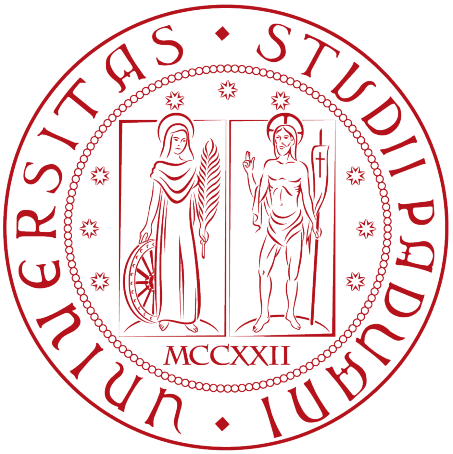
Università degli Studi di Padova

Dipartimento di Matematica "Tullio Levi-Civita"

Corso di Laurea in Informatica



Analisi di motori di ricerca Open Source per siti web informativi

Tesi di laurea triennale

Relatore
Prof.Tullio Vardanega

Laureando
Daniel De Gaspari

Anno Accademico 2016/2017

placeholder con citazione.

— Oscar Wilde

Sommario

Il presente documento descrive il lavoro svolto durante il periodo di stage dal laureando Daniel De Gaspari, della durata di circa trecento ore, presso l'azienda InfoCamere S.C.p.A. di Padova (PD).
Gli obiettivi da raggiungere erano molteplici.
Lo scopo dello stage consisteva nell'analisi delle caratteristiche dei motori di ricerca open source nell'ambito dei siti web di tipo informativo.
In primo luogo era richiesto un approfondimento delle caratteristiche istituzionali dei siti web delle Camere di Commercio.
Successivamente, l'azienda richiedeva di analizzare le potenzialità e specificità di motori di ricerca Sorl e ElasticSearch.
Il passo successivo consisteva nel realizzare un prototipo di un sito web in tecnologia Drupal, con i due motori di ricerca precedentemente citati.
Infine, era richiesta una relazione finale delle potenzialità emerse nell'utilizzo dei due motori di ricerca.
I primi due capitoli del presente documento hanno lo scopo di presentare il contesto aziendale in cui è stato sostenuto lo stage e di spiegare come il progetto di stage si renda utile all'interno della strategia aziendale. Il terzo capitolo documenta invece lo svolgimento dello stage descrivendo le attività che sono state portate a termine, i punti salienti del progetto stesso e le principali scelte attuate. Il quarto ed ultimo capitolo presenta infine una valutazione dello svolgimento dello stage rispetto agli obiettivi aziendali e alle conoscenze acquisite dallo studente.

"Citazione"

— Confucius

Ringraziamenti

Ringraziamenti

Padova, Dicembre 2017 Daniel De Gaspari

Indice

Elenco delle figure

Elenco delle tabelle

Capitolo 1

L'azienda

1.1 Il Profilo Aziendale

1.1.1 Le origini: Cerved

Nata inizialmente come Cerved (Centro Regionale Veneto Elaborazione Dati), InfoCamere S.C.p.A. è stata fondata nel Dicembre del 1974 a Padova dal Professor Mario Volpato, allora Presidente della Camera di Commercio di Padova e Professore di Calcolo delle probabilità all'Università di Padova.
L'obiettivo era di raccogliere e conservare i dati ufficiali anagrafici e amministrativi delle imprese della provincia di Padova in un modo nuovo rispetto a quanto previsto fino ad allora: la conservazione di quei dati su un registro cartaceo, come si faceva dal medioevo ai tempi delle comunità dei mercanti, non bastava più a garantire l'efficienza del mercato e a stimolare lo sviluppo economico.
Le prime tecnologie informatiche aprivano nuovi orizzonti al trattamento massivo e veloce dei dati. Evolveva rapidamente la concezione di una gestione intelligente delle notizie amministrative sulla vita delle imprese, per trasformarle in informazioni rielaborabili ed utilizzabili in modi nuovi da tutti. Nasceva l'idea di valorizzare i dati ufficiali forniti dalle imprese, restituendoli al mercato e alle imprese stesse come informazioni utili per accrescere la propria competitività e progettare lo sviluppo.
Si gettava il seme dell'efficienza nell'organizzazione delle Camere di Commercio, una base nuova per costruire un patto trasparente e vantaggioso tra imprese e Pubblica Amministrazione.

1.1.2 Anni '90: InfoCamere

All'inizio degli anni '90 aumentava sempre più la competizione globale e le sfide per portare l'Italia nella modernità. A tal fine, nel 1993, venne emanata una riforma (legge 29 Dicembre 1993, n. 580) che attribuiva alle Camere un'autonomia rispetto al governo centrale, mediante attribuzione della potestà statutaria e di autonomia finanziaria, oltre al riconoscimento del ruolo finalizzato alla pubblicizzazione delle imprese. Le Camere di Commercio Italiane hanno così modo di vedere un profondo rinnovamento in vari ambiti e in particolar modo nell'ambito tecnologico.
Nel 1995, per scissione da Cerved, nasce InfoCamere che raccoglie la sfida di realizzare il Registro delle imprese. Previsto dal codice civile fin dal 1942 e mai attuato, in due anni, con uno di anticipo sulle previsioni, il risultato è raggiunto: prende vita il primo

esempio in Europa di registro pubblico sulle imprese totalmente telematico; assieme ad un ecosistema di servizi sviluppati attorno al Registro delle imprese, è stato possibile semplificare i processi tra le imprese stesse e la Pubblica Amministrazione.
A quella sfida ne seguono altre che rispondono ai nomi di 'firma digitale', 'posta elettronica certificata', 'comunicazione unica', ecc... .
Attraverso InfoCamere, servizi e tecnologie digitali di frontiera diventavano patrimonio quotidiano della comunità delle imprese e dei professionisti, influendo sulle abitudini di lavoro di migliaia di italiani e stimolando i processi di innovazione nella Pubblica Amministrazione.

Figura 1.1: Logo infocamere

1.1.3 Servizi offerti

InfoCamere S.C.p.A. è la società consortile di informatica delle Camere di Commercio Italiane. Ha realizzato e gestisce il sistema telematico nazionale che collega tra loro tutte le Camere di Commercio, oltre alle rispettive sedi distaccate.
Sua funzione istituzionale è anche la gestione e divulgazione del patrimonio informativo camerale, con particolare riferimento alle informazioni derivanti dal Registro delle imprese.
Le banche dati camerali sono rese disponibili direttamente a imprese, pubbliche amministrazioni, professionisti e cittadini tramite il portale delle Camere di Commercio.
La Società fornisce alle pubbliche amministrazioni l'accesso al Registro Imprese, assicurando loro l'accessibilità dei dati senza oneri, salvo quelli per la fornitura telematica e i servizi a valore aggiunto.
Tramite il sito del Registro delle imprese si può accedere agli strumenti per lo svolgimento delle pratiche telematiche, tra cui la Comunicazione Unica per l'attività d'impresa, valida anche per Agenzia delle Entrate, INPS, INAIL e Albo Artigiani. Il Registro Imprese è inoltre uno strumento di trasparenza amministrativa che fornisce un contributo importante nella lotta contro la criminalità economica. L'azienda ha infatti sviluppato per le autorità investigative alcuni servizi che, in questa direzione, consentono analisi mirate per monitorare fenomeni anomali.
InfoCamere ha realizzato, per conto delle Camere di Commercio, l'infrastruttura tecnologica che garantisce il corretto funzionamento degli Sportelli Unici per le Attività Produttive (SUAP).
Tra le realizzazioni di InfoCamere per il Sistema camerale vi è anche la procedura informatica che consente di gestire il servizio di conciliazione online (Concilia Camera), fornendo così ad imprese, consumatori e professionisti uno strumento che permette di ricevere assistenza specializzata nel raggiungimento di un accordo per risolvere in modo semplice, rapido, economico e sicuro una controversia, evitando di ricorrere alla

giustizia ordinaria.
InfoCamere è inoltre l'Autorità di Certificazione Nazionale che rilascia i certificati digitali delle Carte Tachigrafiche. La società si è dotata di un Sistema di Gestione della Sicurezza delle Informazioni certificato secondo lo standard ISO/IEC 27001, avendo conseguito nel 2012 la prima certificazione di conformità ISO/IEC 27001:2005 e a Marzo 2015 la ricertificazione secondo la nuova versione ISO/IEC 27001: 2013.

1.2 Organizzazione aziendale

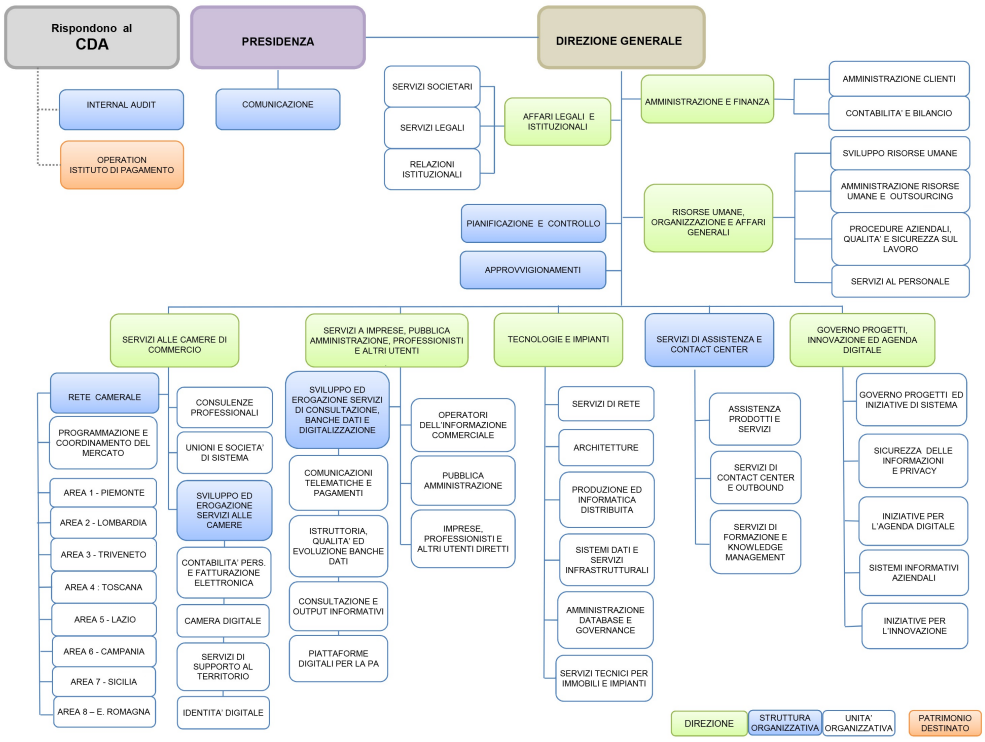


Figura 1.2: Organigramma aziendale

In InfoCamere S.C.p.A. è possibile individuare 4 aree direzionali di maggior interesse:

- Servizi alle Camere di Commercio;
- Servizi a imprese, Pubblica Amministrazione, professionisti e altri utenti;
- Tecnologie e impianti;
- Governo progetti, innovazione ed azienda digitale.

Nello specifico, durante lo stage, ho preso parte all'area direzionale "Servizi alle Camere di Commercio". In quest'ultima è possibile individuare:

- Area commerciale: si occupa degli accordi commerciali con le Camere di Commercio;

- Sviluppo ed erogazione servizi alle Camere: attua quanto accordato dall'area commerciale.

In particolare, sono stato assegnato all'unità organizzativa denominata "Camera Digitale", che risponde allo "Sviluppo ed erogazione servizi alle Camere". Questa unità organizzativa si occupa delle digitalizzazione delle Camere, sia per quanto riguarda la gestione documentale, rispettando le norme riguardanti la conservazione dei documenti, sia per quanto riguarda i siti web informativi delle Camere di Commercio.

1.3 Processi aziendali

In questa sezione presenterò i processi aziendali che riguardano l'unità organizzativa "Camera Digitale", con la quale ho avuto modo di confrontarmi.

1.3.1 La fornitura

L'azienda mette a disposizione dei clienti due differenti tipologie di prodotto e, nello specifico, di siti web:

- Listino: rappresentano prodotti la cui forma, contenuto e funzione siano idonei alla replicazione;

- Commessa: rappresentano prodotti la cui forma, contenuto e funzione vengono fissate dal cliente.

La contrattazione con il cliente viene interamente gestita dall'area commerciale.
Se la tipologia di prodotto scelta dal cliente è di tipo "Commessa", si opera una raccolta dei requisiti a cui segue una proposta commerciale al cliente.
In entrambe le tipologie di prodotto, quando l'offerta è stata concordata, la Camera di Commercio si occuperà dell'approvazione mediante una delibera pubblica.
Una volta avvenuto l'ingaggio, se il prodotto appartiene alla tipologia "Listino", verrà assegnato al nuovo prodotto un codice identificativo generico; in caso contrario, se la tipologia è "Commessa", al prodotto sarà invece attribuito un codice specifico.
A seguito dell'assegnazione di un nuovo codice identificativo del prodotto da realizzare, è possibile iniziare a sviluppare il sito web.

1.3.2 Ciclo di vita dei siti web

Il ciclo di vita dei siti web prodotti da InfoCamere, è così formato:

- Sviluppo: avviene lo sviluppo dei siti web e delle relative funzionalità; comprende file system e database dedicati allo sviluppo, oltre al software che dovrà essere distribuito;

- Test: segue lo sviluppo, dal quale prende in input il software prodotto. Il database è lo stesso che era presente in Produzione il giorno precedente a quello considerato;

- Produzione: contiene il software che ha superato i test, e il database è popolato con dati inseriti direttamente dalla Camera di Commercio.

1.3.3 Auditing

A seguito del caricamento dei contenuti, da parte della Camera, nel sito sviluppato, parteciperanno al collaudo: il commerciale di InfoCamere che ha seguito il lavoro, un referente tecnico di InfoCamere e infine un referente tecnico della Camera di Commercio. Il prodotto di questo processo sarà un verbale del collaudo che dovrà essere firmato e protocollato, oltre a presentare la firma dei due referenti tecnici.

1.3.4 Manutenzione

A seguito della protocollazione del verbale prodotto dal collaudo, si può considerare concluso lo sviluppo del sito e prende avvio l'assistenza al prodotto.

Tipologie di manutenzione

La manutenzione può essere di 3 differenti tipologie:

- Correttiva: vengono corretti bug presenti nel prodotto;
- Adattativa: i siti vengono adeguati a nuove norme in vigore;
- Evolutiva: vengono aggiunte funzionalità al sito; il prodotto subisce un'evoluzione: la natura del prodotto cambia in modo radicale, mantenendo però il prodotto stesso.

Gestione dei ticket

L'assistenza al prodotto viene gestita a più livelli, a cui affluiscono più gruppi di lavoro. Un nuovo ticket, prodotto di una segnalazione di un cliente, può essere gestito in modo autonomo dal primo livello o può essere assegnato ad un gruppo di lavoro appartenente al livello successivo.
Il gruppo di lavoro a cui viene assegnato il nuovo ticket può:

- Decidere di rifiutare il ticket: in questo caso, il ticket dovrà essere assegnato ad un altro gruppo di lavoro del secondo livello;
- Rimandare ad un altro gruppo di lavoro il ticket, nel caso in cui il gruppo identificato sia ritenuto più adatto a risolvere il problema;
- Decidere di utilizzare le proprie competenze per analizzare e cercare di risolvere il problema, fornendo eventualmente una soluzione e un tempo atteso.

I ticket che richiedono una manutenzione evolutiva, comporteranno inoltre uno studio di fattibilità, fatto dal gruppo del secondo livello al quale il ticket è stato assegnato, e un successivo preventivo orario ed economico.
Per i siti web, è garantito ai clienti un pacchetto di ore di supporto specificatamente per i ticket di tipo implementativo. Se il numero di ore preventivato è compreso nelle ore residue del pacchetto garantite al cliente, l'implementazione delle nuove funzionalità può avere luogo. In caso contrario, è necessaria la figura del commerciale per gestire la richiesta di manutenzione evolutiva.

1.4 Tecnologie utilizzate

Questa sezione presenterà le tecnologie utilizzate dall'azienda, ristrette all'ambito in cui ho operato.

1.5 Rapporto con l'innovazione

Questa sezione presenterà il rapporto dell'azienda con l'avanzamento tecnologico.

Capitolo 2

Lo stage

2.1 Gli stage in azienda

Questa sezione presenterà le motivazioni per cui l'azienda ospita stage.

2.2 L'offerta di stage

2.2.1 Presentazione del progetto

Qui verrà presentato il progetto di stage.

2.2.2 Aspettative aziendali

Qui verranno presentate le aspettative aziendali legate all'offerta di stage.

2.2.3 Vincoli

Qui verranno presentati i vincoli legati allo stage.

2.3 Vantaggi personali

In questa sezione descriverò le aspettative personali legate all'accettazione dell'offerta di stage.

Capitolo 3

Resoconto dello stage

3.1 Individuazione dei motori di ricerca

In questa sezione motiverò la selezione di Solr e ElasticSearch come motori di ricerca da studiare.

3.2 Pianificazione

In questa sezione presenterò come è avvenuta la pianificazione dello stage.

3.3 I siti istituzionali delle Camere di Commercio

3.3.1 Funzionalità di ricerca attuali

Questa sezione presenterà la struttura dei siti camerali attualmente in produzione, ponendo l'accento sugli strumenti messi a disposizione all'utente per ritrovare i contenuti in esso presenti.

3.3.2 Possibile evoluzione

Questa sezione presenterà una possibile evoluzione dei siti camerali, contenente funzionalità di ricerca attualmente non presenti.

3.4 Ricerca nativa Drupal

3.4.1 Introduzione a Drupal

Qui verrà introdotto Drupal, spiegando cos'è e come funziona.

3.4.2 Ricerca di base e avanzata

Qui verranno presentate le principali funzionalità offerte dalla prima tipologia di ricerca nativa Drupal.

3.4.3 Ricerca con Search API

Qui verranno presentate le principali funzionalità offerte dalla seconda tipologia di ricerca nativa Drupal.

3.4.4 Considerazioni di Drupal nativo

Questa sezione conterrà conclusioni riguardanti le funzionalità di ricerca offerte globalmente dalla ricerca nativa Drupal.

3.5 Ricerca con Solr

3.5.1 Introduzione a Solr

Qui verrà introdotto Solr, spiegando cos'è e come funziona.

3.5.2 Principali funzionalità di ricerca

Qui verranno presentate le principali funzionalità di ricerca offerte dal motore di ricerca Solr, di possibile interesse per i siti camerali.

3.5.3 Integrazione con Drupal

Qui verranno discusse le funzionalità di ricerca derivanti dall'integrazione tra Solr e Drupal.

Apache Solr

Qui verranno discusse le funzionalità di ricerca derivanti dall'integrazione tra Solr e Drupal mediante il modulo Apache Solr Search.

Search API Solr

Qui verranno discusse le funzionalità di ricerca derivanti dall'integrazione tra Solr e Drupal mediante il modulo Search API Solr Search.

3.6 Ricerca con ElasticSearch

3.6.1 Introduzione a ElasticSearch

Qui verrà introdotto ElasticSearch, spiegando cos'è e come funziona.

3.6.2 Principali funzionalità di ricerca

Qui verranno presentate le principali funzionalità di ricerca offerte dal motore di ricerca ElasticSearch, di possibile interesse per i siti camerali.

3.6.3 Integrazione con Drupal

Qui verranno discusse le funzionalità di ricerca derivanti dall'integrazione tra Elastic-Search e Drupal.

Search API ElasticSearch

Qui verranno discusse le funzionalità di ricerca derivanti dall'integrazione tra ElasticSearch e Drupal mediante il modulo Search API ElasticSearch.

3.7 Considerazioni finali sui motori di ricerca esaminati

Questa sezione conterrà un confronto tra le principali funzionalità, possibilmente di interesse per l'azienda, offerte dalle tecnologie esaminate e quale di queste potrebbe essere la più adatta ai siti camerali.

Capitolo 4

Valutazione retrospettiva

4.1 Bilancio degli obiettivi

4.1.1 Aziendali

Questa sezione descriverà gli obiettivi aziendali soddisfatti, derivanti dallo stage.

4.1.2 Personali

Questa sezione descriverà gli obiettivi personali soddisfatti, derivanti dallo stage.

4.2 Conoscenze acquisite

Questa sezione descriverà le conoscenze acquisite derivanti dallo stage.

4.3 Mondo del lavoro e università a confronto

Questa sezione analizzerà il gap tra gli insegnamenti universitari e il mondo dello stage, specificatamente allo stage svolto.

Glossario

API Insieme di procedure utilizzabili per interfacciarsi con un programma o un sistema informatico in modo standard. Spesso si intendono le librerie software disponibili in un certo linguaggio di programmazione. . 11–13

CMS E' un software per la realizzazione e la gestione di siti dinamici, che possono accrescere e mutare il proprio contenuto continuamente. Un Content Management System consente al committente del sito di occuparsi direttamente della sua gestione senza intermediari esterni. 11, 13

Drupal Drupal è un Content Management System, rilasciato sotto licenza open source, che permette la creazione di siti Internet, blog e portali, gallerie di immagini, forum di discussione, piattaforme intranet e molto altro. Essa è altresì un'applicazione completamente web based e può quindi essere utilizzata attraverso un semplice browser.
E' interamente sviluppato in PHP e utilizza come base di dati MySQL in modo nativo. v, 11

ElasticSearch Piattaforma di ricerca open source, con capacità full text. E' un server di ricerca basato su Java Lucene e supporta architetture distribuite. Tutte le funzionalità sono nativamente esposte tramite interfaccia RESTful; le informazioni sono invece gestite come documenti JSON. v, 11

HTTP Formato adatto all'interscambio di dati fra applicazioni client-server. 12, 13

Java Linguaggio di programmazione ad alto livello, orientato agli oggetti e a tipizzazione statica, specificatamente progettato per essere il più possibile indipendente dalla piattaforma di esecuzione. 11, 12

Java Lucene API gratuita ed open source per il reperimento di informazioni, inizialmente implementata in Java. 11, 12

JSON Formato adatto all'interscambio di dati fra applicazioni client-server. 11–13

Open source Software di cui i detentori dei diritti rendono pubblico il codice sorgente, permettendo ad altri programmatori di apportarvi modifiche. Questo meccanismo è regolato tramite l'applicazione di apposite licenze d'uso. v, 11, 12

MySQL Database relazionale largamente diffuso, composto da un client a riga di comando e un server. 11

PHP Linguaggio di scripting interpretato. 11, 12

REST Stile architetturale che offre la possibilità di manipolare rappresentazioni testuali di risorse Web utilizzando un set predefinito di operazioni. 12

RESTful Le applicazioni basate su REST, si definiscono RESTful e utilizzano le richieste HTTP per inviare i dati (creazione e/o aggiornamento), effettuare query, modificare e cancellare i dati. In definitiva, REST utilizza HTTP per tutte e quattro le operazioni CRUD (Create / Read / Update / Delete). 11, 12

Servlet Oggetti scritti in linguaggio Java che operano all'interno di un server web oppure un server per applicazioni, permettendo la creazione di web applications. 12

Solr Piattaforma di ricerca open source. E' scritto in Java e viene eseguito come server di ricerca full text indipendente all'interno di un contenitore Servlet. Solr usa la libreria di ricerca Java Lucene per la ricerca e l'indicizzazione full text e mette a disposizione chiamate REST come ad esempio HTTP/ JSON e XML API che rendono semplice la comunicazione. v, 12

XML Metalinguaggio che consente la rappresentazione di documenti e dati strutturati su supporto digitale. 12, 13

Acronimi

API Application Programming Interface. 11

CMS Content Management System. 11

HTTP HyperText Transfer Protocol. 11

JSON JavaScript Object Notation. 11

XML eXtensible Markup Language. 12

Bibliografia