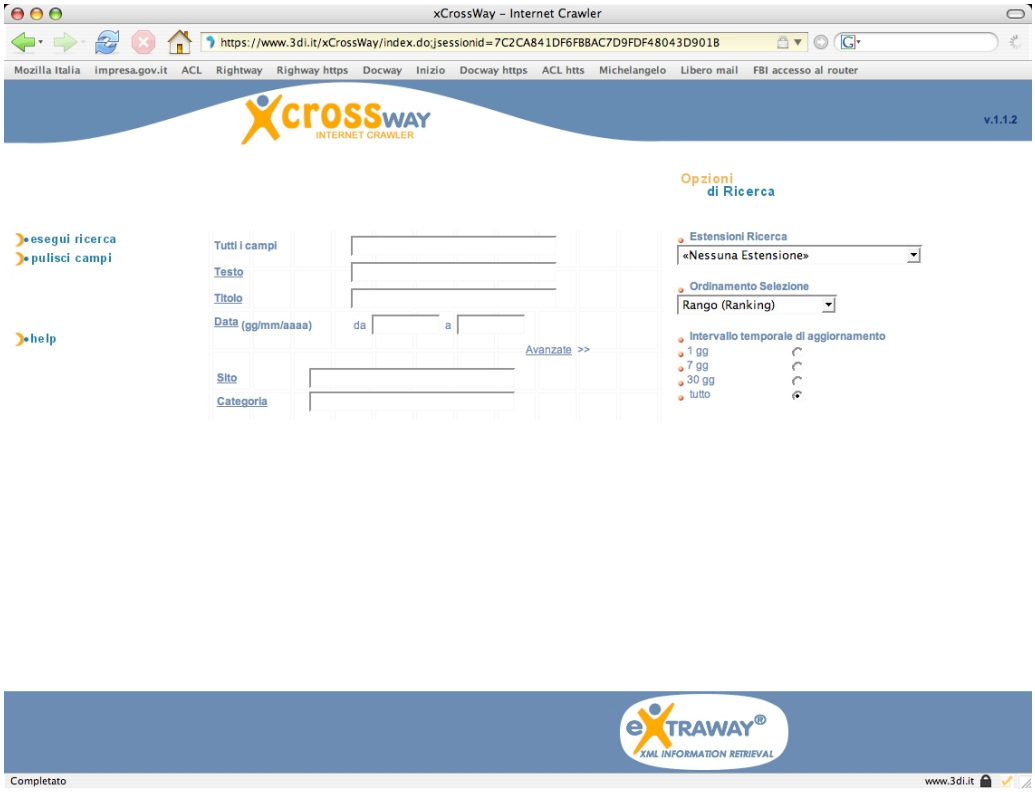
una realtà tecnologica
ITALIANA

INTERNET E INTRANET CRAWLER

3D INFORMATICA
San Lazzaro di Savena
40068 Via Speranza, 35
Tel: +39 051450844
fax: +39 051451942

ROMA
00198 Via di S. Teresa, 23
Tel: +39 068840309
informazioni@3di.it
www.3di.it
www.w4k.it



COSA CONSENTE

E' l'applicativo che si occupa di leggere le pagine di un determinato sito internet (o di più siti internet) o specifici contenuti all'interno di fonti in formato elettronico secondo determinate regole e scorrerle nella loro completezza, alimentando un Information Retrieval Engine.

Opportunamente "istruito" è in grado di raccoglie, indicizzare ed organizzare le informazioni presenti sul web, dal sito istituzionale fino al contenuto di una singola directory di un Server.

XCrossWay è inoltre in grado, qualora lo si renda necessario, di indicizzare e reperire le informazioni strutturate normalmente presenti nei database relazionali. La scansione di questi "target", mediante connessione ODBC, produce e gestisce documenti Xml nell'archivio di **XCrossWay** atti a replicare la distribuzione dei dati presenti nel data base acquisito ed indicizzato.

COME FUNZIONA

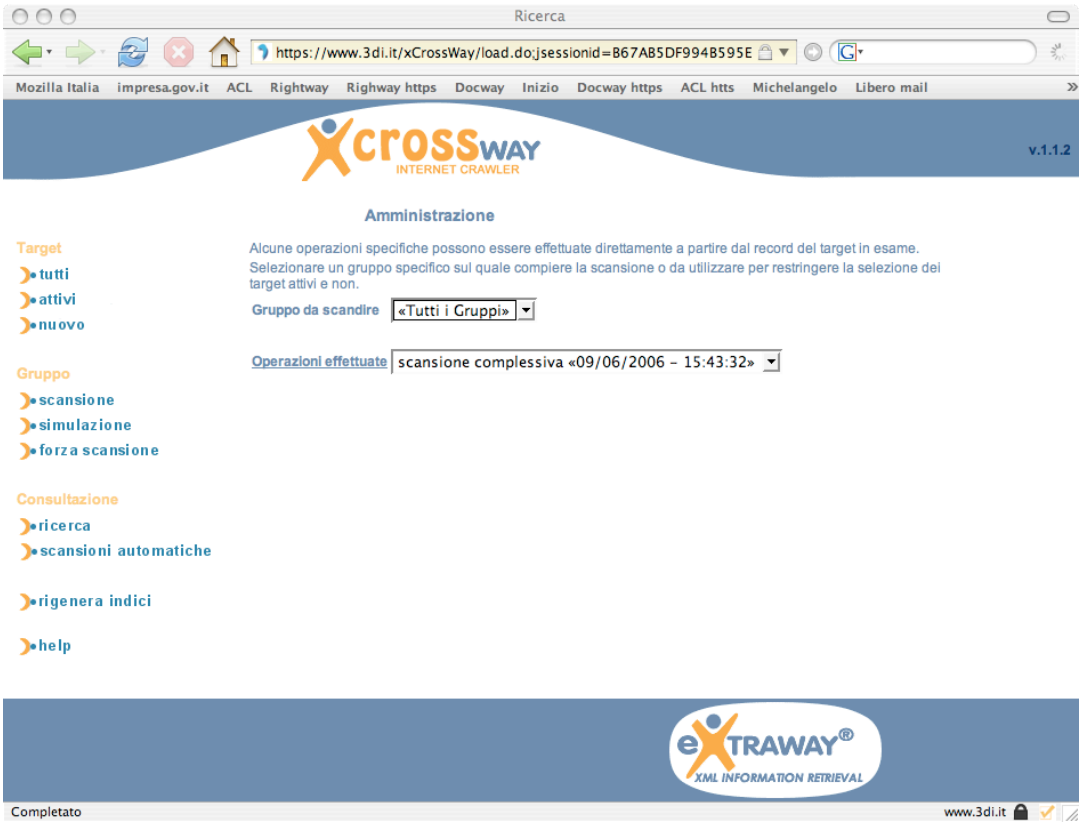
XCrossWay può operare secondo diverse metodologie tese a raggiungere risultati specifici o generici, a seconda dell'obiettivo prefissato. Come detto le informazioni che fanno parte della base documentale possono essere o meno strutturate ed anche su questo fronte si possono differenziare i comportamenti del Crawler.

Il Modulo Crawler deve essere configurato per mezzo di "target" da intendersi come fonti documentali da acquisire e catalogare, sia che esse siano interne sia che esse siano acquisite dal Web.

Detto questo ecco che il profilo del Crawler si delinea secondo alcuni possibili stili:

- ✓ **Il Crawler agisce di propria iniziativa**. Deve essere configurato con alcuni "target" che verranno visitati periodicamente e dai quali, con opportuni accorgimenti, verranno acquisiti genericamente tutti i documenti disponibili ed interessanti. Qualora l'acquisizione interessi siti Internet il grado di configurazione di questi target deve consentire l'acquisizione dei dati evitando tutto il rumore dovuto a dati non necessari.
- ✓ **Il Crawler agisce "on demand"**. Una sorta di protocollo d'intesa tra il produttore dei dati ed il centro che intende gestirli deve rendere edotto il Crawler dell'esistenza e disponibilità di nuovi documenti. Il Crawler a questo punto non deve acquisire con una scansione completa ma può agire nel dettaglio del file o dell'insieme di file indicati. Nulla toglie, comunque, che il Crawler possa eseguire scansione di questi documenti di sua iniziativa, in seguito, per verificare eventuali variazioni ai documenti in esame.
- ✓ **Il Crawler agisce tramite un proprio "agent"** presso la macchina ove i documenti vengono prodotti. Questo agent, non invasivo, può svolgere per conto del Crawler

l'analisi del materiale disponibile selezionando tra esso tutto quello che risulti significativo per i fini preposti. Allo stesso modo può interrogare basi dati locali al server di "produzione" dalle quali riconoscere i documenti pertinenti oppure ancora producendo a sua volta documenti, ad esempio in formato XML, direttamente da "viste" rese disponibili da DB relazionali locali. La scelta del formato XML si presta particolarmente per rispettare la struttura dei dati e renderla significativa anche per il Crawler ed allo stesso modo divenire maggiormente flessibile ed adattabile e future eventuali variazioni nella struttura dati del DB stesso. I "documenti" in esame, fisicamente presenti sul server o dinamicamente prodotti dall'agent, potrebbero anche non essere disponibili al pubblico per mezzo del comune protocollo HTTP ma appositamente inviati al Crawler su sua richiesta. In questo modo il Crawler potrebbe indicizzare anche documenti "privati" per i quali deve vigere un esplicito accordo.



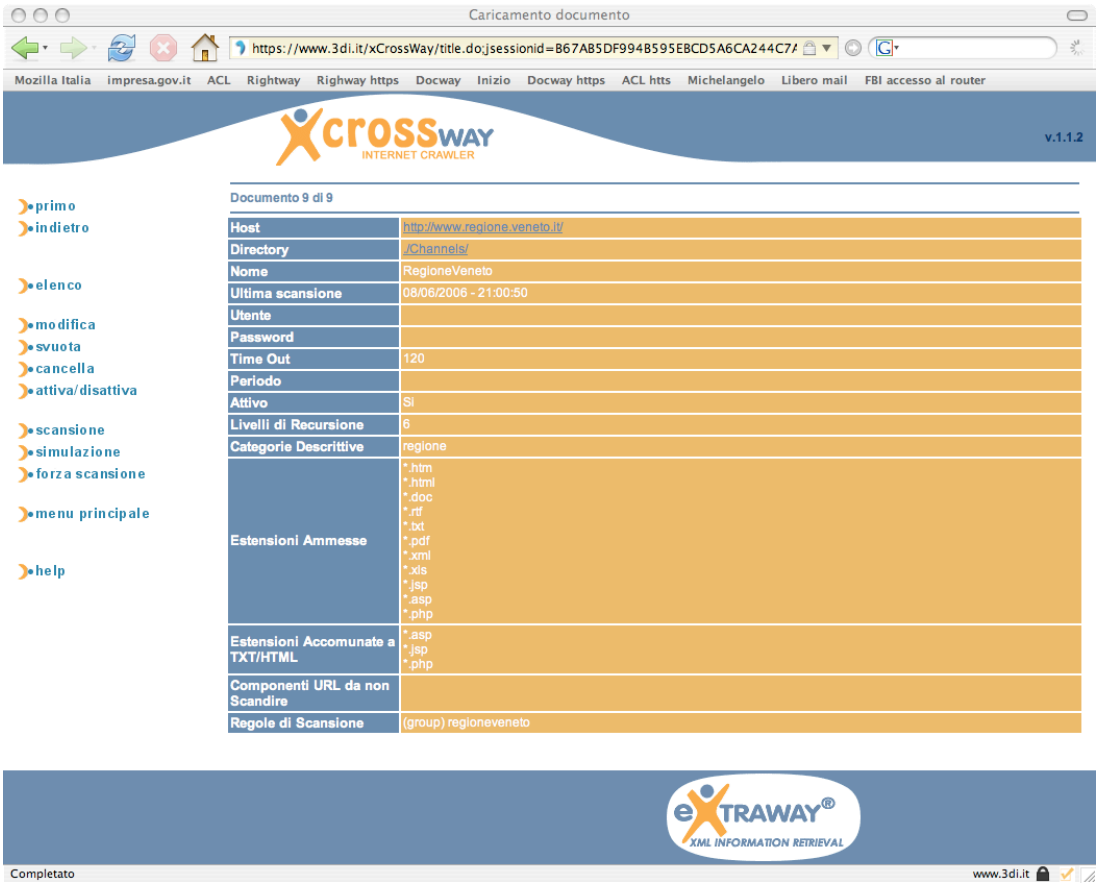
Questi metodi non sono del tutto alternativi e possono essere ipotizzate soluzioni miscelate che uniscano caratteristiche comportamentali proprie delle singole ipotesi in altre maggiormente articolate. La soluzione da scegliere, ovviamente, dipende dal fine che si intende attribuire l'insieme di dati raccolti.

Dopo aver anticipato queste metodiche, il Crawler deve conoscere le "regole" secondo le quali compiere il trattamento dei "documenti" raccolti. Intuitivamente la soluzione più complessa, e che richiede la configurazione più accurata, è quella in cui il Crawler deve "scegliere" il materiale da acquisire ed aggiungere al proprio catalogo. Genericamente, quindi, il Crawler rispetta le seguenti regole:

- ✓ Conoscere, ad esempio tramite una URL o una URN, un "sito" ed al suo interno un punto di partenza entro il quale operare. La scansione procede seguendo tutti i links disponibili nelle singole "pagine" acquisite e non spaziare a 360° per non produrre una inutile dispersione ed un conseguente "rumore", ma si attiene ad URL che "rimangano" nell'ambito dello stesso sito e, potenzialmente, nell'ambito del sott'insieme del sito stesso espresso dal percorso della URL data come punto di partenza. In caso si operi localmente, questa restrizione si può intendere come la directory o il repository documentale da scandire. Il Crawler è configurabile per riconoscere URL già visitate nonostante varianti ad esse applicate ad arte. I documenti disponibili in internet vengono spesso raggiunti tramite URL che vengono modificate di volta in volta con elementi non necessari e di natura randomica aventi il solo scopo di forzare l'acquisizione ex-novo di ogni "pagina". Il Crawler è configurabile per riconoscere e scartare questi elementi così da evitare, quando possibile, l'acquisizione di nuovi "documenti" che non sono altro che copie di quelli già acquisiti. In questo modo si evitano inutili doppioni e si migliorano le performance del Crawler stesso.
- ✓ Riconoscere la tipologia dei "documenti" acquisiti, eleggere al rango di "dati"

quelli noti come utili e scartare tutti quelli considerati non utili attingendo ad una sorta di "white & black list". In questo modo si restringe ai "documenti" effettivamente voluti il campo d'azione del Crawler.

- ✓ Rispettare le regole imposte dai server visitati. Tali regole vengono tipicamente indicate in un file "robots.txt" che i server Web utilizzano per limitare l'operato dei Crawler "rispettosi".
- ✓ Mantenere, se previsto dalla configurazione, una cache dei "documenti" acquisiti. Questa funzionalità ha senso, in particolare, per i dati acquisiti da Internet. Consente di visualizzare, anche se in modo approssimativo, copie dei documenti originali senza dover accedere al sito visitato in precedenza anche perché la natura di alcuni documenti è volatile e gli stessi potrebbero essere disponibili "al pubblico" solo per un tempo limitato. La qualità dell'informazione acquisita, sotto forma di sua rappresentazione esteriore, potrebbe non essere garantita completamente da questa cache, ma la "sostanza" del documento rimarrebbe disponibile anche qualora il produttore degli stessi ne negasse in futuro la visibilità.
- ✓ Classificare i "documenti" acquisiti così come precedentemente esposto o applicando ai singoli "target" le più opportune voci di classificazione.
- ✓ Estrarre da tali documenti, secondo regole da definirsi, eventuali "URN" in grado di rendere univoco il riconoscimento ed il reperimento del documento stesso.



Il trattamento dei documenti, come precedentemente esposto, deve tenere conto del fatto che essi possano essere strutturati o meno.

Per i **documenti non strutturati** (file prodotti con strumenti quali "Acrobat", "Microsoft Office", "Open Office" ed altri che producano file standard (es.: RTF o HTML) possono essere messe a punto regole di riconoscimento che portino ad una distinzione del contenuto (eventualmente per mezzo di meta dati da concordare con chi produce le informazioni) così da poter eleggere parte di esso ad un ruolo differente alimentando quindi diversi "canali di ricerca".

Per i **documenti strutturati** quali sono i file XML (o gli equivalenti file che un agent potrebbe creare dinamicamente in tale formato acquisendo dati da DB Relazionali), vengono applicate le caratteristiche del Motore di Ricerca sul quale **XCrossWay** si basa e che verranno discusse in seguito. In generale, comunque, tutte le componenti del file XML vengono riconosciute e, secondo la configurazione data al catalogo attuale, vanno ad alimentare automaticamente i "canali di ricerca" che il framework metterà a disposizione.

OBIETTIVI E BENEFICI

In conclusione quindi possiamo affermare che **XCrossWay** è una piattaforma di Knowledge Management (KM) che consente una gestione ottimizzata del patrimonio informativo per mezzo delle seguenti funzionalità:

XML Content Repository: i contenuti informativi (oggetti instanziati dal motore) indipendentemente dall'interfaccia con cui sono stati prodotti, sono fisicamente file o parti di file XML.

Content categorization: **XCrossWay** consente di indicizzare e classificare (add-on) automaticamente tutti i documenti indipendentemente dal formato o dal supporto tecnologico (informazioni memorizzate su DB, cartelle e-mail, pagine web, file doc, ppt, XML,…) e parti del web che possono risultare interessanti.

Le informazioni trattate "quotidianamente" possono essere raggruppate in due distinte categorie: quelle strutturate, ovvero quelle storicamente presenti nei database e quelle non strutturate – la maggior parte – che di fatto costituiscono il patrimonio documentale (cartaceo ed elettronico) che l'azienda produce e/o acquisisce nell'esercizio delle sue attività: fax e corrispondenza, modulistica compilata a mano, pratiche e faldoni, fotografie, disegni catastali, ecc.., e con l'avvento di internet, i messaggi di posta elettronica (casella di e-mail istituzionale e non) e i contenuti Web sempre più numerosi.

XCrossWay integra fra loro i dati strutturati e quelli non strutturati organizzando il sistema informativo: qualsiasi documento o informazione può essere acquisito, indicizzato, organizzato, elaborato e distribuito all'interno e all'occorrenza all'esterno dell'organizzazione.

L'informazione è resa disponibile rapidamente ed automaticamente a chiunque ne abbia bisogno per il proprio lavoro, direttamente sul suo PC mediante un comune browser Web.

Grazie alle potenzialità dell'innovazione tecnologica, il sistema KM diviene pertanto un indispensabile servizio di organizzazione e gestione di un contesto lavorativo in quanto strumento flessibile di informazione e comunicazione in grado di garantire:

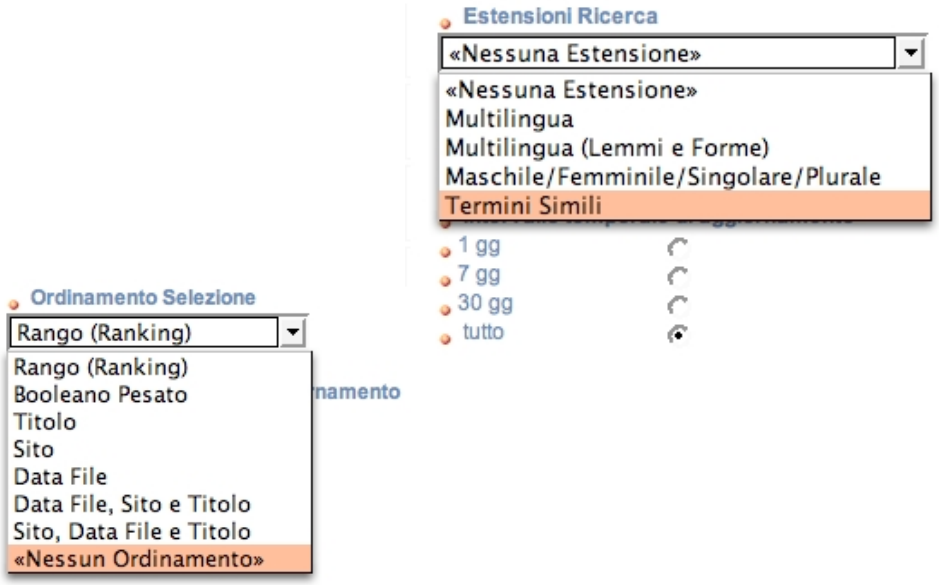
- ✓ L'indicizzazione e la classificazione (catalogazione) automatica delle informazioni
- ✓ L'integrità delle informazioni e dei documenti
- ✓ L'accesso semplice ed immediato ai contenuti
- ✓ Strumenti di ricerca sempre più sofisticati e razionali.

Tutto ciò consente di eliminare carte e tempi morti, errori e/o perdite di documenti, la necessità di continue ricerche e duplicazioni. L'informazione, rese elettroniche, viene così reperita e consultata con il semplice click del mouse. Si riducono i cicli di lavoro ed i tempi decisionali, aumenta la produttività delle persone, cresce l'efficienza gestionale ed amministrativa.

LA RICERCA

Le funzioni di indicizzazione e ricerca sono senza dubbio tra le caratteristiche più importanti di **XCrossWay**:

- ✓ ricerche booleane (AND, OR, NOT);
- ✓ ricerche per similitudine;
- ✓ relevance ranking;
- ✓ prossimità ((Adj, Near, Context);
- ✓ stoplist;
- ✓ wildcard;
- ✓ concatenazione di indici;
- ✓ vocabolari;
- ✓ vocabolari di classificazione;
- ✓ thesauri;
- ✓ albero delle ricerche;
- ✓ analisi spettrali (vocabolario ristretto al set di documenti trovati);
- ✓ modulo opzionale di ricerca con linguaggio naturale;
- ✓ ricerche multilingua (lemmi e forme);
- ✓ ricerche per documenti simili;
- ✓ ricerche ricorsive.



I MODULI ADD-ON

XCrossWay è una piattaforma di ricerca aggiornabile con alcune funzionalità avanzate, che consentono di offrire un ulteriore valore aggiunto oltre ai servizi fondamentali.

- ✓ E-Mail Management
- ✓ Dizionari e contenuti
- ✓ Classificazione automatica
- ✓ Gestione documentale (DOCWAY® e-Document Web Solution)
- ✓ Modulo di interpretazione di script di sessione terminale di tipo 3270, VT100 e simili che permette di gestire connessioni con uno o più host per la cattura di dati da sistemi legacy.

CARATTERISTICHE DELL'ARCHITETTURA

XCrossWay è disponibile per le seguenti piattaforme:

- ✓ Windows
- ✓ Linux (Intel, Sparc, PowerPC)
- ✓ Unix (Intel, Sparc, PowerPC)
- ✓ Mac OSX PPC
- ✓ Mac OSX Intel

Framework

- ✓ **XCrossWay** (XML+mappe e indici)
- ✓ Livello applicativo (JSP, Generic+Specs.JAR)
- ✓ Livello trasformazione file (XSLT ISAPI)
- ✓ Livello di presentazione ed interazione (HTML4, DHTML, CSS, JPG, Applets)

Livello applicativo

Il livello applicativo può essere realizzato in diversi ambienti:

- ✓ Un ambiente Java che utilizza le classi di collegamento base con il motore (**XCrossWay** Java Kit);
- ✓ Un modulo client C/C++ che utilizza le API C;

L'ambiente Java prevede l'utilizzo di classi a vari livello: comunicazione base con il motore (interrogazione, gestione delle selezioni, inserimenti, cancellazioni, modifiche) e gestione di comandi avanzati quali la navigazione, l'esplosione ed aggiornamento di contenitori gerarchici, lookup su altre collezioni e l'utilizzo dei comandi di consultazione ed aggiornamento del thesaurus.

- ✓ E' possibile utilizzare un interprete di pagine JSP quale Tomcat per la produzione di file HTML o XML.

Il livello applicativo comprende moduli specifici per alcuni contesti:

- ✓ Modulo di supporto del protocollo Z39.50 versione 3 che consente di rendere interrogabili le collezioni documentali da client standard Z39.50 e viceversa consente ai client del motore (sia web che desktop Windows) di ottenere dati contemporaneamente da più fonti Z39.50 disponibili sulla rete come se provenissero dall'unico server cui sono collegati.

Livello di trasformazione con fogli di stile

Il risultato del livello applicativo è rappresentato da file in formato HTML o XML. In quest'ultimo caso è necessario operare una trasformazione del file verso il formato HTML. Questa può essere effettuata da un processore XSLT installato, secondo le opportunità in uno dei seguenti ambienti:

- ✓ Nel livello applicativo mediante classi disponibili nella Java Virtual Machine;
- ✓ Sul server web mediante un apposito modulo di Extraway configurato come estensione di Apache HTTP Server o come filtro ISAPI per Microsoft IIS;
- ✓ Sul browser dell'utente.

SEGMENTI DI MERCATO

I segmenti di mercato considerati in questo documento riguardano i settori Industry, Governament (Pubblica Amministrazione), Finance e Media.

Industria

Il segmento Industria

Le esigenze di questo settore che possono essere coperte con **XCrossWay** sono molteplici, ma possono essere ricondotte essenzialmente alle seguenti funzionalità:

- ✓ Semplificazione del reperimento delle informazioni aziendali (con riduzione dei relativi costi);
- ✓ Miglioramento delle performance del search-engine del sito web, con miglioramento della navigabilità e conseguente aumento della customer satisfaction;
- ✓ Ottimizzazione del supporto informativo ottenuto da fonti Web;
- ✓ Ottimizzazione della knowledge sharing all'interno dell'azienda;
- ✓ Attivazione di funzionalità di competitive intelligence per mezzo dell'indicizzazione dei siti web di concorrenti.

Amministrazione Pubblica

Il segmento Amministrazione Pubblica

La Pubblica Amministrazione (PA) ha attivato un piano per la definizione di linee guida alla realizzazione delle soluzioni Web-based:

(Piano e-Gov: http://www.innovazione.gov.it/ita/index.shtml) .

La PA deve interagire con l'utenza (cittadini, imprese, associazioni,…), assicurando sia la trasparenza dell'amministrazione, dei suoi processi e dei suoi atti sia l'accessibilità alle informazioni e ai servizi a tutti, anche ai navigatori meno esperti.

Questo significa offrire agli utenti Web servizi di ricerca adeguati all'interno del portale dell'ente. Il linguaggio della PA (molto legato al linguaggio legislativo) rende oltremodo di importanza fondamentale l'adozione di funzionalità che facilitino il reperimento dell'informazione.

Le esigenze di questo settore che possono essere coperte con **XCrossWay** sono molteplici, ma possono essere ricondotte essenzialmente alle seguenti funzionalità:

- ✓ Semplificazione del reperimento delle informazioni (con riduzione dei relativi costi);
- ✓ Miglioramento delle performance del search-engine del sito web, con miglioramento della navigabilità e conseguente aumento della customer satisfaction;
- ✓ Ottimizzazione della knowledge sharing all'interno dell'ente;
- ✓ Gestione dei riferimenti normativi.

Tale soluzione, attiva presso il CED della Cassazione, consente la ricerca concettuale all'interno degli archivi legislativi.

Finanza

Il segmento Finanza

Il settore Finance, oltre alle evoluzioni relative all'implementazione di siti web a forte interazione con la (banche on-line), ha visto negli ultimi anni lo sviluppo di intranet aziendali finalizzate alla condivisione di documentazione e informazioni.

XCrossWay gestisce in modo nativo il linguaggio tipico delle circolari bancarie e la terminologia d'ambito assicurativo,

Le esigenze di questo settore che possono essere coperte con **XCrossWay** sono molteplici, ma possono essere ricondotte essenzialmente alle seguenti funzionalità:

- ✓ Semplificazione del reperimento delle informazioni aziendali (con riduzione dei relativi costi);
- ✓ Miglioramento delle performance del search-engine del sito web, con miglioramento della navigabilità e conseguente aumento della customer satisfaction;
- ✓ Ottimizzazione del supporto informativo ottenuto da fonti Web;
- ✓ Ottimizzazione della knowledge sharing all'interno dell'azienda;
- ✓ Attivazione di funzionalità di competitive intelligence per mezzo dell'indicizzazione dei siti web di concorrenti.

Tale soluzione, attiva presso i first movers nel campo della finanza on-line in Italia, consente un'ottimizzazione della customer experience degli utenti web.

Editoria

Il segmento Editoria

Le esigenze di questo settore che possono essere coperte con **XCrossWay** sono molteplici, ma possono essere ricondotte essenzialmente alle seguenti funzionalità:

- ✓ Semplificazione del reperimento delle informazioni per i redattori (con riduzione dei relativi costi);
- ✓ Miglioramento delle performance del search-engine del sito web, con miglioramento della navigabilità e conseguente aumento della customer satisfaction;
- ✓ Ottimizzazione della knowledge sharing tra i redattori;
- ✓ Attivazione di funzionalità di pushing a favore degli utenti delle informazioni web.

CONFIGURAZIONE HARDWARE DI MINIMA

Alimentazione elettrica in continuità (300 W) x server

Sistema operativo (a scelta uno dei seguenti):

- ✓ Sparc/Solaris >= 2.6
- ✓ Linux Kernel >=2.2
- ✓ IBM AIX >= 4.1
- ✓ Windows NT 4.0 >=SP3
- ✓ Windows 2000 SP1
- ✓ Windows XP
- ✓ Mac OSX PPC
- ✓ Mac OSX Intel

RAM: >= 512MB

Processore: >=300Mhz

HD: >= 80GB (meglio se in modalit RAID 5 con controller RAID SCSI)

Video: >= 15

Risoluzione: >= 800x600

Unit di backup: DAT

Gruppo di continuità: presente

Connettività di Back-End

Back end firewall condiviso (comprende 1 regola di packet filtering)

Porta LAN attestata ad uno switch con un insieme di indirizzi IP interni

Infrastruttura di back end per accesso tramite ISDN a 64K per la gestione del server

PARAMETRI PRESTAZIONALI

Tempi risposta per una query: da 50ms a 200ms

N.ro di utenti: illimitato

N.ro di documenti indicizzabili: Fino a 16.000.000 per collezione