem analitzar les capçaleres que s'envien i les adreces que es fan servir per respondre a les peticions dels clients.



En l'exemple que ens ocupa, enviarem una petició de la següent manera:

"GET /booking/services/cache-loader/get-no-flights-days?from1=BCN&to1=AMS&months=10&departDate1=20111107&format=~"

Aquí ja podem veure que les ciutats d'origen i destí utilitzen una codificació de tres lletres (BCN= Barcelona) i que la data es passa en format AAAAMMDD.

Quan el servidor rep la petició del client, realitza una sèrie de processos interns orientats a donar-li resposta i li retorna un flux de dades en format HTML que es representarà a la pantalla del navegador. Ens el cas del nostre programa el que es interessa es capturar aquesta resposta no per mostrar-la en el navegador sinó per extraure les dades que hi conté i que son l'objecte de la cerca: el preu, les dates, el nom de la companyia, etc.

Per aconseguir accedir a les dades que necessitem, ens cal analitzar la forma en que es presenten els resultats aquesta pàgina web. Com es lògic pensar, cada empresa tindrà un disseny i disposició diferent, per tant, haurem de realitzar tants processos especifiquis de captura de dades com proveïdors tinguem en el nostre aplicatiu.

En aquest exemple, si el que volem es el preu, l'haurem de buscar dintre d'una taula amb la capçalera "Precio". Això ens obliga a conèixer com esta feta per dins la plana web, es a dir, quin es el codi HTML que conté el preu.

El mateix procediment s'aplicaria en un cas d'hotels. La mecànica general seria la mateixa però canviarien els mètodes d'accés particulars que estarien adaptats a la sintaxi, forma i disseny de la web a "raspar".

Un cop tinguem els resultats dels diferents operadors, els ordenarem per preu i seleccionarem les millors cinc combinacions. A partir d'aquí, l'aplicació haurà de mostrar els resultats en format HTLM en el navegador i generar un arxiu XML per desar en el disc. Aquesta part s'implementarà amb transformacions XSL i les API de Java per generar fitxers XML (JAXP/JAXB).

10. Conclusions

Pel que fa a l'estudi de les tècniques i eines de web scraping, he trobat molt interessant la tasca de recerca i exploració de les tecnologies a emprar, com també la quantitat de possibilitats i potència que ofereixen a l'hora de dissenyar utilitats de cerca dins del mon web, tant en aplicacions pràctiques del dia a dia a nivell particular, com en aplicacions de caire més empresarials o comercials, més enllà de l'abast d'aquest projecte.

En referència als reptes i complexitats que ha representat el projecte, es poden agrupar en dos tipologies. D'una banda, l'obligació d'haver de controlar el grau de desenvolupament de les tasques per ajustar-les a la planificació i prendre tot un seguit de decisions que normalment ja ens venen donades en d'altres assignatures d'aquest estudis. D'altra banda, la desconeixença en l'aplicació tècnica de les eines necessàries, el procés d'aprenentatge, les proves, els errors, la depuració, la pèrdua de temps inherent i l'haver de cercar solucions als problemes que s'han anant presentant, han estat les principals dificultats

En aquest sentit, el fet de que l'enunciat actuï com un marc contenidor dels requisits bàsics, ens dona una llibertat quasi total en l'elecció de les eines i la manera d'implementar la solució, i esdevé en un autèntic repte, ja que deixa a les nostres mans una gran quantitat de, petites i grans, decisions de les quals en dependrà en bona part l'èxit final del projecte.

Ha resultat força enriquidor poder aplicar coneixements, conceptes i tècniques multidisciplinars adquirides al llarg dels estudis, de caire i procedència diversa. Des de temes de disseny formal a tècniques de programació concretes o aspectes de cultura informàtica bàsica, passant per la correcta redacció dels texts o l'aplicació de metodologies de direcció de projectes informàtics.

També trobo interessant comentar la situació de percepció de la manca de temps, encara que es tracti d'un projecte amb setmanes de planificació, acostuma a passar que al tram final s'ajunten les urgències, de tota mena, i sembla que tot estigui pendent de fer i, es llavors, que aquesta sensació de falta de temps material s'aguditza.

Pel que fa al seguiment, planificació i direcció de l'execució del projecte, cal comentar que des de les primeres fases podem observar que les previsions inicials de l'esforç necessari per a la construcció del motor de cerca varen ser massa optimistes, fet que ens demostra la complexitat que comporta aquesta activitat de previsió, especialment, en la fase inicial d'un projecte.

Per compensar aquest risc, en el pla de treball del projecte ja es va incloure a la planificació un marge de maniobra destinat a cobrir imprevistos, fet que ha esmorteït les desviacions produïdes en la consecució d'aquestes tasques, de manera que s'ha evitat que poguessin incidir en el normal desenvolupament de la resta del projecte.

Finalment, m'agradaria destacar com a positiva l'experiència que ha significat tot el procés de realització d'aquest projecte final de carrera en les seves diferents fases: desenvolupant activitats de direcció, gestió i control, implementant el codi, documentant la memòria, fent la presentació o, simplement, cercant informació de quina manera es podia fer cada tasca.

Totes elles molt diferents però alhora necessàries i complementaries, ens ajuden a tenir una visió més completa i real de com funciona un projecte de desenvolupament i de quines son les seves principals dificultats.

11. Línies futures de treball

Durant la implementació d'un projecte apareixen problemes o dubtes que ens obliguen a analitzar i valorar les diferents opcions disponibles i a prendre decissions sobre quina serà la via escollida, descartant d'altres. Es en aquest punt, quan som més conscients i veiem clarament les possibilitats del producte i podem extraure valuoses idees a efectes de millorar-lo.

Ens aquest moments penso que és interessant apuntar algunes futures línies de treball que permetrien afegir millores en el producte desenvolupat i ampliar les seves funcionalitats.

- Millorar interficie gràfica.
- Possibilitat de migració a entorn de navegació web (J2EE).
- Ampliar cerques a més proveidors i/o nous serveis. Per exemple, cotxes de lloguer, bitllets de tren, reserves restaurants, teatres, etc.
- Implementar mòdul de suscripció d'enviament d'ofertes a clients via mail o sms.
- Permetre accedir a la web escollida per formalitzar la reserva.
- Possibilitat de habilitar o no la cerca de només vols o només hotels.
- Robot cercador exhaustiu, sense una llista predefinida, que fos capaç de interactuar amb qualsevol servidor que oferís serveis web per extraure dades de manera automatitzada, utilitzant SOAP i WDSL.
- Parametrització de l'aplicació a nivell d'usuari
- Poder programar i generar cerques en mode desatès (batch) que generin fitxers de resultats.

12. Glossari

Analitzador(parser). Part de programari que analitza les dades retornades per un servidor, en aquest cas.

Back-up (còpia de seguretat). Acció de copiar documents, arxius o fitxers per tal que es puguin recuperar en cas de fallida del sistema o pèrdua de les dades.

Batch/on line. Diferents sistemes d'executar una aplicació. La principal diferència és que en Batch, l'aplicació no interactua amb l'usuari, mentre que en el mode On-line si que es possible que hi hagi interacció.

Caché. Conjunt de dades duplicades d'altres dades originals que presenten l'avantatge d'oferir un accés més ràpid que si es fes a les dades originals. En un entorn web, durant els primers accessos es carreguen una sèrie de dades a memòria que ja no es tornaran a demanar al servidor, agilitzant les càrregues de les planes.

Client/servidor. Arquitectura tècnica que permet que una aplicació s'executi en diferents llocs realitzant peticions a un servidor.

Contingència. Qualsevol succés que es presenti de forma imprevista i que afecti a les activitats habituals de desenvolupament o funcionament d'un sistema.

Cookie *(*galeta*).* És un fitxer de text que grava el servidor a través del navegador amb informació que l'usuari ha fet servir de les seves pàgines.

Criptografia. És la ciència de xifrar i desxifrar informació utilitzant tècniques matemàtiques que facin possible l'intercanvi de missatges de manera que només puguin ser llegits per les persones a qui van dirigits.

DNS (Domain Name System). Sistema de Noms de Domini, es una base de dades distribuïda que gestiona la conversió de direccions d'Internet expressades en llenguatge natural a una adreça numèrica IP.

Escalabilitat. Propietat d'un sistema que indica la seva capacitat per a gestionar el creixement continu de treball de manera fluida i per a estar preparat per a fer-se més gran sense perdre qualitat en els serveis oferts.

Firewall(tallafocs). Elements de seguretat perimetral i interna de xarxa que protegeixen els serveis de xarxa entre els diferents segments que la composen. La seva finalitat és la de restringir l'ús dels sistemes a les adreces IP que estan autoritzades a fer-ho i impedir-ho a la resta.

FTP (Protocol de transferència de fitxers). Protocol utilitzat per proporcionar transferències de fitxers a través d'una extensa varietat de sistemes.

GIF. Format d'intercanvi de gràfics, és un format estàndard per als fitxers d'imatges, que utilitza un mètode de compressió per fer que els fitxers siguin més petits.

HTTP (HyperText Transfer Protocol). Protocol de Transferència de Hypertext, es el protocol utilitzat per transferir fitxers d'hipertext per Internet.

HTTPS (Secure HTTP). Protocol HTTP millorat amb funcions de seguretat amb clau simètrica.

Incidència. És aquell esdeveniment que interromp en més o menys mesura el funcionament dels elements tecnològics.

Java. Llenguatge de programació estàndard i independent de la plataforma on s'ha de fer servir.

Llenguatge (de programació). Conjunt d'instruccions i regles amb el que s'interacciona amb l'ordinador. Permeten escriure aplicacions fent servir expressions mes similars al llenguatge parlat que el sistema 'binari' propi dels processadors.

LOG. Registre, en anglès, que els programes i sistemes creen en fitxers, i en els quals van anotant els passos, missatges o errors que es produeixin durant l'execució d'un procés determinat.

Maquinari(hardware). Conjunt de elements materials que composen un ordinador, que inclou els dispositius electrònics, circuits, cables, targetes, perifèrics i altres elements físics.

Metadates. Descripció i/o definició de les dades.

Organigrama. Esquema d'una organització, d'una tasca o procés.

Pàgina HTML. Qualsevol document d'informació, normalment en format HTML, que és accessible a través de la web per un navegador.

PDF (Portable Document Format). Significa format de document portàtil i és un tipus de format de fitxer que va ser creat per a aquells documents que havien de ser impresos, ja que porta tota la informació necessària per a la presentació final del document. És interpretat pels principals sistemes operatius sense que es modifiqui ni l'aspecte ni l'estructura del document original. És un dels formats més utilitzats en internet i, en general, en l'intercanvi de documents.

Ping (Packet Internet Grouper). Eina de diagnòstic del protocol d'Internet, que bàsicament envia un paquet d'informació demanant al equip de destí que el retorni, i que serveix per comprovar la connectivitat entre dos sistemes.

Pla de Contingència. Document organitzat per implementar respostes d'emergència, operacions de còpies de seguretat, recuperació enfront d'un desastre, manteniment dels Sistemes d'Informació com a part del programa de seguretat, per així poder assegurar la disponibilitat dels processos i actius crítics i facilitar la continuïtat de les operacions enfront d'una emergència.

Plataforma. Entorn tecnològic, tant de maquinari, com de programari, que suporta les aplicacions que realitzen una lògica de negoci i les seves dades.

Programari(Software). Components lògics d'un sistema informàtic, que inclou el sistema operatiu, controladors de dispositiu, programes d'aplicació, utilitats i les dades sobre les quals operen.

Programari maliciós. Software nociu, dissenyat amb l'objectiu de causar problemes a un sistema, com els virus informàtics, cucs de xarxa, troians o bombes lògiques.

Proxy. Programa o dispositiu que realitza una acció en representació d'un altre. La finalitat més habitual és la de permetre l'accés a Internet als equips d'una organització quan només es disposa d'un únic equip connectat amb una única direcció IP.

Raspar(scrap). En sentit figurat, tècnica informàtica orientada a extraure informació de pàgines web.

Recol·lectar (harvest). Acció de recollir les dades de les pàgines web.

Risc. Possibilitat que una amenaça es faci realitat, provocant una sèrie de conseqüències no desitjades sobre un determinat sistema.

Router (encaminador). Dispositiu que administra la connexió entre dues xarxes. La seva funció principal és llegir les adreces dels paquets de dades i canalitzar-los cap al seu destí.

SAI (Sistema d'Alimentació Ininterrompuda). Dispositiu que utilitza bateries per assegurar l'alimentació d'energia elèctrica a tots els dispositius connectats a ell. També regula el flux d'electricitat, controlant les pujades i baixades de tensió i corrent existents en la xarxa elèctrica.

Servidor. Ordinador en el qual s'executa un programa, o servei ,que realitza alguna tasca en benefici d'altres aplicacions anomenades clients.

Sistema operatiu. Conjunt de programari que controla el funcionament bàsic d'un ordinador i sense el qual no podria ni engegar.

TCP/IP (Transmission Control Protocol / Internet Protocol). Protocol de comunicacions de dades a nivells de xarxa i transport.

TIC. Tecnologies de la informació i la comunicació.

Web. De l'anglès "World Wide Web", per aproximació significaria “teranyina mundial”. Rep aquest nom el conjunt de totes les pàgines disponibles a Internet.

WSDL. Inicials de “Web Services Description Language”, un format XML que s'utilitza per descriure Serveis web. WSDL descriu la interfície pública dels serveis web.

Xarxa. Connexió simultània entre diferents equips informàtics. Quan es parla de la Xarxa de xarxes es fa referència a Internet.

13. Bibliografia

Materials de la UOC de les assignatures:

- Direcció estratègica de la tecnologia de la informació
- Interacció humana amb els ordinadors
- Projecte fi de carrera
- Competència comunicativa per a professionals de la informàtica
- Compiladors I i II
- Enginyeria del programari de components i sistemes distribuïts
- Metodologia i gestió de projectes informàtics

Apache. (2011). *HttpComponents*. Consultat el 30 / 10 / 2011, a http://wiki.apache.org/HttpComponents

Apache. (2008). *Lucene*. Consultat el 29 / 10 / 2011, a http://lucene.apache.org/java/docs/index.html

Apache. (2009). *Nutch*. Consultat el 30 / 10 / 2011, a http://nutch.apache.org/

Cloudgarden. (2008). *Jigloo*. Consultat el 05 / 11 / 2011, a http://www.cloudgarden.com/jigloo/

Contadorwap. (2006). *Buscadores-robots*. Consultat el 30 / 10 / 2011, a http://www.contadorwap.com/buscadores-robots.php

Creativecommons. (2011). *Licenses*. Consultat el 20 / 12 / 2011, a http://creativecommons.org/licenses/by-nc/3.0/es/

Debugmodeon. (2009). *Apache-lucene*. Consultat el 30 / 10 / 2011, a http://es.debugmodeon.com/articulo/introduccion-a-apache-lucene-java/3#comments

Half-wit4u. (2008). *Web-scraping-using-java*. Consultat el 29 / 10 / 2011, a http://half-wit4u.blogspot.com/2011/01/web-scraping-using-java-api.html

Harvest. (2008). *Harvest*. Consultat el 01 / 11 / 2011, a http://web-harvest.sourceforge.net/

Javahispano. (2009). *Javahispano*. Consultat el 29 / 10 / 2011, a http://www.javahispano.org/forum/j2se/es/implementacion_open_source_xqj/

Javamex. (2005). *Example_scraping*. Consultat el 02 / 10 / 2011, a http://www.javamex.com/tutorials/regular_expressions/example_scraping_html.shtml

Java-spain.com. (2009). *Jigloo-swing-para-eclipse*. Consultat el 05 / 11 / 2011, a http://java-spain.com/jigloo-un-editor-interfaces-swing-para-eclipse

Larman, C. (2006). *UML y Patrones.* Madrid: Prentice Hall.

Luauf.com. (2008). *Expresiones-regulares-en-java*. Consultat el 30 / 10 / 2011, a http://luauf.com/2008/05/03/expresiones-regulares-en-java/

Microsoft. (2005). *XSLT*. Consultat el 23 / 11 / 2011, a http://msdn.microsoft.com/es-es/library/ms256196%28v=VS.90%29.aspx

Mozenda. (2007). *Web-scraping*. Consultat el 02 / 10 / 2011, a http://www.mozenda.com/web-scraping.aspx

Newprosoft. (01 / 01 / 2010). *Newprosoft*. Consultat el 01 / 11 / 2011, a http://www.newprosoft.com/

Oracle. (2008). *Java*. Consultat el 2 / 11 / 2011, a http://download.oracle.com/javase/tutorial/networking/index.html

Oracle. (2008). *JAXB*. Consultat el 09 / 11 / 2011, a http://docs.oracle.com/cd/E12840_01/wls/docs103/webserv/data_types.html

Oracle. (2006). *regex/Pattern*. Consultat el 22 / 10 / 2011, a http://docs.oracle.com/javase/1.4.2/docs/api/java/util/regex/Pattern.html

Programacion.com. (2005). *Expresiones_regulares_en_java*. Consultat el 29 / 10 / 2011, a http://www.programacion.com/articulo/expresiones_regulares_en_java_127

Scribd. (2009). *Busqueda-a-traves-de-Lucene*. Consultat el 01 / 11 / 2011, a http://es.scribd.com/doc/3013167/Indizacion-y-Busqueda-a-traves-de-Lucene

Sourcecodeonline. (2006). *Web_scraping_java* . Consultat el 02 / 10 / 2011, a http://www.sourcecodeonline.com/list?q=web_scraping_java

Taringa. (2009). *Posicionamiento*. Consultat el 15 / 10 / 2011, a http://www.taringa.net/comunidades/webdesign/1840693/SEO---Posicionamiento-en-Buscadores.html

Taringa. (2009). *Seguridad-web*. Consultat el 01 / 11 / 20011, a http://www.taringa.net/comunidades/webdesign/1620757/tips-SEO-y-seguridad-web-%28Robots_txt%29.html

W3. (2009). *XSLT*. Consultat el 03 / 11 / 2011, a http://www.w3.org/TR/xslt

Webtaller. (2005). *Guia_robots*. Consultat el 28 / 10 / 2011, a http://www.webtaller.com/google/guia_robots.php

Webtaller. (2005). *Robots-arania*. Consultat el 20 / 10 / 2011, a http://www.webtaller.com/maletin/articulos/como-funcionan-robots-arania.php

Wikipedia. (2002). *HTML*. Consultat el 10 / 10 / 2011, a http://ca.wikipedia.org/wiki/HTML

Wikipedia. (2004). *HTTP*. Consultat el 10 / 10 / 2011, a http://ca.wikipedia.org/wiki/HTTP

Wikipedia. (2005). *Java*. Consultat el 02 / 11 / 2011, a http://ca.wikipedia.org/wiki/Java

Wikipedia. (2008). *Nutch*. Consultat el 02 / 10 / 2011, a http://en.wikipedia.org/wiki/Nutch

Wikipedia. (2005). *Regular_expression*. Consultat el 29 / 10 / 2011, a http://en.wikipedia.org/wiki/Regular_expression

Wikipedia. (sense data). *Search_engines*. Recollit de http://en.wikipedia.org/wiki/List_of_search_engines

Wikipedia. (2004). *Socket_programming*. Consultat el 05 / 10 / 2011, a http://en.wikipedia.org/wiki/Socket_programming

Wikipedia. (2006). *Web_scraping*. Consultat el 01 / 10 / 2011, a http://en.wikipedia.org/wiki/Web_scraping

Wikipedia. (2006). *WSDL*. Consultat el 20 / 11 / 2011, a http://ca.wikipedia.org/wiki/WSDL

Wikipedia. (2005). *XML*. Consultat el 02 / 11 / 2011, a http://ca.wikipedia.org/wiki/Xml

Wikipedia. (2008). *Xslt*. Consultat el 02 / 11 / 2011, a http://ca.wikipedia.org/wiki/Xslt

Ywebb. (2009). *Ywebb*. Consultat el 25 / 10 / 2011, a http://www.ywebb.com/

15. Llicències de publicació del projecte

Llicències Creative Commons (CC) [26]

Mitjançant la combinació de diferents principis i restriccions, les llicències autoritzen certs usos lliurement definits pels autors. Les combinacions es generen entorn de quatre condicions bàsiques:

Reconeixement, o *Attribution* (**by**): sempre s'ha de reconèixer l'autoria de l'obra.
No comercial, o *Non Commercial* (**nc**): no es pot utilitzar l'obra ni els seus treballs derivats amb finalitats comercials.

Creative Commons by-nc. Reconeixement – NoComercial: No es permet un ús comercial de l'obra original però sí la generació d'obres derivades.

Llicències GNU de la Free Software Foundation

GNU-GPL: Llicència pública general (en anglès GPL, General Public License) és un tipus de llicència per a programar el que permet la lliure còpia, distribució (comercial o no) i modificació del codi, sempre que qualsevol modificació es continuï distribuint amb la mateixa llicència GPL. La llicència GPL no permet la distribució de programes executables sense el codi corresponent o sense el lloc on es pugui obtenir gratuïtament.

[26] (Creativecommons, 2011)