

Tra editoria e università

**I risultati del gruppo di lavoro
Università di Verona,
CINECA e Aie**



UNIVERSITÀ
di VERONA



Associazione
Italiana
Editori



Giornale
della
libreria

2016

Tra editoria e università

I risultati del gruppo di lavoro
Università di Verona,
CINECA e Aie

Di

Maria Gabaldo (Università di Verona)

Gabriella Scipione (CINECA)

Piero Attanasio (AIE)

Anna Lionetti (mEDRA)

Beatrice Cunegatti (InfoTech Law Firm)

ISBN: 9788899630010

ISBN-A: [10.978.8899630/010](https://www.isbn.it/9788899630010)

© 2016, il testo è rilasciato dagli autori in licenza Creative Commons Attribuzione - Non commerciale

*Questa pubblicazione raccoglie i risultati del progetto **Dottorato congiunto con Enti di Ricerca accreditati VQR e messa in condivisione delle tesi di dottorato tra Università degli Studi di Verona e Atenei partner secondo la politica Open Access con il contributo tecnico di CINECA attraverso la specializzazione del portale PLEIADI**, coordinato dall'Università di Verona con la partnership di CINECA e Associazione Italiana Editori, in collaborazione con mEDRA, e con il contributo di Beatrice Cunegatti (InfoTech Law Firm).*

OpenTesi

Di Gabriella Scipione (CINECA)

ISBN: 9788899630041

ISBN-A: [10.978.8899630/041](https://www.cineca.it/10.978.8899630/041)

© 2016, il testo è rilasciato dagli autori in licenza Creative Commons Attribuzione - Non commerciale

Sommario

1. OpenTesi come aggregatore di tesi di dottorato	4
2. <i>Data quality</i> ed estrazione dei concetti nei dati presenti in OpenTesi.....	7
3. Architettura e funzionamento del sistema	9
4. Pubblicazione in Open Linked Data ed esportazione dei dati relativi alle tesi di dottorato	11

Indice delle figure

Figura 1. Pagina di ricerca per le tesi di dottorato in OpenTesi.....	5
Figura 2. Visualizzazione della scheda informativa di una tesi: oltre agli argomenti principali la scheda include anche argomenti più generici, estratti automaticamente usando le categorie di Wikipedia e contribuendo così ad arricchire i metadati.....	8
Figura 3. Architettura del servizio OpenTesi	9
Figura 4. Home page di OpenTesi: generale	12
Figura 5. Home page di OpenTesi: dataset	13
Figura 6. Home page di OpenTesi: esplora	13

1. OpenTesi come aggregatore di tesi di dottorato

La naturale tendenza che sta emergendo negli ultimi anni è quella di integrare sempre più il catalogo dei prodotti della ricerca di ateneo con sistemi esterni che consentano una maggiore visibilità dei contenuti scientifici di ciascuna università, e quindi un aumento dell'immagine di eccellenza dell'ateneo e dei suoi ricercatori. A questo scopo sono nati aggregatori internazionali dei contenuti scientifici con funzionalità avanzate di ricerca semantica e "social".

Di conseguenza, gli atenei si stanno sempre più dotando di cataloghi dei prodotti della ricerca (archivi istituzionali) sia per censire e valutare la propria produzione scientifica interna, sia per promuoverla esternamente. La realizzazione di un portale per la ricerca e l'accesso alle tesi di dottorato si iscrive in questo contesto ed è stata uno dei principali obiettivi identificati all'interno del progetto.

OpenTesi è di fatto un aggregatore per l'accesso centralizzato alle tesi di dottorato depositate negli archivi istituzionali delle università che partecipano al progetto. La piattaforma consente di importare dati dai cataloghi universitari e dagli archivi aperti.

Il numero di tesi raccolte sino a questo momento è pari a 11.193 e nella tabella che segue è riportata la distribuzione delle tesi per ciascun ateneo partecipante al progetto:

Istituzione	Numero di tesi
UniPI	2993
UniMI	2246
UniVE	1630
UniVR	1587
UniMiB	1350
UniTN	934
UniINSUBRIA	350
UniTO	103

La ripartizione delle pubblicazioni classificate per modalità di accesso presenti su OpenTesi al momento della redazione di questo testo è la seguente:

Modalità di accesso	Numero di tesi
Accesso aperto	5.273
Accesso chiuso	2.990
Accesso riservato	1.118
Accesso con embargo	877

Gli archivi che contengono i documenti e i relativi metadati delle tesi vengono interrogati attraverso un'istanza specializzata del servizio PLEIADI (Portale per la Letteratura scientifica Elettronica Italiana su Archivi aperti e Depositi Istituzionali, che già consente l'accesso centralizzato alla letteratura scientifica depositata da docenti e ricercatori negli archivi aperti

delle università e degli enti di ricerca italiani), tramite il protocollo OAI-PMH¹, funzionale alla raccolta dei dati stessi.

La procedura di raccolta dei dati è seguita da quella di importazione: attraverso una serie di mappature dei diversi formati di dati provenienti dai *repository* istituzionali, che ha richiesto interventi di uniformazione e normalizzazione delle informazioni per aggregarle in un unico punto di accesso, i dati raccolti dai vari *data provider* OAI-PMH sono resi disponibili nell'indice unificato degli archivi.

Per migliorare le funzionalità di ricerca all'interno della piattaforma, sin da subito CINECA ha voluto adottare tecniche avanzate di *semantic web* che consentano ricerche molto più evolute. Inoltre, dal punto di vista della *user experience* la nuova interfaccia di OpenTesi è stata pensata allo scopo di facilitare le modalità di esplorazione delle tesi di dottorato attraverso vari filtri, come mostrato nella figura sottostante. Ad esempio è possibile ricercare le tesi di dottorato in base all'università di afferenza o alle modalità di accesso (aperto, chiuso, con embargo, riservato). L'area di ricerca, specializzata nella presentazione di metadati di tesi di dottorato, è disponibile liberamente agli utenti all'indirizzo <http://esplora.opentesi.cineca.it/>.

Figura 1. Pagina di ricerca per le tesi di dottorato in OpenTesi

OpenTesi

Lingua: Italiano

Entra

Tutti i Campi Q Trova Avanzata

Scorri per Istituzione

UniINSUBRIA
UniMI
UniMib
UniPI
UniTN
UniTO
UniVE
UniVR

Scorri per Rights

Accesso aperto
Accesso chiuso
Accesso con embargo
Accesso riservato

Opzioni di ricerca

- Ultime ricerche
- Ricerca avanzata

Serve aiuto?

- Suggerimenti per la ricerca
- FAQ

¹ <https://www.openarchives.org/pmh/>: OAI-PMH è un protocollo per lo scambio di dati sviluppato come infrastruttura di comunicazione delle informazioni relative a documenti contenuti in archivi digitali.

Si è inoltre pensato di definire e strutturare i metadati descrittivi delle tesi di dottorato sul portale web in modo che il loro significato fosse accessibile non solo a utenti umani, ma anche a programmi che li utilizzano per integrarli e renderli disponibili all'interno di servizi e banche dati esterni. I metadati delle tesi sono stati quindi espressi attraverso i linguaggi di annotazione, arricchiti dall'utilizzo di ontologie e pubblicati in Open Linked Data.

Sintetizzando dunque i passaggi tecnici descritti in dettaglio nei capitoli successivi, lo sviluppo del progetto si è articolato in più fasi:

- il primo passo ha avuto come obiettivo la raccolta delle informazioni relative alle tesi delle università partecipanti, attraverso la predisposizione di un *harvesting* (procedura di raccolta di dati) per reperire i metadati presenti presso gli archivi istituzionali di ogni ateneo.
- Sui metadati provenienti da diverse sorgenti è stata sviluppata una procedura di mappatura e normalizzazione, in modo da tradurre le diverse strutture dei dati in input dagli archivi degli atenei in un unico modello di dati specifico per OpenTesi. Inoltre, sono state realizzate operazioni estese di *data quality* al fine di migliorare la ricerca semantica all'interno della piattaforma.
- Poi si è passati alla realizzazione di un portale che consente la ricerca e l'accesso alle tesi di dottorato attraverso la specializzazione di PLEIADI. Il portale, chiamato OpenTesi, è raggiungibile all'indirizzo <http://opentesi.cineca.it>, mentre la home page di PLEIADI è disponibile all'URL <http://find.openarchives.it>.
- La navigazione delle tesi di dottorato è stata arricchita attraverso l'estrazione dei concetti (dai metadati bibliografici e dagli abstract) in modo automatico, utilizzando il Concept Mapper, uno strumento realizzato da CINECA per identificare ed estrarre i concetti più rilevanti da un documento, annotarli in modo automatico e collegarli alla pagina di Wikipedia che ne fornisce la descrizione.
- L'ultimo step è stato lo sviluppo di un servizio per la pubblicazione in Open Linked Data dei metadati delle tesi di dottorato e per l'esportazione di record.

Depositando quindi i propri lavori negli archivi istituzionali, esposti attraverso OpenTesi, gli autori delle tesi possono beneficiare di una maggiore diffusione e visibilità dei propri prodotti della ricerca.

2. *Data quality* ed estrazione dei concetti nei dati presenti in OpenTesi

Per esprimere in modo formalizzato il significato dei metadati presenti in OpenTesi, si è pensato di utilizzare soluzioni per l'estrazione di concetti e per la loro mappatura utilizzando una ontologia di riferimento. Infatti, l'introduzione di una struttura concettuale che legghi i termini tra loro, mostrandone le relazioni e facendone quindi emergere il significato, è alla base del *semantic web*.

A tal fine si è utilizzato un servizio sviluppato dal CINECA e denominato Concept Mapper². Esso consente di (1) analizzare il contenuto con l'obiettivo di identificare i concetti più rilevanti nel contesto del documento, (2) annotare automaticamente parti del testo con i concetti corrispondenti ed eventualmente con il link alla pagina di Wikipedia che ne fornisce la definizione e la descrizione, (3) associare metadati semantici quali le classi di appartenenza dei concetti più rilevanti (categorie di Wikipedia), (4) mappare il contenuto del documento su un'ontologia specifica, in questo caso sfruttando la conoscenza disponibile in Wikipedia.

Le componenti del Concept Mapper sono quattro moduli che eseguono in sequenza le quattro principali fasi dell'analisi:

- 1) analisi linguistica e individuazione delle frasi nominali;
- 2) identificazione del concetto corrispondente tramite un processo di disambiguazione;
- 3) selezione in base alla rilevanza;
- 4) associazione (eventuale) del concetto a una ontologia specifica.

L'analisi dei concetti è stata effettuata sui singoli abstract e sui metadati descrittivi delle tesi di dottorato. Sfruttando i *tag* semantici prodotti automaticamente da OpenTesi tramite il Concept Mapper è possibile ricercare per concetti che non vengono direttamente citati nel testo ma che sono rilevanti per la tesi.

² <http://conceptmapper.cineca.it/it/home>.

Figura 2. Visualizzazione della scheda informativa di una tesi: oltre agli argomenti principali la scheda include anche argomenti più generici, estratti automaticamente usando le categorie di Wikipedia e contribuendo così ad arricchire i metadati

[* Citazione](#)
[Invia SMS](#)
[Invia email](#)
[Esporta il record](#)
[Aggiungi ai preferiti](#)



Meccanismi molecolari dell'attività antitumorale associata alla modulazione degli ioni zinco in cellule di adenocarcinoma pancreatico

Lo zinco è il secondo metallo maggiormente abbondante nel corpo umano. Questo ione è essenziale in un'ampia varietà di processi cellulari, in quanto ha un ruolo sia funzionale agendo da cofattore per più di 300 enzimi, sia strutturale per la stabilizzazione della struttura terziaria di molte prot...

[Descrizione completa](#)

Autore principale:	Dalla Pozza Elisa
Natura:	Doctoral Thesis
Lingua:	Italian
Pubblicazione:	Università degli Studi di Verona 2008
Argomenti:	Chemioterapia , Apoptosi , Cellula , Cofattore (biologia) , Ciclo cellulare , Necrosi , Caspasi , Molecola , Proteina , Enzima , Mitocondrio , P21 , Tumore , Stress ossidativo , Ossigeno , Terapie farmacologiche , Processi cellulari , Citologia , Enzimologia , Catalisi , Coenzimi , Processi cellulari , Anatomia patologica , Processi cellulari , Biologia molecolare , EC 3.4.22 , Concetti fondamentali di chimica , Fisica molecolare , Specie chimiche , Proteine , Dietetica , Chimica degli alimenti , Enzimologia , Catalisi , Organelli , Biologia molecolare , Proteine , Geni oncosoppressori , Oncologia , Neoplasie , Biochimica , Citologia , Elementi chimici , Ossigeno , Fluidi refrigeranti ,

Accesso
<http://hdl.handle.net/11562/337588>
Accesso chiuso

Documenti analoghi

[Effetti del lisofosfolipide edelfosina e della tricostatina A sulla crescita di cellule di adenocarcinoma pancreatico e meccanismi molecolari associati](#) di: Russignan Anna
Pubblicazione: (2008)

[Effetti della deprivazione di ossigeno e glucosio \(OGD\) sulla fluidità di membrana e sulla modulazione dell'attività di bace1 in cellule endoteliali del microcircolo cerebrale di ratto \(RBE4\)](#) di: Brambilla,
Pubblicazione: (2013)

[Meccanismi cellulari implicati nell'effetto antitumorale del cannabidiolo su cellule di glioma umano U87-MG e caratterizzazione delle sue proprietà antiangiogeniche](#) di: Solinas, Marta
Pubblicazione: (2011)

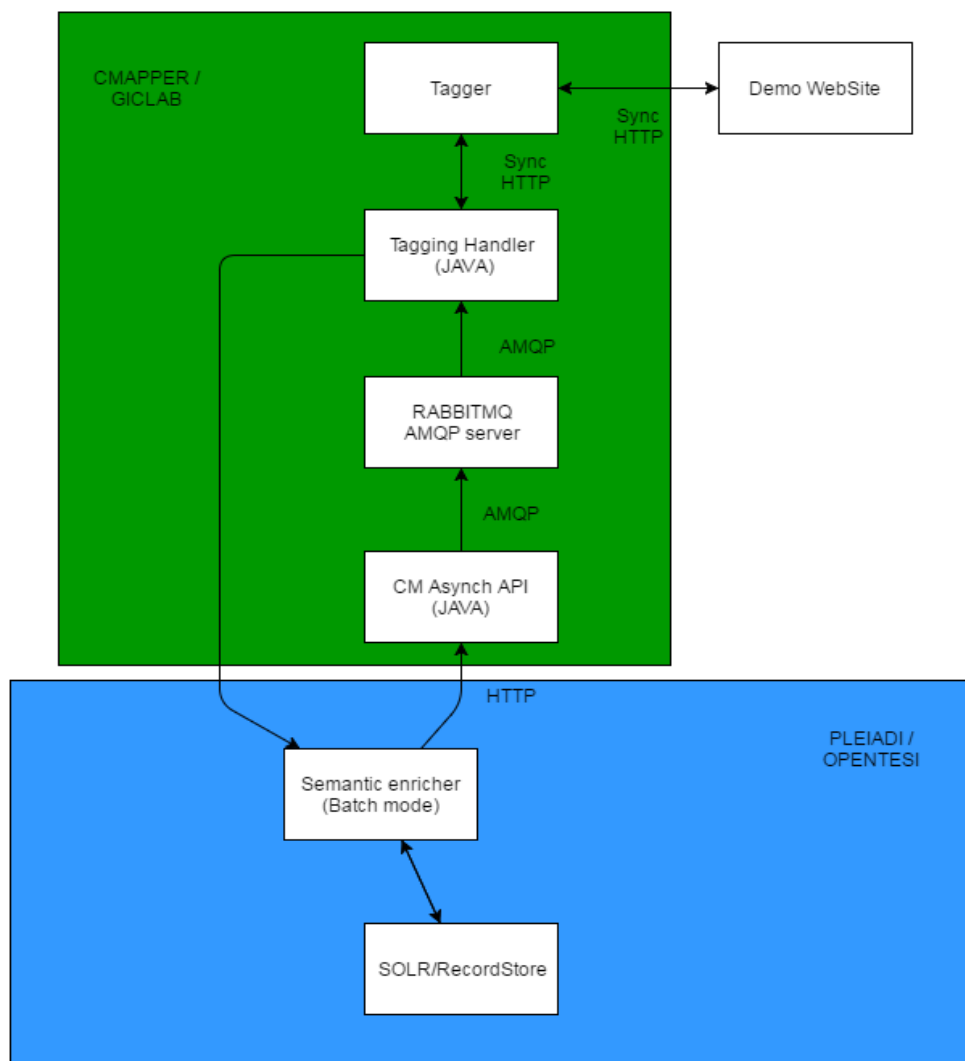
[Effetti modulatori di estratti secchi delle tre specie medicinali di Echinacea sui processi differenziali e maturativi delle cellule dendritiche in presenza o assenza di microambiente tumorale pancreatico](#) di: MUCCI, ILARIA
Pubblicazione: (2009)

[Potenziamento dell'attività immunostimolante delle cellule dendritiche umane: effetti in vitro di Kheyole Limpet Hemocyanin e](#)

3. Architettura e funzionamento del sistema

L'architettura messa a punto allo scopo di migliorare la qualità dei dati e l'indicizzazione semantica della piattaforma di OpenTesi è sinteticamente rappresentata nel diagramma che segue: le frecce mettono in evidenza il flusso dei dati; la parte in verde riguarda l'evoluzione del servizio di *tagging* Concept Mapper, arricchito e migliorato all'interno del progetto; la parte in blu è specifica del progetto OpenTesi e si occupa da una parte di interagire con il servizio Concept Mapper, e dall'altra con il *repository* Solr. OpenTesi è stato infatti pensato per essere indipendente e poter essere utilizzato in altri ambiti allo stesso modo.

Figura 3. Architettura del servizio OpenTesi



Il Semantic Enricher (Batch mode) (nell'area blu della figura) è il componente incaricato di interrogare il *repository* Solr recuperando i metadati e gli abstract relativi alle tesi di dottorato.

Un sistema di code basato su RabbitMQ (nell'area verde della figura) è stato appositamente sviluppato nell'ambito di questo progetto per la gestione delle richieste, potenzialmente di

grandi dimensioni, verso il servizio Concept Mapper: l'utilizzo di questa tecnologia ha permesso l'introduzione di alcune funzionalità importanti per il servizio, ad esempio consente di recuperare le richieste inviate a OpenTesi anche dopo un guasto del sistema di gestione delle code e di gestire diverse richieste in parallelo. In questo modo è possibile un controllo sul bilanciamento del carico di lavoro del sistema, un buon grado di *fault tolerance* (la capacità di un sistema di non subire interruzioni di servizio anche in presenza di malfunzionamenti) e di recupero dei messaggi nel caso in cui alcuni servizi che compongono il Concept Mapper risultassero non raggiungibili o riscontrassero malfunzionamenti.

La componente chiamata Tagging Handler (nell'area verde della figura) preleva le richieste dal sistema di code RabbitMQ contenenti il testo da indicizzare, la lingua e l'indirizzo a cui sottoporre l'indicizzazione. Viene effettuata poi una richiesta verso il servizio Concept Mapper ottenendo l'analisi del testo formata dai concetti estratti dal dominio Wikipedia.

Il Concept Mapper, che supporta le lingue inglese e italiano, consente il riconoscimento automatico dei concetti partendo da un testo in chiaro ed è inoltre in grado di fornire il livello di rilevanza dei concetti estratti rispetto al testo. Ogni concetto è poi collegato a una pagina di Wikipedia permettendo in questo modo di creare un collegamento ipertestuale a questa fonte per ogni frase riconosciuta. Il risultato viene poi restituito al processo *batch* (PLEIADI batch export) per l'inserimento dei *tag* nei relativi campi Solr.

I dati così recuperati e arricchiti vengono nuovamente inviati al server (RabbitMQ) che gestisce le code per il servizio di indicizzazione semantica. Infine, i dati indicizzati vengono inviati alla componente Semantic Enricher che provvede all'aggiornamento dei metadati in Solr con le informazioni semantiche aggiuntive reperite ed elaborate lungo tutto il processo.

4. Pubblicazione in Open Linked Data ed esportazione dei dati relativi alle tesi di dottorato

Per raggiungere l'obiettivo di rendere i dati presenti in OpenTesi semanticamente interoperabili e comprensibili in modo automatico da parte di servizi esterni, è stato necessario utilizzare linguaggi di annotazione (o *markup language*) costruiti a partire dal noto RDF (Resource Description Framework)³, un metalinguaggio di descrizione delle informazioni che consente l'interoperabilità semantica tra applicazioni che condividono risorse sul web. In altri termini le annotazioni servono a rappresentare il significato dei dati annotati, a rendere i dati interconnessi tra di loro, rendendo interoperabili informazioni provenienti da fonti eterogenee.

Lo sviluppo del servizio di pubblicazione in Open Linked Data⁴ dei dati presenti in OpenTesi si è articolato nella definizione dello schema e del formato di esportazione dei dati, nella creazione di una mappatura tra i metadati di Solr e il formato di esportazione, nell'esportazione in formato JSON dei metadati dal *repository* Solr, nella costruzione dei dataset e nel loro deposito sul sito.

Il formato di esportazione dei metadati prevede i seguenti campi:

- identifier: identificativo univoco all'interno di OpenTesi;
- institution: sorgente dei dati e istituzione di riferimento di un autore;
- title: titolo della tesi;
- creator: autore principale della tesi;
- contributor: tutte le persone che hanno contribuito direttamente alla scrittura della tesi come autore, relatore, correlatore;
- description: descrizione della tesi;
- url: pagina web di dettaglio, solitamente sul portale dell'istituzione di riferimento. Da qui è possibile accedere al full-text, quando disponibile;
- language: lingua della tesi, come dichiarata dall'autore. Può differire dalla lingua dei metadati;
- date: anno di pubblicazione della tesi;
- rights: diritti di accesso associati alla tesi. I valori usati fanno riferimento al dizionario e-prints;
- subject: argomenti della tesi. Estratti analizzando titolo, subjectManual (vedi sotto) e descrizione. Oltre agli argomenti principali contiene anche argomenti più generici, ricavati usando le categorie di Wikipedia;
- subjectUri: link alla pagina di Wikipedia di ogni argomento. Ogni argomento è in questo modo disambiguato semanticamente;
- subjectManual: keyword, parole chiave, come inserite manualmente dall'autore.

³ <https://www.w3.org/RDF/>.

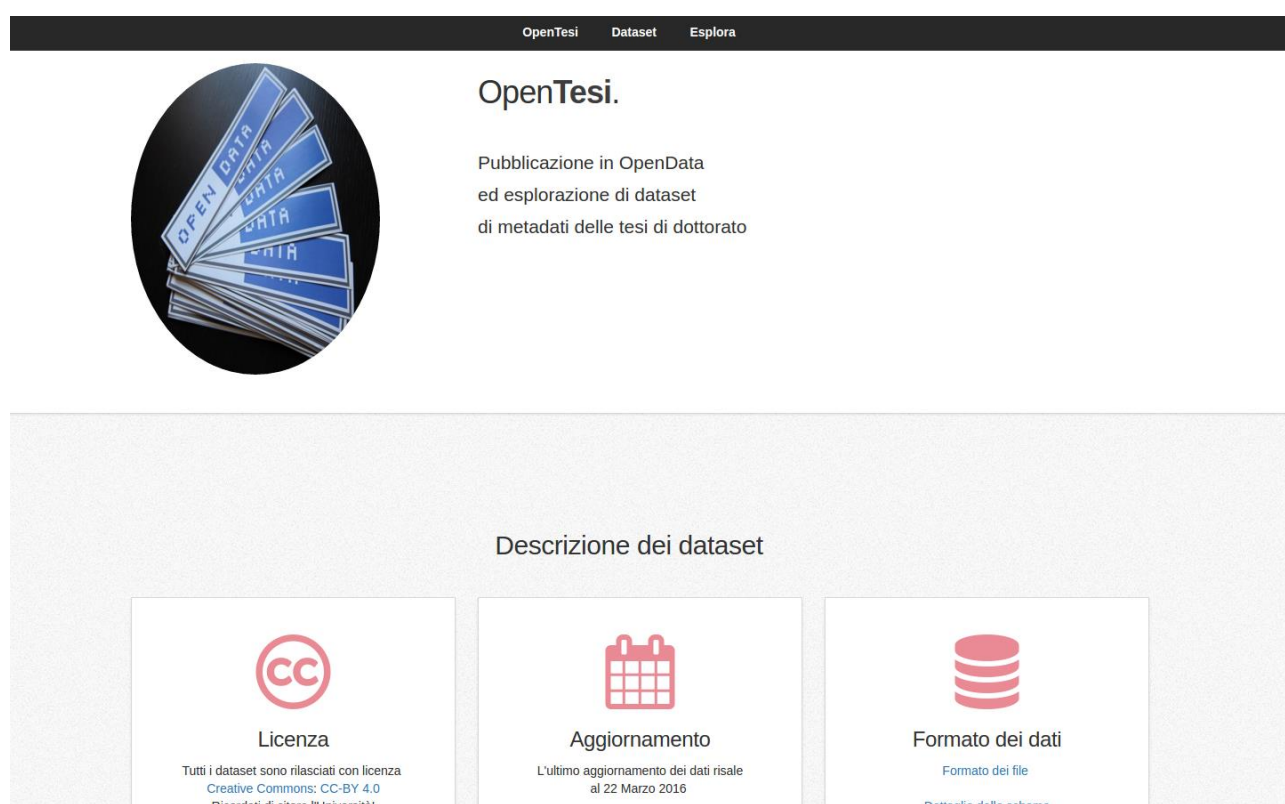
⁴ Pubblicare dati in Open Linked Data significa utilizzare tecnologie che consentano a set di dati strutturati provenienti da diverse sorgenti di essere collegati fra loro e riutilizzabili da parte di servizi esterni.

La home page del progetto (<http://opentesi.cineca.it/>) funge da portale per l'esposizione dei dataset raccolti ed elaborati.

Nella prima parte (Figura 4) vengono descritti i dataset e le caratteristiche comuni:

- la licenza applicata, Creative Commons CC-BY (<https://creativecommons.org/licenses/by/4.0/>), che permette qualunque riutilizzo dei dati con il solo vincolo di citazione della fonte;
- le tempistiche di aggiornamento dei dati stessi;
- il formato dei file scaricabili e il dettaglio dello schema utilizzato, partendo dal contenuto di ogni campo.

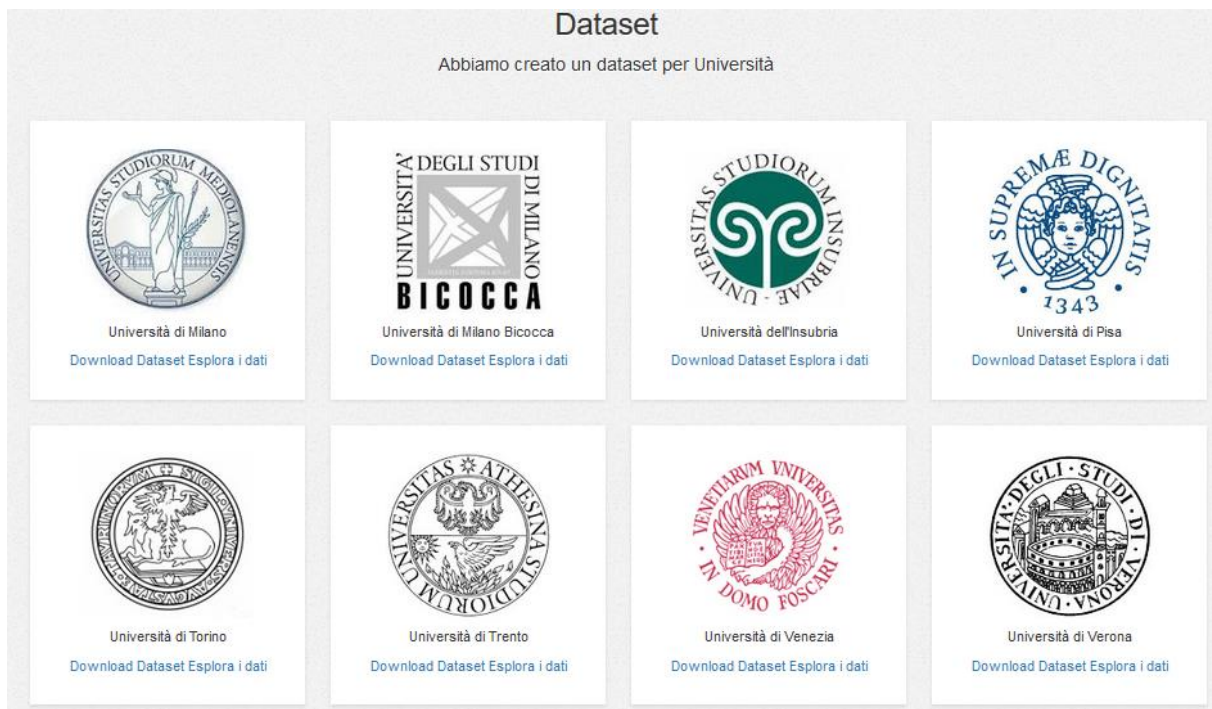
Figura 4. Home page di OpenTesi: generale



Nella seconda parte (Figura 5) è stata creata una sezione per ogni dataset, uno per università. Da qui è possibile:

- scaricare il dataset in modo da poterlo importare in un sistema esterno. Per permettere l'integrazione dei dati con sistemi basati sui Linked Data e il formato RDF, è stato adottato il formato JSON-LD, che consente l'interscambio di Linked Data attraverso una particolare specializzazione del linguaggio JSON, formato di scambio di dati fra applicazioni molto utilizzato;
- fare ricerche all'interno del dataset sfruttando l'interfaccia di navigazione precedentemente descritta e limitando tale ricerca all'università in questione.

Figura 5. Home page di OpenTesi: dataset



Nell'ultima sezione (Figura 6) è messa in evidenza la ricerca per argomento. Sfruttando i *tag* semantici prodotti automaticamente da OpenTesi tramite il Concept Mapper è possibile ricercare per concetti che non vengono direttamente citati nel testo ma che sono rilevanti per la tesi, mettendo a frutto l'apporto innovativo di OpenTesi alla condivisione dei prodotti della ricerca scientifica.

Figura 6. Home page di OpenTesi: esplora

