UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

Relazione

Verso un osservatorio nazionale del turismo - Integrazione automatica degli open data riguardanti le strutture ricettive regionali

Candidato: *Giacomo Gregori*

Relatore: *Prof. Andrea Marchetti*
Ing. Angelica Lo Duca

Correlatore: *Prof. Paolo Macchia*

Anno Accademico 2015-2016

Indice

1. Introduzione

Il settore turistico in Italia è da sempre uno dei più importanti settori economici del nostro paese. Da diversi anni, il turismo ha iniziato a servirsi di un mezzo che ha permesso di amplificare il suo sviluppo: il web. Grazie ad esso, chiunque può ottenere facilmente tutte le informazioni riguardanti una determinata meta turistica o anche prenotare alberghi, ristoranti, visite ai musei, aerei, ecc.

Questo ha permesso lo sviluppo di diversi portali web dove poter usufruire di tali servizi: dai più famosi Booking.com[1], TripAdvisor[2] fino a Expedia[3], Airbnb[4] o agli stessi Google Places[5] e GoogleMaps[6] che incorporano in loro molte, se non tutte, le funzioni dei sopracitati. A questi possiamo aggiungere i social network, in primis Facebook, che contengono tutte le informazioni per permettere al turista di organizzare il proprio viaggio.

Una delle limitazioni più importanti per tutti questi portali, che in molti casi contengono informazioni estremamente complete, è la mancata possibilità di usare liberamente i loro dati, in quanto proprietari. Per risolvere questa problematica, il nostro progetto si prefigge l'obiettivo di creare un portale web di riferimento per il turismo nazionale, utilizzando gli open data forniti dalle regioni e continuamente aggiornato attraverso delle procedure automatiche.

Questi open data sono spesso distribuiti attraverso diversi portali e in formati o strutture dati diverse tra loro. Per questo l'intento è di riunirli in un unico dataset con un unico data model e, oltre alla possibilità di visualizzare le informazioni sul portale, verrà reso disponibile nuovamente sotto forma di open data con una licenza che permetta la rielaborazione e la diffusione.

Nell'ambito del nostro progetto, sono numerosi i dati sui quali poter lavorare, per iniziare si è scelto di raccogliere quelli relativi alle strutture ricettive e di riunirli in

[1] http://www.booking.com
[2] https://www.tripadvisor.it/
[3] https://www.expedia.it/
[4] https://www.airbnb.it/
[5] https://developers.google.com/places/
[6] https://www.google.it/maps

un unico dataset, per poi visualizzarli su una mappa in quella che sarà una prima versione del nostro portale di riferimento sul turismo.

2. Il progetto

2.1 Introduzione

Il nostro progetto prevede la creazione di un portale web che possa aiutare lo sviluppo del turismo in Italia avvalendosi di open data. Per raggiungere tale scopo, abbiamo deciso di impostare il nostro lavoro partendo da un'analisi degli open data riguardanti le strutture ricettive forniti dalla pubblica amministrazione delle regioni e, da esse, creare un'applicazione web che permetta la visualizzazione su mappa della loro posizione e le loro informazioni.

Come vedremo, la concorrenza in questo ambiente è elevata, soprattutto da parte di siti che utilizzano dati proprietari. Ed è proprio questa una delle problematiche che abbiamo affrontato nell'arco della nostra analisi in preparazione allo sviluppo del progetto: la carenza di open data ben formati in contrapposizione ai numerosi siti proprietari.

Oltre a questo, abbiamo anche studiato quali strumenti siano i più utilizzati nell'ambito del turismo e abbiamo analizzato diversi siti web, italiani e non, per capirne la struttura, l'efficienza e l'utilità al turista in cerca di informazioni per il proprio viaggio.

2.2 Il turismo oggi

2.2.1 Il turismo in Italia

Nel 2015, il turismo ha subito un netto miglioramento quantificabile nel numero di arrivi internazionali. Questi infatti hanno avuto un incremento di circa 52 milioni di visitatori, passando da un miliardo e 134 milioni del 2014 a un miliardo e 186 del 2015.

Per l'Italia, il turismo è sicuramente uno dei settori economici di maggior rilievo e che più incide sull'economia del nostro paese. Come possiamo vedere nella Tabella 1, ad oggi l'Italia occupa il quinto posto nella graduatoria degli arrivi internazionali

con 50,7 milioni, dietro la la Francia (84,5), gli Stati Uniti (77,5), la Spagna (68,2) e la Cina (56,9).

Pos.	Nazione	2014	2015	Variazione
1	Francia	83,7	84,5	+0,9%
2	U.S.A.	75,0	77,5	+3,3%
3	Spagna	64,9	68,2	+5,0%
4	Cina	55,6	56,9	+2,3%
5	***Italia***	***48,6***	***50,7***	***+4,4%***
6	Turchia	39,8	n.d.	n.d.
7	Germania	33,0	35,0	+6,0%
8	Regno Unito	32,6	n.d.	n.d.
9	Messico	29,3	32,1	+9,5%
10	Russia	29,8	31,3	+5,0%

Tabella 1. Graduatoria 2015 delle destinazioni turistiche mondiali più frequentate dal turismo straniero (UNWTO World Tourism Barometer, vol.14 - Luglio 2016)

Il settore turistico in Italia ha un impatto economico di rilievo: secondo le stime della WTTC (World Travel & Tourism Council), l'Italia ha incassato solo nel 2015 167,5 miliardi di Euro, circa il 10,2% del Prodotto Interno Lordo nazionale. Anche dal punto di vista occupazionale, ha la sua notevole importanza. Grazie al settore turistico, 2.609.000 lavoratori hanno un impiego, circa l'11,6% di incidenza sull'intera occupazione nazionale.

In questo ambito, possiamo dare uno sguardo al numero di strutture ricettive presenti sul territorio italiano. La Tabella 2 ci mostra la quantità di strutture presenti in Italia, con i relativi posti letto, suddivise per tipologia ricettiva.

	Tipologia ricettiva	Numero di esercizi	Posti letto
Esercizi alberghieri	Alberghi a 5 stelle e stelle lusso	442	70.735
	Alberghi a 4 stelle	5.609	755.629
	Alberghi a 3 stelle	15.355	957.580
	Alberghi a 2 stelle	6.014	191.933
	Alberghi a 1 stella	2.959	68.830
	Residenze turistico alberghiere	2.820	206.011
	Totale esercizi alberghieri	***33.199***	***2.250.718***
Esercizi extra-alberghieri	campeggi e villaggi turistici	2.708	1.365.661
	alloggi in affitto gestiti in forma imprenditoriale	73.075	610.641
	agriturismi	18.525	251.179
	ostelli per la gioventù	592	31.750
	case per ferie	2.325	132.976
	rifugi di montagna	1.091	33.878
	altri esercizi ricettivi n.a.c.	5.819	45.694
	bed and breakfast	30.384	156.836
	Totale esercizi extra-alberghieri	***134.519***	***2.628.615***
	TOTALE	***167.718***	***4.879.333***

Tabella 2. Esercizi alberghieri ed extra-alberghieri 2015 (ISTAT)

I dati della Tabella 2 evidenziano l'elevato numero di strutture ricettive localizzate in Italia: considerando il totale degli esercizi sia alberghieri che extra-alberghieri, abbiamo una media di 5,5 strutture ogni dieci chilometri quadrati o, sotto un altro punto di vista, circa 3 strutture ogni mille abitanti.

Per quanto riguarda gli introiti, come possiamo vedere nella Tabella 3, l'Italia occupa solamente il settimo posto, dietro alla Gran Bretagna, che può vantare solamente un terzo dei nostri siti Unesco, e dietro alla Thailandia che incassa dai visitatori stranieri cinque miliardi più di noi. Possiamo notare anche come l'Italia abbia subito una grande diminuzione degli introiti nel 2015: il 21,9% in meno rispetto all'anno precedente. Ma osservando la tabella, notiamo come sia una tendenza che coinvolge la maggior parte di questi stati, in quanto solo U.S.A., Cina e Thailandia mostrano un guadagno superiore nel 2015 rispetto al 2014.

Pos.	**Nazione**	**2014**	**2015**	**Variazione**
1	U.S.A.	177,2	178,3	+0,6%
2	Cina	105,4	114,1	+8,2%
3	Spagna	65,1	56,5	-13,2%
4	Francia	57,4	45,9	-20,0%
5	Thailandia	38,4	44,6	+16,1%
6	Regno Unito	46,6	42,4	-9,0%
7	***Italia***	***50,5***	***39,4***	***-21,9%***
8	Germania	43,3	36,9	-14,8%
9	Hong Kong (Cina)	38,4	35,9	-6,5%
10	Macao (Cina)	42,6	31,3	-26,5%

Tabella 3. Graduatoria 2015 degli introiti derivati dal turismo internazionale (UNWTO World Tourism Barometer, vol.14 - Luglio 2016)

Osservando i dati ISTAT del 2015 sui flussi turistici stranieri in Italia (Tabella 4) e confrontandoli con quelli degli anni precedenti, notiamo comunque un netto miglioramento. Nel 2008 si sono contati 41.796.724 arrivi, sette anni dopo, nel 2015, il conteggio è salito fino a poco più di 55 milioni con una variazione del 31,7%.

Anno	Arrivi	Variazione arrivi	Entrate (milioni di Euro)	Variazione entrate
2008	41.796.724	n.d.	31.090	n.d
2009	41.124.722	-1,6%	28.856	-7,2%
2010	43.794.338	+6,5%	29.257	+1,4%
2011	47.460.809	+8,4%	30.891	+5,6%
2012	48.738.575	+2,7%	32.056	+3,1%
2013	50.263.236	+3,1%	33.064	+3,6%
2014	51.635.500	+2,7%	34.240	+3,8%
2015	55.033.682	+3,1%	35.556	+3,0%

Tabella 4. Storico degli arrivi stranieri e delle relative entrate in Italia (ISTAT, 2016)

La Tabella 4 ci mostra anche che, come è facile aspettarsi, questo aumento negli arrivi si traduce in un maggior numero di entrate: se nel 2008 queste ammontavano a 31.090 milioni di Euro, nel 2015 sono salite fino a 35.556 milioni con una variazione totale del 14,4%.

E' interessante capire anche in quale percentuale i turisti siano provenienti dall'estero e quanti italiani. Un'elaborazione dell'ENIT, l'Agenzia Nazionale del Turismo, su dati ISTAT del 2015, ci mostra come il 49% dei turisti nel nostro paese sia italiano mentre il restante 51% provenga dall'estero. Il Grafico 1 ci mostra la percentuale di turisti italiani (in rosso) e stranieri (in azzurro) regione per regione.

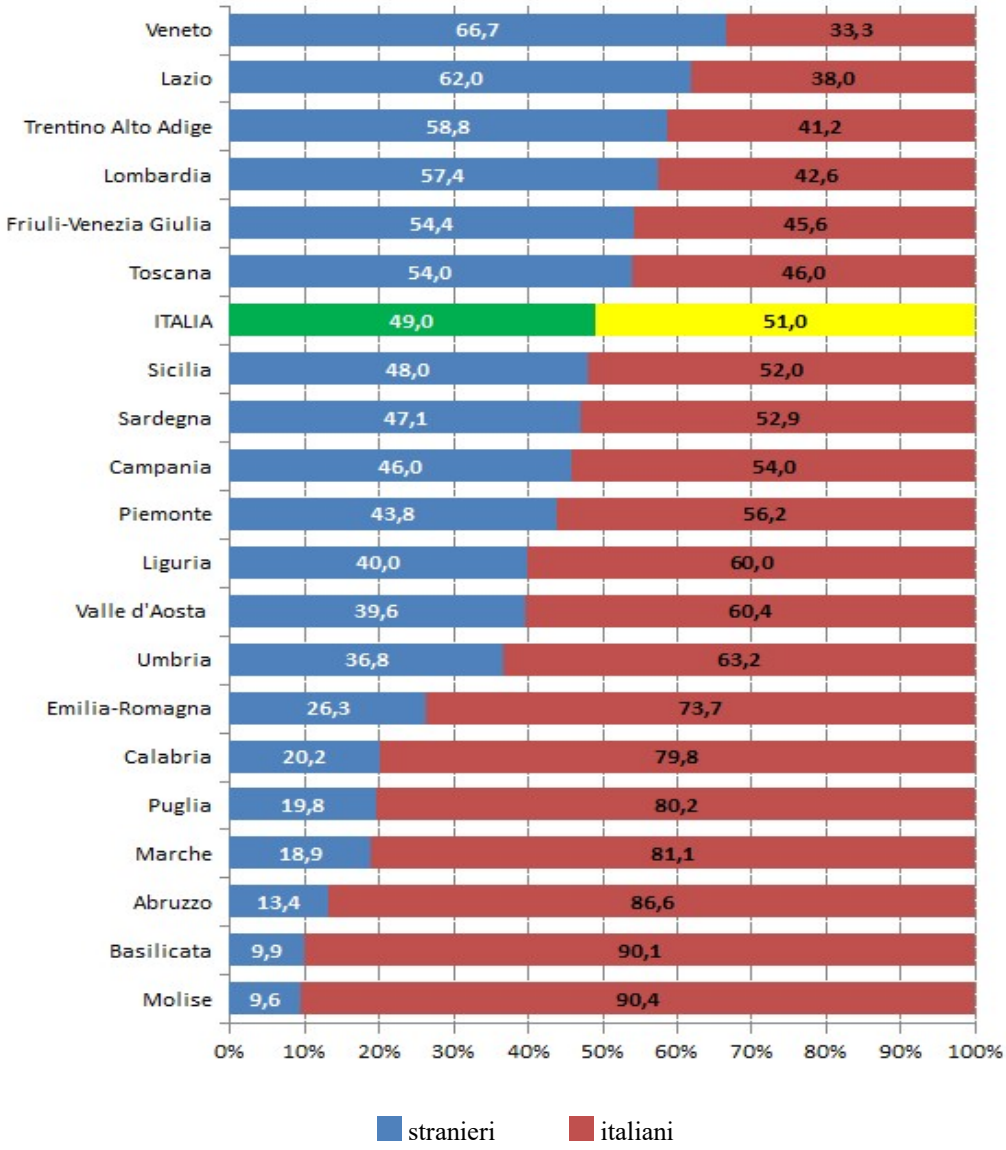

Grafico 1. Percentuale di turisti italiani o stranieri a livello nazionale e regionale

2.2.2 E-tourism

Il web è diventato, ormai da diversi anni, uno strumento molto utilizzato per il turismo. Grazie ad esso è possibile non solo trovare informazioni sull'eventuale meta dell'utente, ma anche prenotare voli, alberghi, noleggio auto, visite ai musei e molto altro.

Una ricerca del 2012 ha dimostrato che in Europa il 36% delle vacanze è stato prenotato online, in America il 39%. Di questi acquisti, il 40% sono stati fatti per

mezzo di OTA (*Online Travel Agency*), cioè agenzie di viaggio online. Solo il 9% delle prenotazioni sono state fatte direttamente agli alberghi, tour operator o attraverso le classiche agenzie di viaggio (Moreno, Hörhager, Schuster, Werthner, 2015).

Un esempio di questa situazione si ha in Austria dove, nel 2011, il 76% delle prenotazioni è stato effettuato online. Nelle nazioni che parlano lingua tedesca, due terzi di tutte le prenotazioni online sono gestite da due portali: Booking.com col 35% e HRS col 28% (Moreno, Hörhager, Schuster, Werthner, 2015). Entrambi i siti hanno una caratteristica che li accomuna e che ha contribuito al loro sviluppo: la possibilità, da parte degli utenti, di assegnare un voto alle strutture e di lasciare una recensione. Questo strumento ha permesso agli albergatori di monitorare la soddisfazione degli ospiti e ha offerto nuovi metodi per determinare le strategie di prezzo, mentre per gli utenti è un metodo per ricavare informazioni e decidere se pernottare in una struttura o meno.

In uno studio condotto dal SAS Institute (SAS Institute, 2014), volto a indagare il comportamento di prenotazione dei viaggiatori online, i risultati principali sono stati che:

- le recensioni online e il prezzo sono i fattori che più influenzano la scelta: sebbene i consumatori prestino attenzione alla categoria dell'hotel e in misura minore al marchio, le recensioni positive influenzano maggiormente la scelta di acquisto del consumatore seguite, a parità di valutazione, dal prezzo più basso.
- Le recensioni negative escludono l'hotel dalle possibili alternative del consumatore: un prezzo basso o un'elevata classificazione a stelle non bastano a rimediare a una reputazione negativa.
- I viaggiatori scelgono l'hotel con il prezzo più basso, a parità di punteggio di valutazione nelle recensioni
- I consumatori prestano attenzione solo al posizionamento e ai punteggi di valutazione elevati. Per questo è fondamentale impegnarsi a migliorare attivamente la propria reputazione online.

Questo studio rende evidente come le recensioni siano diventate uno dei fattori decisivi, se non l'unico, a influenzare le scelte di acquisto e, più in generale, come il web condizioni direttamente l'utente alla ricerca di una meta turistica e il turismo.

2.3 Open data

2.3.1 Definizione

Gli open data, comunemente chiamati col termine inglese open data, sono dei dati liberamente accessibili e rilasciati sotto una licenza, che generalmente ne consente un uso ampio. Qualora vengano utilizzati, l'utente ha l'obbligo, in molti casi, di citare le fonti e non può rendere tali dati proprietari.
Gli open data sono un prodotto derivato direttamente dagli open content. Infatti, mentre l'open content è una qualsiasi opera creativa o contenuto aperto, gli open data sono dati e ricerca scientifica. Altre forme di dati liberamente accessibili sono gli open source, codici informatici il cui accesso è concesso a chiunque, e i software liberi, cioè programmi scaricabili liberamente.
I dati, perché possano essere definiti aperti, necessitano di alcune caratteristiche:

- devono essere indicizzati dai motori di ricerca,
- devono essere disponibili in un formato aperto, standardizzato e leggibile da un'applicazione informatica per facilitare la loro consultazione ed incentivare il loro riutilizzo anche in modo creativo,
- devono essere rilasciati attraverso licenze libere che non impediscano la diffusione e il riutilizzo da parte di tutti i soggetti interessati.

La definizione di "dato aperto" è stata anche formalmente riconosciuta dalla legge italiana con la legge n.221 del 17 Dicembre 2012 inserendola all'interno dell'art. 68 del Codice dell'Amministrazione Digitale. Tale legge definisce come dati di tipo aperto quei dati che:

- a) sono disponibili secondo i termini di una licenza che ne permetta l'utilizzo da parte di chiunque, anche per finalità commerciali, in formato disaggregato;

- b) sono accessibili attraverso le tecnologie dell'informazione e della comunicazione, ivi comprese le reti telematiche pubbliche e private, in formati aperti ai sensi della lettera a), sono adatti all'utilizzo automatico da parte di programmi per elaboratori e sono provvisti dei relativi metadati;
- c) sono resi disponibili gratuitamente attraverso le tecnologie dell'informazione e della comunicazione, ivi comprese le reti telematiche pubbliche e private, oppure sono resi disponibili ai costi marginali sostenuti per la loro riproduzione e divulgazione.

2.3.2 Licenza

Una delle caratteristiche degli open data è la possibilità, da parte di tutti, di poter prelevare e utilizzare i dati che vengono messi a disposizione. Esistono diverse livelli di libertà nell'utilizzo di questi dati che vengono gestiti attraverso le licenze. Quelle più diffuse sono le Creative Commons.

Le licenze Creative Commons offrono sei diverse articolazioni dei diritti d'autore per artisti, giornalisti, docenti, istituzioni e, in genere, creatori che desiderino condividere in maniera ampia le proprie opere secondo il modello "alcuni diritti riservati". Il detentore dei diritti puo' non autorizzare a priori usi prevalentemente commerciali dell'opera (opzione Non commerciale, acronimo inglese: NC) o la creazione di opere derivate (Non opere derivate, acronimo: ND); e se sono possibili opere derivate, puo' imporre l'obbligo di rilasciarle con la stessa licenza dell'opera originaria.
Le licenze Creative Commons sono strutturate in due parti: le libertà e le condizioni di utilizzo dell'opera.

Le prime si suddividono a loro volta in due tipologie:

- libertà di condivisione: libertà di copiare, distribuire o trasmettere l'opera,
- libertà di rielaborazione: libertà di riadattare l'opera.

Le condizioni di utilizzo dell'opera sono invece di quattro diversi tipo:

- Attribuzione: permette che altri copino, distribuiscano, mostrino ed eseguano copie dell'opera e dei lavori derivati da questa a patto che venga indicato l'autore dell'opera, con le modalità da questi specificate,
- Non commerciale: permette che altri copino, distribuiscano, mostrino ed eseguano copie dell'opera e lavori derivati da essa o sue rielaborazioni, solo per scopi non commerciali,
- Non opere derivate: permette che altri copino, distribuiscano, mostrino ed eseguano soltanto copie identiche dell'opera; non sono ammesse opere derivate o sue rielaborazioni,
- Condividi allo stesso modo: permette che altri distribuiscano lavori derivati dall'opera solo con una licenza identica (non maggiormente restrittiva) o compatibile con quella concessa con l'opera originale.

Ognuna di queste quattro clausole individua una condizione particolare a cui il fruitore dell'opera deve sottostare per poterne usufruire liberamente. Combinandole si ottengono sedici possibili combinazioni, di cui undici sono licenze CC valide mentre le altre cinque non lo sono. Di queste ultime, quattro includono sia la clausola ND (Non opere derivate) sia quella SA (Condividi allo stesso modo) che sono mutuamente esclusive, mentre una non è valida perché non contiene né la ND né la SA.

Delle undici combinazioni valide, le cinque che non presentano la clausola BY (Attribuzione) sono state ritirate perché richieste da meno del 3% degli utenti; rimangono tuttavia disponibili per la consultazione sul sito di Creative Commons. Quindi le licenze Creative Commons in uso sono sei più la CC0 (o pubblico dominio) e sono spiegate nella Tabella 5.

Licenza	Descrizione
CC BY	Permette di distribuire, modificare, creare opere derivate dall'originale, anche a scopi commerciali, a condizione che venga riconosciuta una menzione di paternità adeguata, fornito un link alla licenza e indicato se sono state effettuate delle modifiche.

CC BY-SA	Permette di distribuire, modificare, creare opere derivate dall'originale, anche a scopi commerciali, a condizione che venga riconosciuta una menzione di paternità adeguata, fornito un link alla licenza e indicato se sono state effettuate delle modifiche inoltre alla nuova opera deve essere attribuita la stessa licenza dell'originale.
CC BY-ND	Permette di distribuire l'opera originale senza alcuna modifica, anche a scopi commerciali, a condizione che venga riconosciuta una menzione di paternità adeguata e venga fornito un link alla licenza. Quindi non possono essere distribuite opere modificate o basate sull'opera licenziata con questa licenza.
CC BY-NC	Permette di distribuire, modificare, creare opere derivate dall'originale, ma non a scopi commerciali, a condizione che venga riconosciuta una menzione di paternità adeguata, fornito un link alla licenza e indicato se sono state effettuate delle modifiche. Chi modifica l'opera originale non è tenuto ad utilizzare le stesse licenze per le opere derivate.
CC BY-NC SA	Permette di distribuire, modificare, creare opere derivate dall'originale, ma non a scopi commerciali, a condizione che venga riconosciuta una menzione di paternità adeguata, fornito un link alla licenza, indicato se sono state effettuate delle modifiche e che alla nuova opera venga attribuita la stessa licenza dell'originale.
CC BY-NC ND	Questa licenza è la più restrittiva: consente soltanto di scaricare e condividere i lavori originali a condizione che non vengano modificati né utilizzati a scopi commerciali, sempre attribuendo la paternità dell'opera all'autore.

Tabella 5. Tipologie di licenza Creative Commons

Oltre alle Creative Commons, nell'arco del nostro progetto, abbiamo incontrato più di un dataset coperto dalla Italian Open Data License. Concettualmente questo tipo di licenza è molto simile alla CC-BY, descritta nella Tabella 5. L'Italian Open Data License è un contratto di licenza che ha lo scopo di consentire agli utenti di condividere, modificare, usare e riusare liberamente la banca di dati, i dati e le informazioni con essa rilasciati, garantendo al contempo la stessa libertà per altri. La licenza mira a facilitare il riutilizzo delle informazioni pubbliche nel contesto dello

sviluppo della società dell'informazione. L'utente è libero di riprodurre, distribuire al pubblico, presentare e dimostrare in pubblico, mettere a disposizione del pubblico e creare un Lavoro derivato a patto di:

- indicare la fonte delle Informazioni e il nome del Licenziante,
- non riutilizzare le Informazioni dando loro un carattere di ufficialità o inducendo l'utente a credere che il Licenziante approvi l'uso che fai delle Informazioni,
- prendere ogni misura ragionevole affinché gli usi innanzi consentiti non traggano in inganno altri soggetti e le Informazioni medesime non vengano travisate.

2.4 Stato dell'arte

Lo stato dell'arte è un'analisi accurata di applicazioni web che trattano la stessa tematica del nostro progetto. In preparazione al nostro progetto, abbiamo studiato i servizi forniti da queste applicazioni per capire quale fossero i punti di forza e le mancanze da colmare in questa tipologia di siti web.

2.4.1 Portali turistici

Per prima cosa abbiamo analizzato il sito ufficiale italiano del turismo, www.italia.it. Il sito si presenta ben curato graficamente. Al suo interno troviamo una sezione per ogni regione in cui sono descritte le particolarità e cosa può proporre ognuna di esse. Oltre a questo, troviamo delle idee di viaggio suddivise per tematiche, purtroppo sono tutte molto vaghe e prive di particolari. Nella sezione "info" è possibile trovare diverse informazioni sul nostro paese, sia per quanto riguarda mezzi di trasporto, documenti, clima, sia per la prenotazione online di musei. Infine abbiamo la sezione "mappe" dove è possibile visualizzare la posizione di diverse attrazioni turistiche suddivise per categorie di appartenenza. Purtroppo nel sito non vi è nessuna traccia di informazioni riguardo a dove poter alloggiare o ristoranti, mancanza non da poco visto il ruolo di riferimento per il turismo in Italia che vorrebbe ricoprire il portale. In

definitiva il sito non risulta essere adeguato nei contenuti tralasciando informazioni necessarie.

2.4.2 Portali open

Dal punto di vista dei contenuti, OpenStreetMap[7] è estremamente completo. In questo portale è possibile trovare un elevato numero di strutture ricettive, ristoranti, bar, pizzerie, ecc, e di visualizzare la loro posizione su mappa effettuando una ricerca. Purtroppo le informazioni riguardanti i punti di interesse italiani non sono complete in quanto solamente quelli più famosi delle maggiori città italiane contengono dati oltre al nome. Questa è una mancanza importante poiché è probabile che l'utente voglia, una volta trovata la struttura, conoscerne il numero di telefono o l'indirizzo e-mail per stabilire un contatto o altre informazioni quali il numero di stelle di un albergo, l'orario di apertura di un bar, ecc. OpenStreetMap ottiene i dati direttamente dalla community, infatti chiunque può aggiungere o modificare quelli presenti, e li rende disponibili per il riutilizzo, a patto di attribuirli a loro e, in caso di modifica, di distribuirli mediante la stessa licenza. Essendo questi dati forniti dagli utenti, è più facile che ricevano un aggiornamento continuo avendo una community attiva. Anche Wikidata, in quanto piattaforma in stile "Wiki", si serve dei dati forniti e modificati direttamente dagli utenti. Wikidata[8] è un database libero, collaborativo e multilingue che raccoglie dati strutturati. I dati pubblicati sono pubblicati sotto la licenza Creative Commons Public Domain Dedication 1.0 che ne permette il riutilizzo in numerosissimi ambiti. Abbiamo analizzato questo portale in quanto fornitore di open data e come ulteriore esempio di community attiva che contribuisce allo sviluppo dei dati anche se quest'ultimi rimangono non ufficiali. Dal punto di vista del turismo, Wikidata non fornisce molto dati. Ad esempio, cercando gli alberghi italiani troviamo pochi riferimenti a strutture con, al massimo, inseriti il nome, l'indirizzo e qualche foto.

[7] https://www.openstreetmap.org/
[8] https://www.wikidata.org/

Un progetto simile al nostro è quello di TouriNet[9]. Esso ha lo scopo di condurre ricerca e sviluppo finalizzati alla definizione di nuove tecnologie per migliorare il business delle imprese del turismo, valorizzando la loro reputazione sul web e promuovendo la creazione di sinergie tra attività di diverso tipo. Queste nuove tecnologie sfrutteranno la grande mole di informazioni che è presente sul web in modo disorganizzato ed eterogeneo.

TouriNet sfrutterà tecniche di *data extraction* automatiche e manuali sfruttando anche open data e API. Una volta ottenuti i dati passeranno alla fase di *data integration* per migliorarne la qualità.

Il progetto è attualmente in sviluppo e la sua conclusione è prevista per Giugno 2017. Ad oggi non è quindi possibile verificare l'effettiva solidità del progetto e come esso sia sviluppato.

Nel progetto TouriNet verranno sfruttate sorgenti come TripAdvisor e Booking.com che sono proprietari dei dati. Quindi TouriNet utilizzerà dati proprietari la cui licenza ristretta non permette un ampio utilizzo, limitandone le possibilità. Il nostro progetto, al contrario, prevede l'utilizzo di dati aperti.

2.4.3 Agenzie di viaggio online

TripAdvisor[10] è invece un esempio di portale appositamente creato per i turisti. In esso sono contenuti i dati di un gran numero di strutture ricettive e ristoranti. Per ognuno di essi l'utente può lasciare una recensione scritta e un voto da uno a cinque a seconda del livello di gradimento. Di ogni punto d'interesse vengono forniti tutti i dati necessari comprese, talvolta, le tariffe, e viene data la possibilità di visualizzare su mappa la loro posizione. Sicuramente TripAdvisor può essere un importante punto di riferimento per il nostro progetto, nonostante questo utilizzi dati proprietari e quindi non aperti, caratteristica fondamentale del nostro portale. Lo stesso si può dire di tutti i siti di prenotazione online come Booking.com, Expedia, Airbnb che, in quanto possessori dei dati che pubblicano, non li mettono liberamente a disposizione per l'utilizzo e la rielaborazione.

[9] http://www.tourinet.it/

[10] https://www.tripadvisor.it/

Un discorso a parte va fatto per GoogleMaps[11] in quanto non è un portale per la prenotazione online, ma un'applicazione web che dà la possibilità di visualizzare su mappa i dati relativi a diversi punti di interesse. Le informazioni reperibili, anche in questo caso proprietarie, sono molto dettagliate, come per i siti analizzati precedentemente, vi è la possibilità, da parte dell'utente, di scrivere recensioni e lasciare un voto, oltre a collegarsi con portali come Booking.com per effettuare direttamente le prenotazioni. GoogleMaps è sicuramente il miglior punto di riferimento per il nostro progetto in quanto racchiude in sé tutte le caratteristiche che vorremmo proporre agli utenti, con in più la possibilità di utilizzare e rielaborare i nostri dati.

2.4.4 Social network

Infine abbiamo analizzato i social network, i quali sono tutti possessori dei dati pubblicati, in cui i gestori di strutture inseriscono i propri dati per pubblicizzare la propria attività. Abbiamo preso in esame Facebook, Foursquare e Google Places. All'interno di Facebook[12] sono diverse le pagine dedicate alle strutture ricettive ma, non essendo la funzione principale di questo social network, la ricerca non è facile all'interno di esse. Infatti, a meno che non si conosca il nome della struttura da ricercare, è difficile trovare quelle di una determinata città o regione, in quanto la ricerca si basa semplicemente sul nome della pagina. In ogni caso, una volta trovata la struttura desiderata, le informazioni non mancano in quanto sono i gestori in prima persona ad inserirle. Utile è anche la possibilità, da parte degli utenti, di inserire una recensione basata sulla loro esperienza personale nella struttura. Inoltre vi è la possibilità di seguire la pagina mettendo un "mi piace", così che l'utente possa venire a conoscenza di eventuali offerte e informazioni qualora il gestore decidesse di pubblicarle.

Foursquare[13] è un'applicazione Web, utilizzabile tramite browser o via cellulare, che consente di segnalare la propria posizione alle persone del nostro network. Chi è

[11] https://www.google.it/maps
[12] https://www.facebook.com/
[13] https://it.foursquare.com/

iscritto a Foursquare può comunicare in tempo reale l'esatta posizione in cui si trova, con particolare riferimento a negozi, attività commerciali, luoghi di interesse artistico o culturale. Individuato il luogo, gli iscritti a Foursquare possono ricevere consigli e suggerimenti dagli altri utenti che lo hanno visitato: se siete in un ristorante riceverete consigli sui piatti da mangiare, se siete in un museo quelli su quali opere vedere e così via. Una caratteristica interessante è la possibilità di essere eletti "sindaco" di un punto d'interesse essendo l'utente che vi ha effettuato più volte il check-in (l'attività di registrarsi in un luogo). In America, dove Foursquare è molto diffuso, colui che è stato eletto "sindaco" di un luogo, può, in alcuni casi, ricevere sconti particolari. I luoghi registrati sono molti, ma dal punto di vista delle strutture ricettive è carente. Per fare un esempio, gli alberghi di Firenze registrati in Foursquare sono solo 90, contro i 384 di Booking.com. Va detto, però, che in caso di assenza di un punto d'interesse, l'utente può aggiungerlo per colmare la mancanza. Google Places[14] è uno strumento a disposizione dei gestori di attività per permettere agli utenti di Google di far conoscere il proprio business. Grazie ad esso è possibile creare una scheda della propria attività con tutte le informazioni necessarie, comprese foto (della struttura, ma anche street view), posizione, orario di apertura (con anche la possibilità di visualizzare se l'attività è aperta o meno), giorni di chiusura, ecc. Gli utenti possono lasciare recensioni e votare il loro livello di gradimento. Interessante è la possibilità di creare un tour virtuale per mostrare, attraverso una serie di foto dentro il quale ci si può muovere virtualmente, gli interni dei locali. Diversi sono anche gli strumenti a disposizione dei gestori che possono visualizzare una serie di statistiche (visualizzazioni, clic, numero di telefonate ricevute...) di aiuto per la gestione dell'attività. Sicuramente Google Places mostra diverse caratteristiche innovative e interessanti che aiutano sia l'utente che il gestore nel loro intento.

[14] https://developers.google.com/places/

2.4.5 Portali statistici

I portali statistici offrono un esempio di analisi e diffusione delle statistiche riguardanti il settore turistico. Nel nostro studio, abbiamo analizzato i siti dell'Agenzia Nazionale del Turismo - ENIT[15] e dell'ISNART - Istituto Nazionale Ricerche Turistiche[16].

Il primo mostra due aspetti strettamente collegati al nostro progetto: prima di tutto la sezione "Studi" dove è possibile trovare una serie di dati interessanti sulla situazione del settore turistico italiano. Questi dati sono supportati da grafici e tabelle che aiutano l'utente a comprendere meglio la situazione.

L'altra sezione strettamente collegata al nostro progetto è quella dell'Osservatorio Nazionale del Turismo. In questo portale è possibile trovare degli articoli riguardanti il settore turistico e diverse statistiche tra cui il trend di presenze, spese, entrate legate al turismo, mappe e grafici che mostrano la ricettività e altri aspetti del settore turistico. I dati utilizzati sono prelevati dal sito ISTAT e dalla Banca d'Italia per poi essere elaborati e diffusi. Tuttavia i dati non possono essere riutilizzati liberamente in quanto proprietari.

Inoltre, nonostante sia possibile trovare un buon quantitativo di dati statistici, non vi è traccia di dataset riguardanti strutture ricettive, punti d'interesse, ecc.

Il sito dell'Istituto Nazionale Ricerche Turistiche contiene al suo interno una banca dati che raccoglie una serie di ricerche ed indagini riguardanti il settore turistico. E' possibile fare una ricerca attraverso una selezione dell'argomento che l'utente vuole approfondire e una serie di altri campi per effettuare una ricerca maggiormente mirata. La banca dati contiene uno storico che permette di accedere, in alcuni casi, anche a dati risalenti a più di quindici anni fa.

Tuttavia, alcuni di questi dati sono scaricabili solo dopo aver effettuato un abbonamento annuale dal costo di 5.000€ che permetterà il completo accesso ai dati, ai dossier e alle ricerche presenti sul portale.

[15] http://enit.it/it/

[16] http://www.isnart.it/

Nome	*Tipologia*	*Tipologia Dati*	*Pro*	*Contro*
Agenzia Nazionale del Turismo - ENIT	Portale statistico	Proprietari	Buon quantitativo di dati statistici sul settore turistico	I dati sono proprietari, assenza di dataset riguardanti strutture o simili
Booking	Social Media	Proprietari	Prenotazione online, recensioni	I dati sono proprietari
Facebook	Social Network	Proprietari	Recensioni, contatto diretto con i gestori, dati di buona qualità	I dati sono proprietari, difficoltà di ricerca
Foursquare	Social Network	Proprietari	Interattività da parte degli utenti, possibilità di aggiungere dati da parte degli utenti	I dati sono proprietari, assenza di molti punti d'interesse
Istituto Nazionale Ricerche Turistiche - ISNART	Portale statistico	Proprietari	Buono storico delle statistiche sul settore turistico	I dati sono proprietari e solo parzialmente accessibili gratuitamente
Italia.it	Portale turistico	Proprietari	Informazioni generali sul paese, itinerari di viaggio	I dati sono proprietari e scarsi
GoogleMaps	Social Media	Proprietari	Elevato numero di punti d'interesse visualizzabili su mappa con dati eccellenti	I dati sono proprietari
Google Places	Social Network	Proprietari	Molte funzionalità innovative	I dati sono proprietari

OpenStreetMap	Portale open	Aperti	Numero elevato di punti d'interesse	Pochi dati relativi ai punti d'interesse, dati forniti dagli utenti
TouriNet	Portale open	Aperti	Open data, recensioni, strumento per turisti e imprese	Utilizzo di dati proprietari, progetto ancora in fase di sviluppo
TripAdvisor	Social Media	Proprietari	Prenotazione online, recensioni	I dati sono proprietari
Wikidata	Portale open	Aperti	Open data	Dati forniti dagli utenti, scarsità di dati sul turismo

Tabella 6. Riassunto dei siti analizzati nello stato dell'arte

3. Estrazione dei dati

3.1 Introduzione

L'estrazione dei dati (o *data mining*) è comunemente definita come il processo che porta alla scoperta di modelli o conoscenze da dati come database, testi, immagini, il web, ecc. Il modello deve essere valido, potenzialmente utile e comprensibile. Il data mining è un campo multidisciplinare che comprende l'apprendimento automatico (*machine learning*), la statistica, i database, l'intelligenza artificiale, il recupero delle informazioni e la loro visualizzazione.

Un'applicazione di data mining di solito inizia con la comprensione del dominio applicativo da parte dell'analista dei dati (*data miner*) che ricercherà i dati adatti per lo scopo. Una volta ottenuti i dati necessari, è possibile eseguire il data mining che si svolge in tre passi:

- Precompilazione: il dato puro solitamente non può essere utilizzato per eseguire il mining per diversi motivi. Può essere necessario una pulizia per rimuovere imperfezioni o piccoli errori. I dati potrebbero essere troppi grandi

con caratteristiche irrilevanti allo scopo finale, in tal caso è necessaria una semplificazione,

- Data mining: i dati elaborati vengono forniti ad un algoritmo di data mining che produrrà modelli e conoscenza,
- Postcompilazione: in molte applicazioni, non tutti i modelli trovati sono utili. In questo passo verranno identificati quelli utili allo scopo finale.

Questo processo è quasi sempre iterativo e molto spesso necessita di diverse applicazioni prima di ottenere un risultato soddisfacente.

Con lo sviluppo del Web e dei documenti di testo, sono diventate molto importanti le tecniche di *Web mining* e *text mining*.

Nel nostro caso, il data mining consiste nel raccogliere i dati relativi alle strutture ricettive da alcuni open data forniti dalle regioni per creare un dataset nazionale aperto. Pertanto, rispetto ai classici problemi di data mining, il nostro caso è leggermente semplificato, in quanto le regioni già forniscono i dataset, pronti per essere scaricati. Tuttavia, in un'ottica di integrazione, tali dati non possono essere utilizzati così come sono, in quanto molto diversificati tra di loro: essi necessitano di una procedura di mapping ed integrazione.

Il problema dell'estrazione dei dati si è articolato in quattro fasi:

1) raccolta dei dati: ricerca del portale relativo ad ogni regione e successiva ricerca all'interno del portale stesso di un eventuale dataset relativo alle strutture ricettive,
2) analisi statistica dei dati, culminata con la creazione di un portale web contenente le statistiche riassuntive relative alla situazione italiana,
3) disseminazione dei risultati, attraverso un contatto diretto con le regioni al fine di instaurare una collaborazione per integrare al meglio i dati ricavati,
4) aggiornamento dati: creazione di un software che permetta l'aggiornamento quotidiano del dataset nazionale.

Nella parte restante del presente capitolo, ogni fase verrà analizzata separatamente.

3.2 Raccolta dei dati

Come abbiamo detto, per prima cosa è stato necessario ricercare sul web i portali di open data delle diverse regioni italiane e, se esistenti, scaricare i dataset contenenti i dati sulle strutture ricettive. Per trovare tali portali si è utilizzato il motore di ricerca Google e analizzato i risultati per vedere quanto corrispondessero a ciò che stavamo cercando. Delle venti regioni la grande maggioranza ha un portale di open data di riferimento: solamente le Marche e la Sicilia non ne hanno uno ufficiale. Purtroppo, delle altre diciotto regioni, solamente dodici mettono a disposizione un dataset sulle strutture ricettive e con qualche limitazione: in quello della Basilicata è possibile trovare solo i dati riguardanti la provincia di Matera, mentre nel portale del Trentino-Alto Adige solamente quelli della provincia autonoma di Trento. Come vedremo in seguito, contattando la regione lucana, siamo riusciti a trovare una banca dati più completa che comprendesse anche informazioni sulle strutture ricettive dell'altra provincia, nonché capoluogo di regione, Potenza. Per quanto riguarda la Sardegna, abbiamo trovato solo un file in formato .pdf che, non essendo in formato *machine readable*, siamo stati costretti a scartare. La Tabella 7 riassume i risultati della fase di raccolta di dati:

Regione	Portale Open Data	Disponibilità Dataset	Formato
Abruzzo	opendata.regione.abruzzo.it	No	-
Basilicata	http://www.aptbasilicata.it	Sì	.xsl
Calabria	dati.reggiocal.it	No	-
Campania	opendatacampania.it	No	-
Emilia-Romagna	dati.emilia-romagna.it	Sì	.csv
Friuli-Venezia Giulia	dati.friuliveneziagiulia.it	Sì	.csv
Lazio	dati.lazio.it	No	-
Liguria	www.regione.liguria.it	Sì	.csv

Lombardia	dati.lombardia.it	Sì	.csv
Marche	goodpa.regione.marche.it	Sì	.csv
Molise	n/a	-	-
Piemonte	dati.piemonte.it	Sì	.csv
Puglia	dati.puglia.it	Sì	.csv
Sardegna	opendata.sardegna.it	Sì	.pdf*
Sicilia	n/a	-	-
Toscana	dati.toscana.it	Sì	.csv
Trentino-Alto Adige	dati.trentino.it	Solo prov. di Trento	.xml
Umbria	dati.umbria.it	Sì	.csv
Valle d'Aosta	www.regione.vda.it	No	-
Veneto	dati.veneto.it	Sì	.csv

* scartato in quanto non machine readable

Tabella 7. Risultati della raccolta dati

3.3 Analisi statistica dei dati

Una volta ottenuti la maggior quantità di dati possibili, è stato sviluppato un portale dove si riassumesse la situazione a livello italiano degli open data relativi alle strutture ricettive.

Per la grafica del sito è stato utilizzato un template di Bootstrap, modificato secondo le esigenze.

All'interno del portale è possibile trovare le sezioni "Numeri" e "Confronto" che mostrano, attraverso dei grafici, i risultati della ricerca statistica effettuata. Il Grafico 1 mostra la distribuzione delle varie strutture ricettive presenti, normalizzata ogni 1000 abitanti, mentre il Grafico 2 mostra la stessa distribuzione, normalizzata ogni 10 chilometri quadrati.

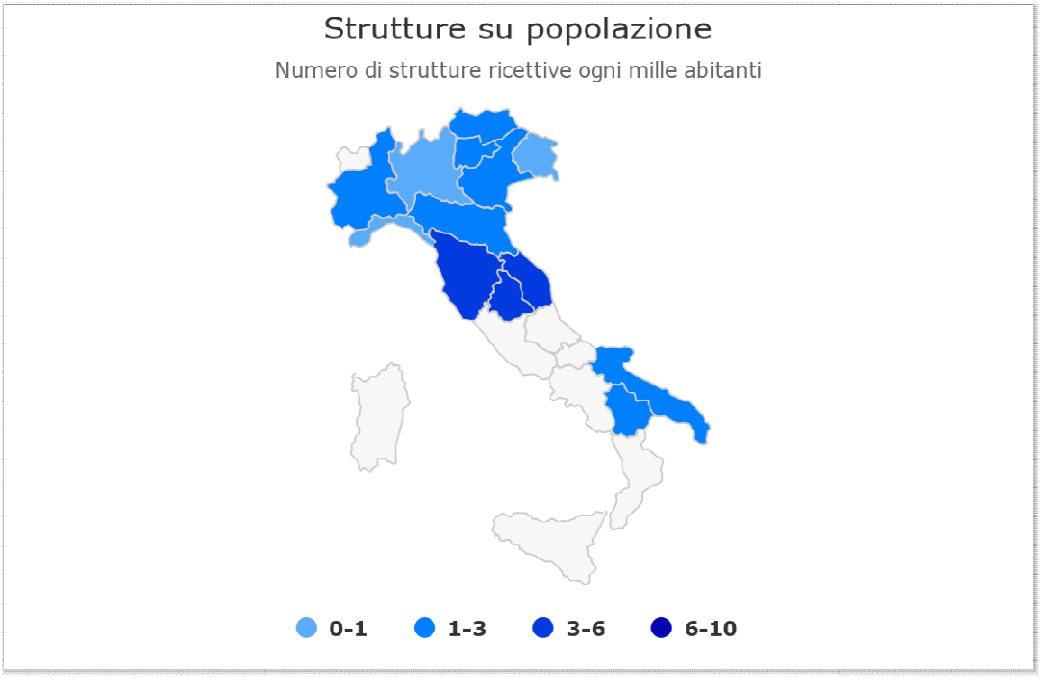

Grafico 1. Strutture ricettive negli open data ogni mille abitanti

Secondo il Grafico 1, le situazioni migliori sono quelle di Toscana, Umbria e Marche, mentre la Lombardia, la Liguria e il Friuli-Venezia Giulia sono quelle che mostrano una distribuzione minore.

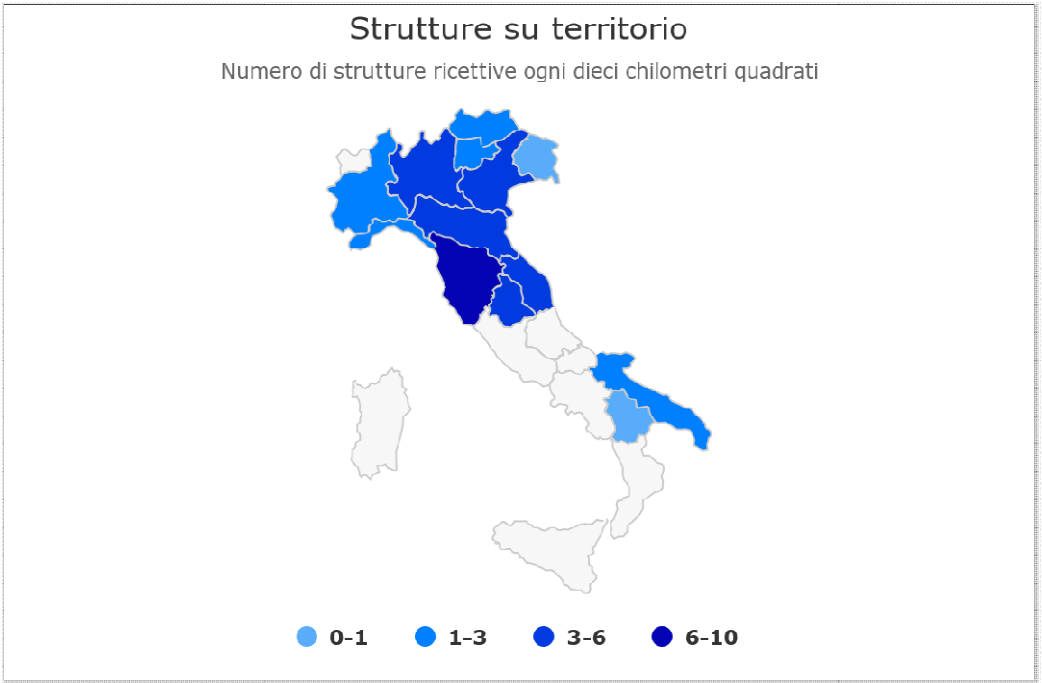

Grafico 2. Strutture ricettive negli open data ogni dieci chilometri quadrati

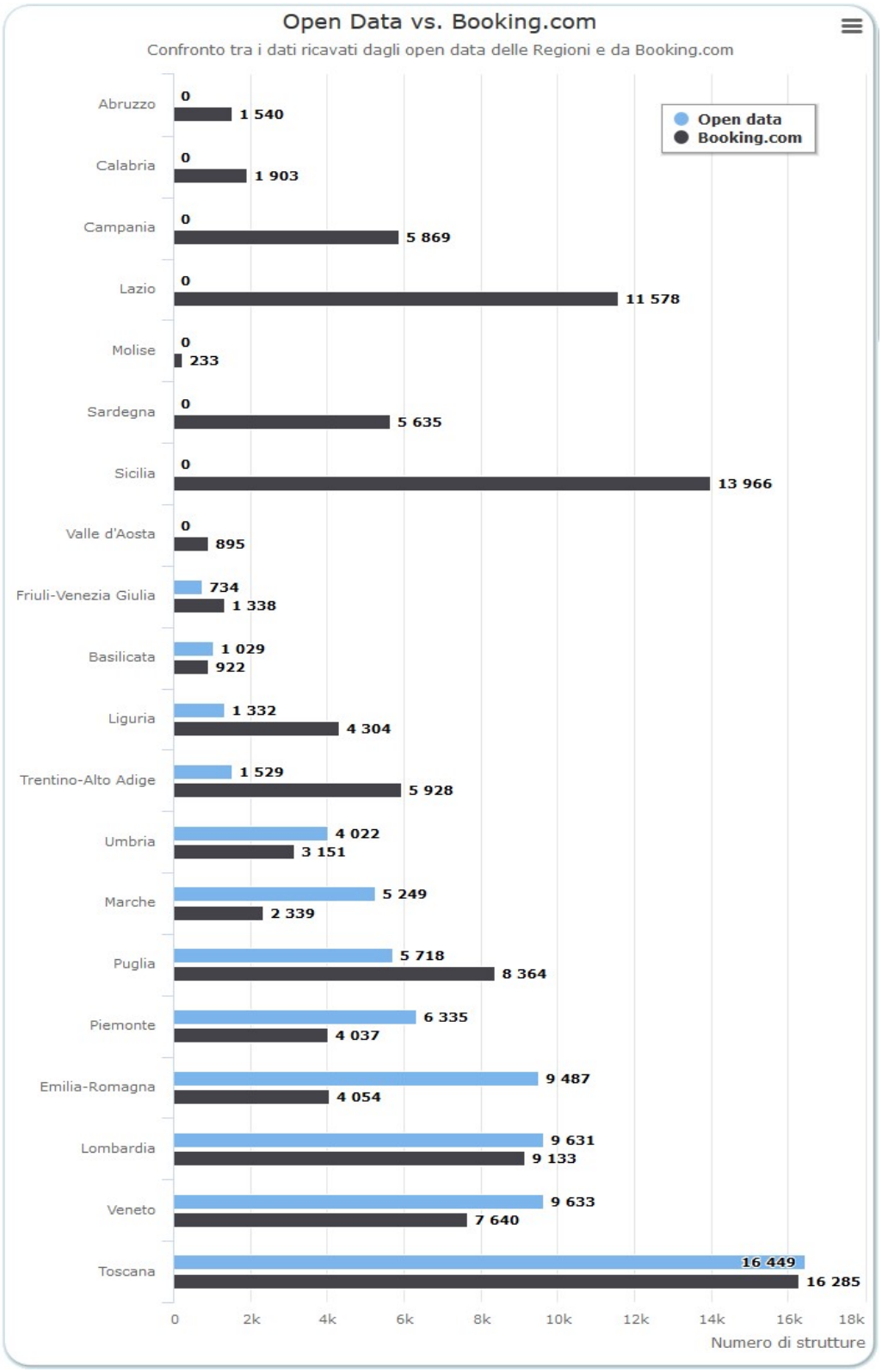

Grafico 3. Confronto tra le strutture presenti negli open data e quelle presenti su Booking.com

Stando al Grafico 2 la situazione cambia leggermente: la Toscana rimane la regione con la migliore densità, al contrario di Umbria e Marche che scendono nella fascia inferiore. Il Friuli-Venezia Giulia rimane una delle situazioni peggiori, ma stavolta è accompagnato in ultima fascia dalla Basilicata anziché Liguria e Lombardia che vedono migliorare la loro situazione.
Al fine di analizzare la completezza dei vari dataset, è stato fatto un confronto con i dati estratti dal portale Booking.com (Grafico 3). In particolare, per ogni regione, è stato estratto il numero totale di strutture ricettive presenti su Booking.com.
Utilizzando Booking.com come metro di paragone, possiamo valutare quanto attendibili siano gli open data delle regioni. Otto di questi contengono un maggior numero di informazioni riguardo alle strutture ricettive rispetto al famoso portale. Allo stesso tempo, Friuli-Venezia Giulia, Liguria, Puglia e Trentino-Alto Adige sembrano mostrare un notevole scarto tra le due statistiche rivelando una inadeguatezza di informazioni contenute negli open data. Le restanti regioni non hanno un open data.

3.4 Disseminazione dei risultati

Il portale web è stato creato per essere uno strumento da mettere a disposizione delle regioni per conoscere il progetto e sensibilizzarle ad un contributo nello sviluppo. Il passo successivo è stato quello di contattare le regioni o i responsabili degli open data alla ricerca di una collaborazione che permettesse di ottenere dati più aggiornati o, qualora mancassero completamente, un portale dove poterli scaricare.
Riportiamo, a titolo di esempio, la e-mail inviata al Dipartimento Turismo, Cultura e Paesaggio dell'Abruzzo:

Salve,
sono Giacomo Gregori, tesista presso **l'Istituto di Informatica e Telematica del CNR di Pisa**, sotto la supervisione del prof. Andrea Marchetti e dell'ing. Angelica Lo Duca, che leggono in CC.

Nell'ambito della mia attività di ricerca, sto cercando di costruire un **portale sul turismo** che contenga tutte le strutture ricettive delle varie regioni italiane.

L’obiettivo principale è quello di creare un **osservatorio nazionale sul turismo che contenga anche un portale open** che sia una valida alternativa a tutti i portali proprietari (booking.com, tripadvisor ecc), dove poter trovare un database open collettivo di tutta Italia e dei dati statistici riguardanti il turismo a livello regionale e nazionale.

Fino ad oggi ho raccolto gli open data di ogni regione (quando disponibile) e ho fatto un piccolo studio sulla loro reperibilità. Molti di questi dati si sono rivelati incompleti o non aggiornati. Tale studio può essere consultato sul portale http://tour-pedia.org/it.

Per quanto riguarda la Vostra Regione, sul portale opendata.regione.abruzzo.it, non sono riuscito a trovare un dataset contenente un elenco dettagliato delle varie strutture ricettive presenti sul territorio, ma solo uno contenente i dati statistici sulla capacità delle strutture ricettive.

Nell’ambito di questo progetto, mi domandavo se la Vostra Regione non disponesse dei dati relativi alle strutture ricettive e non fosse disponibile a condividerli con noi.

RingraziandoVI in anticipo per la risposta, Vi porgo cordiali saluti.

Giacomo Gregori

In totale le e-mail inviate sono state trentadue, ma le risposte sono state solamente dodici. Il Grafico 4 riassume i risultati dell'indagine.

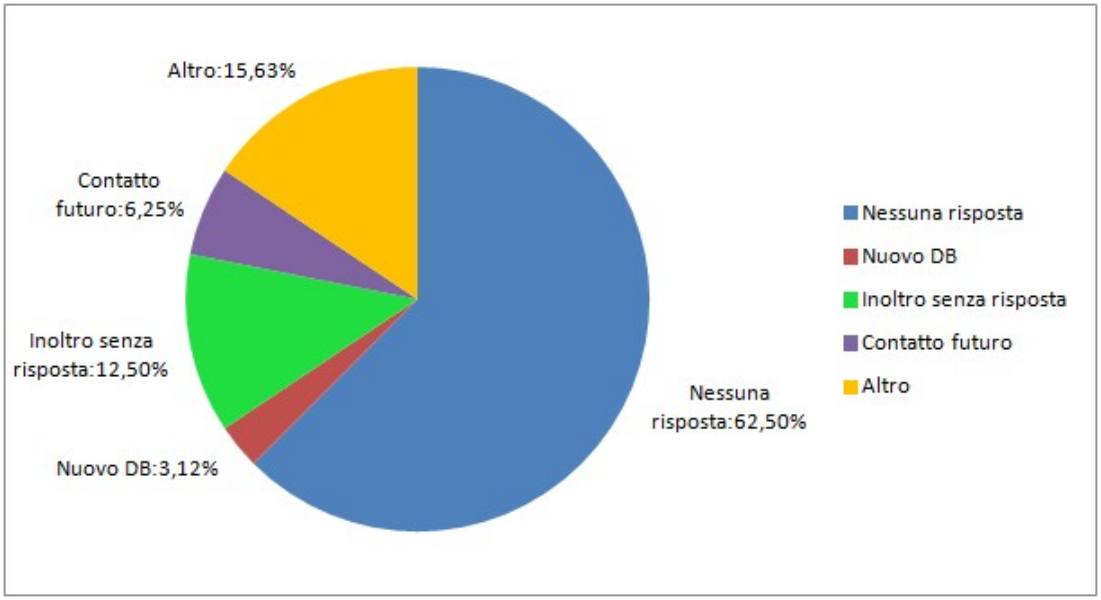

Grafico 4. Risposte ricevute

- la maggior parte di queste sono state solo delle notifiche di inoltro per il seguito di competenza ma non hanno mai avuto ulteriore risposta,
- altre regioni hanno dichiarato che ci avrebbero risposto con maggiore calma nei giorni successivi, ma non lo hanno mai fatto,
- il Veneto, essendo, nel periodo dello sviluppo del progetto, in attesa di un aggiornamento dell'anagrafe, ha posticipato ai primi del mese un eventuale contatto,
- la Valle d'Aosta ci ha comunicato la possibilità di trovare, sul portale del turismo, tutte le strutture ricettive sul territorio. Trattandosi di un motore di ricerca online, non è stato possibile scaricare un dataset, rivelandosi così inutile allo sviluppo del nostro progetto,
- la Basilicata ci ha fornito l'URL di un altro portale open dove poter scaricare un dataset che contenesse i dati che cercavamo. Questo ha migliorato nettamente quello precedentemente in nostro possesso che conteneva solamente i dati relativi alla sola provincia di Matera,
- undici regioni non hanno fornito alcuna risposta.

In definitiva, solamente la Basilicata ha contribuito concretamente a migliorare lo sviluppo del nostro progetto. Deludente è stata la mancata risposta della maggior parte delle regioni, in particolare di quelle, come Abruzzo, Calabria, Campania e Molise, che non avevano un open data già disponibile.

3.5 Creazione del dataset unificato a livello nazionale

Dopo aver raccolto tutti i dataset e aver contattato le regioni per ottenerne altri o versioni migliori, abbiamo potuto iniziare la procedura di creazione del dataset a livello nazionale.

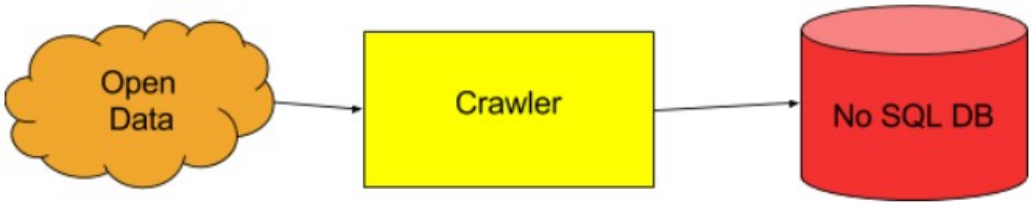

Figura 1. Architettura del software per la creazione del dataset a livello nazionale

Per raggiungere tale scopo è stato implementato un software la cui architettura è mostrata in Figura 1. Il sistema consiste essenzialmente di un crawler, che scarica i vari dataset dagli open data delle varie regioni e da un database no SQL (nella fattispecie MongoDB), su cui tali dati vengono memorizzati.

3.5.1 Data model

Il primo problema da affrontare preliminarmente è concettuale: quale data model utilizzare per il nuovo dataset? Infatti ogni open data ha una struttura differente rispetto alle altre. Alcuni si differenziano semplicemente per il nome dei campi, ma altri anche per i tipi di dati contenuti: si va dal più semplice, la Toscana, che ha solo quattordici campi, al più articolato, la Lombardia, che ne ha trentacinque. Per risolvere il problema si è scelto di rinominare i campi comuni sotto un'unica etichetta in lingua inglese e di mantenere i campi che potessero fornire informazioni utili per lo scopo prefissatoci.
Ma come comportarsi con quei record che non contengono alcuni campi presenti in altri dataset? Utilizzando un dataset no SQL, ci viene fornita la possibilità di includere tali campi senza problemi, infatti ci permette di avere una struttura "variabile" senza che si incorra in conflitti e senza la necessità di impostare a NULL i valori non presenti nell'open data originale. I campi che troviamo all'interno del dataset nazionale sono spiegati nella Tabella 8.

Campo	Descrizione	Open data che contengono il dato
_id	ID	12
name	Nome della struttura	12
description	Descrizione	12
category	Categoria di appartenenza (es. Bed and breakfast, hotel, ecc..)	1
address	Indirizzo	12

postal-code	CAP	12
city	Comune	12
province	Provincia (sigla)	11
hamlet	Frazione	6
locality	Località	6
region	Regione	12
latitude	Latitudine	5
longitude	Longitudine	5
number of stars	Numero di stelle	10
telephone	Recapito telefonico	12
telephone2	Recapito telefonico alternativo	2
cellular phone	Numero di cellulare	3
fax	Numero di fax	11
web site	Link al sito web	11
email	Indirizzo e-mail	12
beds	Numero di letti	8
rooms	Numero di stanze	7
suites	Numero di suite	1
toilets	Numero di bagni	4

Tabella 8. Data model

3.5.2 Crawler

Una volta stabilito il data model del nostro dataset unico, è stato possibile lavorare sulla procedura di inizializzazione e aggiornamento del dataset.

Il crawler implementa il codice che scarica i vari dataset e, dopo aver adeguato il data model, li inserisce nel nostro dataset. Esso è suddiviso in quattro fasi principali:

1) Connessione al sito e scaricamento del dataset: per ogni regione, il crawler si connette al sito regionale ed esegue il download automatico del dataset,
2) Mapping dei dati ottenuti nel data model: gli open data vengono analizzati e ristrutturati secondo il data model del nuovo dataset nazionale,
3) Salvataggio dei dati: le informazioni riguardanti le strutture vengono inseriti in MongoDB,
4) Aggiornamento delle informazioni di corredo: vengono inseriti in una collezione di log i dati riguardanti la data di ultimo aggiornamento di ogni singolo open data e la quantità di strutture ricettive presenti in essi.

Mentre la prima fase è stata relativamente semplice, le altre hanno richiesto maggiori attenzioni, per cui verranno descritte con maggiore dettaglio, nelle sezioni successive.

3.5.3 Mapping dei dati

Una volta risolte tutte le problematiche riguardanti il download dei dataset, sono stati affrontati quelli sulla gestione dei dati. La maggior parte degli open data ricavati è in formato .csv, ma non la totalità. Dei dodici dataset a nostra disposizione, dieci sono in tale formato (ma anche tra questi vi sono delle differenze), uno, la Basilicata, in formato .xls e uno, il Trentino, in formato .xml.

Per la gestione dei dati aventi il formato .csv, prima di tutto abbiamo analizzato le strutture degli open data originali in modo da capire quali dati si potessero prelevare. Dopodiché i dati selezionati sono stati mappati al data model mostrato nella Tabella 8 e sono stati inseriti all'interno di MongoDB.

Un caso particolare è rappresentato dal dataset dell'Emilia-Romagna: questo, innanzitutto, è diviso in otto file diversi, uno per provincia, quindi è stato necessario creare una funzione che li analizzasse uno dopo l'altro.

Diversa è stata la gestione dei dataset di Basilicata e Trentino essendo in formati diversi da .csv. In entrambi i casi, è stato necessario implementare del codice aggiuntivo, basato sull'uso di librerie esterne.

3.5.4 Inserimento dei dati

Una volta eseguito il mapping dei dati per modificare il data model e farlo coincidere col nostro, abbiamo inserito i dati all'interno del nostro dataset attraverso MongoDB. Il dataset è stato organizzato in tre collezioni: nuovo, vecchio e temp:

- NUOVO: la versione più recente del dataset,
- VECCHIO un backup della scorsa versione,
- TEMP una copia della versione precedente a VECCHIO.

Ogni volta che viene fatto un aggiornamento, la collezione nuovo viene spostata in vecchio e nella collezione log vengono aggiunte le informazioni di corredo. Da notare, che di volta in volta, si ha una sola copia di backup del dataset. L'aggiornamento del dataset è fatto quotidianamente attraverso una procedura automatica. Nel caso in cui l'aggiornamento fallisca, i dati sono recuperati dalla collezione VECCHIO.

3.5.5 Log

In un primo momento si era pensato di mantenere nel dataset tutte le collezioni che venivano create quotidianamente dallo script così da poter studiare e monitorare i dati inseriti. Questa possibilità è stata poi accantonata a favore di quella descritta nel paragrafo precedente per evitare che il dataset occupasse eccessivo spazio sul server. Così facendo, però, avremmo perso la possibilità di studiare l'evoluzione sia del nostro dataset, sia degli open data forniti dalle regioni. Per ovviare a tale perdita, si è optato per la creazione della collezione LOG nella quale troviamo, oltre l'id identificativo, la data dell'aggiornamento e, per ogni regione, la quantità di strutture inserite nel database e l'ultima data di aggiornamento dell'open data. Per ricavare la data di aggiornamento, per la maggior parte delle regioni è bastato ricavare il campo "last modify" attraverso la funzione `stream_get_meta_data` che riporta le informazioni del file, in altri casi è stato necessario ricorrere allo scraping della pagina web di download del dataset.

3.6 Licenza

Tutti gli open data che abbiamo utilizzato per la creazione del nostro dataset unificato a livello nazionale sono protetti da una licenza. Come è possibile vedere nella Tabella 9, la maggior parte di queste sono Creative Commons, le restanti sono Italian Open Data License.

Regione	Licenza
Basilicata	Non specificata
Emilia-Romagna	CC-BY-RER (CC-BY della Regione Emilia-Romagna)
Friuli-Venezia Giulia	Italian Open Data License
Liguria	CC-BY
Lombardia	Italian Open Data License
Marche	Non specificata
Piemonte	CC0 1.0 Universal
Puglia	Italian Open Data License
Toscana	CC-BY
Trentino-Alto Adige	CC-BY
Umbria	CC-BY
Veneto	CC-BY

Tabella 9. Licenza degli open data forniti dalle regioni.

Avendo utilizzato dati coperti da questo tipo di licenza, dovremo riportare le fonti dalle quali sono stati forniti e, volendo a nostra volta mettere a disposizione il nostro dataset unificato, dovremo farlo coprendolo con quella più restrittiva tra quelle utilizzate dalle regioni che hanno fornito gli open data. Considerando che la CC0 1.0 Universal non ha vincoli di utilizzo e che l'Italia Open Data License e la CC-BY permettono le stesse libertà, abbiamo potuto liberamente scegliere tra le due e abbiamo optato per la CC-BY essendo più facilmente riconosciuta a livello internazionale.

4. Arricchimento dei dati

4.1 Introduzione

L'arricchimento dei dati (o *data enrichment*, utilizzando il termine inglese) è un processo utilizzo per accrescere, rifinire o migliorare i dati puri. Questa idea, insieme ad altri concetti simili, contribuiscono a rendere i dati una risorsa preziosa per quasi tutte le aziende o i business moderni.

Sebbene l'arricchimento dati possa funzionare in molti modi differenti, molti degli strumenti utilizzati per questo scopo implicano un perfezionamento di dati che potrebbe includere piccoli errori. Un comune processo di arricchimento dei dati potrebbe essere, ad esempio, correggere possibili errori di battitura o tipografici in un database tramite l'uso di algoritmi di precisione. Gli strumenti di arricchimento dati possono anche aggiungere informazioni a semplici tabelle di dati.

Un altro modo in cui l'arricchimento dei dati può funzionare è attraverso l'estrazione di nuovi dati da integrare con quelli già in possesso. Attraverso metodologie come la logica fuzzy (una logica che parte dal concetto di logica binaria modificandola), gli ingegneri sono in grado di estrapolare un maggior numero di dati partendo da una versione più grezza.

Come abbiamo visto nel paragrafo 2.5.1, ogni open data fornito dalle regioni che abbiamo incontrato nel nostro progetto, ha un data model differente rispetto agli altri. Per questo abbiamo cercato di adattarli ad uno schema comune, ma alcuni di essi mostrano delle mancanze, talvolta importanti, nella compilazione del dataset. Nello sviluppo del nostro progetto abbiamo quindi valutato come colmare tali mancanze e quali strumenti utilizzare.

4.2 Geocoding

I primi campi che abbiamo ritenuto fondamentali da integrare, sono la latitudine e la longitudine delle strutture ricettive. Questo in previsione dello sviluppo di un'applicazione web che mostrasse su mappa la loro posizione. Dei dodici open data

a nostra disposizione solamente cinque (Emilia-Romagna, Lombardia, Marche, Puglia e Toscana) contengono le coordinate geografiche, mentre sette ne sono sprovvisti. Per raggiungere il nostro scopo abbiamo utilizzato la Geocoding API di Google che, dati in ingresso indirizzo e comune della struttura, fornisce in uscita le coordinate geografiche.
Purtroppo la versione gratuita delle API ha delle limitazioni di utilizzo:

- 50 richieste al secondo,
- 2500 richieste giornaliere.

Questo ci impedisce di calcolare le coordinate ad ogni esecuzione dello script poiché solamente una piccola parte dei record riuscirebbero ad essere integrati essendo più di 24000 le strutture sprovviste di tali informazioni.

4.2.1 Inizializzazione del dataset

Per risolvere questo problema si è fatto in modo che il software implementato per il geocoding, in caso di assenza di latitudine e longitudine nell'open data fornito dalle regioni, prelevasse le coordinate dalla versione precedente del nostro dataset nazionale: la collezione VECCHIO.
Infatti il primo passo è quello di eseguire un software che inizializza il dataset e poi, man mano, in rispetto delle politiche di utilizzo dell'API di Google, ne viene eseguito un secondo che esegue il geocoding fino ad esaurire tutte le coordinate. Lo script, a causa delle limitazioni giornaliere dell'API, dovrà essere eseguito per più giorni fino alla completa integrazione dei dati mancanti. Una volta fatto ciò e quindi inizializzato il dataset, viene copiato nella collezione VECCHIO e sarà possibile aggiornarlo quotidianamente, mantenendo le coordinate, senza doverle ricalcolare. Se vengono aggiunte delle nuove strutture, lo script prova a trovarle nella collezione VECCHIO. Non riuscendoci, esegue il geocoding, stavolta evitando problemi di limitazioni giornaliere visto che difficilmente verranno aggiunte più di 2500 strutture nello stesso giorno.

4.3 Previsioni di arricchimento

Nel corso del nostro progetto è stato implementato l'arricchimento relativo alle coordinate geografiche, ma in seguito sarà possibile e necessario svolgere tale operazione anche per altre informazioni.

Innanzitutto potrebbe essere importante cercare di integrare le informazioni delle strutture dove mancano rispetto al nostro data model. Per fare ciò la soluzione principale sarebbe quella di sfruttare le collaborazioni instaurate con le regioni chiedendo loro se è possibile aggiungere i campi mancanti all'interno del loro dataset così che l'inserimento nel nostro possa essere automatico.

Oltre a questo, potrebbe essere utile inserire il link alla scheda social di ogni struttura ricettiva. Il primo social network da prendere in considerazione è Facebook, avendo in sé un gran numero di pagine create dai gestori delle strutture. Inoltre potrebbero essere interessanti anche i profili Instagram, Twitter, Pinterest e YouTube, anche se questi sono meno diffusi nel campo della ricettività.

Un'altra valida alternativa potrebbe essere contattare direttamente le strutture ricettive, giacché di alcune di esse è fornito anche l'indirizzo email. Oppure si potrebbe pensare ad un portale in cui ogni singola struttura possa modificare e aggiungere le proprie informazioni.

5. L’applicazione web

5.1 Introduzione

Tutto quello che abbiamo esposto fino ad ora è stato fatto con uno scopo finale: la creazione di un’applicazione web che mettesse a disposizione degli utenti il nuovo dataset nazionale. Al fine di rendere il sito user-friendly, si è utilizzato un layout simile a quello di GoogleMaps che visualizzi su una mappa le strutture ricettive che vengono selezionate attraverso una barra di ricerca dove l’utente possa inserire una città o una regione dove vuole trovare un alloggio.

5.2 L’applicazione

Per lo sviluppo dell’applicazione abbiamo utilizzato la Google Maps API che ci ha permesso di utilizzare le mappe di Google per visualizzare la posizione delle strutture ricettive contenute nel nostro dataset.
Al momento dell’apertura viene visualizzata la cartina, centrata sull’Italia, a tutto schermo con, in alto a sinistra, il form, nel quale inserire il nome della città nel quale ricercare le strutture ricettive, e il pulsante che esegue la ricerca.
Una volta cliccato il pulsante, viene eseguita una ricerca che trova tutti i dati di ogni struttura localizzata nella città ricercata dall’utente e sulla mappa compaiono i marker abbinati ad esse. Passando il puntatore sopra ognuno di essi verranno visualizzati il nome, l’indirizzo, il numero di telefono e il sito web della struttura.

5.3 Sviluppi futuri

Quella presentata fino ad ora è solo una versione iniziale dell’applicazione web. Questa dovrà essere sviluppata in modo da includere un maggior numero di funzionalità.
Innanzitutto la possibilità di ricercare le strutture non solo per la città in cui si trovano, ma anche per la regione, provincia, località, frazione o anche per nome, categoria di appartenenza o numero di stelle.

Potrebbe essere utile implementare, una volta effettuata la ricerca, una presentazione ad elenco dei risultati per rendere più chiare e complete le informazioni riguardanti gli alloggi trovati. Per quanto riguarda la visualizzazione delle strutture sulla mappa, potrebbero essere utilizzati i *cluster marker*, cioè un raggruppamento dei marker presenti in zone limitrofe, per impedire che questi si accumulino sulla cartina impedendo una facile comprensione.

Se il nostro dataset venisse arricchito anche con le schede social, potrebbe essere interessante anche mostrare nell'elenco delle foto per mostrare all'utente come sia la struttura.

Un'altra funzionalità implementabile all'interno dell'applicazione è un sistema di recensioni che permetta agli utenti di condividere le impressioni sul loro soggiorno nelle strutture.

6. Conclusioni

Nel corso dello sviluppo del nostro progetto abbiamo avuto modo di entrare in contatto con il mondo degli open data. Questo ha presentato sia lati positivi che negativi. Infatti, se da una parte abbiamo trovato alcuni dataset estremamente completi e aggiornati con frequenza, dall'altra abbiamo visto come alcune regioni non abbiano messo a disposizione i dati riguardanti le strutture ricettive e, nel caso del Molise, non è stato nemmeno possibile trovare un portale open data. Un esempio di carenza dei dati ci è fornito dalle coordinate geografiche, fornite solo da cinque delle dodici regioni di cui abbiamo raccolto i dati. Nel nostro progetto abbiamo sopperito a tale mancanza utilizzando il geocoding fornitoci mediante l'API di Google. Per risolvere altre mancanze riteniamo che la soluzione migliore sia quella di interpellare direttamente le regioni o i responsabili degli open data per collaborare verso una completa integrazione dei dati.

Analizzando la situazione finale, il progetto ci ha permesso di creare un nuovo dataset unificato a livello nazionale che possa essere riutilizzato da terzi. E' sicuramente un servizio che va a colmare una mancanza nel mondo degli open data. Il lavoro che è stato fatto lato server con la procedura di aggiornamento automatico svolta quotidianamente, nonché la parte più corposa del nostro progetto, ci ha fornito la possibilità di avere un dataset continuamente aggiornato e ci garantisce l'opportunità di studiare l'evoluzione di ogni singolo open data distribuito dalle regioni. In futuro, qualora altri dataset aperti venissero resi disponibili dalle autorità amministrative, basterà aggiornare il software prendendo come esempio la procedura utilizzata fino ad ora.

Il lavoro svolto sull'applicazione web prepara il territorio per futuri sviluppi che permetteranno l'implementazione di nuove funzionalità. In futuro potranno essere inseriti anche i dati riguardanti bar, ristoranti, attrazioni turistiche, monumenti, ecc. La loro ricerca su mappa potrebbe diventare solo una sezione di un portale in cui si possano trovare ancora più informazioni riguardo a itinerari turistici, prenotazioni, mezzi di trasporto e quant'altro. Interessante potrebbe essere la possibilità di allargare il bacino di utenza agli stati stranieri, visto il gran numero di turisti

provenienti dagli stati esteri, traducendo tutto ciò in più lingue. Questa applicazione ha le potenzialità per riservarsi un suo ruolo di importanza all'interno del panorama turistico, soprattutto grazie alle sue funzionalità open che permetteranno agli sviluppatori o ai ricercatori nel settore del turismo di poter usufruire di un dataset unificato a livello nazionale.

Lo sviluppo di questo portale di riferimento per il turismo italiano è ai nastri di partenza, ma ha già fatto una buona parte del lavoro necessario per entrare in questo mondo così affollato e pieno di concorrenza, con dalla nostra parte la volontà di offrire servizi che possano contribuire per lo sviluppo del settore turistico in Italia.

7. Bibliografia

UNWTO World Tourism Barometer, vol.14 - July 2016. http://cf.cdn.unwto.org/sites/all/files/pdf/unwto_barom16_04_july_excerpt_.pdf (31/01/2017)

Sito ufficiale dell'Agenzia Nazionale del Turismo - ENIT. Studio sul turismo straniero in Italia. http://www.enit.it/it/studi.html (02/02/2017)

Corriere della sera. Il turismo in Italia è cresciuto (Ma restiamo solo quinti). http://www.corriere.it/cronache/16_luglio_26/turismo-italia-cresciuto-d649f92e-52a5-11e6-9335-9746f12b2562.shtml (02/02/2017)

(ISTAT, 2016) - *Sito ufficiale dell'Istituto Nazionale di Statistica.* https://www.istat.it/ (02/02/2017)

(Moreno, Hörhager, Schuster, Werthner, 2015) - María del Carmen Calatrava Moreno, Gernot Hörhager, Rainer Schuster and Hannes Werthner. *Strategic E-Tourism Alternatives for Destinations. Information and Communication Technologies in Tourism 2015.* Springer. Lugano, Svizzera. 2015.

Digital Marketing Turistico. http://digitalmarketingturistico.it/5333/il-revenue-management-nellera-delle-recensioni-online/ (31/01/2017)

Wikipedia. Voce Dati aperti. https://it.wikipedia.org/wiki/Dati_aperti. (27/01/2017)

Open definition. Definizione di Conoscenza aperta. http://opendefinition.org/okd/italiano/ (27/01/2017)

Simone Aliprandi. *Il fenomeno open data. Indicazioni e norme per un mondo di dati aperti.* Ledizioni. Italia. Febbraio 2014.

Pagina ufficiale Creative Commons. http://creativecommons.it/ (31/01/2017)

Pagina ufficiale Italian Open Data License v2.0 http://www.dati.gov.it/iodl/2.0/ (31/01/2017)

Wikidata. https://www.wikidata.org/wiki/Wikidata:Main_Page (29/01/2017)

Bing Liu. *Web Data Mining. Exploring Hyperlinks, Contents and Usage Data. Second Edition.* Springer. Berlin. 2011.

Techopedia. Definizione di data enrichment. https://www.techopedia.com/definition/28037/data-enrichment (01/02/2017)

8. Ringraziamenti

Il cammino che mi ha portato a tagliare questo traguardo è stato lungo e, come tale, ha visto diverse persone dare il proprio apporto per permettermi, alla fine, di poter dire “ce l’ho fatta”.

Prima di tutto vorrei ringraziare coloro che mi hanno seguito nello sviluppo di questo progetto: il Prof. Andrea Marchetti e l’Ing. Angelica Lo Duca. Lavorare con loro è stato un piacere, mi hanno permesso di apprendere nuove nozioni che, sono sicuro, mi serviranno molto nel mondo del lavoro. Il ringraziamento va anche e soprattutto per il lato umano che hanno mostrato: sempre disponibili, pronti ad aiutarmi e comprendermi. Grazie.

Un grazie va anche al Prof. Paolo Macchia che ha accettato di ricoprire il ruolo di correlatore in tempi piuttosto stretti aiutandomi a redigere questa tesi.

Il grazie più grande va alla mia famiglia. Aspettare pazientemente che arrivassi a questo traguardo con tutto quello che comporta, anche economicamente, è segno di grande fiducia e amore. Grazie per le spintarelle nei momenti di blocco e per gli incoraggiamenti in quelli di ripartenza.

Un grazie grandissimo agli amici. Tutti.

Grazie a chi c’è da sempre, Bene e Dani. Il tempo ci ha allontanati, ma mai separati del tutto. Grazie per il tempo condiviso, per l’affetto che tutt’oggi si manifesta negli eventi importanti delle nostre vite.

Grazie a Woody, Giulia, Andrew, Pierca, Ele, Chiara, Marti, Luca, Anna e Alessio. Grazie non solo per le serate passate insieme ma per tutto il supporto che mi hanno dato nei momenti difficili. Mi hanno sempre sostenuto, fatto riflettere su ogni decisione e spinto a fare quello che fosse più giusto per me. Hanno sopportato le mie continue irruzioni nei momenti bui e mi hanno sempre accolto con un sorriso.

Grazie ai “miei bimbi”. Quando ho iniziato il cammino universitario, per loro occupavo il ruolo di educatore. Questo ruolo col tempo si è evoluto, sono cresciuto con loro e, oggi, sono per me degli amici e confidenti.

Un grazie particolare va a Sara. Un’amicizia nata molto tempo fa che, ad un certo punto, poteva anche sembrare persa, ma che è rispuntata nel momento più

inaspettato. Se oggi sono qui, una buona parte è per lei e la ringrazio per ciò che ha fatto per me.

Il grazie più “numeroso” va a Bohnanza. Un gruppo di più di quaranta persone che quattro anni fa ha sfondato la porta ed è entrato nella mia vita sconvolgendola. Grazie perché mi ha accolto come se fossi stato uno di loro da sempre. Con loro è stato più facile concludere ciò che avevo iniziato. Mi hanno donato sorrisi quando più ne avevo bisogno, ma anche affetto che, forse, quotidianamente non si vede ma compare sempre al momento giusto. Con il loro entusiasmo, la loro semplicità, il loro esempio e la loro amicizia ho deciso di non mollare e sono arrivato alla fine di questo percorso.

Infine grazie ad Angela. La persona che più di tutti mi è stata vicino in questo periodo di tirocinio e tesi. Con lei ho condiviso la fatica e le gioie dei progressi, lo studio intenso e i momenti di meritato riposo. In questi nostri sei mesi abbiamo già condiviso diversi momenti felici e difficoltà più o meno grandi e lo abbiamo sempre fatto con amore. Quando la guardo negli occhi non riesco a non immaginare il futuro. Un futuro tutto da scoprire insieme.

Se oggi sono qui è perché ognuno di loro ha messo una parte di sé nella mia vita.

Grazie. Grazie a tutti voi.