Alma Mater Studiorum · Università di Bologna

FACOLTÀ DI SCIENZE MATEMATICHE, FISICHE E NATURALI
Corso di Laurea Magistrale in Informatica

CONVERSIONE PER IL SEMANTIC WEB E PUBBLICAZIONE NEL LINKED DATA DI DATI RELATIVI A BENI ARTISTICI E CULTURALI DELLA REGIONE EMILIA ROMAGNA

Tesi di Laurea in Tecnologie Web/Internet

Relatore:
Prof.
FABIO VITALI

Presentata da:
GIOVANNI DE MARCO

Correlatore:
Dott.
GIUSEPPE FRANGIAMONE

Sessione I
Anno Accademico 2010/2011

Alla mia famiglia.

Ad Anna.

Introduzione

La rappresentazione della conoscenza in ambito culturale, con particolare riferimento alle operazioni legate alla catalogazione dei dati relativi ai beni culturali, tramite l'utilizzo di strumenti informatici risulta una problematica ampiamente discussa e per la quale, nel corso degli anni, sono stati proposti diversi approcci e soluzioni. Il presente lavoro di tesi si propone di identificare le tecnologie legate al mondo del *Semantic Web* come uno degli strumenti attualmente più utili e adatti al perseguimento di tale obiettivo, individuando in particolare nel modello basato sui principi del *Linked Data* le caratteristiche più affini e valide per l'ottenimento di una modellazione dei dati completa e rispondente alle esigenze richieste. I beni culturali, intesi in quest'ambito sia come insieme di artefatti in senso lato (quali oggetti, pitture, sculture o edifici che le contengono) sia come tradizioni e costumi che in qualche modo costituiscono il 'patrimonio culturale' e a cui la letteratura di settore fa spesso riferimento con il nome di '*cultural heritage*', hanno la caratteristica di richiedere una conoscenza completa di diverse informazioni ad essi correlate per poter ottenere una piena comprensione del loro valore. Una corretta rappresentazione della conoscenza ad essi legata richiede, pertanto, una modellazione in cui i dati possano essere intrecciati tra loro, rivelando un valore informativo che altrimenti resterebbe nascosto e permettendo di descrivere in modo approfondito tutti gli aspetti informativi rilevanti.
Il progetto *OpenERCH* (Open Emilia Romagna Cultural Heritage), oggetto di questo lavoro di tesi, si propone proprio di mettere in evidenza i vantaggi ottenibili da una rappresentazione formale e incentrata sulla semantica

relativa a diversi beni facenti parte del patrimonio artistico e culturale della regione Emilia Romagna. Il lavoro svolto consiste principalmente nella creazione di un *dataset*, rispondente alle caratteristiche tipiche del Semantic Web, ottenuto dalla conversione di dati di vario genere relativi a musei e altri edifici e contenitori dislocati nel territorio della regione, facendo riferimento ad uno specifico modello ontologico e sfruttando i principi e le tecnologie del Linked Data per ottenere un insieme di informazioni il più possibile correlate e arricchite. L'obiettivo finale è quello di ottenere una sorgente informativa in cui i dati sono intrecciati non solo tra loro ma anche con quelli presenti in sorgenti esterne, dando vita ad un percorso di collegamenti che permette di ottenere una ricchezza informativa utilizzabile per la creazione di un valore aggiunto che altrimenti non sarebbe possibile rivelare. Tale aspetto è messo in evidenza, in questo lavoro di tesi, tramite la creazione di un applicativo di esempio che utilizza il dataset prodotto, e i collegamenti che esso presenta con la rete del Linked Data, come punto di partenza per il reperimento di informazione aggiuntiva, dando vita ad un *mash-up* informativo multidominio.

L'attività di catalogazione dei beni culturali, soprattutto se legata agli approcci tradizionali non basati sull'utilizzo di strumenti semantici, presenta diversi limiti e problemi la cui analisi risulta fondamentale per individuare quali possono essere le tecnologie in grado di affrontarli. L'operazione di catalogazione dei beni o, in generale, degli oggetti di interesse effettuata tramite applicazioni informatiche consiste spesso nella realizzazione di grandi basi di dati che hanno il compito di raccogliere le diverse informazioni. L'utilizzo pervasivo delle tecnologie informatiche orientate a questi scopi ha portato alla generazione di una vasta varietà di dati digitali relativi ai beni artistici e culturali, caratterizzata principalmente da una grande eterogeneità relativa sia alla tipologia di media e formati usati per la pubblicazione, sia in riferimento alla dimensione, al livello di accessiblità e ai modelli logici e strutturali di rappresentazione utilizzati. Una delle più importanti conseguenze portate dalla presenza di una mole così elevata di dati eterogenei consiste nella nasci-

ta di numerosi raccoglitori di informazioni che risultano non interoperabili in alcun modo e danno vita a vere e proprie isole di dati difficilmente collegabili tra loro. Tale conseguenza risulta evidentemente spiacevole se si tiene conto del fatto che, come prima affermato, le informazioni relative ai beni culturali hanno, per loro natura, un forte bisogno di correlazione e ricchezza per poter esprimere pienamente il proprio valore.
Un altro aspetto di fondamentale importanza e spesso fonte di problemi è la difficoltà di reperire con precisione le informazioni di interesse tramite i metodi tradizionali, tipici soprattutto del Web, basati sull'utilizzo di parole chiave per l'identificazione dei contenuti. Tutte queste difficoltà diventano ancora più evidenti e profonde se l'attore interessato alla ricerca e al consumo dei dati è rappresentato da un processo automatico anziché da una persona, essendo, soprattutto nell'ambito del Web, le informazioni pensate e presentate soprattutto per un consumo umano. La gestione del patrimonio culturale tramite l'uso degli strumenti informatici risulta quindi un'operazione molto complessa sia a causa della presenza di formati e schemi di rappresentazione eterogenei e spesso ambigui sia per le difficoltà incontrate dalle attività di ricerca e filtraggio rendendo difficile l'integrazione dei dati presenti in sorgenti diverse che permetterebbe di ottenere un valore aggiunto notevole.
Diversi sono stati, nel corso degli anni, gli approcci e le tecnologie utilizzate per far fronte a queste problematiche. Uno dei primi modelli di raccolta di informazioni utilizzato in quest'ambito, forse artefice di gran parte dei problemi e limitazioni di cui si parlava precedentemente, è rappresentato dall'utilizzo di 'schede di catalogazione' degli oggetti, tramite un approccio di memorizzazione simile a quello bibliografico. Le informazioni, in questi contesti, erano rappresentate principalmente da descrizioni testuali di ogni singolo aspetto dell'oggetto considerato, impedendo operazioni complesse sui dati e risultando inadeguate a coprire pienamente l'alta complessità intrinseca degli oggetti e delle loro relazioni. Diverse organizzazioni e commissioni, a livello nazionale e internazionale, si sono preoccupate, nel corso degli anni, di creare standard e procedure per la catalogazione e la documentazione

del patrimonio artistico portando alla creazione di modelli che indicassero in maniera precisa i criteri di strutturazione dei beni. Ciò ha portato alla nascita di modelli di dati in cui le informazioni sono frammentate in diverse entità di base, descritte tramite un elenco di proprietà ad esse associate, dando vita a diverse applicazioni informatiche basate per lo più su grandi basi di dati relazionali. Il principale limite, in questo caso, continua ad essere il ruolo centrale attribuito agli oggetti: la descrizione delle loro caratteristiche tramite un elenco di proprietà ha portato, infatti, all'origine di informazione spesso ridondante e troppo semplificata, mettendo in evidenza l'incapacità, tramite tale modello di dati, di rappresentare associazioni semantiche tra gli oggetti e con altri pezzi di conoscenza pertinenti ad altre discipline.
La grande diffusione del web ha cambiato il modo tradizionale di intendere e recuperare le informazioni, rendendo sempre più importanti gli aspetti legati all'interoperabilità. I principali veicoli per la diffusione e la pubblicazione di informazioni relative ai beni culturali diventano i siti web, creati e gestiti da diverse entità per rendere accessibili facilmente i propri cataloghi. La necessità di permettere l'interoperabilità tra diverse sorgenti di dati ha portato alla creazione di diversi approcci pensati per lo scambio di metadati, come quello proposto da *Dublin Core*. Tutti questi approcci, seppur maggiormente orientati all'interoperabilità, hanno il difetto di rimanere comunque incentrati sugli oggetti e le informazioni provenienti da diverse sorgenti, e inerenti a diversi oggetti o discipline, possono essere correlate solo tramite interventi manuali. Diventa pertanto essenziale l'utilizzo di un'infrastruttura tecnologica che permetta di identificare i concetti e le equivalenze in modo univoco e automatico, facendo riferimento a una rappresentazione della conoscenza altamente formale che permetta di eseguire ragionamenti e inferenze. Le ontologie rappresentano il principale strumento utilizzato in questo senso, in quanto permettono di rappresentare gli oggetti non solo tramite la descrizione delle proprie caratteristiche ma anche tramite la descrizione delle relazioni prsenti tra loro e ponendo attenzione sul significato dei termini e sulla struttura e la natura del dominio informativo. Il *CIDOC Conceptual*

Reference Model, ontologia alla base del progetto OpenERCH, rappresenta uno dei principali esempi di questo tipo, in quanto fornisce una sorta di linguaggio altamente specifico per la rappresentazione di informazioni sui beni culturali relative anche ad ambiti spaziali e temporali.
Tutti i problemi e gli approcci fino ad ora presentati sono trattati in modo più approfondito nel primo capitolo, in cui sono descritti i principali limiti e difficoltà relativi all'attività di catalogazione (sezione 1.1), le diverse tecniche utilizzate per farne fronte (sezioni 1.2 e 1.3) fino ad una breve descrizione di alcuni approcci e progetti orientati principalmente alla condivisione della conoscenza e alla creazione di una sorta di architettura orientata ai servizi, come quella proposta dal progetto Europeana (sezione 1.4). L'analisi e le considerazioni esposte nel primo capitolo portano ad individuare nelle tecnologie legate al Semantic Web una possibile soluzione per l'ottenimento di una modellazione dei dati rispondente alle diverse esigenze prima elencate, in cui assume un ruolo fondamentale la semantica intrinseca presente nei dati. Semantica significa essenzialmente significato, e il significato permette un utilizzo migliore e più efficace delle informazioni, generando un processo di comunicazione in grado di esprimere la conoscenza in modo formale, preciso e comprensibile anche da agenti automatici.
L'idea del Semantic Web è basata fondamentalmente su un uso massiccio di strumenti che permettano di esprimere esplicitamente il significato dei dati, rendendolo comprensibile alle macchine e permettendo il collegamento di diverse risorse informative presenti nella rete. Le ontologie rappresentano il più importante tra questi strumenti, dando la possibilità di creare una sorta di schema sul quale basare l'intero modello organizzativo delle informazioni caratterizzato da un livello di formalizzazione molto elevato che consente di esprimere in modo completo il significato ontologico dei dati, spiegando come questi dovono essere interpretati. L'obiettivo finale è quello di ottenere una rete di dati collegati tra loro in base al significato ottenuta creando delle associazioni tra diversi raccoglitori informativi dislocati nella rete. Il Linked Data rappresenta uno dei risultati più tangibili e interessanti in quest'ambi-

to, fornendo una vera e propria 'rete di dati' espressi formalmente e tra loro correlati secondo una serie di principi e pratiche il più possibile basate sugli standard. Tale strumento (la cui utilizzazione pratica più importante è fornita dal *Linking Open Data Project*) rappresenta lo spazio ideale in cui i dati relativi al patrimonio artistico e culturale possono assumere un significato e un valore più ricco, diventando parte integrante di una rete informativa che ne permette la correlazione con diverse discipline rivelando pienamente il loro valore. La descrizione più approfondita di queste tecnologie è contenuta nel secondo capitolo, in cui sono esposti sia i principi alla base del web semantico e dei suoi approcci di strutturazione dell'informazione (sezioni 2.1 e 2.2) sia le caratteristiche e le pratiche su cui si basa il Linked Data (sezione 2.4).
Tutte le considerazioni fino ad ora esposte sono alla base del progetto OpenERCH, realizzato in questo lavoro di tesi su proposta dell'azienda NSI (*Nier Soluzioni Informatiche*) di Bologna, che ha lo scopo primario di convertire dati relativi all'ambito artistico e culturale, riferiti nello specifico a musei ed edifici storici presenti nella regione Emilia Romagna. L'attività di conversione, realizzata tramite la creazione di specifiche componenti software, ha tenuto conto delle regole e dei principi proposti dal Linked Data, esprimendo il significato dei dati tramite una loro rappresentazione basata sul modello ontologico proposto da CIDOC-CRM. Il principale obiettivo è quello di ottenere un dataset che possa essere reso disponibile in formato aperto e accessibile, caratterizzato dalla presenza di collegamenti in uscita verso altre sorgenti informative presenti nel *Web of Data*. Lo scopo primario non è dunque quello di permettere una consultazione visuale delle descrizioni di ogni oggetto, ma piuttosto esporre i dati per eventuali utilizzi anche da parte di applicazioni o agenti software automatici, proprio nell'ottica tipica del Semantic Web. Il lavoro ha compreso anche la creazione di una serie di applicazioni web che permettono di accedere al dataset ottenuto dalla conversione secondo le modalità standard indicate dal Linked Data inerenti sia la navigazione delle informazioni sia la loro interrogazione tramite linguaggi specifici. Nell'ambito del progetto è stata sviluppata, infine, l'applicazione *ERCH-Mashup*,

che permette di effettuare ricerche tra i metadati relativi ai musei e agli altri oggetti contenitori, aggiungendo dove possibile arricchimenti ottenuti da informazioni recuperate da altri dataset esterni. Tale applicazione ha il compito di fornire un esempio di come i dati, espressi in modo formale tramite l'ontologia, possano essere facilmente filtrati, correlati e arricchiti ottenendo vantaggi informativi rilevanti. La descrizione delle principali funzionalità del progetto OpenERCH è presentata nel terzo capitolo, in cui sono esaminati i dati oggetto di lavoro (sezione 3.2) e le caratteristiche relative sia all'attività di modellazione e alle metodologie di accesso ai dati (sezione 3.3) che all'applicativo di mash-up (sezione 3.4). Gli aspetti tecnici e implementativi inerenti il progetto OpenERCH e gli strumenti utilizzati per la sua realizzazione sono invece descritti e approfonditi nel quarto capitolo.
La valutazione dei risultati ottenuti, contenuta nel quinto capitolo, mette in evidenza come l'attività svolta rispetti a pieno tutti i principi e le tecniche di conversione e accesso ai dati proposte dal Linked Data. L'aderenza a tali principi permette di garantire che i dati prodotti rispettino quanto richiesto per poter entrare pienamente a far parte del Web of Data e di conseguenza essere resi disponibili, in una sorta di formato aperto ed accessibile, a chiunque abbia necessità di utilizzarli per scopi che vanno anche al di là della semplice consultazione (come mostrato nell'applicazione di mash-up di esempio). Tale valutazione porta quindi a ritenere la modellazione di OpenERCH come un'attività che apporta un importante miglioramento qualitativo all'accesso e all'utilizzo dei dati messi a disposizione dalla regione Emilia Romagna, fornendo nuovi scenari di utilizzo e ampie possibilità di creazioni applicative su essi basate. Tale attività presenta, quindi, numerose possibilità di sviluppi futuri legati sia alle componenti realizzate che ad eventuali nuove e ampie possibilità che potranno essere offerte dai progressi relativi a quanto legato al mondo del Semantic Web. Oltre ad eventuali attività che possono riguardare l'estensione dei dati oggetto di conversione o il miglioramento di diversi aspetti relativi sia alla loro pubblicazione on-line che al loro utilizzo nell'applicativo creato, esistono infatti diverse attività future attualmente difficili

da prevedere nello specifico. L'allargamento della rete di dati in cui i risultati ottenuti sono stati inclusi, pur non rappresentando un aspetto direttamente controllabile, potrebbe, infatti, fornire diverse possibilità di estensione e di correlazione con i dati prodotti in OpenERCH, consentendo di arricchire ulteriormente e rendere ancora più evidente e utile il valore aggiunto ottenibile e la cui esistenza è alla base di tutte le attività raccontate in questo documento.

Indice

Introduzione **i**

1 Problemi e approcci informatici alla gestione del patrimonio culturale **1**

1.1 **Problematiche nella catalogazione del patrimonio culturale** 1

1.1.1 *La catalogazione dei beni culturali* 1

1.1.2 *Limiti e problematiche principali* 3

1.2 **Primi approcci per la rappresentazione della conoscenza in ambito culturale** 6

1.3 **Soluzioni ed esigenze per la rappresentazione dei beni culturali nel Web** 9

1.3.1 *Ricerca di interoperabilità nello scambio dei metadati* . 9

1.3.2 *La rappresentazione della conoscenza mediante modelli formali* 10

1.4 **Approcci basati sulla condivisione della conoscenza** . . 13

1.4.1 *Beni culturali e reti P2P* 13

1.4.2 *Soluzioni basate su architetture orientate ai servizi* . . 14

2 Semantic Web e Linked Data **17**

2.1 **Il Semantic Web e l'importanza delle relazioni semantiche** 17

2.1.1 *Basi e principi del Semantic Web* 17

2.1.2 *La definizione di semantica* 19

2.2 **Nuovi approcci alla strutturazione delle informazioni** . 20
2.2.1 *Da web a semantic web* 20
2.2.2 *Knowledgebase e Database* 23
2.2.3 *Il modello a grafo e il concetto di risorsa* 26
2.3 **Le ontologie** 28
2.4 **Linked Data e Web of Data** 33
2.4.1 *Il Linked Data* 33
2.4.2 *I principi del Linked Data* 35
2.4.3 *Il Linking Open Data Project* 36

3 Il progetto OpenERCH **41**
3.1 **Descrizione generale** 41
3.2 **Collezioni di dati gestite in OpenERCH** 44
3.3 **Attività di modellazione e conversione semantica dei dati** 47
3.3.1 *Tipologie di risorse in OpenERCH* 48
3.3.2 *Disponibilità e accesso ai dati* 51
3.3.3 *Collegamenti con dataset esterni* 55
3.4 **Applicazione mash-up di esempio** 57
3.4.1 *Reperimento di informazioni aggiuntive su categorie e sottocategorie* 59
3.4.2 *Ricerca di informazioni aggiuntive relative a persone* . 62
3.4.3 *Ricerca di informazioni aggiuntive su città e luoghi di interesse* 63

4 Il progetto OpenERCH: aspetti tecnici e implementativi **67**
4.1 **Strumenti utilizzati** 67
4.1.1 *L'ontologia CIDOC-CRM* 67
4.1.2 *Principali dataset di collegamento* 72
4.2 **Architettura generale** 78
4.3 **Attività di conversione dei dati** 82
4.3.1 *Modellazione ontologica e rappresentazione dei dati* . . 82

4.3.2 *Il trasformatore* 91
4.4 **Accesso ai dati e deferenziazione degli URI** 95
4.5 **Applicazione di mash-up: aspetti tecnici e dettagli implementativi** 98

5 Valutazione **103**
5.1 **Analisi quantitativa** 103
5.2 **Aderenza ai principi del Linked Data e scoperta informativa** 107

Conclusioni **111**

Bibliografia **115**

Elenco delle figure

2.1 Ontology Spectrum 31
2.2 Diagramma della Linking Open Data Cloud 38

3.1 Scheda di un museo nel portale 'Samira' 43
3.2 Esempio di rappresentazione in HTML delle informazioni di un museo in OpenERCH 53
3.3 Interfaccia web per le interrogazioni SPARQL proposta da OpenERCH 54
3.4 Applicazione ERCH-Mashup: interfaccia per la ricerca 58
3.5 Applicazione ERCH-Mashup: esempio di informazioni aggiuntive per categoria 60
3.6 Applicazione ERCH-Mashup: esempio di informazioni aggiuntive per musei 61
3.7 Applicazione ERCH-Mashup: esempio di informazioni aggiuntive su persone 62
3.8 Applicazione ERCH-Mashup: esempio di informazioni aggiuntive su città 64
3.9 Applicazione ERCH-Mashup: esempio di informazioni aggiuntive su luoghi vicini 66

4.1 Classi e proprietà di CIDOC-CRM per la descrizione spaziale . 70
4.2 Classi e proprietà di CIDOC-CRM per la descrizione temporale 71
4.3 Architettura generale del progetto OpenERCH 79
4.4 Diagramma di modellazione di un oggetto *Site* 84

4.5 Diagramma di modellazione di un oggetto *Place* 87
4.6 Diagramma di modellazione di un oggetto *Person* 90
4.7 Deferenziazione degli URI in OpenERCH 97

Capitolo 1

Problemi e approcci informatici alla gestione del patrimonio culturale

1.1 Problematiche nella catalogazione del patrimonio culturale

Il settore relativo ai beni artistici e culturali rappresenta sicuramente una delle realtà in cui l'utilizzo delle tecnologie, in particolare quelle legate all'ICT (*Information and Communication Technologies*), ha trovato una intensa diffusione e utilizzazione, nonché una delle discipline in cui vi è maggiore dibattito relativamente a come queste devono essere utilizzate e a quali specifiche tecnologie possano permettere di raggiungere gli obiettivi ritenuti di maggiore importanza.

1.1.1 *La catalogazione dei beni culturali*

All'interno di questo documento di tesi verrà utilizzato, dove non diversamente specificato, il concetto di 'beni artistici e culturali', o più genericamente di 'beni culturali', in un'ottica abbastanza ampia, a cui in letteratura

si fa spesso riferimento con il termine inglese '*cultural heritage*'. In accordo a quanto definito in [PDA05], per 'beni culturali' s'intendono l'insieme di artefatti in senso lato (oggetti, pitture, sculture, edifici) ma anche tradizioni, costumi, che in qualche modo costituiscono il nostro 'patrimonio culturale', nello specifico: parchi archeologici, musei, oggetti di valore storico-artistico, collezioni, edifici rilevanti, monumenti, chiese o teatri. Come indicato in [PDA05], sono diversi gli utilizzi che possono essere fatti delle tecnologie ICT per i beni culturali. Particolarmente rilevante risulta, in quest'ambito, l'aspetto relativo alla gestione del patrimonio culturale, termine generico con il quale si vuole indicare la possibilità di conoscere quali sono i vari beni, la loro collocazione nonché la loro descrizione, a beneficio di diversi soggetti quali regioni, provincie, musei o privati proprietari di beni culturali rilevanti, che necessitano di applicazioni che li aiutino a facilitare e mantenere tale conoscenza.

Le attività necessarie per preservare e dare accesso a tale patrimonio culturale consistono principalmente nella raccolta di informazioni riguardo i beni e gli oggetti, memorizzati e organizzati in sistemi informativi sotto forma di cataloghi e archivi che forniscano strumenti di ricerca. Una delle principali difficoltà legate a tale attività risiede nel fatto che gli oggetti e i fenomeni presenti nelle collezioni del patrimonio culturale non sono direttamente accessibili, pertanto tali sistemi devono fare riferimento a delle loro rappresentazioni, in genere in formato testuale. Senza la presenza di tali descrizioni, organizzate secondo precise modalità, gli oggetti risulterebbero sostanzialmente inaccessibili. [MJV09]

Principale compito delle applicazioni informatiche risulta, quindi, la creazione di una 'catalogazione' di tutti gli immobili e gli oggetti di interesse, messi a disposizione di chiunque ne abbia necessità. Tale operazione è realizzata in genere tramite la creazione di grandi basi di dati che raccolgono le informazioni dando vita alla creazione di numerose 'schede di catalogazione', spesso però di difficile consultazione e il cui utilizzo rimane in genere limitato solo a chi ha interessi specifici.

Il principale risultato di questa pervasiva utilizzazione delle tecnologie informatiche, nonché della creazione di applicazioni di questo genere, è rappresentato dalla generazione di un'enorme quantità di dati relativi ai beni artistici e culturali in formato digitale, realizzati da soggetti diversi spesso tramite modalità e sistemi completamente indipendenti. Come descritto in [CAN05], principale caratteristica di tali dati è infatti la vasta varietà relativa sia alla tipologia dei media e dei formati utilizzati per la loro pubblicazione, nonché alla loro dimensione, al livello di accessibilità e ai modelli logici e strutturali utilizzati per la loro rappresentazione. Questa attività di digitalizzazione ha, difatti, portato alla produzione di diverse tipologie di dati che variano da documenti di testo creati con diversi word processor, a presentazioni multimediali pensate per essere usufruite on-line piuttosto che off-line, a database organizzati secondo schemi logici diversi e basati su differenti terminologie fino a documenti HTML di vario genere distribuiti sul web. Grande varietà e molte differenze possono essere individuate anche relativamente alle strutture dati utilizzate per la memorizzazione di tali informazioni, alcune delle quali presentano un livello di formalizzazione alto (facendo utilizzo, ad esempio, di schemi relazionali alla base di *database record oriented* piuttosto che basandosi su dati espressi tramite linguaggi dichiarativi come XML) mentre altri risultano completamente grezzi o non strutturati (come immagini o disegni). Anche le tecnologie utilizzate non sono esenti da differenze ed eterogeneità, risultando spesso dissimili e includendo diversi sistemi di gestione di database, sistemi di rappresentazioni della conoscenza piuttosto che semplici server web.

1.1.2 *Limiti e problematiche principali*

La presenza di una mole così elevata di dati eterogenei ha spesso come conseguenza la nascita di sistemi chiusi, definiti come come '*stovepipe system*' [1], ma soprattutto di una grande varietà di dataset che risultano non

[1]Nel campo ingegneristico ed informatico, uno *stovepipe system* è un sistema che risulta come un assemblaggio di vari elementi collegati tra loro in modo così stretto che i singoli

interoperabili in alcun modo e finiscono per creare vere e proprie isole di dati completamente sconnesse e distanti. Tale conseguenza risulta particolarmente sconveniente se si pensa al fatto che, per la loro natura, le informazioni relative ai beni culturali si adattano molto bene ad essere interconnesse fra loro, e tale interconnessione potrebbe portare un valore aggiunto di rilevante importanza per i possessori delle diverse informazioni isolate. Dati relativi alla stessa opera, o ad uno stesso autore possono, ad esempio, essere presenti nei sistemi informativi di diversi musei, ognuno dei quali potrebbe avere informazioni specifiche o di rilevante interesse per gli altri: la possibilità di correlare tali informazioni tra di loro risulta però spesso impossibile a causa delle diversità di formati e organizzazioni dei dati, rendendo difficile sia la scoperta dell'esistenza vera e propria di tali correlazioni che l'effettivo recupero e integrazione dei dati dalle diverse sorgenti. Proprio per questo negli ultimi anni si è assistito alla creazione di numerosi gruppi di lavoro per lo sviluppo di standard e vocabolari specifici per i beni culturali.

Oltre alle difficoltà di integrazione e standardizzazione dei contenuti appena descritte, la gestione informatica del patrimonio culturale deve far fronte ad un'altra importante problematica, comune a tutti i campi caratterizzati da una vasta eterogeneità e vastità di dati, legata alle difficoltà di reperire le informazioni di interesse. Specie dopo la diffusione del *World Wide Web*, e del relativo modello ad ipertesto tipico di tale realtà, i principali metodi di ricerca sono basati sull'utilizzo di parole chiave per l'identificazione dei contenuti. Tali paradigmi di ricerca e navigazione rendono arduo il compito di localizzare informazioni specifiche: la presenza di omonimie o grandi quantità di sinonimi possono far diventare frustrante, ad esempio, l'attività di ricerca di dati relativi ad una specifica opera o autore. Ancora più problematico può diventare effettuare ricerche o interrogazioni particolarmente complesse o articolate: la ricerca di 'monumenti risalenti all'epoca medievale

elementi che lo compongono non possono essere differenziati, aggiornati o riprodotti. Si tratta di sistemi che, quindi, possono essere rimpiazzati solo tramite la costruzione di sistemi completamente nuovi, e sino a quel momento devono essere mantenuti in vita.

che non si trovino in Italia né in Austria' potrebbe, ad esempio, portare a numerosi risultati irrilevanti (dovuti all'incapacità del motore di ricerca di comprendere la negazione) oppure all'omissione di risultati rilevanti (dovuti all'esclusione, a causa della negazione, di contenuti utili ma in cui sono menzionati anche monumenti italiani o austriaci). Di conseguenza operazioni ancora più complesse, come l'inclusione di informazioni in sorgenti dati non accessibili direttamente dal motore di ricerca o ottenibili intrecciando dati e informazioni presenti in dataset diversi, risultano ancora più difficili, se non impossibili, basandosi esclusivamente su questi modelli.
Se tali difficoltà si presentano per i normali utenti, intesi come esseri umani, il problema diventa ancora più profondo se rapportato a processi automatizzati. Soprattutto nell'ambito del Web, le informazioni e le descrizioni dei beni culturali sono infatti in genere pensate principalmente per essere presentate e consumate da utenti umani, ponendo molta attenzione ad aspetti puramente descrittivi, utilizzando sovente descrizioni testuali o immagini per la presentazione dell'informazione. Se l'essere umano è in grado di interpretare il significato di questi elementi, e quindi comprendere l'informazione che viene presentata, molte difficoltà possono invece essere incontrate da un processo automatizzato o un 'agente software'. Si crea quindi un divario profondo tra l'informazione disponibile per strumenti che hanno il compito di risolvere tali problematiche, e quella mantenuta in forma leggibile per l'uomo.
La gestione del patrimonio culturale tramite l'utilizzo di mezzi informatici risulta pertanto un'operazione molto complessa a causa della grande ampiezza di dati disponibili, della loro natura e dell'importanza di ogni singolo dettaglio ad essi legato. Il principale problema è, in conclusione, rappresentato dal fatto che i dati sono spesso presentati in formati molto eterogenei e spesso secondo degli schemi di rappresentazione ambigui o troppo diversificati. Questo rende difficile sia attività di ricerca e filtraggio specifica, essenziale in un mondo così vasto e variegato, sia l'integrazione delle informazioni presenti in sorgenti dati diverse, che permetterebbe ai dati, data la loro natura, di ottenere un valore aggiunto notevole intrecciandoli e correlandoli tra loro.

1.2 Primi approcci per la rappresentazione della conoscenza in ambito culturale

L'attività di catalogazione dei beni artistici e culturali ha delle radici molto profonde e antiche: sin dall'inizio del Novecento infatti in Italia diversi studiosi e storici hanno sentito l'esigenza di creare un quadro di riferimento per la descrizione delle opere d'arte, evidenziando in particolare l'importanza del contesto in cui tali opere sono state prodotte. Come precedentemente descritto, l'avvento dei computer e in generale delle tecnologie informatiche e di comunicazione ha influenzato notevolmente il modo in cui tale attività è stata svolta, portando alla nascita di alcuni problemi legati principalmente alla grande varietà tipica del patrimonio culturale italiano. Le tecnologie e gli strumenti utilizzati per far fronte a tali difficoltà, e più in generale per approcciare il problema della catalogazione, sono diversi e si sono evoluti nel corso del tempo, risentendo notevolmente degli sviluppi e delle scoperte tecnologiche.

Il principio guida, alla base del processo di catalogazione, è rappresentato dalla rappresentazione della conoscenza, sia in riferimento allo specifico oggetto che deve essere rappresentato sia riguardo a tutte le altre informazioni che possano aiutare a comprendere le complesse relazioni semantiche che tale oggetto presenta con altri beni o con altre risorse appartenenti ad altre discipline. La completa conoscenza del contesto storico o religioso risulta, ad esempio, spesso l'unico modo per comprendere pienamente il valore e il messaggio di un oggetto. Come descritto in [SIG09] diverse tecnologie sono state utilizzate nel corso degli anni per cercare di trovare delle modalità sempre più adatte e automatizzate per la rappresentazione di tale conoscenza.

Le prime applicazioni informatiche nell'area della gestione dei beni culturali facevano utilizzo, in molti casi, di un approccio per la memorizzazione delle informazioni di catalogo molto simile a quello utilizzato in campo bibliografico. Gli oggetti venivano infatti descritti tramite delle 'schede di catalogazione' in cui le informazioni erano organizzate in diverse sezioni semanticamente

coerenti che descrivevano, ad esempio, l'autore di una determinata opera, il periodo o eventuali note critiche e storiche. I principali utilizzi delle tecnologie informatiche consistevano quindi nell'utilizzo di diverse metodologie di memorizzazione usate per ottenere documenti di vario genere per lo più consistenti di contenuti testuali organizzati in qualche modo. Tali metodologie di memorizzazione si basano innanzitutto sull'identificazione di un insieme ridotto di diverse schede di catalogazione, corrispondenti a diverse tipologie di oggetti (oggetti d'arte, archeologici, quadri, architettura, musei, ecc.) e su un successivo raggruppamento delle informazioni in diverse categorie molto generali (come autori, luoghi, materiali, informazioni storiche, ecc.). Tale approccio, seppur sembri adattarsi alle diverse esigenze, presenta alcune debolezze, principalmente legate al fatto che il patrimonio culturale è molto diverso e più complesso rispetto a quello bibliografico, risultando spesso largamente interdisciplinare. Le schede di catalogazione inoltre, presentando principalmente contenuti testuali, sono pensate essenzialmente per le persone, che sono in grado di leggerle, comprenderle e dedurre ragionamenti e connessioni riguardo al loro contenuto. Operazioni più complesse quali ricerche e filtraggi di dati specifici da parte dei calcolatori risultano spesso molto ardue o semplicistiche facendo riferimento solo alle corrispondenze di termini. Per tali ragioni nel corso degli anni numerosi sono stati i tentativi di creare degli schemi per la classificazione e la memorizzazione delle informazioni relative ai beni culturali, basati sulla definizione dei principi fondamentali per la creazione di regole di catalogazione ben precise da adottare come modelli di riferimento. Diverse organizzazioni e commissioni, nazionali e internazionali, tra cui si possono citare ICCD[2] (Istituto Centrale per il Catalogo

[2]L'*Istituto Centrale per il Catalogo e la Documentazione*, all'interno del Ministero per i Beni e le Attività Culturali (MIBAC), si occupa di definire standard, procedure e strumenti per la Catalogazione e la Documentazione del patrimonio archeologico, architettonico, storico artistico e demoetnoantropologico nazionale in accordo con le Regioni e svolge funzioni di formazione e ricerca nel settore della catalogazione. `http://www.iccd.beniculturali.it/`

e la Documentazione) e CIDOC[3] (International Committee for Documentation of the Council of Museums), hanno focalizzato i propri sforzi verso la creazione di standard che indicassero in maniera sempre più precisa e approfondita i criteri di strutturazione e catalogazione dei beni (alcuni esempi di schede di catalogazione organizzate secondo questi schemi sono contenuti in [MMS10]). I principali approcci tecnologici usati per la rappresentazione degli oggetti erano basati, almeno inizialmente, sull'utilizzo di modelli concettuali standard tipici dei database, con particolare riferimento al modello Entità-Relazione. Tali modelli consistono nell'identificazione, innanzitutto, di alcune entità di base (quali oggetti, autori o locazioni) e nella successiva individuazione di relazioni tra tali entità. Questo processo ha portato alla creazione di un modello concettuale semplice e coerente, in cui il ruolo centrale è assunto dall'oggetto rappresentato. Le principali applicazioni informatiche sono, in questo contesto, basate su database relazionali che raccolgono informazioni secondo degli schemi standard ben precisi, che hanno la pretesa di permettere la rappresentazione di oggetti che spaziano in diversi campi e risultano spesso differenti tra loro. La maggiore limitazione di questo approccio consiste nel fatto che esso rimane altamente incentrato sui singoli oggetti rappresentati, utilizzando un modello in qualche modo simile a quello delle schede di catalogazione, anche se gli oggetti sono modellati in modo più elaborato. Proprio questa visione incentrata sugli oggetti, che si occupa solo di descriverne le caratteristiche tramite una serie di proprietà, è origine di informazione spesso ridondante e non permette la rappresentazione di associazioni semanticamente più complesse. È importante sottolineare come tali limitazioni non siano legate alle specifiche tecnologie di memorizzazione utilizzate, piuttosto che al modello di rappresentazione. Questo approccio infatti non fornisce un livello di formalizzazione abbastanza alto tale da per-

[3] L'*International Council of Museums* (ICOM) è un'organizzazione internazionale che si occupa di diversi aspetti relativi principalmente alla documentazione e conservazione del patrimonio culturale; CIDOC è una commissione che fa parte di ICOM che si occupa nello specifico dei requisiti di documentazione e standard relativi a musei, archivi o organizzazioni simili. `http://cidoc.icom.museum/home(en)(E1).xml`

mettere alle macchine di effettuare compiti di comprensione e deduzione di nuova informazione o collegamento tra le diverse risorse.

1.3 Soluzioni ed esigenze per la rappresentazione dei beni culturali nel Web

1.3.1 *Ricerca di interoperabilità nello scambio dei metadati*

L'esplosione e la grande diffusione del Web ha portato ad un nuovo e notevole cambiamento nel modo di presentare e memorizzare le informazioni. Una delle principali caratteristiche di questo fenomeno, come indicato in [COY07], consiste nel fatto che ha condotto ad una radicale trasformazione dei metodi utilizzati per l'accesso alle informazioni. Se prima i dati potevano essere ottenuti solo a partire da *repository ufficiali*, quali librerie o cataloghi di musei, grazie all'avvento del Web la maggior parte delle informazioni possono essere raggiunte a partire da interrogazioni generiche che, tramite una serie di collegamenti, permettono di ricercare i dati di interesse. I siti web iniziano a diventare pertanto uno dei principali veicoli per la raccolta e la pubblicazione delle informazioni relative ai beni culturali: siti in genere creati e gestiti da entità diverse con lo scopo di rendere pubblici e facilmente accessibili i cataloghi del patrimonio culturale. Per tutte queste ragioni nell'attività di catalogazione diventa sempre più centrale il ruolo dell'interoperabilità, intesa sia in senso tecnico (e garantita in questo caso dai protocolli di comunicazione del Web) che in senso semantico, ossia orientata alla possibilità di combinare la conoscenza disponibile in sorgenti dati diverse. I principali strumenti tecnologici utilizzati per garantire lo scambio dei dati sono rappresentati dall'utilizzo di linguaggi dichiarativi, in genere basati su XML, facenti riferimento a specifici schemi che si rifanno a loro volta agli standard di catalogazione.

La diffusione del Web e la sempre maggiore esigenza di interoperabilità nello

scambio delle informazioni è stata affrontata tramite la creazione di sistemi e meccanismi che permettano il raggruppamento e lo scambio dei metadati relativi ai diversi oggetti. Uno dei sistemi di raccolta e scambio dei metadati più utilizzati in quest'ambito è rappresentato da *Dublin Core*, una sorta di formato di scambio per la condivisione di dati presenti in diverse collezioni. Come spiegato in [BAK00], Dublin Core costituisce in qualche modo una forma moderna delle schede di catalogo, in cui sono utilizzati una serie di elementi per la descrizione delle risorse. Tali elementi, chiamati anche qualificatori, sono rappresentati da identificatori unici formati da un nome preceduto da un indirizzo che fa riferimento al namespace in cui tali elementi sono definiti. In questo contesto un namespace è un vocabolario, formalmente pubblicato nel Web, che descrive gli elementi e i qualificatori tramite definizioni, etichette o qualsiasi altro tipo di documentazione. Dublin Core, e in generale i vari formati di scambio utilizzati in questo contesto, permettono di esprimere diverse proprietà relative ai vari oggetti rappresentati, in modi più o meno specifici e rigorosi a seconda delle necessità. Seppur questi approcci siano maggiormente orientati all'interoperabilità e allo scambio delle informazioni, il modo con cui le risorse vengono rappresentate, ancora una volta facendo assumere un ruolo centrale agli oggetti, continua a presentare forti limitazioni riguardo la possibilità di interconnessione semantica tra le risorse presenti nelle diverse raccolte di dati. La mera descrizione delle caratteristiche degli oggetti non permette infatti di esprimere pienamente la semantica dei dati, e le informazioni relative a diversi oggetti, o addirittura inerenti a discipline diverse, possono essere collegate solo tramite un intervento umano, manuale, ottenendo i diversi dati e filtrandoli appropriatamente. [SIG09]

1.3.2 *La rappresentazione della conoscenza mediante modelli formali*

La principale problematica legata alle tecnologie fino ad ora esposte è rappresentata, in definitiva, dall'incapacità del modello strutturale utilizzato di esprimere in modo chiaro e univoco i legami e le relazioni tra i diversi oggetti.

Uno dei principali inconvenienti per gli utenti finali consiste nella difficoltà di reperire informazioni specifiche, in quanto i motori di ricerca legati al web sono spesso poveri di semantica sia in fase di indicizzazione che in fase di ricerca. Come indicato in [SMM05], si rende pertanto sempre più necessario l'utilizzo di proposizioni più ricche dal punto di vista espressivo, che permettano una rappresentazione dell'informazione più accessibile e comprensibile rendendo più facile ed efficace l'attività di ritrovamento delle informazioni necessarie e consentendo di indicizzare le risorse esistenti in maniera più ricca rispetto al tradizionale metodo basato sulla corrispondenza di parole chiave o concetti. In presenza di documentazione strutturata, l'utente è in grado di effettuare ricerche più precise indicando non solo i termini da ricercare, ma anche quale ruolo tali termini assumono nella struttura generale del documento. Un approccio di questo genere richiede tuttavia che l'utente stesso sia a conoscenza della struttura dell'informazione e che questa sia quindi in qualche modo comprensibile e indicata in maniera chiara. Si rende perciò necessario l'utilizzo di un'infrastruttura tecnologica in cui i concetti sono identificati in maniera univoca e in cui, possibilmente, il compito di realizzare le associazioni e identificare le equivalenze sia effettuato da agenti software, che fanno riferimento a una rappresentazione delle conoscenza altamente formalizzata e abbiano la capacità di eseguire ragionamenti e inferenze.
Il principale strumento usato per avere una descrizione formale, condivisa ed esplicita delle informazioni, anche nel campo dei beni culturali, è rappresentato dall'utilizzo di particolari schemi chiamati ontologie. La principale differenza rispetto alle tecniche di modellazione descritte precedentemente consiste nel fatto che le ontologie permettono di rappresentare gli oggetti non solo tramite una descrizione delle proprie caratteristiche ma anche tramite una descrizione dettagliata delle relazioni esistenti tra essi, fornendo una sorta di linguaggio formale e specifico per la rappresentazione delle risorse e della conoscenza ad esse associata. Le ontologie si distinguono profondamente dai semplici meccanismi di classificazione: mentre questi ultimi si concentrano prevalentemente sulle esigenze di accesso all'informazioni trami-

te criteri predeterminati e codificati mediante elementi sintattici, le ontologie pongono l'attenzione sul significato dei termini e sulla struttura e la natura del dominio informativo.

Numerosi sono stati pertanto gli sforzi messi in atto per la creazione di una '*core ontology*', il cui ruolo risulta essenziale in un contesto ampio e decentralizzato come quello del patrimonio culturale e del web, in cui l'integrazione dell'informazione assume particolare importanza. L'obiettivo principale è quello di avere un modello globale ed estendibile in cui possono essere messi in corrispondenza i dati provenienti da fonti diverse ed eterogenee. In quest'ambito è molto importante, come evidenziato in [DHL03], sottolineare la distinzione tra una core ontology e gli altri meccanismi di strutturazione che vengono definiti '*core metadata*' (come Dublin Core). I metadati sono infatti pensati per essere utilizzati e compresi principalmente da un lettore umano, mentre un'ontologia è un modello formale usato da strumenti spesso automatici che provvedono all'integrazione dei dati. Se nel primo caso assumono particolare importanza fattori umani quali la leggibilità, le ontologie possono invece privilegiare la completezza e la correttezza logica anche a fronte di un livello di complessità maggiore. Nell'ambito dei beni culturali il principale modello ontologico di riferimento che può essere classificato come core ontology è rappresentato dal *CIDOC Conceptual Reference Model* (*CIDOC-CRM*) [CDG10]. Tale ontologia, diventata standard ISO nel 2006, si pone come modello di riferimento per il patrimonio culturale e permette di rappresentare, in modo formale e altamente specifico, informazioni sui beni culturali relative anche agli ambiti spaziali e temporali, supportando quindi operazioni di ragionamento ed inferenza anche in queste direzioni. Tramite ontologie di questo genere è possibile convertire in un linguaggio formale tutte le informazioni normalmente disponibili in formati testuali e necessarie per la definizione del contesto dei vari oggetti. Se i vantaggi, come precedentemente descritto, sono notevoli, le principali problematiche in questi contesti sono rappresentate dalla complessità delle operazioni di conversione e rappresentazione dei dati presenti in altri formati, attività che spesso risultano

specifiche e che difficilmente possono essere automatizzate in modi generici. Data la complessità di tali operazioni è inoltre importante che le ontologie vengano utilizzate in modo corretto.

1.4 Approcci basati sulla condivisione della conoscenza

La descrizione delle tecnologie presentate fino ad ora, che rappresentano i principali approcci descritti in letteratura per far fronte al problema della rappresentazione della conoscenza legata al patrimonio culturale, ha messo in luce come il principale obiettivo perseguito sia quello di ottenere una rete di dati fortemente interconnessi tra loro. Per perseguire tale obiettivo gli approcci presentati fino ad ora si concentrano sulla ricerca di un modello di rappresentazione il più possibile ricco, completo e formale, come quello ottenibile tramite l'utilizzo di ontologie.

1.4.1 *Beni culturali e reti P2P*

Una proposta alternativa e abbastanza particolare è fornita in [CAN05], in cui è sottolineato come le metodologie finalizzate allo sviluppo di pratiche, strumenti e formati standard che in qualche modo devono essere accolti e adottati passivamente tramite un'accettazione gerarchica possa portare in qualche modo ad un indebolimento della loro stessa diffusione. La soluzione proposta da [CAN05] consiste invece nell'adozione di tecnologie di condivisione quali il *Peer-to-Peer* anche nel campo dei beni culturali per la diffusione e lo scambio delle informazioni. L'utilizzo di questi sistemi può portare alla creazione di archivi virtuali di contenuti che in qualche modo eliminano l'esigenza di archiviazioni strutturate, permettendo un accesso all'informazione direttamente alla sua fonte, e quindi nel proprio formato nativo [GAR01]. Un sistema basato su questo approccio dovrebbe permettere pertanto la possibilità di condividere i dati sugli oggetti di interesse e, tramite delle interfacce di

ricerca, permettere di recuperare eventuali informazioni messe a disposizione in una sorta di comunità di scambio.
Sebbene sia stato ripetuto più volte che l'idea della condivisione dei dati sia importante nell'ambito dei beni culturali, e nonostante questa idea sia alla base delle tecnologie legate al Peer-to-Peer, la principale limitazione di un approccio di questo genere consiste proprio nel fatto che la totale assenza di linguaggi e modelli di descrizione rappresenta un freno troppo grande all'utilizzo stesso delle informazioni. L'idea di avere diversi fornitori di dati che permettano un'interconnessione tra i loro repository può risultare vincente e adatta per il raggiungimento degli obiettivi prima elencati, tuttavia puntare l'attenzione sui meccanismi tecnici relativi allo scambio dei dati piuttosto che al modo con cui questi possono essere rappresentati risulta senz'altro un approccio sbagliato, considerando anche il fatto che , per far fronte a questo aspetto, esistono già degli standard fortemente consolidati, come quelli tipici del Web.

1.4.2 *Soluzioni basate su architetture orientate ai servizi*

Diversi progetti ed iniziative hanno cercato di far fronte al problema della condivisione facendo riferimento ad un modello di servizi per la raccolta e la fruizione delle informazioni. Tale approccio ha portato alla creazione di vere e proprie librerie digitali, che si assumono il compito di raccogliere, secondo dei modelli ben definiti, le informazioni messe a disposizione da diversi fornitori, presentando interfacce di ricerca e di accesso in genere basate sulle tecnologie web. Si possono citare, a tal proposito, progetti quali *D-Space* [TBS03] e *Fedora* [LPS05], che si prefiggono l'obiettivo di raccogliere e preservare contenuti eterogenei, fornendo le funzionalità necessarie per definire l'intero flusso di lavoro di ogni collezione disponibile. Come descritto in [BCC09], tali approcci si basano sull'idea di un'architettura '*Service-Oriented*' per la rappresentazione dei contenuti, ossia che metta a disposizione una serie di servizi per il recupero e l'integrazione dei dati. Tra i numerosi progetti di que-

sto genere presenti in letteratura, una menzione merita sicuramente DECHO [ABV11], una sorta di framework per l'esplorazione digitale degli oggetti del patrimonio culturale. Tale framework si occupa sia del recupero dei dati da diversi database (per lo più appartenenti al settore archeologico), sia della loro gestione e rappresentazione tramite approcci ontologici di diversi livelli (a seconda degli utenti e delle esigenze a cui le varie rappresentazioni dei dati sono rivolte) nonché della rappresentazione visuale e tridimensionale degli oggetti veri e propri.

Nell'ambito dei progetti relativi alla creazione di piattaforme di raccolta dati sui beni culturali, l'iniziativa più estesa e avanzata in ambito europeo è sicuramente rappresentata dal progetto Europeana[4]. Europeana è una sorta di piattaforma dedita alla raccolta di diverse sorgenti dati (in particolare relative a musei, archivi, collezioni audiovisive e librerie), messe a disposizione da diverse istituzioni europee che operano nell'ambito della gestione del patrimonio culturale, permettendone la ricerca e l'esplorazione. Tale progetto fa utilizzo di un modello di rappresentazione dei dati (descritto in [AMD10]) molto ricco ed espandibile fondato su basi ontologiche. Tale modello, che riutilizza alcuni concetti proposti dal modello CIDOC-CRM, permette di effettuare operazioni di arricchimento dei dati messi a disposizione dalle diverse parti. Un oggetto digitale fornito da un determinato fornitore di dati può essere, ad esempio, contestualmente arricchito dai metadati di un altro fornitore che si riferiscono allo stesso oggetto o ad altre risorse ad esso inerenti, tramite un meccanismo di collegamento che permette di mostrare chiaramente la provenienza delle informazioni. In particolare, il modello dati di Europeana fa una netta distinzione tra la creazione tecnica e intellettuale presentata da un fornitore di dati (intesa come l'insieme di risorse associate agli oggetti curati da quel fornitore), gli oggetti a cui tale struttura si riferiscono e la loro rappresentazione digitale, che può essere acceduta via Web. Tale modello non si pone inoltre come uno schema fisso che detta in modo

[4]Il portale del progetto Europeana, da cui è possibile effettuare ricerche sulla vasta mole di dati raccolti, è disponibile all'indirizzo `http://www.europeana.eu/portal/`.

univoco il modo di rappresentare i dati, ma piuttosto come un ancora a cui vari modelli più specifici possono attaccarsi, rendendoli almeno parzialmente interoperabili a livello semantico, facendo comunque mantenere ai dati la ricchezza e l'espressività originali.

Le iniziative e i progetti sviluppati e appena descritti mettono chiaramente in evidenza come la direzione percorsa attualmente sia quella di esprimere i dati secondo modelli ontologici ben definiti per riuscire a intrecciarli fra loro. In un'ottica di valorizzazione dei dati relativi al patrimonio culturale, risulta molto importante tuttavia effettuare dei collegamenti non solo tra informazioni correlate presenti in diversi dataset sui beni culturali, ma anche con sorgenti dati relative a discipline diverse. Le informazioni relative ad un museo potrebbero acquistare un valore aggiunto molto forte se si arricchissero, ad esempio, di dati specifici relativi alla posizione geografica e che permettono di conoscere eventuali altri posti interessanti presenti nelle vicinanze. Problematiche di questo genere sono tipicamente affrontate dalle soluzioni tecnologiche che vengono indicate con il nome di '*Semantic Web*' e, in particolare, tramite l'utilizzo della relativa rete di dati aperti e disponibili che costituisce il *Linked Data*. Scopo del progetto di questo lavoro di tesi è proprio dimostrare come l'utilizzo di queste tecnologie consenta di esprimere ed arricchire fortemente le informazioni relative a beni artistici e culturali, permettendo sia la scoperta continua di nuova informazione che la relativa creazione di applicazioni più ricche dal punto di vista informativo.

Capitolo 2

Semantic Web e Linked Data

2.1 Il Semantic Web e l'importanza delle relazioni semantiche

2.1.1 *Basi e principi del Semantic Web*

Nel capitolo precedente sono stati presentati i principali approcci e problemi relativi all'utilizzo di strumenti informatici per la catalogazione e l'interconnessione di dati relativi all'ambito artistico e culturale. L'analisi presentata ha messo in evidenza come le limitazioni principali della maggior parte delle tecnologie utilizzate a tale scopo risiedano nella difficoltà di organizzare una mole molto vasta di dati, caratterizzati da forte eterogeneità, tramite delle strutture che ne mettano in evidenza i legami semantici. Si è inoltre sottolineato come molti degli approcci finora adottati abbiano portato alla creazione di diverse sorgenti di dati, messi a disposizione da soggetti diversi, in cui operazioni di integrazione o arricchimento risultano spesso complicate, specie se con dati inerenti a discipline diverse.
Tutte queste difficoltà non si sono, tuttavia, delineate esclusivamente nel settore dei beni culturali, ma sono invece condivise e sentite in tutti i contesti informativi caratterizzati dalla presenza di dati vasti ed eterogenei. La grande diffusione delle tecnologie legate al Web ha, inoltre, permesso una diffusione

delle informazioni molto più veloce e semplice, ampliando e moltiplicando in qualche modo la quantità di dati disponibili e facendo crescere ancora di più l'esigenza di strutturarli in modo formale per migliorarne l'accesso. Specie negli ambiti governativi, inoltre, si sta affermando sempre di più il concetto di *Open Data*, che si basa sulla volontà da parte di enti e amministrazioni pubbliche di rendere disponibili e accessibili tutti i dati di interesse generale (tra cui anche quelli relativi al patrimonio culturale) [BCE11].
Proprio in riferimento a queste problematiche generali, condivise da diversi settori, nasce l'esigenza di superare i limiti posti dai meccanismi di rappresentazione della conoscenza tradizionali, specie quelli legati al Web e pensati esclusivamente per essere compresi dai soli esseri umani. Si rende necessario quindi un Web Semantico, che aggiunga uno strato di contenuti accessibili anche a processi automatizzati, creando una conoscenza espressa esplicitando la semantica implicitamente contenuta all'interno di dati, pagine, programmi e altre risorse del Web e fornendo la possibilità di creare servizi di livello qualitativo maggiore e completamente nuovo. Solo offrendo supporti che vanno in questa direzione si permetterebbe al Web di "esprimere il proprio vero potenziale", realizzando quel sogno che *Tim Berneers-Lee*, padre e inventore del Semantic Web, descrive come un mondo in cui "i meccanismi quotidiani del commercio, della burocrazia, e delle nostre vite quotidiane saranno gestiti da macchine che interagiscono con macchine, lasciando agli umani il compito di fornire l'ispirazione e l'intuizione". [BF00]
La presenza di servizi automatizzati permette di migliorare la capacità di assistere gli utenti nella realizzazione dei loro obiettivi, 'comprendendo' maggiormente il contenuto del Web e fornendo operazioni di filtraggio, categorizzazione e ricerca delle informazioni sempre più accurate. Tale processo ha lo scopo di portare a un sistema estremamente basato sulla conoscenza che fornisce vari tipi di servizi di ragionamento specializzato, che supportino l'uomo in quanti più aspetti possibili della vita, rendendo l'accesso alle informazioni sempre più mirato, pervasivo e necessario. Come esposto in [SAA08], a causa dell'abbondanza di dati con un valore semantico intrinseco molto elevato,

della loro particolare attitudine ad essere intrecciati e del fatto che si tratta prevalentemente di informazioni di pubblica accessibilità, l'ambito dei beni culturali si dimostrana un candidato ideale per l'applicazione delle tecnologie legate al Semantic Web.

2.1.2 *La definizione di semantica*

L'obiettivo principale perseguito dalle tecnologie del Semantic Web consiste nel permettere di costruire una base di conoscenza e un linguaggio interpretabile da agenti automatici. L'approccio su cui si basa per costruire un meccanismo dalle pretese così importanti è incentrato sulla Semantica, una 'parola magica' la cui definizione diventa a questo punto essenziale.
Semantica significa fondamentalmente significato, e il significato permette un utilizzo migliore e più efficace dei dati. La semantica è anche definita come un processo di comunicazione che permette di esprimere una quantità di conoscenza tale da generare un azione. Una sequenza di simboli può, ad esempio, essere utilizzata per comunicare qualcosa che ha un significato, derivato appunto dalla conoscenza, e tale comunicazione può poi influenzare il comportamento. Spendere un po' di energie per rendere le relazioni semantiche tra i dati esplicite all'interno di un'applicazione permetterebbe di ottenere un comportamento del sistema che cambi in base al significato dei dati e, non di meno, l'utilizzo intelligente di tali informazioni anche ad altri programmi creati da persone diverse. Allo stesso modo scrivere programmi che siano in grado di comprendere la semantica permette di creare applicazioni che fanno uso di insiemi di dati esterni, eventualmente anche non previsti quando il sistema era stato progettato.
Sotto quest'ottica applicazioni in grado di combinare informazioni in modi nuovi, permettendo agli utenti di dare origine a connessioni e comprendere relazioni che prima erano nascoste, creando un valore aggiunto altrimenti invisibile, si rivelano molto potenti e dal crescente potenziale. Da semplici applicazioni in grado, ad esempio, di mostrare su una mappa informazioni statistiche ad altre tanto informative da permettere di conoscere tutti i risto-

ranti che sono disponibili nelle vicinanze di un museo che si intende visitare. Basandosi sulle architetture tradizionali si è però visto come la costruzione di applicazioni di questo tipo richiede spesso dei processi specializzati e ristretti, che utilizzano tecniche di rilevamento e integrazione delle informazioni ah-hoc e troppo specifiche a causa della difficoltà di interpretare la natura semantica dei dati presenti nel web. Il significato è, difatti, spesso assente all'interno delle sorgenti di informazioni, lasciando all'utente o a complesse istruzioni di programmazione il compito di comprenderlo e fornirlo.

2.2 Nuovi approcci alla strutturazione delle informazioni

2.2.1 *Da web a semantic web*

L'assenza di una formale strutturazione dell'informazione e la conseguente impossibilità di creare legami semantici porta alla creazione di sistemi isolati e incompleti. Il link ipertestuale rappresenta il principale mezzo con cui il Web ha cercato di collegare i diversi raccoglitori di dati, un mezzo però che si è rivelato molto fragile in quanto incapace di esprimere il significato semantico del collegamento che esso rappresenta.
Alcune pagine web aggiungono una base di semantica per i motori di ricerca tramite l'uso del tag <META>, che racchiude una serie di parole chiave. Tali parole chiave sono, tuttavia, caratterizzate dal difetto di essere isolate in quanto risultano assenti dei collegamenti tra esse che permettano di fornire un contesto più completo e significativo. Una semantica di questo tipo si rivela pertanto debole e si limita solamente alla corrispondenza esatta dei termini. Altrettanto debole risulta anche la semantica espressa tramite l'utilizzo tradizionale dei database, il cui significato dei dati è suggerito esclusivamente dalla scelta dei nomi assegnati alle tabelle e alle colonne. Ciò che non permette alla semantica di essere pienamente espressiva, in questi casi, è la mancanza di relazioni tra le parole chiave. Si può pensare, ad esempio,

ad una pagina web dedicata ai musei in cui, all'interno del tag <META> è riportata la parola 'innovazione' per sottolinearne l'importanza; a causa della difficoltà di esprimere la semantica tramite la lingua italiana, solo l'interpretazione dell'utente potrà permettere di capire se tale pagina parla di musei particolarmente innovativi o di musei che riguardano l'innovazione. L'esistenza di uno standard formale sulla disposizione dei termini permette a queste parole, che a loro volta formano un linguaggio, di aderire a delle regole di grammatica. Più questa rete di regole grammaticali e termini presi da un linguaggio si espande tramite delle relazioni, maggiore è l'arricchimento di significato e semantica che si ottiene.
Il Semantic Web rappresenta una rete di dati descritti e collegati tra loro in modo da stabilire un contesto o una semantica che aderiscono a degli specifici costrutti grammaticali e di linguaggio. Essenziale, per la costruzione di una rete semantica di questo tipo, diventa creare delle connessioni standardizzate tra informazioni correlate. Affinché ciò sia possibile è necessario, innanzitutto, che i dati siano individuati in modo unico ed indirizzabile.
Nel World Wide Web i contenuti sono collegati tra loro tramite l'utilizzo di *Universal Resource Locator* (URL), collegamenti che si basano essenzialmente sul contesto circostante (se presente) per comunicare il proprio scopo o significato. Il compito di inferire la semantica è lasciato pertanto all'utente. L'attenzione è incentrata principalmente sulla presentazione dei contenuti per renderla, ancora, più digeribile ad un consumo umano, a scapito dell'utilizzo di costrutti logici formali. Ciò che manca è una strutturazione precisa delle informazioni che indichi in modo univoco il significato delle relazioni presenti tra di esse: in un'unica parola diventa essenziale l'utilizzo di uno *schema* in grado di definire come devono essere organizzati tali dati e collegamenti. I link ipertestuali, principale strumento di collegamento tra le informazioni nel mondo del Web, sono lasciati all'iniziativa dell'utente e spesso non fanno riferimento a nessuno schema preciso, oltre a nascondere completamente alla macchina il significato di tale connessione.
L'idea di un web semantico si basa, invece, massicciamente sull'utilizzo di

diversi schemi, le ontologie, ognuno dei quali ha il compito di descrivere i diversi domini di informazione. Questo permette di creare link e vere e proprie affermazioni comprensibili alle applicazioni. Tali affermazioni (o *statements*) sono collegate tra loro tramite costrutti che sono in grado di esprimere una semantica, rappresentata dal significato del collegamento, portando alla creazione di percorsi ben definiti e significativi. Tali statement, contenendo una propria logica espressa dall'ontologia, permettono anche interpretazioni più profonde e la possibilità di inferire nuove affermazioni e, in definitiva, nuova informazione. La flessibilità e la possibilità di creare statement di diverso tipo permette inoltre la definizione di espressioni più ricche, semplificando l'integrazione e la condivisione oltre a permettere delle attività di estrazione di informazione più significative e precise, lasciando comunque i dati liberi, distribuiti e dinamici. [HFB09, SET09]

In definitiva, volendo trarre una morale, il web semantico non ha tanto la pretesa di sostituire l'attuale World Wide Web ma piuttosto quella di estenderlo, rendendolo più accessibile in modo logico e preciso per raggiungere l'obiettivo di avere una condivisione dell'informazione ubiqua e automatizzata. Da una rete fatta di pagine e collegamenti ipertestuali creati in base a parole chiave, si passa a una rete fatta di risorse collegate tra loro in base al significato, e ancora a grandi raccoglitori di informazioni, espresse tramite ontologie, collegati in base a relazioni semantiche di similitudine tra le risorse in esse contenute.

Caratteristica	Web Tradizionale	Semantic Web
Semantica	Semantica implicita	Significato espresso esplicitamente
Destinatari	Realizzato per la fruizione degli umani	Realizzato per rendere comprensibile alle macchine la semantica dei dati
Componente fondamentale	Contenuto semanticamente non strutturato	Statement formali
Collegamenti	Indicano una locazione	Indicano locazione e significato
Vocabolario primario	Istruzioni di formattazione e struttura del documento	Semantica e logica
Logica	Informale	Logica descrittiva

Tabella 2.1: *Tabella di confronto fra Web tradizionale e Semantic Web*

2.2.2 *Knowledgebase e Database*

La principale ragione per cui si parla di potenziare l'infrastruttura Web consiste nel permettere a delle applicazioni 'intelligenti' di esprimere il proprio potenziale. Anche l'applicazione più dinamica e furba è tanto intelligente quanto lo sono i dati che ha a disposizione, fornendo risultati confusi e disconnessi a fronte di input contradditori o inconsistenti. La sfida del web semantico non è quella di creare un'infrastruttura intelligente, piuttosto quella di renderla il più possibile appropriata all'integrazione di informazioni, connettendo i dati ad applicazioni web 'intelligenti' in modo che l'intera esperienza sia migliorata. Ciò che si rende necessario, è che le applicazioni possano ottenere esattamente i dati di cui hanno bisogno.
Il principale mezzo utilizzato dal Semantic Web per raccogliere e mettere a disposizione le informazioni, sotto forma di statements, consiste nell'or-

ganizzazione delle basi di dati in basi di conoscenza (o *knoledgebase*) che offrono una memorizzazione dinamica ed estendibile, simile a quella offerta dai database relazionali. Esistono, tuttavia, delle importanti differenze tra la memorizzazione organizzata all'interno di basi di conoscenza e quella tipica utilizzata dai DBMS relazionali, come riassunto in tabella 2.2.
All'interno dei database relazionali che non fanno utilizzo di particolari meccanismi per la rappresentazione della conoscenza, lo schema è espresso essenzialmente tramite la struttura, in base alla composizione e ai nomi attribuiti ai componenti delle tabelle. Le relazioni tra dati sono, invece, limitate solo ad un'unica tipologia rappresentata dall'utilizzo di chiavi esterne di riferimento (*foreign key*). Organizzare i dati in una base di conoscenza significa, invece, utilizzare dei particolari tipi di schema, le ontologie, per stabilire la struttura dei dati; le ontologie risultano molto precise e descrittive e, oltre ad essere specifiche per ogni singolo dominio di applicazione, hanno la pretesa di essere riutilizzate o combinate. L'approccio basato sulla conoscenza permette, inoltre, di avere relazioni multidimensionali e di vario genere, quali ereditarietà, appartenenza, associazione nonché relazioni logiche o che esprimono vincoli.
Ciò che però forse maggiormente distingue il modo di raccogliere i dati utilizzato nel semantic web è rappresentato dal fatto che il linguaggio adoperato per esprimere la struttura e quello usato per la descrizione delle istanze dei dati è lo stesso. I database relazionali utilizzano invece un linguaggio specifico, il *Data Description Language* (DDL), per la creazione dello schema; aggiungere una tabella o una colonna risulta, in tal caso, molto diverso dall'aggiungere una riga all'interno di una tabella, ossia dall'inserimento vero e proprio dei dati. All'interno delle basi di conoscenza invece l'uso di statement regolari serve sia per definire la struttura dello schema che per la dichiarazione di specifiche istanze di dati o individui.
Il semantic web si configura, sotto questo aspetto, come un insieme di grandi '*silos* di dati' collegati tra loro secondo delle ontologie ben definite ed accessibili, che forniscono dati integrati e connessi. Un'applicazione può accedere direttamente a tali sorgenti di informazioni e ottenere immediatamente ac-

cesso ad ampie e dinamiche basi di conoscenza, permettendo la creazione di interrogazioni che vanno ben oltre il semplice meccanismo delle tag o della ricerca per parole chiave ma che, nello stesso tempo, forniscono un'estrazione di informazioni più mirata e precisa.
È importante sottolineare, tuttavia, che risulta sicuramente sbagliato pensare al Semantic Web come ad una nuova tecnologia di memorizzazione dei dati che sostituisca l'utilizzo dei database relazionali. L'uso di una rete semantica può servire per estendere e integrare le attuali applicazioni informative. Raccogliere i dati in basi di conoscenza non implica l'utilizzo di nuove modalità di memorizzazione, bensì un'organizzazione dei dati tale da permetterne l'adesione a specifiche ontologie per rendere chiare le relazioni semantiche e rendere i dati stessi più accessibili, indipendentemente dai meccanismi tecnici a cui si fa riferimento per l'archiviazione vera e propria.

Caratteristica	**Database Relazionali**	**Knowledgebase**
Struttura	Schema semplice	Schema definito da un'ontologia
Dati	Righe di tabella	Istanze degli oggetti definiti nell'ontologia
Relazioni	Definite tramite *foreign key*	Multidimensionali e tipate
Logica	Espressa o realizzata esternamente	Realizzata tramite statement di logica formale
Unicità	Ottenuta tramite chiavi per le tabelle	Ottenuta tramite gli URI

Tabella 2.2: *Tabella di confronto fra Web tradizionale e Semantic Web*

2.2.3 *Il modello a grafo e il concetto di risorsa*

Come accennato precedentemente all'interno del Semantic Web le informazioni sono rappresentate come un insieme di asserzioni dette statement. Tali statement sono composti essenzialmente da tre parti: soggetto, predicato e oggetto; proprio per tale ragione essi vengono spesso indicati con il termine di triple. Il significato degli elementi che compongono una tripla è analogo a quello attribuito dalla grammatica linguistica. Il soggetto di uno statement indica l'oggetto rispetto al quale si sta facendo una descrizione, mentre il predicato descrive la relazione esistente tra il soggetto e l'oggetto. Il *Resource Description Framework* (RDF) rappresenta il principale strumento tecnologico usato dal Semantic Web per esprimere il significato delle informazioni e le relazioni tra i dati. Un documento RDF può delineare precisamente le relazioni tra degli oggetti facenti parte di un vocabolario costruendone una sorta di rappresentazione grammaticale; le asserzioni fatte in diversi documenti RDF possono, inoltre, essere combinate per fornire un'informazione più completa, fornendo una struttura di interrogazione molto potente e flessibile. [AS06]
La ricerca di informazioni all'interno di grafi RDF è resa possibile grazie ad un linguaggio di interrogazione standard chiamato SPARQL (*SPARQL Protocol and RDF Query Language*). Una funzionalità particolarmente importante messa a disposizione da questo linguaggio consiste nella possibilità di specificare, direttamente nelle query, il grafo al quale le interrogazioni devono essere dirette. Ogni query SPARQL è sempre riferita ad almeno un grafo di default, tuttavia è possibile specificare, tramite degli URI di riferimento, uno o più grafi di destinazione, che possono essere ottenuti anche tramite un processo di unione. SPARQL permette anche di tenere traccia del grafo da cui proviene una specifica risposta a un'interrogazione, tenendo conto anche di eventuali grafi che non appartengono o restringono il grafo di default.
Nel precedente capitolo si è visto come il problema più importante legato ai modelli di rappresentazione della conoscenza in ambito culturale utilizzati dai principali approcci proposti in letteratura consista nella centralità attribuita

agli oggetti, descritti tramite una mera esposizione delle loro caratteristiche. Il modello a grafo dell'RDF permette, invece, di individuare e integrare tra loro diverse risorse, consentendo una descrizione formale per ognuna di esse e creando quindi un contesto informativo più ricco e utilizzabile.
La principale caratteristica del modello a grafo appena descritto e del linguaggio RDF è rappresentata dal fatto che qualsiasi oggetto del dominio di interesse è concettualizzato e inteso come una risorsa. In RDF una risorsa rappresenta qualsiasi cosa a cui può essere assegnato un nome; la risorsa stessa anzi non è altro che un nome che rappresenta un oggetto, un'azione o un concetto. Affinché tale modello di rappresentazione non contenga delle ambiguità è necessario che le risorse siano rappresentate univocamente. Lo strumento utilizzato per ottenere tale importante caratteristica è l'URI[1]. I soggetti, predicati e oggetti degli statement RDF sono, a parte casi particolari, delle risorse identificate tramite URI. Gli URI costituiscono un aspetto fondamentale per l'infrastruttura di condivisione e collegamento dei dati in quanto esistono all'interno di un unico namespace universale: ciò significa che qualsiasi statement che ha come soggetto una specifica risorsa identificata con uno specifico nome descrive in modo univoco e senza possibilità di ambiguità quella specifica risorsa, indipendentemente dal luogo in cui tale affermazione viene fatta. Grazie ai meccanismi attraverso i quali le diverse organizzazioni creano dei nomi per le cose alle quali sono interessate, meccanismi ormai consolidati e parte integrante del Web, si evita il problema che due gruppi diversi scelgano lo stesso nome per entità diverse. Gli URI forniscono, quindi, alle risorse un nome che risulta universalmente unico e rimane valido in qualsiasi contesto, permettendo alle informazioni espresse in RDF di essere portabili e di descrivere sostanzialmente qualsiasi cosa in qualsiasi dominio, da oggetti concreti a concetti astratti.
È inoltre importante sottolineare che gli URI utilizzati all'interno dell'RDF

[1]L'*Uniform Resource Identifier* (URI) risulta una parte essenziale dell'infrastruttura del World Wide Web, e rappresenta un meccanismo standard per indicare e denominare le risorse. Alcuni URI includono anche le informazioni che indicano come accedere alle risorse che rappresentano in Internet, e vengono denominati URL (*Uniform Resource Locator*).

hanno lo scopo di identificare una generica rappresentazione della risorsa, mentre il compito di localizzare una specifica rappresentazione serializzata delle informazioni disponibili su tale risorsa è lasciato agli URL. L'URI di una risorsa, in questo contesto, non si riferisce, quindi, ad una sua specifica rappresentazione fisica; in altre parole, tale URI può non produrre nessuna informazione se digitato all'interno di un web browser. Tale rappresentazione, la cui esistenza e disponibilità è considerata una buona pratica (i principi legati alla distribuzione delle risorse nel Linked Data sono descritti nella sezione 2.4.2), è invece identificata da uno specifico URL. [HFB09]

2.3 Le ontologie

La rappresentazione delle risorse a grafo costituisce un metodo flessibile e funzionale per rappresentare la conoscenza. L'analisi di entità e relazioni tra gli oggetti rappresenta tradizionalmente uno degli aspetti più discussi nell'ambito della filosofia. Molti filosofi si sono spesso preoccupati di conoscere cosa esiste, come poterlo descrivere e, soprattutto, dopo aver capito l'esistenza di un oggetto, come poterlo collocare tra tutti gli altri già esistenti. Tali sforzi ricadono nel dominio chiamato, in ambito filosofico, con il termine di ontologia. Dato che anche i sistemi software che si basano sulla semantica hanno bisogno di un modello che descriva la realtà da cui estrarre conoscenza relativa al dominio in cui stanno operando, anche in quest'ambito si sente il bisogno di definire delle ontologie. Al contrario dei filosofi, che si preoccupano di capire il funzionamento dell'intero Universo, le ontologie del Semantic Web si devono preoccupare solo di descrivere le relazioni degli oggetti che sono di interesse per le specifiche applicazioni.
La costruzione di modelli per la descrizione dei dati non rappresenta qualcosa di nuovo o di cui si è iniziato a parlare parallelamente alla nascita dei concetti relativi al web semantico: i progettisti di sistemi usano da sempre modelli di questo tipo, da quelli utilizzati per definire schemi per i database a quelli per la specifica delle relazioni tra oggetti. Tuttavia all'interno di sistemi che

esplicitano la semantica, è necessario esprimere questi modelli in modo che i vari sistemi distribuiti nel web siano in grado di leggere e comprendere con precisione come questi devono essere adoperati. [SET09]
Esistono vari concetti e metodi che si occupano di rappresentare, classificare e disambiguare i contenuti semantici per estrarne conoscenza, tutti concetti che possono essere fatti ricadere nella definizione di ontologia. In figura 2.1 è rappresentato quello che in [DOS03] viene chiamato *Ontology Spectrum*, un diagramma che racchiude alcuni di questi concetti e modelli classificandoli in una sorta di spazio ontologico, mostrandone le relazioni e comparandoli in base alla ricchezza semantica offerta. Nel diagramma sono riportati sia i principali modelli ontologici concettuali (dai vocabolari alle teorie logiche) che linguaggi e soluzioni pratiche, da quelli utilizzati nell'ambito dei database, come il linguaggio entità relazione (ER) e il modello *Extended Entity-Relational* (EER), che altri utilizzati nell'ambito *Object-Oriented* come il linguaggio *Unified Modeling Language* (UML). Salendo all'interno del diagramma la ricchezza semantica cresce: la parte più bassa racchiude le 'semantiche deboli', ossia quelle che permettono di esprimere significato molto semplice e basilare, salendo si arriva a 'semantiche forti', in grado di esprimere significato arbitrariamente complesso. Di seguito si propone una breve descrizione di alcuni di questi modelli ontologici:

- Un *vocabolario* è una collezione di termini definiti in modo non ambiguo e utilizzati per comunicare; i termini del vocabolario non dovrebbero essere ridondanti a meno di specifiche identificazioni di ridondanza, inoltre tali termini devono avere un significato consistente in tutti i contesti di dominio. Il vocabolario rappresenta una maniera di base per tentare di esprimere significato.

- Una *tassonomia* è un vocabolario in cui i termini sono organizzati in maniera gerarchica. Ha un potere descrittivo maggiore in quanto si occupa non solo di definire gli oggetti del dominio ma anche di indicarne delle relazioni. Tali relazioni sono di tipo 'padre-figlio', tuttavia la

semantica di tale rapporto rimane spesso mal definita o non definita affatto; in genere si tratta di relazioni di sottoclasse o di contenimento.

- Il *tesauro* prende in considerazione le relazioni tra termini (parole o frasi), a loro volta strutturati in una tassonomia; si tratta pertanto di una tassonomia con delle relazioni semantiche aggiuntive tra termini. Tali relazioni possono essere di diverso tipo: gerarchiche, associative o di equivalenza. In altre parole, un tesauro è una sorta di vocabolario controllato che supporta nell'attività di recupero di informazione sia chi ha assegnato e creato i metadati che chi intende ricercare delle informazioni, permettendo loro di utilizzare gli stessi termini per gli stessi concetti. Un tesauro assicura che i concetti siano descritti in modo consistente e che gli utenti esperti possano facilmente affinare le proprie ricerche per localizzare l'informazione di interesse.

- Un *modello concettuale* è un modello relativo ad un'area di conoscenza, in genere chiamata dominio, che ne rappresenta le entità primarie (oggetti del dominio) e le relazioni tra esse, gli attributi ed i rispettivi valori (in genere chiamati proprietà) relativi a tali entità e relazioni e, a volte, delle regole che associano entità, relazioni e attributi in modo più complesso. Le regole permettono di generare nuova conoscenza basandosi su quanto espresso dal modello concettuale e dalle proprietà assegnate alle varie entità e relazioni. Il principale tipo di relazione utilizzato in quest'ambito è il concetto di 'sottoclasse', che permette alle entità di ereditare proprietà e attributi fornendo una base per la creazione di regole logiche. Il principale strumento proposto dal Semantic Web per esprimere modelli concettuali è il linguaggio *RDF-Schema*.

- Le *teorie logiche* rappresentano il punto più alto dell'ontology spectrum. Le ontologie rappresentate come teorie logiche sono direttamente interpretabili semanticamente dal software, e sono basate su assiomi (degli statement, da semplici a complessi, che si assume siano veri) e regole di inferenza (ossia regole che, date delle assunzioni, giungono

a delle conclusioni). In quest'ambito oltre alle relazioni di sottoclasse sono presenti relazioni più complesse e ricche, che possono esprimere la disgiunzione piuttosto che la transitività. Lo strumento utilizzato dal Semantic Web per esprimere ontologie complesse che rappresentano teorie logiche è il linguaggio OWL (*Web Ontology Language*).

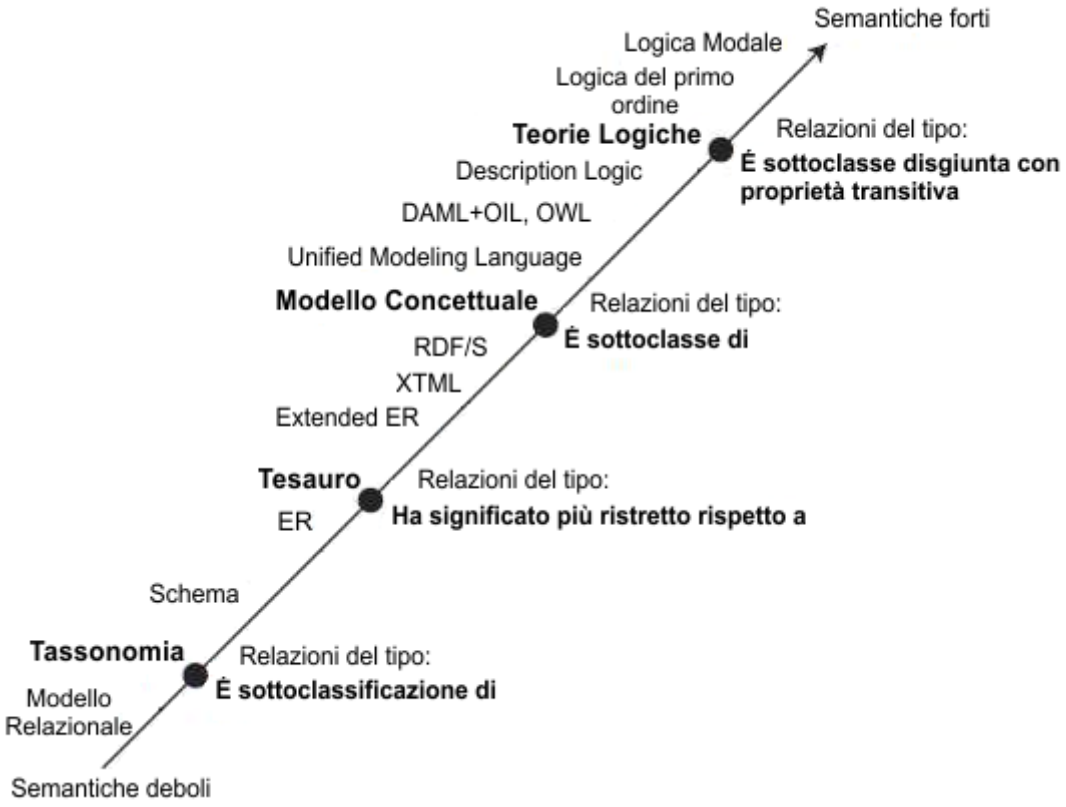


Figura 2.1: *Ontology Spectrum: diagramma dei vari tipi di ontologia in base alla forza semantica, come proposto in [DOS03].*

Le tecnologie messe a disposizione dal Semantic Web si collocano tra i punti più alti di questo diagramma, permettendo di esprimere in modo completo il significato ontologico dei dati spiegando come questi devono essere interpretati. L'ontologia può, sotto quest'ottica, essere pensata come ad una sorta di contratto sociale tra il fornitore e il consumatore dei dati. I principali

strumenti pensati per la definizione di questi modelli sono rappresentati dai linguaggi RDFS e OWL. Essi permettono la creazione di documenti RDF strutturati esattamente come qualsiasi altro generico documento RDF, ma facendo utilizzo di una serie di costrutti aggiuntivi identificati da uno specifico namespace. L'utilizzo di tali costrutti permette la creazione di triple che si occupano di definire la struttura di un vero e proprio modello concettuale; tali triple si differenziano da quelle che si riferiscono, invece, ai dati veri e propri, che ne rappresentano una sorta di istanza e che dovranno essere espressi tramite degli statement RDF che devono aderire al modello concettuale definito. Mentre RDFS permette di definire le ontologie prevalentemente tramite la definzione di classi e proprietà tra esse (i cui dati veri e propri ne rappresenteranno delle istanze), OWL aggiunge anche funzionalità più avanzate quali possibilità di definire catene e restrizioni di proprietà, *datatype* specifici o definire veri e propri insiemi di regole per l'asserzione di nuova informazione (tramite il linguaggio *SWRL - Semantic Web Rule Language* [2]). [HFB09]

In molti sistemi di modellazione esiste una netta divisione tra i dati e il loro schema. Lo schema di un database relazionale, ad esempio, non è espresso in una forma tabellare, così come un DTD non ha la forma di un documento XML valido. Spesso le versioni moderne di questi sistemi tendono a modellare lo schema nello stesso formato dei dati. Nel caso dell'RDF invece il linguaggio per la definizione dello schema è stato pensato e definito in RDF sin dall'inizio. Le relazioni tra le risorse 'semplici' e le risorse che fanno parte dello schema sono infatti definite tramite triple, proprio come le relazioni tra tutte le altre tipologie di risorse. Questa eleganza nella modellazione permette di rendere particolarmente facile la fruizione di una descrizione formale della semantica, semplicemente fornendo delle regole di inferenza che lavorano su dei pattern di triple. Non si tratta però solo di una buona pratica

[2] Il *Semantic Web Rule Language* (SWRL) è un linguaggio proposto dal W3C per esprimere sistemi di regole che si basano sulla veridicità di determinati *statement* per asserirne degli altri; tale possibilità permette di modellare eventuali situazioni in cui le classi e le proprietà già definite da OWL non siano sufficienti.

ingegneristica. In RDF tutto è espresso tramite triple: anche il significato della nuova informazione asserita in base alle regole è espresso tramite delle nuove triple. La struttura che guida questa inferenza, che descrive il significato dei dati, è a sua volta espresso in triple. Ciò significa che tale processo può continuare per quanto è necessario e le informazioni relative allo schema che permettono di fornire un contesto per l'informazione del Semantic Web possono, a loro volta, essere distribuite nel Semantic Web stesso. [AH08]

2.4 Linked Data e Web of Data

2.4.1 *Il Linked Data*

Sono stati fino ad ora descritti i principi e le tecnologie che sono alla base del Semantic Web. In particolare si è visto come queste tecnologie possano essere usate per mettere in risalto ed esprimere il significato e la conoscenza presente nelle informazioni. L'obiettivo principale di questa evoluzione, tuttavia, non è solo quello di esporre la semantica dei dati attraverso modi e formati nuovi, ma anche quello di renderli fruibili in maniera più intelligente e diretta, valorizzandoli e permettendo di utilizzarli in applicazioni e modalità sempre più interessanti. Alla base di questa evoluzione si collocano infatti una serie di pratiche e principi relativi alla pubblicazione e connessione di dati strutturati nel Web che danno vita a ciò che viene chiamato Linked Data.
Ci sono diversi modi per creare una rete di 'dati collegati' tra loro, tuttavia si è visto come, specie nel campo dei beni culturali, in assenza di pratiche standard tali informazioni finiscano per essere circoscritte a specifici e ristretti confini creando delle isole di dati spesso di fatto isolate e poco accessibili. Il termine Linked Data, creato da Tim Berners-Lee, viene spesso utilizzato quasi come sinonimo di Semantic Web, quando si vuole però mettere in evidenza una vera e propria rete di dati che connette informazioni appartenenti ai più disparati domini e permette la creazione di nuove applicazioni che possano accederci, creandone liberamente un valore aggiunto. [BHB09]

Il presupposto che sta alla base di questa idea si fonda sul principio che il valore e l'utilità dei dati cresce quanto più questi sono interconnessi tra loro. In sintesi quindi il Linked Data non è altro che l'utilizzo del Web per creare degli specifici collegamenti tra i dati di diverse sorgenti, alle quali si fa spesso riferimento con il nome di silos. I dogmi di tale approccio consistono nell'utilizzare l'RDF come modello per la pubblicazione di dati strutturati nel Web e i collegamenti che l'RDF mette a disposizione per interconnettere i dati dei diversi silos, nonché l'utilizzo del protocollo HTTP per effettuarne l'accesso. L'applicazione di tali principi porta alla creazione di una vasta rete di dati, uno spazio in cui persone e organizzazioni possono inserire ed usufruire di informazioni riguardanti qualsiasi cosa; tale spazio viene spesso definito come *Web of Data*, altro termine spesso utilizzato per fare riferimento al Semantic Web.

Il Web of Data può essere acceduto utilizzando dei particolari browser chiamati *Linked Data Browser* che, proprio come fanno i browser HTML nel web tradizionale, permetto agli utenti di navigare tra le diverse sorgenti di dati seguendo i link RDF da esse proposti. Ciò permette di iniziare la consultazione da un determinato *data source* per poi navigare spostandosi all'interno di una rete potenzialmente infinita di informazioni connesse tra loro. Per fare un esempio, durante la consultazione dei dati riguardanti uno specifico museo all'interno di un determinato silos di dati, un utente potrebbe essere interessato alle informazioni riguardanti la città in cui esso si trova: seguendo l'appropriato link RDF, l'utente può navigare e ottenere i dati riguardanti tale città, contenuti in un altro dataset. Allo stesso modo esistono anche dei motori di ricerca basati sul Linked Data che analizzano l'intera rete di dati seguendo i link tra le diverse sorgenti e permettono di effettuare ricerche e interrogazioni sui nuovi dati aggregati. [BCH07]

Utilizzando gli URI per l'identificazione delle risorse, il modello RDF per la loro descrizione e il protocollo HTTP come meccanismo di accesso, il Linked Data si basa essenzialmente sull'architettura generale del Web. Il Web of Data può pertanto essere visto come uno strato aggiuntivo che è strettamente

intrecciato con il Web tradizionale.

2.4.2 *I principi del Linked Data*

Come già più volte asserito, il termine Linked Data si riferisce ad una serie di pratiche da utilizzare per la pubblicazione di contenuto strutturato nel Semantic Web. Tali tecniche sono state definite da Tim Berners-Lee in un articolo riguardante l'architettura del Web [BER07], e possono essere riassunte nei seguenti quattro principi, a cui ci si riferisce spesso con il nome di *Linked Data Principles*:

1. Utilizzare gli URI per identificare le risorse;
2. Utilizzare gli URI in conformità al protocollo HTTP, in modo che possano essere consultati;
3. Utilizzare gli standard (RDF e SPARQL) come modello per la rappresentazione dei dati;
4. Includere link agli URI che identificano altre risorse, in modo che possano essere scoperte nuove informazioni.

Come precedentemente descritto (sezione 2.2.3), il linguaggio RDF si basa sull'utilizzo degli URI come strumento per l'identificazione delle risorse. Tali risorse possono essere sia documenti web o contenuti digitali, che oggetti del mondo reale piuttosto che concetti astratti. Affinché i dati possano aderire a questi principi, è però necessario che venga fornita una sorta di rappresentazione di tali risorse, obiettivo che non si raggiunge in automatico semplicemente facendo utilizzo del linguaggio RDF.
Il protocollo HTTP rappresenta un meccanismo di accesso universalmente utilizzato all'interno del Web. Utilizzare gli URI in conformità a tale protocollo, significa combinare un'identificazione univoca globale con un meccanismo semplice e conosciuto di recupero delle informazioni. Seguire il secondo principio significa pertanto identificare le risorse facendo utilizzo di particolari URI che possano essere deferenziati tramite il protocollo HTTP, ossia

permettano di ottenere una sorta di descrizione del concetto o dell'oggetto identificato da tale indirizzo.
Il quarto principio chiede di utilizzare i link come strumento per la connessione delle risorse. Nell'ottica del Linked Data, tale principio indica di connettere tramite link in particolare risorse che si trovano in silos di dati differenti, in modo da creare un percorso che permetta di spostarsi tra le diverse sorgenti di dati, rendendo sempre più interconnessa quella rete di dati di cui si è parlato precedente. Il principale strumento che permette di stabilire come tali interconnessioni devono essere adottate è l'utilizzo e l'esposizione delle ontologie: ogni sorgente di dati in genere utilizza infatti una propria ontologia, specifica per il proprio dominio, che permette di comprendere come le risorse sono rappresentate e quindi che tipo di interconnessioni possono avere.

2.4.3 *Il Linking Open Data Project*

Come precedentemente descritto, quando si parla di Linked Data si fa spesso riferimento ad una visione del Semantic Web più pratica e orientata alla condivisione vera e propria di informazioni organizzate secondo specifici principi. Adottare le tecniche proposte dal Linked Data permette di entrare a far parte di quella rete di dati le cui caratteristiche sono state appena descritte e i cui vantaggi appaiono chiari. Il Web of Data, tuttavia, non è solo un concetto teorico dalle grandi potenzialità, esistono infatti diversi progetti basati sulle tecnologie e sulle pratiche fino ad ora descritte che hanno portato alla nascita di una vera e propria rete di dati aperta e consultabile.
Uno dei più visibili esempi di adozione e applicazione dei principi del Linked Data si può trovare nel *Linking Open Data Project*[3], un progetto sviluppato da una comunità fondata nel Gennaio del 2007 e supportata dal '*W3C Semantic Web Education and Outreach Group*'. Lo scopo iniziale del progetto,

[3]Il sito web del progetto, presentato come wiki, che contiene varie informazioni sulle attività svolte e gli individui e organizzazioni partecipanti, è disponibile all'indirizzo: `http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData`.

e quello tuttora supportato, consiste nel costruire una rete di dati concreta identificando eventuali dataset già esistenti e disponibili tramite licenze aperte, convertendoli in formato RDF secondo i principi del Linked Data e pubblicandoli liberamente sul Web.
Inizialmente i partecipanti a questo progetto erano essenzialmente ricercatori o sviluppatori che lavoravano nell'ambito di lavori di ricerca universitari o piccole aziende. Nel corso degli anni, però, il progetto è cresciuto notevolmente, arrivando a coinvolgere anche grandi aziende come la *BBC* o la *Thomson Reuters*. Tale crescita è stata permessa grazie alla natura aperta del progetto, in cui ognuno può partecipare semplicemente pubblicando i propri dati, attenendosi alle pratiche relative al Linked Data, e interconnettendoli con quelli dei dataset già esistenti.
Il diagramma in figura 2.2 mostra l'attuale dimensione dei dati raccolti nell'ambito di questo progetto aggiornata a Settembre 2010. Ogni nodo del diagramma rappresenta uno specifico dataset pubblicato come Linked Data, mentre gli archi indicano i collegamenti esistenti tra oggetti presenti in due dataset connessi. Gli archi più spessi corrispondono a un numero di collegamenti più elevato tra le due sorgenti a cui fanno riferimento, mentre quelli bidirezionali indicano che esistono dei collegamenti in uscita da entrambi i dataset.
Il contenuto di questo diagramma, che viene spesso paragonato ad una 'nuvola di dati' (*Linked Data Cloud*), ha una natura molto variegata ed eterogenea, e comprende dati che riguardano locazioni geografiche, persone, aziende, libri, pubblicazioni scientifiche, film, musica, programmi televisivi e radiofonici, informazioni su geni, proteine, droghe e trattamenti clinici, comunità online, dati statistici nonché varie tipologie di dati governativi o pubblici. Tutti i silos, ognuno dei quali incentrato su specifici domini, sono collegati tra loro permettendo di navigare in questa sorgente così variegata e vasta di informazioni.
Come si può vedere in figura 2.2, alcuni dataset raccolgono un gran numero di collegamenti, formando una sorta di fulcro di raccolta dei vari link. I più

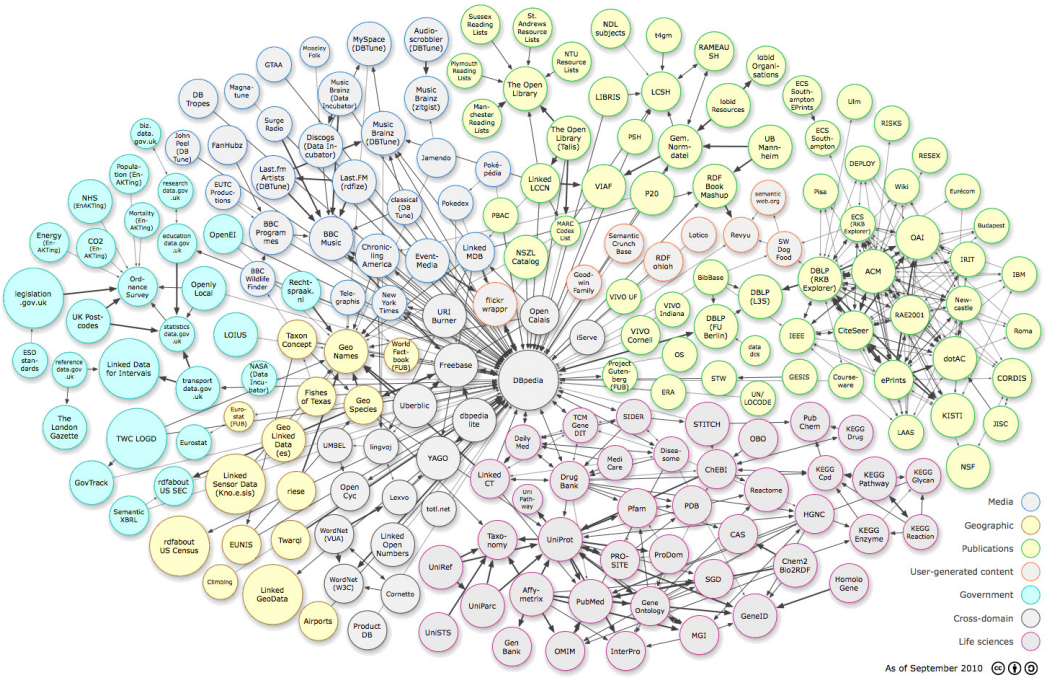


Figura 2.2: *Linking Open Data cloud diagram, diagramma che racchiude tutti i dataset del Linking Open Data Project presenti fino a Settembre 2010 (fornito da Richard Cyganiak and Anja Jentzsch, http://lod-cloud.net/)*

evidenti casi di questo genere sono rappresentati dai dataset di *DBpedia*, che raccoglie triple RDF estratte dalle informazioni presenti negli articoli di *Wikipedia*, e di *Geonames*, che fornisce invece descrizioni RDF relative a milioni di locazioni geografiche. La ragione dell'importanza attribuita a queste sorgenti di dati consiste nel fatto che, oltre ad essere state tra le prime ad essere inserite nel progetto, raccolgono e forniscono URI e descrizioni RDF relative a moltissime entità o concetti molto comuni, e vengono pertanto referenziati molto frequentemente da altri dataset che raccolgono, invece, informazioni più specializzate. [BHB09]

La rete di dati ottenuta grazie a questo progetto, e il concetto di Linked Data in generale, si candidano quindi come modello ideale da utilizzare per

l'arricchimento e l'integrazione dei dati sui beni culturali. Si è più volte sottolineato, infatti, come in tale ambito risulti particolarmente importante fornire informazioni ampie e dettagliate su tutti i diversi aspetti inerenti un determinato oggetto, ottenendo un valore aggiunto notevole specie quando tali informazioni appartengono a discipline diverse. Il Web of Data rappresenta pertanto lo spazio ideale in cui i dati relativi al patrimonio artistico e culturale possono assumere un significato e un valore molto più ricco, diventando veramente utilizzabili e disponibili per chiunque ne abbia necessità di impiego negli ambiti e nei modi più disparati. Dati relativi a musei o ad opere in essi contenute possono essere arricchiti tramite le informazioni enciclopediche proposte da DBpedia, piuttosto che tramite quelle geografiche e di localizzazione messe a disposizione da GeoNames, o ancora valorizzati da eventuali nuovi dataset che in futuro popoleranno questa nuvola di dati. Come descritto in [BCE11] tutte queste tecnologie rappresentano quindi il principale riferimento del quale bisogna tenere conto quando si intende pubblicare e mettere a disposizione dati aperti e di interesse generale. Tutte queste considerazioni hanno portato numerose amministrazioni pubbliche, a partire da quella inglese e americana, fino all'esempio italiano della regione Piemonte, a pubblicare numerosi dati di pubblica utilità (da quelli culturali a quelli di bilancio) inserendoli all'interno di tale rete informativa. Le stesse considerazioni hanno portato allo sviluppo del progetto presentato in questa tesi, orientato a convertire nei formati precedentemente descritti informazioni relative a beni artistici e culturali della regione Emilia Romagna, rendendoli pronti per la divulgazione all'interno del Linked Data.

Capitolo 3

Il progetto OpenERCH

3.1 Descrizione generale

Nei capitolo precedenti è stata presentata un'analisi delle principali caratteristiche e problematiche legate all'attività di catalogazione e divulgazione dei dati riguardanti beni artistici e culturali. Tale analisi ha portato all'individuazione delle tecnologie legate al Semantic Web e, più nello specifico, al Linked Data, come gli strumenti attualmente più adatti per far fronte alla pubblicazione e rappresentazione della conoscenza relativa a tali dati. Il progetto *OpenERCH* (*Open Emilia Romagna Cultural Heritage*), oggetto di questo lavoro di tesi, ha proprio lo scopo primario di convertire dati relativi all'ambito artistico e culturale, in particolare riferiti a musei ed edifici storici, della regione Emilia Romagna nei formati e nei modi proposti dal Semantic Web e precedentemente descritti. Tale attività di conversione ha tenuto conto delle regole e dei principi proposti dal Linked Data, permettendo ai dati di esprimere formalmente il proprio significato semantico, aderendo ad una specifica ontologia, e consentendo la creazione di collegamenti in uscita verso altri dataset presenti nel Web of Data. Il progetto è stato realizzato su proposta dell'azienda *Nier Soluzioni Informatiche*[1] di Bologna con la colla-

[1]NSI è un'azienda bolognese molto attiva, nata nel 2002, che si occupa prevalentemente dell'integrazione di sistemi informativi e dello sviluppo software di vario tipo, risultando

borazione dell'*Istituto per i Beni artistici, Culturali e naturali della regione Emilia Romagna* (IBC)[2], che ha fornito i dati di interesse.

I dati di riferimento, oggetto del lavoro di conversione, sono presenti nel portale *Samira*[3], realizzato dall'IBC: un sistema informativo che contiene il catalogo del patrimonio culturale della regione. Tale catalogo è consultabile liberamente tramite un apposito sito web che permette attività di ricerca, navigazione e consultazione delle informazioni sui beni. Le descrizioni dei vari oggetti sono presentate essenzialmente nella forma tipica delle schede di catalogazione: essa è infatti affidata a schede informative di vario genere che ne espongono, prevalentemente tramite contenuti testuali, le varie caratteristiche, come mostrato in figura 3.1. I dati presenti all'interno del catalogo sono essenzialmente di due tipi: edifici ed oggetti contenitori (quali musei, teatri storici, luoghi d'arte contemporanea, ecc.) e beni di catalogo veri e propri in essi contenuti (opere ed oggetti d'arte, reperti archeologici, fotografie, ecc.). Il progetto OpenERCH fa riferimento principalmente alla prima tipologia di dati, basandosi su un dataset in cui sono contenuti dati relativi ad oltre mille siti tra musei, teatri e altri luoghi di interesse artistico.

L'obiettivo principale dell'attività svolta è quello di rendere i dati disponibili in formato aperto, consentendone l'utilizzo non solo a chi ha interessi generici di consultazione ma anche a chi intende sfruttare tali informazioni correlandole con le altre reperibili nel Linked Data per ottenere un arricchimento informativo e un valore aggiunto o per dare vita ad applicazioni di vario genere. È stato ampiamente descritto come tale obiettivo sia molto

---

nel contempo molto attenta alle nuove tecnologie e ai nuovi modelli proposti dal mondo informatico. `http://www.nsi-online.it/web/guest`

[2]L'Istituto per i beni artistici, culturali e naturali della Regione Emilia-Romagna (IBC) è un organo che si occupa di svolgere attività conoscitiva ed operativa, di indagine e di ricerca, per la valorizzazione ed il restauro del patrimonio storico e artistico e per lo svolgimento di ogni funzione relativa ai beni artistici, culturali e naturali. Esso esercita le funzioni amministrative di competenza regionale relative alla materia 'musei e biblioteche di enti locali'. `http://www.ibc.regione.emilia-romagna.it`

[3]Il portale *Samira* è raggiungibile all'indirizzo: `http://bbcc.ibc.regione.emilia-romagna.it/samira/v2fe/index.do`

Figura 3.1: *La scheda con la descrizione di un museo così come presentata dal portale 'Samira'*

difficile da raggiungere con le tecniche e gli approcci informatici tradizionali che non fanno uso di modelli formali di rappresentazione. L'esistenza di una vera e propria rete di dati aperti molto ampia e già fruibile, quale quella offerta dal Linking Open Data Project (sezione 2.4.3), con i quali intrecciare le informazioni culturali disponibili rappresenta inoltre un contesto ideale per l'esposizione dei contenuti presi in considerazione da OpenERCH. È quindi importante sottolineare come lo scopo non sia quello di permettere una consultazione visuale delle descrizioni di ogni oggetto, attività tra l'altro già

portata a termine efficacemente dal portale Samira, bensì quello di esporre i dati per eventuali utilizzi anche da parte di applicazioni o agenti software automatici, proprio nell'ottica tipica del Semantic Web. Al fine di dimostrare eventuali utilizzi di questo genere, nell'ambito del progetto è stata sviluppata anche l'applicativo di esempio *ERCH-Mashup*, un'applicazione di mash-up[4] che permette di effettuare ricerche tra i metadati relativi ai musei e agli altri oggetti contenitori, aggiungendo dove possibile arricchimenti ottenuti da informazioni recuperate da altri dataset esterni presenti nel Linked Data. Tale applicazione rappresenta, quindi, solo un esempio di utilizzo dei dati messi a disposizione da OpenERCH, e la sua realizzazione va pensata in modo indipendente rispetto all'attività di conversione semantica e collegamento con il Web of Data. Di seguito si espongono, dopo la descrizione dei dati di base presenti nel catalogo, le principali caratteristiche e utilizzi dei risultati ottenuti tramite tali due diverse attività.

3.2 Collezioni di dati gestite in OpenERCH

I dati oggetto di conversione si riferiscono a luoghi di interesse artistico e culturale dislocati in tutta la zona della regione Emilia Romagna. Nel catalogo proposto in Samira sono raccolte informazioni che fanno riferimento a circa un migliaio di siti/contenitori di varie tipologie, nello specifico catalogati come: musei, teatri storici, luoghi del percento per l'arte, luoghi d'arte contemporanea, nuclei e collezioni o altri contenitori generici.
Ogni oggetto presenta una serie di informazioni di base necessarie per la sua identificazione e localizzazione. Dal portale è possibile reperire questi metadati generali relativi ad ogni oggetto catalogato effettuando una ricerca: il sito permette infatti di ottenere i risultati in diversi formati. Di seguito si pre-

[4]Un *mash-up*, in termini informatici, indica un'applicazione che usa contenuti ottenuti da più sorgenti per creare dei servizi completamente nuovi. Nell'ambito del Semantic Web un mash-up rappresenta l'esempio tipico di applicazione realizzabile, in cui vengono sfruttati i collegamenti tra i vari silos di dati per consentire la scoperta di nuova informazione e per arricchire i dati iniziali intrecciandoli in qualche modo con quelli esterni.

sentano i dati generali associati ad ogni oggetto, così come presenti nel portale Samira, che forniscono l'insieme di metadati generale d'individuazione di ognuno di essi:

- Denominazione del sito o contenitore, in formato testuale; in alcuni casi è possibile che diversi oggetti abbiano lo stesso nome: si pensi ad esempio ai musei civici o archeologici presenti in diverse città, a meno di specifiche denominazioni (magari dedicate a qualche personaggio illustre), tali musei sono indicati con lo stesso appellativo (come 'Museo Civico' o 'Museo Archeologico') in quanto il nome della città di appartenenza non è incluso nella denominazione.

- Provincia di appartenenza, indicata tramite la sigla a due caratteri.

- Città di appartenenza, nell'ambito della provincia prima indicata.

- Indirizzo di ubicazione del sito o contenitore, espresso in formato testuale in genere indicando il nome della via (o piazza) e il numero civico separati da una virgola (senza indicazione della città o provincia); l'unione di questi ultimi tre elementi permette di ottenere un indirizzo completo utile per la localizzazione univoca.

- Macrocategoria di riferimento (indicata come 'classe'): ad ogni oggetto è assegnata un'unica categoria di questo tipo, non è quindi possibile che un oggetto appartenga a più macrocategorie distinte; le macrocategorie presenti nel catalogo sono le seguenti: archeologia, arte, arte e archeologia, composito, etnografia e antropologia, scienza e storia naturale, scienza e tecnica, specializzato, storia e territorio. Alcuni oggetti non hanno nessuna categoria assegnata.

- Sottocategorie specifiche (indicate come 'sottoclassi'): oltre alla macrocategoria ogni oggetto può avere associate uno o più sottocategorie più specifiche (nell'ambito della macrocategoria 'arte', ad esempio, alcuni

oggetti possono avere sottocategorie più specifiche come 'arte figurativa', 'arte astratta' o 'arte sacra'); esiste un piccolo numero di siti ai quali non è assegnata nessuna sottocategoria.

- Catalogo: tale valore indica la tipologia del sito tra quelle prima elencate; è possibile che alcuni oggetti rientrino in più categorie e in tal caso sono associati più valori di catalogo; ad ogni oggetto ne è comunque assegnato almeno uno.

Oltre ai metadati generali, per ogni oggetto è possibile ottenere un insieme di informazioni più dettagliate. Tali dati costituiscono una vera e propria scheda di catalogazione, e nello specifico sono rappresentati da:

- Descrizione generale, in formato puramente testuale, del museo o dell'oggetto specifico di riferimento;
- Storia dell'edificio, anche questa presentata tramite una descrizione testuale;
- Titolarità dell'edificio, intesa come titolarità pubblica o privata in base a chi è assegnata la gestione o la proprietà del sito;
- Data di apertura/costruzione dell'edificio, in genere espressa tramite l'anno;
- Tipologie degli eventuali oggetti contenuti: possono essere diversi e fanno riferimento in particolar modo ai musei, è tuttavia possibile che anche in altri tipi di siti siano contenuti degli oggetti di interesse;
- Attività interne: indicano le attività svolte all'interno di quell'edificio (ad esempio ricerche scientifiche o prestito di materiali per mostre o studio);
- Eventuali orari di apertura e tariffe per la visita, entrambi espressi tramite contenuto testuale non strutturato.

Tutti i dati appena descritti, sia quelli generali che quelli più specifici per ogni oggetto, sono stati presi in considerazione per l'attività di modellazione e conversione di OpenERCH. L'ontologia adottata si è dimostrata abbastanza ampia e ricca per permettere la rappresentazione della conoscenza ottenuta tramite ognuna di queste informazioni. Per la descrizione generale dell'ontologia si rimanda alla sezione 4.1.1, mentre la descrizione più dettagliata di come ognuna di queste informazioni è stata modellata in OpenERCH è presentata in sezione 4.3.1.

3.3 Attività di modellazione e conversione semantica dei dati

Come già descritto, il principale obiettivo del progetto OpenERCH consiste nell'ottenere, a partire dai dati grezzi relativi a diversi luoghi di interesse artistico e culturale, un dataset in cui tali informazioni sono espresse in modo formale in base ad uno specifico modello semantico. Il modello di riferimento adottato è rappresentato dall'ontologia CIDOC-CRM che, come già esposto (sezione 1.3.2), rappresenta il principale modello ontologico standard per la rappresentazione formale della conoscenza in ambito culturale. Tale ontologia permette di esprimere delle asserzioni riguardanti diverse tipologie di risorse, come oggetti fisici o concetti astratti, risorse temporali, di localizzazione o che fanno riferimento ad attori e persone fisiche. CIDOC-CRM mette inoltre a disposizione un'altrettanto vasta tipologia di relazioni tra queste risorse, che permettono di modellare in modo formale anche concetti semanticamente complessi quali, ad esempio, le diverse tipologie di attività proposte da un museo o eventi specifici relativi alla creazione di un determinato oggetto o alla nascita di un determinato attore.
Tutte queste informazioni sono rappresentate da vere e proprie risorse, che a loro volta sono descritte tramite relazioni con altre risorse sempre più specifiche, secondo il modello a grafo proposto da RDF. Questo mette in evidenza come tale modello di rappresentazione si discosti da quelli utilizzati dalle

tradizionali schede di catalogazione, in quanto ogni informazione di interesse rappresenta una vera e propria risorsa informativa a parte, facendo perdere in qualche modo quel ruolo di centralità assegnato agli oggetti che, secondo quanto proposto dalla letteratura (e come descritto in sezione 1.2), rappresenta uno dei principali limiti degli approcci tradizionali per la rappresentazione della conoscenza in ambito culturale.

3.3.1 *Tipologie di risorse in OpenERCH*

Il dataset ottenuto dalla conversione oltre a fare riferimento all'ontologia CIDOC-CRM è aderente ai principi del Linked Data, descritti nella sezione 2.4.2. Ogni risorsa in OpenERCH è innanzitutto individuata univocamente da specifici URI. Tali indirizzi di riferimento sono stati scelti seguendo le linee guida proposte in [SCA08]. Come prima descritto, inoltre, l'ontologia utilizzata permette di esprimere concetti anche molto elaborati tramite l'uso di una vasta gamma di entità e proprietà; per tale ragione esistono diverse tipologie di risorse alle quali sono stati assegnati dei namespace di riferimento specifici. Più precisamente, tutti gli URI che identificano delle risorse sono creati tramite uno specifico schema ottenuto tramite l'utilizzo di un prefisso del tipo ' *http://dominioIBC /resource/*[5]', al quale segue l'indicazione specifica della tipologia di risorsa e il nome, ottenendo quindi un suffisso del tipo '*tipologia:nome_risorsa*'.
Le principali tipologie di risorse presenti nel dataset di OpenERCH sono le seguenti:

- Risorse relative a siti e contenitori: rappresentano le risorse di base alle quali fanno riferimento i dati dell'intero catalogo. Tali risorse rappresentano i musei e le altre tipologie di siti nell'ottica di edifici

[5]Dato che al momento in cui si scrive il dataset non è ancora stato ufficialmente pubblicato in un dominio web raggiungibile pubblicamente, gli URI utilizzano una struttura generica come quella indicata. Nel momento in cui i dati saranno raggiungibili nel Web e nel Linked Data, e quindi pubblicati su uno specifico dominio Web, la parte indicata come '*dominioIBC*' sarà sostituita con l'indirizzo di dominio specifico assegnato.

o vere e proprie costruzioni che possono essere identificate, descritte e localizzate. Le risorse di questo tipo fanno quindi riferimento, ad esempio, ai musei intesi come spazi di raccolta e di interesse artistico, e non come luoghi geografici locati e identificati secondo specifici indirizzi. Queste risorse rappresentano quindi gli oggetti di base del catalogo e gli URI utilizzati per la loro identificazione sono del tipo: '*http://dominioIBC/resource/Museo_Civico_Il_Correggio*'.

- Risorse di appellazione: rappresentano le denominazioni relative ad altre generiche risorse e sono utilizzate per la loro identificazione, secondo quanto indicato da CIDOC-CRM. Tali entità sono pertanto collegate tramite apposite relazioni con le risorse che identificano e permettono quindi prevalentemente di conoscerne il nome. Per identificare la risorsa che, ad esempio, identifica e fornisce il nome del museo civivo 'Il Correggio' è utilizzato il seguente URI: '*http://dominioIBC/resource/appellation:Museo_Civivo_Il_Correggio*'.

- Risorse di classificazione: indicano tipologie e catalogazioni che possono essere assunte dagli oggetti presenti nel dataset. Fanno parte di questa categoria, ad esempio, le risorse che rappresentano i vari tipi di categorie e sottocategorie assegnate ai siti piuttosto che i diversi tipi di attività svolte al loro interno o le tipologie di oggetti contenuti in un museo. La categoria arte, per fare un esempio, è identificata dal seguente URI: '*http://dominioIBC/resource/classification:Arte*'.

- Risorse relative a luoghi: sono di questo tipo tutte quelle entità che fanno riferimento a luoghi intesi, questa volta, come vere e proprie locazioni geografiche che possono essere identificate dal nome del luogo (ad esempio il nome di una città o provincia) piuttosto che da coordinate più precise (indirizzi geografici o coordinate di geolocalizzazione spaziale). Oltre alle città, in questa categoria rientrano anche le risorse che rappresentano i siti nell'ottica di luoghi veri e propri. Esempi di URI che fanno riferimento a risorse di questo tipo sono

'*http://dominioIBC/resource/place:City_of_Correggio*' o '*http://dominioIBC/resource/place:Museo_Civico_Il_Correggio*'.

- Risorse temporali: fanno riferimento a eventi temporali di vario genere. Alcune di esse sono utilizzate, ad esempio, per la rappresentazione di specifici istanti temporali mentre altre indicano periodi di tempo più generici, come periodi storici. Sono utilizzate prevalentemente per identificare temporalmente specifici eventi e hanno URI del tipo '*http://dominioIBC/resource/time:Year_1981*' o '*http://dominioIBC/resource/time:Period_Medioevo*'.

- Risorse relative ad attività: rappresentano vere e proprie attività quali, ad esempio, la visita di un museo o particolari attività svolte all'interno di esso. L'URI che identifica la risorsa che rappresenta la visita del Museo Civico 'Il Correggio' avrà, ad esempio, il seguente URI: *http://dominioIBC/resource/activity:Visit_of_Museo_Civico_Il_Correggio*'.

- Risorse relative a persone, per lo più artisti di vario tipo le cui opere sono contenute all'interno di alcuni dei siti rappresentati. Gli URI che fanno riferimento a queste risorse contengono, oltre al nome dell'attore, anche un codice identificativo utilizzato all'interno del catalogo di Samira per la corretta identificazione (dato che, in tal caso, nome e cognome non assicurano l'unicità); un esempio di questo tipo è: '*http://dominioIBC/IBCBologna/resource/people:30680253_Normanno_Gobbi*'.

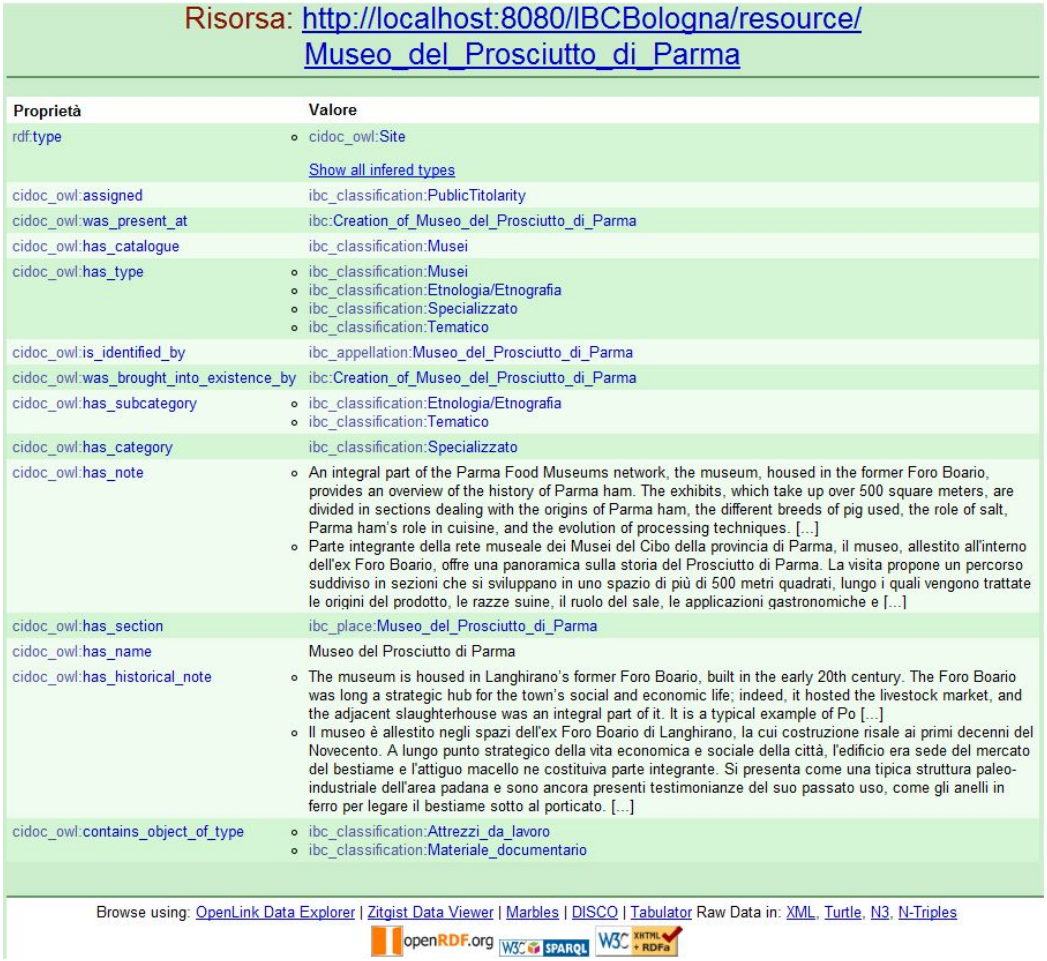
- Risorse relative a definizioni ontologiche: rientrano in questa categoria tutte le risorse che fanno riferimento a classi e proprietà descritte in CIDOC-CRM, ad esempio '*http://dominioIBC/IBCBologna/owl/Place*'.

3.3.2 *Disponibilità e accesso ai dati*

In conformità alle linee guida proposte in [SCA08] e ai principi del Linked Data, in OpenERCH esiste una netta distinzione tra le risorse vere e proprie e i documenti che le descrivono. Ad ogni risorsa sono pertanto associati diversi URI in aggiunta a quelli a cui si è fatto riferimento prima e che indicano la risorsa vera e propria; ognuno di questi nuovi indirizzi indica infatti specifiche rappresentazioni documentali. In altre parole, deve essere possibile ottenere la descrizione di ogni risorsa presente nel dataset digitandone l'URI relativo all'interno di un web browser.
L'attività appena descritta rappresenta una delle metodologie di base per l'accesso e la navigazione dei dati presenti all'interno di in una sorgente pubblicata nel Linked Data. Affinché i dati di OpenERCH siano raggiungibili e consultabili via web, essi sono mantenuti all'interno di una sorta di raccoglitore semantico dei dati, chiamato *RDF Store*, che si occupa di memorizzare e mantenere le informazioni, analogamente a quello che fa un DBMS. Affinché tali dati siano consultabili fornendo la propria rappresentazione documentale, è stata necessaria la creazione di un'applicazione web che si occupi di recuperare e fornire tali documenti. Ciò mette in evidenza come l'attività svolta da OpenERCH, e in generale quanto richiesto per rendere delle informazioni disponibili nel Web of Data, non si limiti alla sola attività di conversione ma anche alla realizzazione di quanto necessario per ottemperare ai principi del Linked Data.
Gli URI delle specifiche rappresentazioni delle risorse possono essere ottenuti a partire dall'indirizzo di identificazione della risorsa, e aggiungendo la tipologia di documento desiderata. Per ottenere, ad esempio, un documento HTML che descrive la risorsa relativa al Museo Civico 'Il Correggio' è possibile utilizzare l'URI '*http://dominioIBC/resource/Museo_Civico_Il_Correggio/html*', ottenendo, tramite l'inserimento di tale indirizzo all'interno del browser, un risultato come quello proposto in figura 3.2. Come si può notare, tale rappresentazione mette in evidenza tutte le relazioni che la risorsa indicata possiede con altre risorse presenti nel sistema, fornendo quindi una completa descri-

zione delle caratteristiche del museo presenti nel catalogo. Cliccando su altre eventuali risorse oggetto delle relazioni indicate, si ottiene la loro descrizione. Tramite l'URI '*http://dominioIBC/resource/Museo_Civico_Il_Correggio/rdf*' è possibile invece ottenere un documento RDF contenente tutti gli statement in cui compare la risorsa indicata; oltre alla serializzazione in XML (considerata come il modello di rappresentazione RDF standard), tali informazioni possono essere ottenute anche in altri formati di serializzazione (*Turtle*, *N3*, ecc.). Tramite queste attività è possibile, in definitiva, ottenere la descrizione dei dati in diverse tipologie a seconda delle esigenze e della natura del richiedente (un essere umano sarà sicuramente più interessato alla rappresentazione HTML, mentre un agente automatico sarà in grado di comprendere più facilmente quella in RDF). Dalla figura 3.2 si può notare come la descrizione degli oggetti in OpenERCH differisca in modo sostanziale da quella ottenibile, ad esempio, nel sito di Samira. Entrambe sono pensate per l'essere umano, tuttavia esse sono state create per esigenze diverse: mentre la seconda propone una descrizione testuale delle varie caratteristiche degli oggetti, l'altra permette di ottenere, in un formato più comprensibile e leggibile rispetto all'RDF, la descrizione formale di tali oggetti aderente all'ontologia CIDOC-CRM e nella forma a grafo tipica del Semantic Web. Quest'ultima può essere utile, ad esempio, per chi intende navigare tra le risorse presenti nel dataset, eventualmente per utilizzarle successivamente per lo sviluppo di qualche applicazione o per correlarle con quelle presenti in una propria sorgente dati.

I principi del Linked Data richiedono che le informazioni, oltre ad essere descritte nei modi appena delineati, siano accessibili tramite le tecnologie standard proposte dal Semantic Web, facendo riferimento in particolare ai linguaggi RDF e SPARQL. Il primo rappresenta il linguaggio di base con cui i dati sono descritti e memorizzati in OpenERCH; per ottemperare a tali principi il sistema permette inoltre di accedere ai dati anche tramite delle interrogazioni SPARQL. A tale scopo il sistema mette a disposizione uno *SPARQL Endpoint*, ossia una sorta di servizio, conforme al protocol-



Risorsa: http://localhost:8080/IBCBologna/resource/Museo_del_Prosciutto_di_Parma

Proprietà	Valore
rdf:type	∘ cidoc_owl:Site Show all infered types
cidoc_owl:assigned	ibc_classification:PublicTitolarity
cidoc_owl:was_present_at	ibc:Creation_of_Museo_del_Prosciutto_di_Parma
cidoc_owl:has_catalogue	ibc_classification:Musei
cidoc_owl:has_type	∘ ibc_classification:Musei ∘ ibc_classification:Etnologia/Etnografia ∘ ibc_classification:Specializzato ∘ ibc_classification:Tematico
cidoc_owl:is_identified_by	ibc_appellation:Museo_del_Prosciutto_di_Parma
cidoc_owl:was_brought_into_existence_by	ibc:Creation_of_Museo_del_Prosciutto_di_Parma
cidoc_owl:has_subcategory	∘ ibc_classification:Etnologia/Etnografia ∘ ibc_classification:Tematico
cidoc_owl:has_category	ibc_classification:Specializzato
cidoc_owl:has_note	∘ An integral part of the Parma Food Museums network, the museum, housed in the former Foro Boario, provides an overview of the history of Parma ham. The exhibits, which take up over 500 square meters, are divided in sections dealing with the origins of Parma ham, the different breeds of pig used, the role of salt, Parma ham's role in cuisine, and the evolution of processing techniques. [...] ∘ Parte integrante della rete museale dei Musei del Cibo della provincia di Parma, il museo, allestito all'interno dell'ex Foro Boario, offre una panoramica sulla storia del Prosciutto di Parma. La visita propone un percorso suddiviso in sezioni che si sviluppano in uno spazio di più di 500 metri quadrati, lungo i quali vengono trattate le origini del prodotto, le razze suine, il ruolo del sale, le applicazioni gastronomiche e [...]
cidoc_owl:has_section	ibc_place:Museo_del_Prosciutto_di_Parma
cidoc_owl:has_name	Museo del Prosciutto di Parma
cidoc_owl:has_historical_note	∘ The museum is housed in Langhirano's former Foro Boario, built in the early 20th century. The Foro Boario was long a strategic hub for the town's social and economic life; indeed, it hosted the livestock market, and the adjacent slaughterhouse was an integral part of it. It is a typical example of Po [...] ∘ Il museo è allestito negli spazi dell'ex Foro Boario di Langhirano, la cui costruzione risale ai primi decenni del Novecento. A lungo punto strategico della vita economica e sociale della città, l'edificio era sede del mercato del bestiame e l'attiguo macello ne costituiva parte integrante. Si presenta come una tipica struttura paleo-industriale dell'area padana e sono ancora presenti testimonianze del suo passato uso, come gli anelli in ferro per legare il bestiame sotto al porticato. [...]
cidoc_owl:contains_object_of_type	∘ ibc_classification:Attrezzi_da_lavoro ∘ ibc_classification:Materiale_documentario

Browse using: OpenLink Data Explorer | Zitgist Data Viewer | Marbles | DISCO | Tabulator Raw Data in: XML, Turtle, N3, N-Triples

Figura 3.2: *Esempio di rappresentazione in HTML delle informazioni presenti nel sistema relative alla risorsa che rappresenta il museo del prosciutto di Parma*

lo proposto dal Semantic Web, che è in grado di accettare delle richieste di interrogazione in linguaggio SPARQL restituendone gli eventuali risultati. OpenERCH accetta tutte le tipologie di query possibili e permette di ottenere i risultati in diversi formati. Le richieste SPARQL possono essere effettuate anche attraverso una semplice interfaccia web, creata appositamente, tramite la quale, come mostrato in figura 3.3, è possibile inserire la propria interrogazione, scegliendo lo specifico linguaggio (oltre a SPARQL

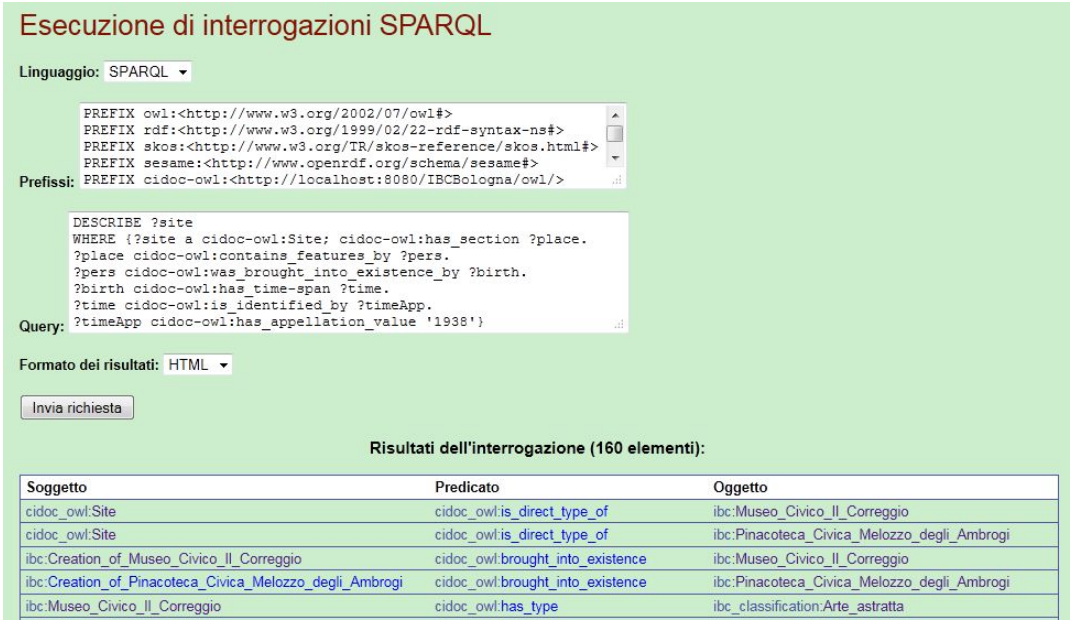


Figura 3.3: *Interfaccia web proposta da OpenERCH per effettuare interrogazioni in SPARQL.*

è possibile utilizzare linguaggi proprietari simili) e il formato di ritorno dei risultati. In particolare è possibile scegliere l'HTML se si intende ottenere una visualizzazione dei risultati direttamente nella pagina in forma tabellare (come mostrato in figura 3.3), oppure specifiche serializzazioni RDF per ottenere il relativo documento con i risultati. I formati di ritorno cambiano automaticamente a seconda della tipologia di query inserita nell'apposita area. Ovviamente le interrogazioni possono essere effettuate anche senza fare riferimento a tale interfaccia grafica, interrogando l'endpoint SPARQL tramite richieste HTTP o tramite linguaggi di programmazione, utilizzando le funzioni messe a disposizione da tutte le librerie per il Semantic Web.
L'accesso ai dati tramite l'utilizzo di interrogazioni SPARQL rappresenta un metodo alternativo per visualizzare le informazioni di interesse rispetto a quello effettuato tramite la deferenziazione degli URI prima descritto. Quest'ultimo può essere utilizzato, ad esempio, per ottenere la descrizione com-

pleta di una risorsa e per comprendere quindi come questa è rappresentata, ottenendo informazioni utili su come strutturare eventuali interrogazioni. La creazione di query richiede infatti una conoscenza precisa del modo in cui i dati sono descritti, e quindi dell'ontologia utilizzata, ma risulta un meccanismo molto potente per filtrare e individuare precisamente le informazioni di interesse. Tramite specifiche interrogazioni è possibile ottenere, ad esempio, tutti i musei che contengono oggetti di artisti nati in un certo anno piuttosto che le città in cui sono presenti edifici storici che contengono oggetti d'arte astratta e sono gestiti da un privato. L'individuazione di informazione così precisa risulta, in questo modo, molto più semplice ed efficace rispetto a quella ottenibile basandosi sui meccnismi standard del web costruiti sulla corrispondenza di termini e parole chiave.

3.3.3 *Collegamenti con dataset esterni*

Alcune delle risorse presenti nel dataset di OpenERCH hanno, sempre in conformità a quanto richiesto dai principi del Linked Data, dei collegamenti con risorse di dataset esterni. La tipologia principale di collegamenti di questo tipo consiste in associazioni di similitudine: si tratta cioè di link che relazionano le risorse interne con quelle di altre sorgenti dati che però hanno lo stesso significato. La presenza di questi collegamenti permette, ad esempio, di navigare da un dataset all'altro, ottenendo la descrizione di una stessa risorsa sotto diversi punti di vista, in quanto le diverse sorgenti di dati possono contenere informazioni completamente diverse e nuove relativamente ad una specifico oggetto. Per fare un esempio, le risorse che in OpenERCH fanno riferimento a delle città presentano dei link in uscita verso le risorse che rappresentano la stessa città all'interno di altri dataset (come quello di DBpedia o GeoNames). Ciò permette, seguendo questo collegamento, di ottenere maggiori informazioni sulla città: in DBpedia, ad esempio, si potranno reperire informazioni generali come la descrizione piuttosto che la popolazione o ancora l'URI di altre risorse che rappresentano individui nati in quella città, mentre da GeoNames è possibile ottenere dati geografici quali

le coordinate spaziali piuttosto che altre città e luoghi geografici vicini.
Le risorse che presentano collegamenti con l'esterno sono essenzialmente le seguenti:

- Risorse che rappresentano categorie e sottocategorie di siti; esse presentano collegamenti con risorse di altri dataset, in particolare quelli di DBpedia e *Freebase*. Nello specifico queste risorse sono connesse, tramite due diverse tipologie di relazioni, sia con risorse che in questi dataset rappresentano lo stesso concetto inteso come categoria (e sono quindi utilizzati anche lì come forma di classificazione), che con quelle che lo rappresentano come concetto generico, e forniscono quindi maggiori informazioni e descrizioni di vario tipo su esso. La categoria Arte, per fare un esempio, è connessa con una relazione di similitudine alla risorsa di DBpedia che indica, anche in tale dataset, l'Arte intesa come categoria vera e propria, mentre è associata alle risorse di DBpedia e Freebase che indicano l'arte in generale tramite una relazione che specifica queste ultime come fonte di informazione addizionale.

- Risorse relative a città; esse sono collegate con risorse che rappresentano la stessa città in altri dataset. Come spiegato prima, nei diversi dataset è possibile reperire informazioni di diverso genere relative allo stesso luogo, da quelle più generiche e descrittive a quelle più specifiche e di ordine geografico.

- Risorse che rappresentano persone. Come spiegato nella sezione 3.3.1, esse fanno riferimento prevalentemente ad artisti (pittori, scultori, ecc.) le cui opere sono contenute nei musei e presentano dei collegamenti con altre risorse (se esistenti) che si riferiscono allo stesso individuo dalle quali è possibile eventualmente recuperare informazioni più dettagliate.

- Risorse che rappresentano periodi storici. Analogamente ai casi precedenti sono presenti, dove possibile, link di similitudine con risorse esterne che rappresentano quello stesso periodo storico.

La presenza di collegamenti esterni permette in qualche modo di arricchire il dataset di OpenERCH tramite dati che originariamente non sono disponibili, lasciando agli utilizzatori di tali informazioni la possibilità di creare applicazioni o forme di utilizzo che individuino e sfruttino il valore aggiunto ottenibile tramite l'intreccio dei dati.

3.4 Applicazione mash-up di esempio

L'idea alla base del Linked Data, e di conseguenza del lavoro effettuato per la realizzazione del dataset di OpenERCH, non è quella di creare una sorta di base di dati pronta per essere utilizzata secondo degli scopi specifici. Tale attività, in altre parole, non è stata svolta tenendo in mente una specifica applicazione di riferimento che avrebbe potuto utilizzare i dati, ma piuttosto con lo scopo di ottenerne una rappresentazione quanto più possibile formale e arricchita, lasciando agli utilizzatori il compito di decidere come i dati potranno essere usati. La particolare struttura adottata e la conoscenza dell'ontologia di riferimento consentono infatti di ottenere un modello di rappresentazione della conoscenza formale facilmente accessibile anche in maniera automatica. La sorgente dati di OpenERCH si presta quindi ad essere utilizzata da chiunque abbia necessità di includere in una propria applicazione i dati messi a disposizione, magari intrecciandoli con quelli presenti invece negli altri dataset esterni con i quali sono correlati.
Allo scopo di testare i meccanismi di accesso per usufruire dei dati creati, e per dare un esempio di come questi possono essere facilmente arricchiti, è stata creata l'applicazione ERCH-Mashup, che fa utilizzo delle informazioni di OpenERCH intrecciandole, dove possibile, con quelle recuperate all'interno del Web of Data. Tale applicazione presenta un'interfaccia web che consente di effettuare delle ricerche tra i metadati dei musei, permettendo di individuare, ad esempio, tutti i musei d'arte che presentano oggetti medievali presenti nella provincia di Bologna.
L'interfaccia proposta da ERCH-Mashup permette, come si può vedere in

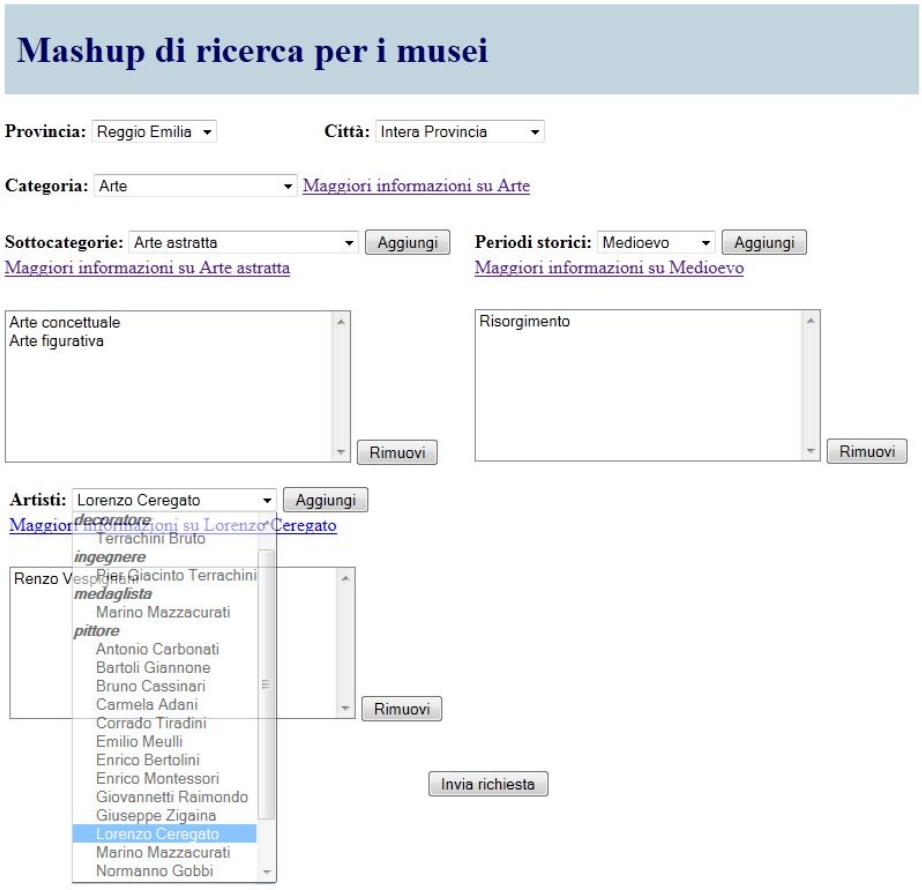


Figura 3.4: *Interfaccia proposta dall'applicazione di mash-up per la ricerca dei musei.*

figura 3.4, di scegliere città e provincia in cui si intende effettuare le ricerche, permettendo eventualmente di fare riferimento a tutte le città appartenenti ad una provincia, senza ulteriori specificazioni. In aggiunta ai dati di localizzazione geografica è possibile specificare anche la categoria all'interno della quale i musei devono essere ricercati, indicando anche eventuali sottocategorie più specifiche. A seconda della categoria selezionata, il sistema permette di aggiungere nei criteri di ricerca anche possibili periodi storici ai quali i musei devono fare riferimento o artisti ai quali devono essere correlati. Il sistema si occupa di utilizzare i parametri di ricerca indicati per la creazione

di opportune query SPARQL, fornendo come risultato la lista dei musei che aderiscono a quanto richiesto.
Lo scopo principale dell'applicazione, tuttavia, non è quello di proporre un'interfaccia di ricerca, quanto quello di arricchire le informazioni con le quali questa è effettuata. Di seguito si presentano le principali tipologie di informazioni aggiuntive che essa è in grado di reperire.

3.4.1 *Reperimento di informazioni aggiuntive su categorie e sottocategorie*

Un primo tentativo di reperimento di nuova informazione è effettuato da ERCH-Mashup in riferimento ai vari valori di classificazione assegnati ai diversi musei. Durante la fase di selezione dei parametri di ricerca prima descritta l'applicazione permette infatti di ottenere maggiori informazioni riguardo agli specifici valori selezionati. Ogni volta che dagli appositi menù a tendina viene selezionato un particolare valore, accanto ad esso viene visualizzato un link, al click del quale viene avviata l'attività di ricerca di notizie aggiuntive ad esso relative.
Il risultato di tale attività di ricerca varia a seconda della tipologia di valore al quale si fa riferimento. Nel caso in cui si cerchino informazioni aggiuntive riguardo una determinata categoria o sottocategoria, si ottiene un risultato simile a quello presentato in figura 3.5, che mostra le informazioni reperite dall'applicazione relativamente alla categoria 'Arte'. La pagina ottenuta presenta, come si può vedere, un'immagine correlata al concetto di arte e una descrizione testuale generica. Questi dati sono reperiti dal dataset di DBpedia, e rappresentano infatti l'immagine e l'introduzione presenti nell'articolo di Wikipedia che si riferisce all'arte. Appositi link a fondo pagina permettono infatti di consultare tale articolo per intero o di accedere alla pagina di Freebase dedicata alla stessa risorsa. Oltre a queste informazioni generiche sono presentati degli elenchi relativi a persone correlate all'argomento e a musei e collezioni internazionali e italiani che si occupano di arte. Le prime due tipologie di informazioni sono estratte ancora una volta dal dataset di

Figura 3.5: *Esempio di informazioni aggiuntive reperite in riferimento alla categoria Arte.*

DBpedia, mentre quelle relative ai musei italiani sono recuperate da un altro dataset dedicato proprio alla descrizione di tutti i musei presenti nel paese. Tali dataset esterni sono descritti nel capitolo seguente alla sezione 4.1.2, mentre per una descrizione dettagliata di come queste informazioni vengono reperite si rimanda alla sezione 4.5.
Cliccando sui nomi delle persone o dei musei presenti negli elenchi, viene avviata in automatico la ricerca di informazioni aggiuntive più specifiche su ognuno di essi. Nel caso dei musei internazionali, reperiti da DBpedia, viene ricercata e presentata, in modo analogo a quanto esposto precedentemente, una descrizione generica del museo e l'eventuale immagine, oltre al riferimen-

Figura 3.6: *Esempio di informazioni aggiuntive reperite in riferimento a musei internazionali (in alto) o nazionali (in basso).*

to all'articolo di Wikipedia e alla homepage del museo. Nel caso di musei italiani invece l'applicazione tenta comunque una ricerca analoga (anche se, trattandosi di musei molto specifici, difficilmente loro informazioni sono presenti in DBpedia), aggiungendo alle eventuali informazioni reperite anche quelle presenti nel dataset da cui tali musei sono stati ottenuti, tra cui l'indirizzo, il sito internet e altre informazioni generali inerenti la struttura. Un esempio delle due tipologie di risultati ottenuti è presentato in figura 3.6 .
Oltre che per le categorie e sottocategorie, è possibile ottenere informazioni aggiuntive in modo del tutto analogo a quello descritto anche sui periodi storici.

3.4.2 *Ricerca di informazioni aggiuntive relative a persone*

Figura 3.7: *Esempio di informazioni aggiuntive reperite in riferimento a persone esterne (in alto) o interne (in basso) al dataset di OpenERCH.*

Esistono essenzialmente due tipologie di persone le cui informazioni possono essere recuperate da ERCH-Mashup. La prima è rappresentata da tutti gli attori recuperati tramite la ricerca di informazioni aggiuntive relative a specifiche categorie prima descritta: si tratta quindi di persone i cui dati

possono non esistere all'interno del dataset di OpenERCH. Informazioni aggiuntive relative a tali individui sono, quindi, ricercate sempre all'interno del dataset di DBpedia. Oltre ad eventuali foto e alla descrizione generale, in questo caso viene ottenuta anche una lista di tutte le aree di interesse a cui tale individuo è collegato. Queste possono, a loro volta, essere oggetto di nuova ricerca informativa, effettuata allo stesso modo di quanto descritto precedentemente per le categorie. Ciò significa che potranno essere identificati ulteriori individui attinenti a quell'area, che a loro volta possono essere correlati ad ulteriori aree, dando vita ad una sorta di navigazione aperta potenzialmente infinita.
La seconda tipologia di persone per le quali è possibile ricercare notizie è rappresentata dagli artisti del catalogo, ossia quelli le cui opere sono presenti all'interno di alcuni dei siti in esso descritti: gli stessi utilizzati per raffinare i parametri di ricerca e che sono quindi presenti nel dataset di OpenERCH. Così come per le categorie e sottocategorie, è possibile avviare la ricerca di informazioni aggiuntive relative a tali individui cliccando sull'apposito link presentato ogni volta che uno di essi è selezionato nel menù di scelta iniziale. In questo caso la ricerca è simile a quella prima descritta per gli attori esterni, tuttavia in aggiunta sono presentate anche le informazioni interne quali l'attività svolta e l'anno e la città di nascita e morte. Cliccando sulle città si possono ottenere le informazioni ad essa relative e descritte di seguito. La figura 3.7 mostra un esempio di risultati ottenuti relativamente alle due tipologie di persone descritte.

3.4.3 *Ricerca di informazioni aggiuntive su città e luoghi di interesse*

In ERCH-Mashup i risultati finali ottenuti da una ricerca sui musei consistono nell'elenco, proposto in forma tabellare, di tutti i musei che corrispondono ai parametri indicati. Per ognuno di essi, oltre che ad un riepilogo dei principali metadati di identificazione, è possibile ottenere maggiori informazioni relativamente sia alla città in cui essi sono ubicati che ad eventuali altri

luoghi di interesse nelle vicinanze.

Figura 3.8: *Esempio di informazioni aggiuntive reperite in riferimento alla città di Bologna.*

La ricerca di informazioni su una città permette di ottenere, come negli altri casi, l'immagine e la descrizione generica ad essa relativa e il collegamento all'articolo di Wikipedia o alla risorsa di Freebase. In tal caso, tuttavia, maggiori dati sono reperiti facendo riferimento ai dataset che contengono informazioni geografiche, come GeoNames. Grazie ad esso è infatti possibile ottenere una lista di altri luoghi e città vicine. Cliccando su ognuna di esse, o richiedendo la visualizzazione in GeoNames della città in base alla quale la ricerca è stata effettuata, è possibile accedere al servizio da esso messo a disposizione che permette di visualizzare tale luogo su una mappa geografica interattiva. Per ogni città sono, infine, indicati eventuali artisti del catalogo

che in essa sono nati o morti. Un esempio di informazioni ottenute relativamente alla città di Bologna è presente in figura 3.8.
Per ogni museo presentato tra i risultati di una ricerca è possibile conoscere anche la lista di eventuali altri luoghi di interesse ad esso vicini. Essi sono individuati facendo riferimento all'indirizzo del museo e ricercando tutti i luoghi presenti nel raggio di un chilometro. Sono considerati luoghi di interesse eventuali bar, ristoranti, o altri tipi di negozi, chiese o luoghi di culto oppure teatri e monumenti. Questa ricerca è effettuata grazie ai dati messi a disposizione da un altro dataset di natura geografica presente nel Web of Data chiamato *LinkedGeoData*. I luoghi individuati sono presentati all'interno di una mappa e, cliccando su uno di essi, è possibile ottenere le indicazioni stradali da seguire per raggiungerlo a partire dal museo di riferimento. La figura 3.9 fornisce un esempio di questo tipo (le mappe presentate nelle pagine sono ottenute grazie ai servizi di *GoogleMaps*).

Figura 3.9: *Esempio di informazioni aggiuntive che indicano le diverse tipologie di luoghi di interesse vicini al museo 'Casa Carducci' (in alto) e le istruzioni per raggiungere uno di questi (in basso).*

Capitolo 4

Il progetto OpenERCH: aspetti tecnici e implementativi

4.1 Strumenti utilizzati

4.1.1 *L'ontologia CIDOC-CRM*

Come descritto nel capitolo precedente, OpenERCH utilizza l'ontologia CIDOC - Conceptual Reference Model (CRM) come modello ontologico di riferimento per la rappresentazione dei dati. Come indicato in [CDG10], CIDOC-CRM è un'ontologia creata per l'integrazione, la mediazione e lo scambio di informazioni eterogenee riguardanti i beni culturali ed è stata ottenuta da un lavoro di oltre 10 anni da parte del *CIDOC Documentation Standards Working Group*. Dal Dicembre 2006 tale ontologia è diventata uno standard ISO, e rappresenta quindi il modello di riferimento per la rappresentazione di informazione relativa al patrimonio culturale in ambito semantico. L'obiettivo principale è quello di permettere lo scambio e l'integrazione dell'informazione tra sorgenti eterogenee di dati relativi ai beni culturali, fornendo le definizioni semantiche necessarie per trasformare sorgenti informative diverse e localizzate in una sorta di risorsa globale coerente. CIDOC-CRM utilizza pertanto una prospettiva sovra-istituzionale e astratta da ogni specifi-

co contesto locale, su tale prospettiva si basano e sono determinati i costrutti e il livello di dettaglio utilizzati. Nello specifico, questo modello ontologico definisce principalmente la semantica implicita presente negli schemi e nelle strutture dei database e dei documenti utilizzati in ambito culturale, con particolare riferimento alla documentazione museale, definendole in termini di un'ontologia formale. Non fa quindi utilizzo della terminologia che tipicamente costituisce i dati di queste strutture, ma piuttosto prevede la definizione delle relazioni necessarie per il loro utilizzo. Lo scopo non è quindi quello di proporre cosa le istituzioni dovrebbero documentare, ma piuttosto quello di rappresentare la logica che sta dietro a ciò che è attualmente documentato, permettendone l'interoperabilità semantica.

L'ontologia è composta da 86 classi, che indicano le varie tipologie di entità che possono essere rappresentate, organizzate secondo una specifica gerarchia, e 137 proprietà che tra loro possono essere definite. Ognuno di questi elementi è identificato tramite un nome in lingua inglese (l'ontologia prevede che i nomi possano essere tradotti in diverse lingue, mantenendo la compatibilità con il CRM) preceduto da un codice univoco assegnato ad ognuno di essi. Le classi messe a disposizione permettono di rappresentare concetti astratti quali entità temporali, periodi o eventi e oggetti veri e propri come luoghi, persone o elementi fisici e tangibili. Le informazioni sono pertanto descritte tramite statement che fanno riferimento ad oggetti indicati come specifiche istanze di queste classi e proprietà che li mettono in relazione. L'ontologia è inoltre basata sul principio di *monotonia*, nel senso che tutti i costrutti e le deduzioni effettuate da essa rimangono sempre validi anche a fronte dell'aggiunta di nuovi costrutti o eventuali estensioni al modello.

Una delle più importanti caratteristiche di questa ontologia consiste nel fatto che, essendo molto generica e pensata per racchiudere un vasta tipologia di descrizioni, mette comunque a disposizione dei meccanismi per integrare eventuali tesauri o strutturazioni specifiche. Le descrizioni strutturate che si riferiscono a musei, ad esempio, includono in genere informazioni riguardo la loro tipologia, in genere descritta tramite l'indicazione di classificazioni, cate-

gorie o classi di appartenenza. Il principale meccanismo messo a disposizione da CIDOC-CRM per esprimere informazioni di questo genere è rappresentato dall'uso della classe *E55 Type*, le cui istanze rappresentano concetti che possono essere utilizzati per indicare le classificazioni assegnate ai vari oggetti. Tutte le classi sono infatti definite come gerarchicamente derivate dalla generica classe *E1 CRM Entity* (ogni oggetto sarà quindi una *CRM Entity*) che a sua volta presenta la proprietà *P2 has type* con la classe E55 Type. Tale proprietà sarà quindi ereditata da tutte la classi, fornendo un meccanismo di specializzazione valido per qualsiasi oggetto descritto, collegandolo ad eventuali vocabolari, tesauri od ontologie esterne. Questi possono infatti essere rappresentati tramite la creazione di entità indicate come sottoclassi di E55 Type, collegate tra loro sia con eventuali proprietà esterne addizionali che con quelle messe a disposizione da CIDOC-CRM come *P127 has broader term*, che permette di creare gerarchie tra i termini. Oltre a specificare le tipologie degli oggetti nei modi appena descritti, CIDOC-CRM permette di descrivere e identificare il processo stesso che porta alla definizione di tali categorizzazioni, tramite l'utilizzo della classe *E83 Type Creation*.
Alternativamente alle modalità appena descritte, è possibile assegnare delle categorie agli oggetti estendendo le altre classi dell'ontologia. Per caratterizzare una persona come 'artista', ad esempio, invece che utilizzare la proprietà P2 has type può essere estesa la classe *E21 Person* tramite una sottoclasse *E21.1 Artist*. Seppur il risultato finale sia pressoché lo stesso, l'indicazione del tipo tramite la creazione di una nuova sottoclasse dovrebbe avvenire solo nei casi in cui il concetto risulti sufficientemente stabile e associato a delle proprietà addizionali specifiche ed esplicitamente modellate. La creazione di un'istanza di E55 Type risulta invece più semplice e permette di ottenere maggiore flessibilità.
La figura 4.1 mostra come alcune classi e proprietà di CIDOC-CRM possono essere utilizzate per descrivere la conoscenza in riferimento ad informazioni spaziali. Le relazioni di ereditarietà sono mostrate tramite delle frecce a linea doppia, mentre le proprietà tra le varie entità sono indicate tramite freccie

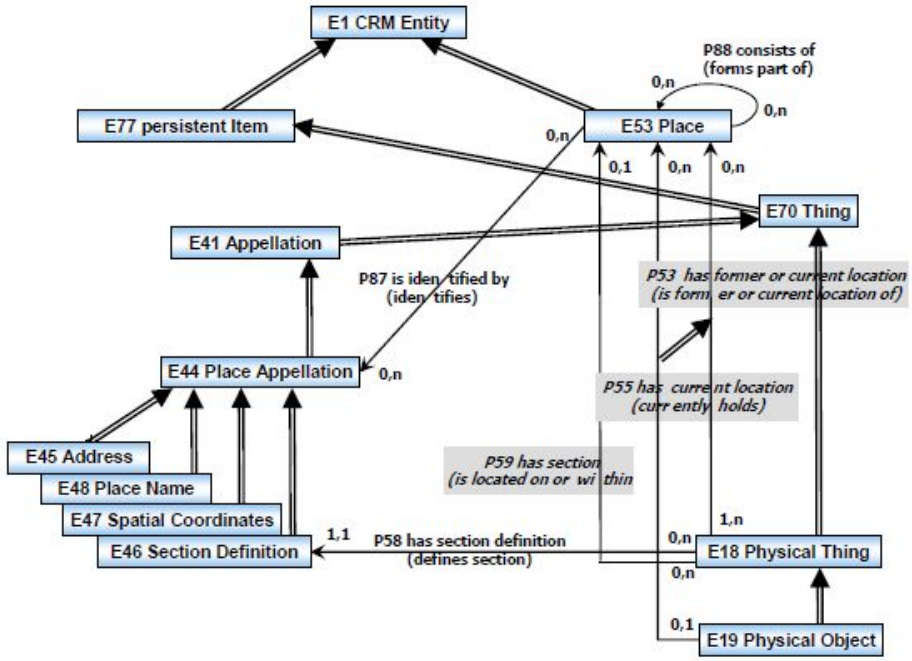


Figura 4.1: *Rappresentazione in CIDOC-CRM di descrizioni relative a informazioni spaziali.*

semplici. Come si può notare la classe *E53 Place*, che indica un luogo, è identificata tramite istanze della classe *E44 Place Appellation*, che a sua volta è estesa da altre classi più specifiche che indicano vari tipi di identificativi riferiti a un luogo (indirizzo, nome, coordinate spaziali o specifiche parti del luogo stesso). La proprietà *P58 has section definition* permette di associare agli oggetti un'istanza della classe *E46 Section Definition*, fornendo quindi un metodo indiretto per indicare la locazione di tali entità. Esistono anche una serie di proprietà dirette tra le classi che si riferiscono agli oggetti e quelle che indicano il luogo: queste proprietà costituiscono delle 'scorciatoie' per indicare, secondo significati diversi, l'appartenenza di un oggetto ad uno specifico luogo. Tramite la proprietà *P58 has section definition* è possibile definire un'istanza di E53 Place come una specifica sezione di un'istanza di un altro oggetto, permettendo di definire situazioni in cui si conosce la spe-

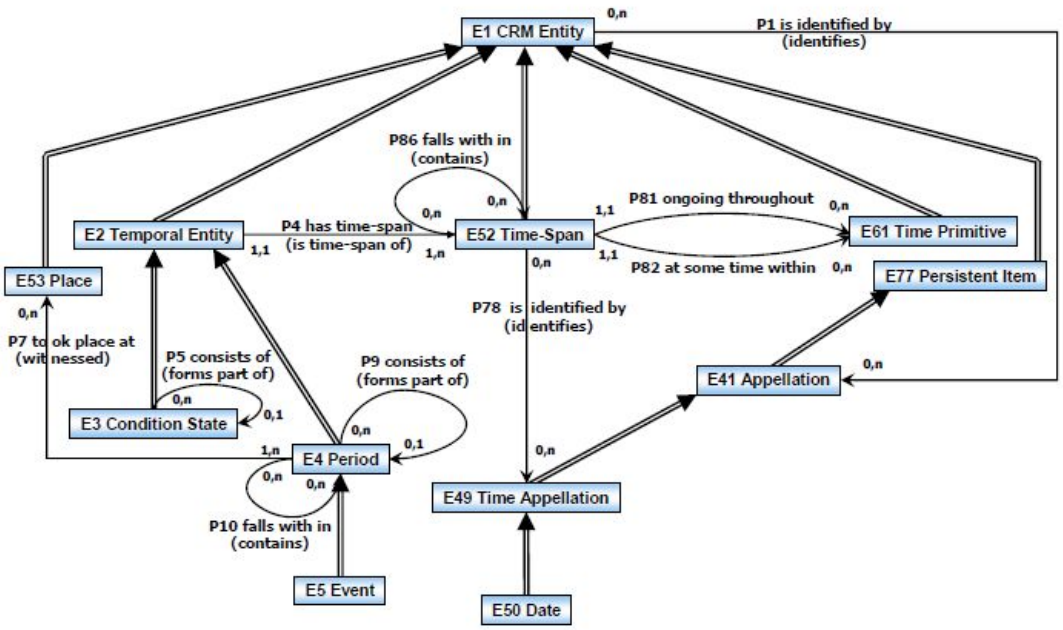


Figura 4.2: *Rappresentazione in CIDOC-CRM di descrizioni relative a informazioni temporali.*

cifica locazione di un oggetto all'interno di un altro, pur non conoscendo la posizione di quest'ultimo.

La figura 4.2 indica, invece, come informazioni di tipo temporale possono essere modellate tramite CIDOC-CRM. La classe *E2 Temporal Entity* racchiude una serie di sottoclassi che contengono un contenuto temporale, come quelle che indicano un evento o un periodo, quest'ultimo a sua volta può essere collegato al luogo in cui è avvenuto. Un'istanza della classe *E52 Time-Span* indica, invece, un vero e proprio intervallo temporale che non fa riferimento ad uno specifico contenuto geografico. Una serie di proprietà permettono di associare a quest'ultimo sia un'identificativo, in genere rappresentato da una data, che, tramite primitive, i limiti temporali entro i quali si estende.

CIDOC-CRM permette, in definitiva, di esprimere una vasta gamma di informazioni molto variegate e dettagliate in modo altamente formale, permettendo di individuare facilmente i dati di interesse e di rispondere a specifiche

domande di ricerca. I modelli di rappresentazione appena descritti sono stati utilizzati all'interno della modellazione di OpenERCH per la descrizione di tutti gli oggetti e i siti presenti nel dataset.

4.1.2 *Principali dataset di collegamento*

Una delle caratteristiche più importanti della modellazione dei dati attuata tramite OpenERCH, e in generale tipica dei dataset esposti nel Linked Data, consiste nel fatto che questi presentano dei collegamenti con risorse descritte in sorgenti di dati esterne. Come descritto nella sezione 3.3.3, OpenERCH presenta dei link di similitudine con risorse presenti in alcuni dataset pubblicati nel Web of Data, in particolare quelli forniti dal Linking Open Data Project. Come descritto in sezione 2.4.3, tale rete di dati è molto vasta e variegata; di seguito si presenta una breve descrizione dei dataset con cui sono stati effettuati i collegamenti.

DBpedia

Il dataset di DBpedia, nato nel 2007 da un progetto della *Free University of Berlin* e della *University of Leipzig*, rappresenta uno dei punti di interconnessione più importanti del Web of Data e consiste in una gigantesca sorgente di dati ottenuti estraendo informazioni strutturate da Wikipedia e rendendole accessibili nel Web. Il knowledgebase messo a disposizione da DBpedia copre, pertanto, diversi domini, raccoglie informazioni in diverse lingue e viene costantemente aggiornato in modo che si evolva automaticamente in base ai cambiamenti di Wikipedia. Attualmente tale dataset raccoglie dati relativi a più di 3,5 milioni di oggetti di vario tipo (persone, luoghi, film, album musicali, malattie, ecc.) descritte secondo una specifica ontologia tramite circa 672 milioni di triple RDF estratte dalle varie edizioni, in diverse lingue, di Wikipedia. Data la vasta copertura di domini e l'elevato grado di sovrapposizione delle informazioni con altri dataset già presenti nel web, un numero sempre maggiore di sorgenti di dati hanno iniziato a creare dei collegamenti, tramite link RDF, con il dataset di DBpedia, rendendolo uno

dei punti centrali di intercollegamento all'interno del Web of Data.
Gli articoli di Wikipedia consistono principalmente di testo libero, ma contengono anche varie tipologie di informazione strutturata, in genere espressa tramite un preciso linguaggio di markup tipico delle piattaforme wiki. Tali informazioni includono tabelle informative di riepilogo, informazioni di categorizzazione, immagini, coordinate geografiche, link a pagine web esterne, pagine di disambiguazione e link tra le varie edizioni in lingue diverse di Wikipedia. Le varie risorse di DBpedia vengono identificate tramite un'URI ottenuto estraendo il nome della risorsa a partire dall'URL dell'articolo di Wikipedia corrispondente. Ogni entità è descritta sia tramite una serie di proprietà generali che tramite proprietà specificamente estratte dalle tabelle informative presenti negli articoli (dette *infobox*), se presenti. Tali due tipologie di proprietà vengono ricavate ed espresse tramite due modalità differenti. Un primo approccio, più generico, consiste nel creare, a partire dai dati presenti nell'infobox, delle triple i cui predicati e oggetti sono stabiliti facendo riferimento ai nomi e valori degli attributi della tabella informativa così come sono. In questo modo ci si assicura di raccogliere tutte le possibili tipologie di dati presenti, evitando perdite di informazioni. Il secondo approccio mappa invece i template di Wikipedia in base all'ontologia di DBpedia, per cercare di risolvere il problema di sinonimie nei nomi degli attributi o l'utilizzo di diversi template per le stesse cose. Questa ontologia è stata costruita tenendo conto dei template standard e maggiormente utilizzati nell'edizione inglese di Wikipedia; una serie di regole di mappaggio permettono di indicare i più disparati attributi presenti nelle infobox tramite specifiche classi e proprietà descritte nell'ontologia. I dati raccolti tramite questa seconda tipologia di estrazione sono meno vasti rispetto al primo approccio, ma hanno una strutturazione più chiara e precisa. Le informazioni estratte tramite le due diverse tipologie di estrazioni sono, tra l'altro, indicate tramite predicati diversi che fanno utilizzo di due namespace differenti. [BLK09]
I dati di DBpedia possono essere interrogati tramite query SPARQL facendo riferimento ad uno specifico *endpoint* che permette l'interrogazione sia trami-

te librerie di programmazione che tramite specifiche applicazioni web messe a disposizione per gli utenti. Il dataset di DBpedia è utilizzato in OpenERCH per creare collegamenti tra risorse che fanno riferimento a concetti per i quali si intende recuperare informazione aggiuntiva generica (categorie, periodi storici, artisti, ecc.): tramite una serie di interrogazioni vengono infatti individuate le risorse che si riferiscono allo stesso oggetto. I dettagli di questo meccanismo sono descritti in sezione 4.3.2.

FreeBase

Freebase è "un grafo di entità che rappresentano persone, luoghi e cose, costruito da una comunità di persone che amano l'Open Data", o per lo meno così viene definito dai suoi creatori (`http://www.freebase.com/`). In effetti Freebase non è altro che un enorme database libero, accessibile sia per il recupero di dati relativi ai più svariati domini, che per l'immissione di nuove informazioni. Nato nel 2007 da un progetto dell'azienda *Metaweb* (recentemente acquisita da Google) con lo scopo di fare da ponte tra la visione di fondo basata sulla condivisione tipica del Web 2.0 e il mondo più preciso e strutturato proposto dal Semantic Web, a partire dall'anno successivo Freebase consente di accedere a tutte le risorse in esso contenute tramite i meccanismi standard del Linked Data fornendone una completa descrizione in RDF.
Freebase raccoglie informazioni importate da moltissime sorgenti di dati, oltre che quelle inserite liberamente dagli utenti della comunità manualmente o tramite i tool di importazione messi a disposizione. Sotto il punto di vista dell'eterogeneità dei contenuti, Freebase è molto simile a DBpedia, tuttavia esistono molte differenze riguardanti lo schema utilizzato per la descrizione dei dati. Freebase, infatti, pur avendo una struttura di organizzazione delle risorse a grafo, adotta uno schema di memorizzazione e identificazione molto particolare, che ne permette l'accesso e il recupero anche tramite modalità diverse oltre a quelle tipiche del Linked Data.
Freebase raccoglie informazioni riguardo circa 20 milioni di risorse, che ven-

gono chiamate *Topic*. Un topic rappresenta un singolo concetto o una vera e propria entità del mondo reale, alla quale viene riservata una specifica pagina web in Freebase. Ad ognuna di queste entità è assegnato un particolare *ID*, che deve risultare unico e ha lo scopo di permettere di distinguere diverse entità che possono avere nomi simili. In genere i topic vengono identificati con degli ID rappresentati da URI, che ne consentono un'identificazione univoca e stabile, tuttavia le entità possono essere identificate anche tramite altre tipologie di identificatori. La creazione di diversi identificatori che individuano la stessa risorsa è ottenuta tramite l'utilizzo di *chiavi* e *namespace*. Il namespace è una sorta di *directory* gerarchica che raccoglie una serie di oggetti che, all'interno di quello specifico percorso, hanno un nome univoco; in particolare all'interno di ogni namespace gli oggetti sono identificati tramite chiavi e la gerarchia formata dal namespace e le relative chiavi associate formano un percorso che permette di identificare univocamente un singolo nodo del grafo di dati. Tale approccio deriva dal fatto che le informazioni in Freebase sono ottenute tramite un'attività di fusione dei diversi dati raccolti dai vari data source e inseriti anche successivamente dagli utenti: i diversi indirizzi ottenuti tramite l'accoppiamento di namespace e chiavi permette, in un certo senso, di vedere la risorsa sotto diversi aspetti. Ai topic che contengono informazioni importata da Wikipedia, ad esempio, sono associati, tra gli altri, anche gli ID basati sul numero e sul titolo dell'articolo di Wikipedia relativo.

Ad ogni risorsa di Freebase sono assegnati uno o più *tipi* e una serie di *proprietà* addizionali, che nel complesso formano lo *schema* dei dati. Le varie tipologie di ID assegnati ai vari topic, ad esempio, non sono altro che delle proprietà. I vari tipi sono organizzati in forma gerarchica facendo utilizzo di namespace e una collezione di tipi che condividono lo stesso namespace viene chiamata *dominio*. L'insieme dei vari tipi e delle proprietà che vengono inserite nello schema di ogni topic costituisce una sorta di ontologia; la particolarità offerta da Freebase si basa sul fatto che gli utenti, oltre ad inserire i dati, possono costruire un proprio schema, modificando essenzial-

mente l'ontologia dei dati stessi. Ovviamente tali attività sono sottoposte ad una serie di linee guida e limitazioni, per evitare una strutturazione troppo confusionaria dei dati.
I dati in Freebase possono essere interrogati tramite uno specifico linguaggio chiamato *MQL* (*Metaweb Query Language*). Come prima specificato, tuttavia, dal 2008 Freebase fornisce anche una rappresentazione delle risorse in RDF e mette a disposizione un endpoint SPARQL per il recupero e l'interrogazione tramite query. In quest'ambito le risorse sono identificate tramite degli specifici URI, pensati appositamente per la rappresentazione RDF delle risorse, e le varie proprietà assegnate ad un topic sono rappresentate tramite predicati, che si rifanno allo specifico schema di Freebase. In OpenERCH i collegamenti con questo dataset sono effettuati per scopi e in modi del tutto analoghi a quelli utilizzati per i link con DBpedia.

GeoNames e LinkedGeoData

Nel Linked Data i principali strumenti che mettono a disposizione dati per la geo-localizzazione delle risorse e dei luoghi sono rappresentati dai dataset di GeoNames e LinkedGeoData.
GeoNames è un database geografico che contiene oltre 10 milioni di nomi geografici relativi a luoghi. Lo scopo è proprio quello di individuare univocamente tali luoghi, assegnando loro un nome nelle varie lingue, una sorta di categorizzazione (la tipologia dei luoghi è individuata tramite delle categorie chiamate *feature class*, che indicano se si tratta, ad esempio, di una nazione o di un lago) e identificando la loro posizione tramite delle coordinate spaziali di latitudine e longitudine. Una delle principali caratteristiche dei dati raccolti da GeoNames consiste nel fatto che esistono molti collegamenti tra le diverse risorse, che rappresentano le varie posizioni geografiche e permettono di individuare eventuali altri luoghi vicini, confinanti o compresi all'interno di un altro luogo più grande. L'idea alla base dell'ontologia consiste nel considerare i vari luoghi come delle risorse, alle quali sono associati dei documenti detti *feature*, che ne indicano le caratteristiche e i legami con eventuali altri

luoghi. In particolare, per ogni risorsa è possibile ottenere diverse tipologie di documenti di feature, come quelli che racchiudono i luoghi geografici contenuti o confinanti con la risorsa considerata piuttosto che quelli geograficamente vicini. L'accesso alle informazioni di GeoNames è possibile principalmente per mezzo di una serie di *Web Service* messi a disposizione tramite delle API o accedendo direttamente ai documenti di feature; non è presente invece un vero e proprio endpoint SPARQL.
Una delle principali limitazioni del dataset di GeoNames è rappresentata dal fatto che contiene informazioni riguardo luoghi geografici tipici (quali continenti, nazioni, montagne, laghi, ecc.), mentre è difficile individuare nodi che fanno riferimento, ad esempio, a luoghi commerciali o di interesse più specifici (quali ristoranti, pub o bar). In quest'ambito risulta, invece, molto più ricco il dataset di LinkedGeoData. Tale sorgente dati è infatti ottenuta convertendo in formato semantico i dati raccolti dall'*OpenStreetMap Project*, un progetto collaborativo che ha lo scopo di creare una mappa libera ed editabile del mondo, in cui gli utenti hanno la possibilità di inserire nuovi punti di interesse tramite una serie di tool messi loro a disposizione. Tale progetto ha permesso la raccolta di una mole molto ampia di dati geografici, che fanno riferimento anche a specifici punti di interesse settati dagli utenti oltre che a quelli classici che si riferiscono a nazioni o luoghi geografici veri e propri. La trasformazione di tali dati in RDF è stata effettuata secondo una specifica ontologia le cui classi e proprietà sono state create in base ai valori degli specifici attributi assegnati ai vari oggetti dell'OpenStreetMap. Tali classi permettono, ad esempio, di sapere se una determinata risorsa rappresenta una montagna, piuttosto che un lago o un ristorante. Le informazioni sono raggiungibili tramite interrogazioni SPARQL, in particolare è possibile effettuare le query sia su un endpoint statico, che fa riferimento all'intera mole di dati estratta dall'OpenStreetMap entro una certa data, sia su un endpoint dinamico che effettua eventuali sincronizzazioni con i dati dell'OpenStreetMap per raccogliere nuove informazioni.
I dataset geografici di GeoNames e LinkedGeoData sono utilizzati in OpenERCH

per effettuare collegamenti con risorse che rappresentano posti o luoghi geografici, per ottenere maggiori informazioni, ad esempio, su altri luoghi di interesse vicini.

4.2 Architettura generale

L'obiettivo del progetto OpenERCH, come discusso nel capitolo precedente, consiste nella creazione di un dataset che racchiude la descrizione delle informazioni relative a beni artistici e culturali (musei, teatri e siti in generale) in modo formale secondo l'ontologia CIDOC-CRM. Lo scopo finale è quello di pubblicare tali dati all'interno della rete Linked Data, rendendoli disponbili all'utilizzo da parte anche di applicazioni software in relazione a tutte le altre informazioni presenti nel Web of Data. La creazione di tale dataset è avvenuta a partire da un insieme di dati non strutturati e tenendo conto dei principi proposti dal Linked Data. Al fine di proporre un esempio di utilizzo combinato delle informazioni ottenute, il progetto comprende anche un'applicazione mash-up di esempio.
La figura 4.3 presenta l'architettura generale di OpenERCH, illustrandone le principali componenti e le relazioni tra esse presenti. I dati iniziali, oggetto di conversione, sono presenti all'interno di un database in cui sono conservati prevalentemente sottoforma di schede di catalogazione: ad ogni oggetto è assegnata una lista di proprietà che ne descrivono le varie caratteristiche. La descrizione degli oggetti è pertanto effettuata, all'interno di tale base di dati, prevalentemente tramite un elenco di valori testuali associati a ciascuno di essi. Come descritto nella sezione 3.1, tutte le informazioni presenti in questo database sono presentate all'interno del portale Samira grazie al quale possono essere consultate e ricercate tramite un apposito sito web. L'attività di conversione è effettuata su due tipologie di dati estratte dal database: un insieme di metadati generali di identificazione e localizzazione di ogni oggetto e un'altra collezione di dati più specifici di descrizione che permettono di ottenere le informazioni più dettagliate relative ad ognuno di essi (si rimanda alla

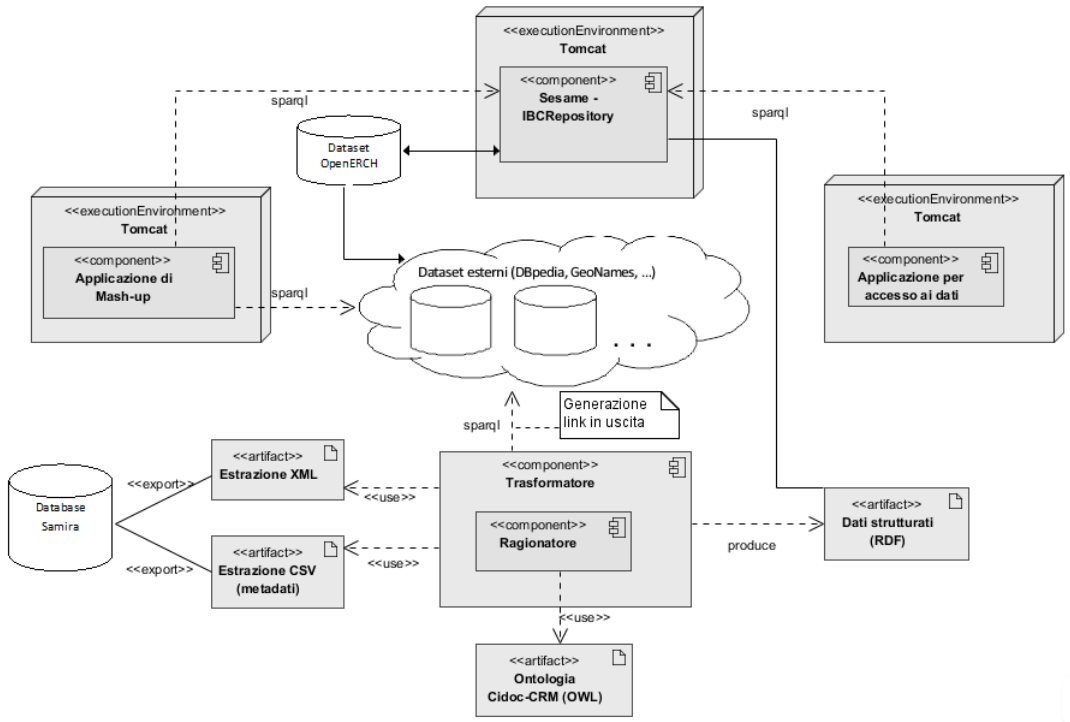


Figura 4.3: *Diagramma di Deploy che mostra l'architettura generale di OpenERCH indicando le componenti principali e le loro connessioni.*

sezione 3.2 per una descrizione dettagliata di entrambe le categorie). I primi sono ottenibili direttamente dal portale Samira in formato CSV[1], mentre gli altri sono estratti dal database in formato XML secondo uno specifico schema aderente agli standard di catalogazione previsti e descritti in [ACG98]. L'attività di conversione vera e propria è effettuata dalla componente software chiamata *Trasformatore* che si occupa, a partire dai file di dati appena indicati estratti dal database, di leggere le singole informazioni e creare a

[1]Il CSV (*comma-separated values*) è un formato basato su file di testo utilizzato per l'importazione ed esportazione (ad esempio da fogli elettronici o database) di una tabella di dati. In questo formato le informazioni relative ai diversi oggetti sono presentate su righe diverse (ad esempio una riga per ogni record di un database), che a loro volta sono divise in campi (ad esempio le singole colonne di una record di un database) separati da un apposito carattere separatore, in genere rappresentato dalla virgola.

partire da esse degli statement RDF che le descrivano in base al modello ontologico proposto da CIDOC-CRM. La trasformazione comprende anche un'attività di inferenza effettuata da un *ragionatore*. Tale componente si occupa di ottenere, a partire dagli statement di base prodotti inizialmente dal trasformatore, ulteriori triple eventualmente non presenti ma che possono essere dedotte in base a quanto indicato nell'ontologia, descritta a sua volta tramite il linguaggio OWL. Il ragionatore si occupa di inferire nuova informazione anche in base ad una serie di regole di inferenza fornitegli in input secondo uno specifico linguaggio. L'utilizzo del ragionatore permette di semplificare e rendere più facilmente comprensibile e modificabile l'attività di conversione, permettendo al trasformatore di occuparsi solo della conversione di base prevista per ogni singola tipologia di informazione, lasciando ad eventuali regole esterne o al modello ontologico il compito di fornire tutte le triple aggiuntive che possono essere dedotte da questa informazione di base. Man mano che gli statement vengono prodotti e che le risorse vengono create, il trasformatore ha anche il compito di individuare, tramite una serie di interrogazioni in linguaggio SPARQL rivolte ai dataset esterni presenti nel Linked Data (descritti nella sezione precedente), eventuali altre risorse con le quali effettuare dei collegamenti.

L'attività di conversione e inferenza effettuata dal trasformatore permette di ottenere le informazioni strutturate nella forma di triple RDF. Affinché queste possano essere correttamente raccolte ed accedute, formando un vero e proprio dataset, è necessario che vengano raccolte all'interno di un *RDF Store*, ossia un componente che si occupa di mantenere i dati creando un *repository* che permette di ottenerli tramite diverse modalità: una sorta di DBMS semantico. Lo store di dati utilizzato nell'ambito di OpenERCH è rappresentato da *Sesame*, un framework gratuito che mette a disposizione diversi strumenti per la gestione di dati in RDF, tra cui soluzioni per la memorizzazione di dati e API e librerie per il loro reperimento. Nello specifico lo store utilizzato è rappresentato da un'applicazione web che funziona facendo utilizzo dell'*application server Tomcat*; tale applicazione permette, trami-

te interfacce grafiche di configurazione, di creare dei repository che possono essere popolati caricando, da file o tramite modalità alternative, i dati in formato RDF. Questi possono poi essere acceduti sia tramite query SPARQL che tramite metodi di accesso alternativi messi a disposizione da spacifiche librerie per *Java*. Come mostrato nell'architettura di figura 4.3 le triple RDF ottenute dalla trasformazione vengono caricate in Sesame ottenendo un vero e proprio dataset che a sua volta presenta dei collegamenti con altri dataset esterni, proprio come previsto dalle specifiche Linked Data.
Come spiegato nella sezione 3.3.2, il dataset di OpenERCH può essere acceduto sia richiedendo specifiche rappresentazioni delle varie risorse in esso contenute che tramite l'utilizzo di un endpoint SPARQL. Tale risultato è garantito da un'applicazione web, anch'essa basata sull'ambiente di esecuzione Tomcat, che permette di accedere ai dati nelle modalità appena indicate. Questa applicazione si occupa prevalentemente di garantire la deferenziazione degli URI associati alle varie risorse del dataset: a seconda della risorsa a cui l'URI si riferisce e della rappresentazione richiesta essa recupera, tramite apposite interrogazioni, i dati ad essa associati e li converte per ottenere un documento che li esponga nella modalità indicata. È, inoltre, tale applicazione a rappresentare l'endpoint per le interrogazioni SPARQL: essa infatti riceve le query, si occupa di eseguirle sullo store e restuire i risultati così ottenuti nei formati richiesti. Essa contiene inoltre l'interfaccia grafica che permette di effettuare le interrogazioni direttamente via web. Maggiori dettagli sulle attività appena indicate sono riportati in sezione 4.4.
L'ultima componente nel diagramma di figura 4.3 è rappresentata dall'applicazione ERCH-Mashup (descritta in sezione 3.4). Anch'essa non è altro che un'applicazione web che gira su Tomcat e che racchiude una serie di servizi e pagine web create dinamicamente in base alle ricerche effettuate e alle visualizzazioni di dati aggiuntivi richieste. Le informazioni da essa presentate sono pertanto reperite tramite query SPARQL sia sul repository interno che sugli endpoint messi a disposizione dai dataset esterni del Linked Data. È importante sottolineare ancora una volta che tale applicazione rappresenta

solo un esempio di utilizzo dei dati creati e delle potenzialità del Web of Data, pertanto non è parte essenziale dell'architettura di OpenERCH, almeno per quanto concerne la messa a disposizione dei dati.

4.3 Attività di conversione dei dati

La conversione dei dati in base al modello ontologico di riferimento rappresenta il cuore dell'attività svolta in OpenERCH, grazie alla quale è ottenuto il dataset su cui si basano tutte le altre attività e le varie componenti prima presentate. Di seguito si descrivono nel dettaglio alcuni aspetti tecnici e implementativi relativi a tale operazione, indicando in che modo i dati sono stati modellati e rappresentati tramite CIDOC-CRM e fornendo alcuni dettagli sull'attività svolta dal trasformatore relativa alla conversione dei dati e al loro collegamento con i dataset esterni.

4.3.1 *Modellazione ontologica e rappresentazione dei dati*

Nella sezione 4.1.1 è stata presentata l'ontologia CIDOC-CRM, i principi su cui si basa e alcuni esempi di come può essere utilizzata per la modellazione di particolari tipi di informazione. Durante l'attività di modellazione dei dati si è cercato di attenersi il più possibile al modello ontologico da essa proposto, evitando l'utilizzo di costrutti o tipologie di modellazione non previste. Sono, tuttavia, state apportate delle piccole modifiche e aggiunte all'ontologia di base proposta da CIDOC-CRM, le cui classi e proprietà sono state descritte tramite l'utilizzo del linguaggio OWL. Come indicato negli esempi in sezione 4.1.1, tutte le entità e proprietà nel modello CRM includono, nel nome, un identificativo composto da una lettera ('E' per le classi e 'P' per le proprietà) e un numero, seguito dal nome vero e proprio. In tal modo tutti gli oggetti dell'ontologia sono identificati univocamente, tuttavia ciò rende più complicata la lettura delle informazioni e, in particolare, tale scelta

risulta molto fastidiosa soprattutto durante la creazione di query SPARQL, specie se costruite manualmente, in quanto è difficile ricordare tutti i codici associati alle varie entità. Per tale ragione nell'ontologia OWL utilizzata in OpenERCH sono stati rimossi dai nomi gli identificativi, pur aggiungendo, ad ogni elemento ontologico, una proprietà *ID* grazie alla quale è possibile ottenere l'identificativo ad esso associato. Dato che si è reso necessario l'utilizzo solo di un sottoinsieme di tutte le entità proposte da CIDOC-CRM, non si sono presentati casi in cui lo stesso ID potesse essere assegnato a più entità. Un'altra aggiunta all'ontologia di base è rappresentata dall'inserimento di alcune proprietà aggiuntive che fungono da 'scorciatoia'. In alcuni casi, come si è visto negli esempi di figura 4.1 e 4.2, la modellazione di determinati concetti richiede infatti la creazione di numerose entità intermedie collegate con varie proprietà: in questi casi il modello proposto dall'ontologia è stato comunque rispettato, tuttavia sono state aggiunte delle proprietà più dirette che permettono di navigare più velocemente tra le risorse, facilitando specialmente la creazione di query.
Come spiegato in 3.3.1, la modellazione di ogni sito (museo, teatro storico, ecc.) presente nel catalogo è effettuata tramite la creazione di due tipologie di risorse: una che lo rappresenta come vero e proprio edificio, ossia come spazio fisico che raccoglie al suo interno diversi oggetti, e una che lo rappresenta come luogo geografico. Per ogni oggetto vengono pertanto create due risorse, corrispondenti alle due tipologie appena descritte: la prima viene rappresentata tramite un'istanza della classe ontologica *Site*, mentre la seconda fa riferimento alla classe *Place*[2]; tali risorse sono collegate tra loro tramite la proprietà *has_section*, che ha il compito di indicare, appunto, il luogo geografico in cui si trova l'edificio che rappresenta il contenitore considerato. Il diagramma in figura 4.4 mostra le principali proprietà attribuite alla risorsa di tipo Site, utilizzate per descrivere determinate caratteristiche

[2]In base al modello RDF, la rappresentazione del fatto che una risorsa è un'istanza di una determinata classe ontologica è ottenuta utilizzando il predicato *rdf:type* tra la risorsa vera e propria e la risorsa che rappresenta la classe nell'ontologia.

dell'oggetto. I nodi rappresentati in forma ovale si riferiscono a risorse, all'interno è indicato l'URI utilizzato per la loro identificazione abbreviato tramite l'utilizzo di namespace (che variano a seconda della tipologia di risorsa, come spiegato in sezione 3.3.1). Le proprietà tra le diverse risorse sono indicate tramite delle frecce, quelle tratteggiate indicano che la proprietà a cui fanno riferimento è una scorciatoia; i nodi di forma rettangolare indicano invece valori letterali. Di seguito si presenta una descrizione di come le varie caratteristiche dell'oggetto sono modellate tramite le proprietà mostrate in figura 4.4.

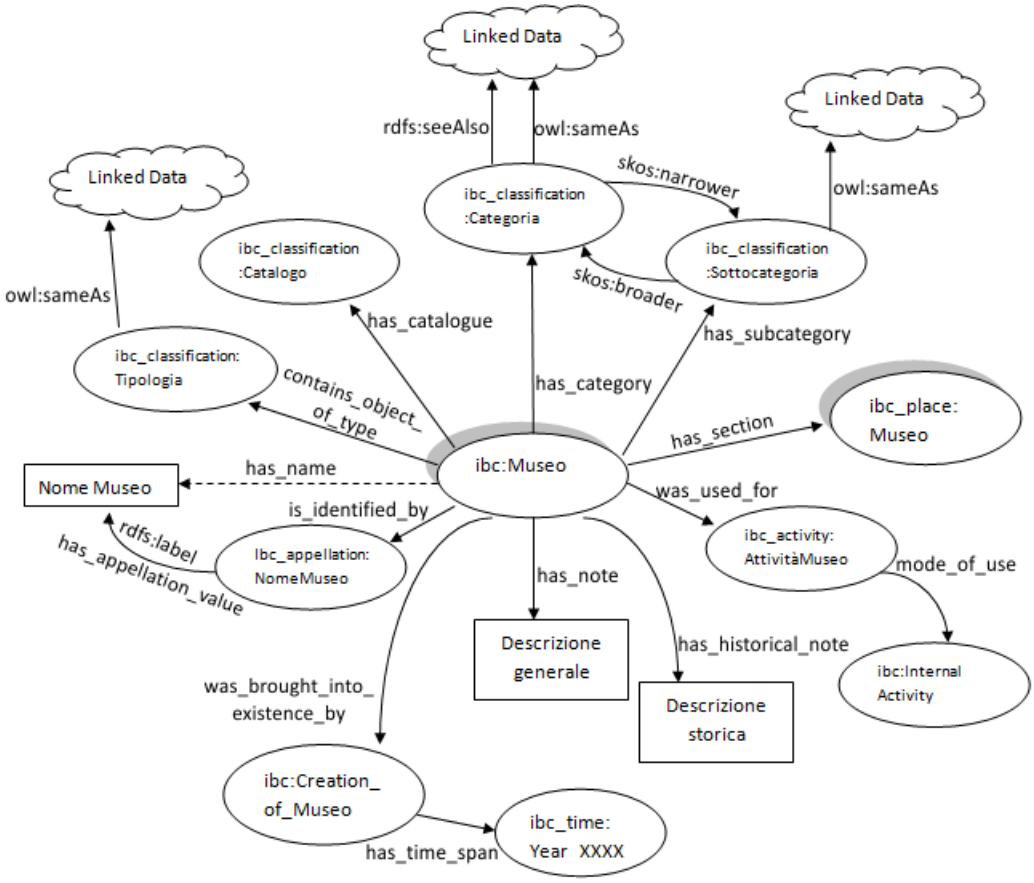


Figura 4.4: *Diagramma di modellazione che indica tutte le proprietà e le entità collegate alla risorsa di tipo 'Site' che indica un oggetto del catalogo inteso come edificio.*

- Classificazioni relative al sito: sono ottenute tramite la creazione di tre risorse che fanno riferimento, rispettivamente, al valore di catalogo, categoria e sottocategoria. Tali risorse sono istanze di specifiche classi create appositamente (*Catalogue*, *Category* e *Subcategory*) che a loro volta sono state aggiunte nell'ontologia come sottoclassi della classe *Type*. Tra le risorse che fanno riferimento alla categoria e sottocategoria sono, inoltre, utilizzate le proprietà dell'ontologia *SKOS* che consentono di indicare relazioni di contenimento, permettendo sostanzialmente di ottenere una sorta di tassonomia legata alle categorie. Queste risorse sono associate a quella che rappresenta l'oggetto tramite la proprietà *has_type*, conformemente a quanto previsto dal modello di CIDOC-CRM per la creazione di tipologie e tassonomie personalizzate (come descritto in sezione 4.1.1). Al fine di avere un'indicazione più precisa del tipo di classificazione espressa, sono state create anche delle proprietà più specifiche, definite nell'ontologia come sottoproprietà di has_type (mantenendo, in questo modo, anche il collegamento generico tramite il predicato has_type come richiesto da CIDOC-CRM).
  Le risorse che fanno riferimento ad una categoria presentano anche relazioni con risorse esterne di altri dataset del Linked Data. Nello specifico si tratta di relazioni di somiglianza, ossia che le collegano con altre risorse che, in questi nuovi dataset, rappresentano lo stesso concetto. La proprietà utilizzata in questi casi per ottenere tale relazione è *owl:sameAs*[3]; nel caso delle categoria viene utilizzato anche il predicato *rdfs:seeAlso*. Come spiegato in sezione 3.3.1, il dataset di DBpedia presenta sia risorse che fanno riferimento ad una determinata categoria come concetto generico che altre che la rappresentano come vera e

[3]Dato che in RDF i predicati sono espressi tramite URI, l'URI completo relativo al predicato indicato è il seguente: "http://www.w3.org/2002/07/owl#sameAs". Per ottenere una leggibilità maggiore tutti gli URI ai quali sono assegnati namespace noti (come quelli che fanno riferimento a risorse di *rdf* o *rdfs*) saranno indicati con i rispettivi namespace piuttosto che nella forma completa. Anche le risorse interne di OpenERCH sono indicate utilizzando specifici namespace piuttosto che gli URI completi.

propria forma di categorizzazione (per il concetto di arte, ad esempio, esiste sia una risorsa che la considera come categoria vera e propria che un'altra che la descrive come concetto generale). Nel caso delle categorie, pertanto, il collegamento con la prima tipologia di risorse è effettuato tramite il predicato rdfs:seeAlso, mentre il collegamento con le specifiche categorie di DBpedia è effettuato tramite la proprietà owl:sameAs.

- L'identificazione di ogni oggetto è ottenuta tramite la creazione di una risorsa di tipo *Appellation*, collegata con la risorsa Site, che a sua volta presenta la proprietà *has_appellation_value* con un letterale che contiene il nome effettivo dell'oggetto considerato. Il letterale è associato alla risorsa di appellazione anche con il predicato più generale *rdfs:label*. La proprietà scorciatoia *has_name* permette invece di collegare il nome direttamente alla risorsa che rappresenta l'oggetto.

- Per indicare la data di creazione di un museo, il modello CRM prevede l'utilizzo di una risorsa di tipo *Creation* che indica la creazione vera e propria dell'edificio, a sua volta collegata ad una risorsa temporale di tipo *Time-Span* che fa riferimento all'istante temporale (in genere espresso tramite l'anno) in cui tale evento è avvenuto. L'identificazione di tale istante temporale è ottenuta con una risorsa Appellation tramite un meccanismo del tutto simile a quello descritto per gli oggetti nel punto precedente.

- L'indicazione delle attività interne svolte in un museo avviene tramite la creazione di una risorsa di tipo *Activity* per ognuna di esse, il cui URI contiene il nome dell'attività considerata. La proprietà *mode_of_use* permette infine di indicare che si tratta di attività interne.

- Le informazioni relative alla descrizione, generale e storica, dell'edificio sono indicate tramite dei letterali (differenziati in base alle diverse lingue presenti) associati tramite le proprietà *has_note* e *has_historical_note*.

In figura 4.5 sono presentate invece le principali proprietà tra la risorsa di tipo Place, che rappresenta l'oggetto di riferimento come luogo geografico, e altre entità che ne permettono ulteriori descrizioni non possibili, invece, facendo riferimento all'oggetto Site.

- L'identificazione del luogo geografico in cui è situato l'oggetto considerato è modellata facendo utilizzo di diversi identificativi, rappresentati

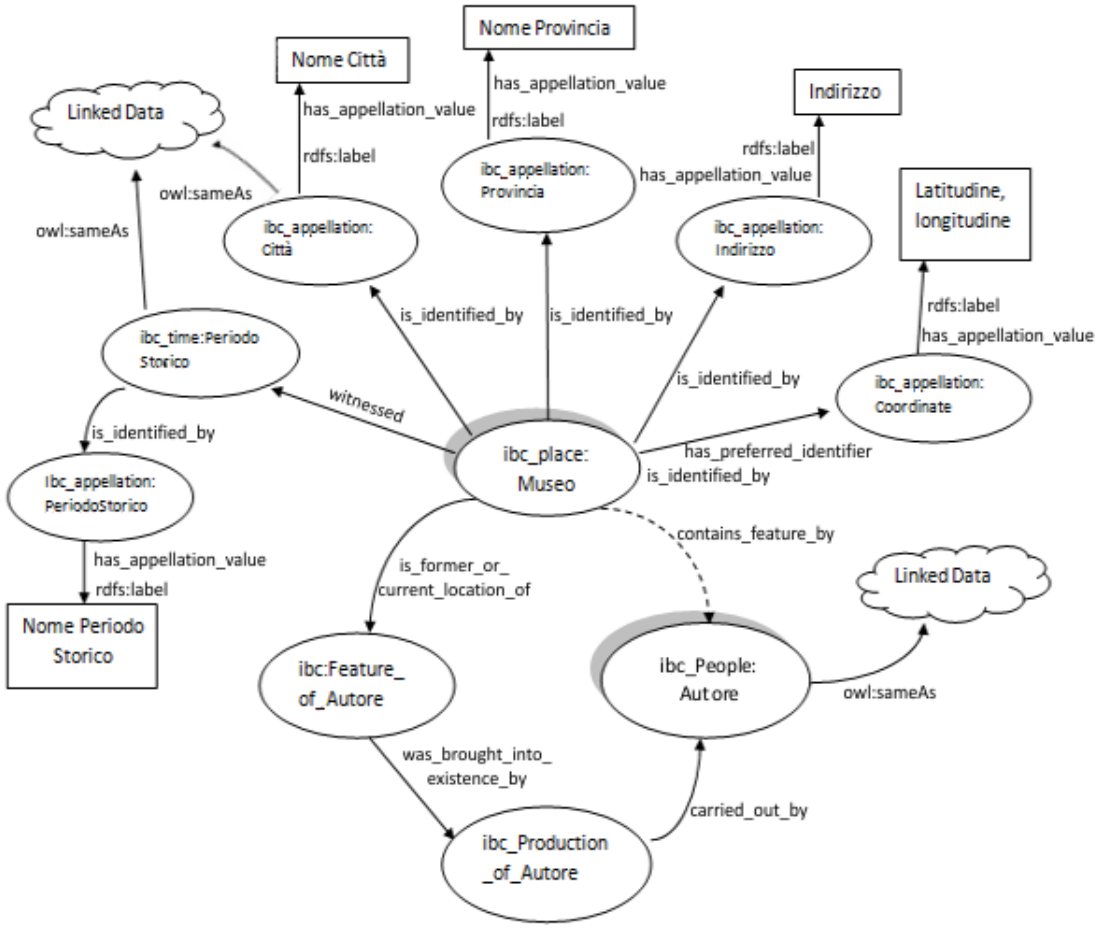


Figura 4.5: *Diagramma di modellazione che indica tutte le proprietà e le entità collegate alla risorsa di tipo 'Place' che indica un oggetto del catalogo inteso come luogo geografico.*

da varie tipologie di risorse di appellazione. Nello specifico sono create: una risorsa di tipo *City* che indica la città, una di tipo *Province* che fa riferimento alla provincia, una di tipo *Address* che indica l'indirizzo e una di tipo *SpatialCoordinates* che indica più nello specifico le coordinate di geolocalizzazione (tramite l'indicazione di latitudine e longitudine). Ognuna di esse presenta un ulteriore predicato per indicare il valore letterale dell'identificativo considerato (il nome della città, della provincia, l'indirizzo e le coordinate di latitudine e longitudine). Tutte queste entità sono sottoclassi della più generica classe Appellation. In base all'ontologia CIDOC-CRM un determinato luogo può infatti essere identificato anche da diversi appellativi; uno di essi può essere inoltre scelto come identificativo 'preferito', in base alle esigenze e preferenze del modellatore, tramite la proprietà *has_preferred_identifier*. In questo caso tale scelta è ricaduta sulle coordinate, che rappresentano l'identificativo che permette di ottenere le informazioni più precise e specifiche per l'identificazione del luogo. Le risorse che rappresentano le città sono inoltre collegate, tramite il predicato owl:sameAs, con altre presenti sia nei dataset informativi di DBpedia e FreeBase che in quelli geografici di GeoNames e LinkedGeoData.

- La proprietà *witnessed* permette di collegare la risorsa Place con una risorsa di tipo *Period* che rappresenta il periodo storico. Tramite il solito meccanismo di identificazione quest'ultima è associata al letterale che contiene il nome vero e proprio del priodo storico considerato.

- La rappresentazione delle informazioni riguardanti gli autori collegati ad un museo è effettuata tramite un meccanismo che richiede la creazione di diverse risorse intermedie. La modellazione utilizzata, come mostrato in figura 4.5, fa riferimento innanzitutto a una risorsa che indica la presenza, all'interno del museo, di oggetti che contengono caratteristiche legate all'autore considerato: la classe proposta da CIDOC-CRM per indicare tali risorse è *Man-Made Feature*; è creata

un'entità di questo tipo per ogni autore collegato al museo cosiderato. Ognuna di queste è collegata a un'ulteriore risorsa di tipo *Production* che fa riferimento all'evento di produzione delle caratteristiche legate ad uno specifico artista; quest'ultima è infine associata, tramite la proprietà *carried_out_by*, alla risorsa di tipo *Person* che si riferisce all'autore inteso come persona. Data la profondità di relazioni necessaria per la descrizione degli autori, è stata creata la proprietà scorciatoia *contains_feature_by* che collega l'oggetto Place direttamente con la risorsa che si riferisce alla persona. Quest'ultima presenta a sua volta dei collegamenti con risorse che si riferiscono alla stessa persona in DBpedia e FreeBase.

Il diagramma in figura 4.6 presenta infine la modellazione utilizzata in OpenERCH per descrivere le informazioni relative alle persone, ossia agli autori/artisti collegati con l'oggetto considerato. Come prima spiegato le risorse che si riferiscono ad ognuno di essi sono associate alla risorsa di tipo Place riferita al sito considerato. Di seguito una descrizione delle relazioni presentate in figura 4.6.

- L'identificazione dell'autore, che fa riferimento al suo nome e cognome, avviene tramite l'utilizzo di una risorsa Appellation secondo il meccanismo utilizzato in tutti gli altri casi e descritto precedentemente.

- Per descrivere l'occupazione dell'autore (ad esempio se si tratta di uno scultore, piuttosto che di un pittore) è utilizzato il meccanismo delle tipologie fornito dal modello CRM. Le risorse che fanno riferimento ad ogni occupazione sono infatti istanze della classe Type e sono associate alla risorsa che rappresenta l'autore tramite il predicato has_type.

- La descrizione delle città di nascita e morte dell'autore avviene tramite una modellazione che prevede la creazione di risorse che indicano, rispettivamente, la nascita e la morte dell'autore. CIDOC-CRM prevede infatti l'esistenza delle classi *Birth* e *Death* proprio per indicare

Figura 4.6: *Diagramma di modellazione che indica tutte le proprietà e le entità collegate alla risorsa di tipo 'Person' che indica un autore del catalogo.*

entità che fanno riferimento agli eventi di nascita e morte di una persona. Ad essi è associata la risorsa che indica la città in cui tali eventi sono avvenuti; quest'ultima è identificata tramite due risorse di appellazione riferite alla città e alla provincia, e presenta dei collegamenti con dataset esterni in modo del tutto analogo a quanto già descritto precedentemente per le altre risorse che fanno riferimento alle città.

- L'indicazione della data di nascita e morte avviene tramite la creazione di risorse temporali che rappresentano l'anno di nascita, inteso come

istante nel tempo. Esse sono collegate sia ai rispettivi eventi di nascita e morte degli autori, che alle specifiche risorse di appellazione che le identificano tramite dei letterali che indicano l'anno di riferimento.

4.3.2 *Il trasformatore*

Come mostrato nell'architettura propostra in figura 4.3, il trasformatore è il componente che si occupa di trasformare i dati grezzi iniziali in base al modello ontologico di CIDOC-CRM ottenendo la modellazione appena descritta. Tale coponente è stato realizzato in linguaggio Java facendo utilizzo delle librerie open source *Jena*, che mettono a disposizione un ambiente di programmazione per la creazione e gestione di grafi RDF, permettendo anche l'utilizzo di ontologie espresse in RDFS o OWL e del linguaggio di interrogazione SPARQL.
L'attività di trasformazione è incentrata sostanzialmente su un meccanismo di conversione progressivo basato sulla tipologia di informazione; ogni tipo di dato viene infatti rappresentato tramite la creazione di diverse risorse e proprietà secondo i modelli di rappresentazione esposti nella sezione precedente. I dati grezzi iniziali, che, come spiegato in 4.2, sono contenuti in file XML e CSV che espongono un elenco di proprietà assegnate ad ogni singolo oggetto, sono innanzitutto letti e memorizzati in alcune strutture e interfacce interne tramite la creazione di appositi oggetti Java, sui quali viene lanciata poi la vera e propria attività di trasformazione. Tale attività intermedia si è resa necessaria in quanto, all'inizio del lavoro legato ad OpenERCH, non tutti i dati erano disponibili nella loro completezza: l'utilizzo di interfacce dati ha permesso quindi di implementare le operazioni necessarie per la conversione nella loro completezza anche su un set iniziale di dati non completo, ottenendo comunque un componente generico che è poi stato utilizzato in un secondo momento anche per ottenere la trasformazione dell'intero dataset.
Nella sezione precedente si è mostrato come il modello ontologico utilizzato richieda la creazione di numerose risorse e proprietà, ad ognuna delle quali deve essere assegnato un URI. Esistono sostanzialmente due tipolo-

gie di indirizzi all'interno di OpenERCH: quelli usati per identificare le risorse ontologiche (classi e proprietà di CIDOC-CRM) e quelli relativi alle risorse vere e proprie. I primi, ai quali, nella rappresentazione RDF prodotta, è assegnato il namespace '*cidoc-owl*', sono ottenuti aggiungendo al prefisso '*http://dominio/IBCBologna/owl/*' il nome dell'entità a cui si fa riferimento; ai secondi sono, invece, assegnati diversi namespace a seconda delle tipologie di risorse rappresentate (per le risorse di appellazione, ad esempio, è utilizzato il namespace '*ibc-appellation*', per quelle di classificazione '*ibc-classification*', e così via), tali URI condividono comunque il prefisso '*http://dominio/IBCBologna/resource/*'. Per ogni oggetto del catalogo, come spiegato, sono create due risorse (una di tipo Site e una di tipo Place), il cui URI è ottenuto concatenando al prefisso prima indicato il nome vero e proprio del museo (apportando eventuali modifiche per eliminare caratteri non consentiti negli URI). Il 'Fotomuseo Giuseppe Panini', ad esempio, sarà rappresentato tramite la creazione delle due risorse '*http://dominio/IBCBologna/resource/Fotomuseo_Giuseppe_Panini*', che rappresenta il museo in generale come edificio, e '*http://dominio/IBCBologna /resource/place:Fotomuseo_Giuseppe_Panini*', che lo rappresenta come luogo geografico. Allo stesso modo le risorse che si riferiscono invece alle categorie piuttosto che ai periodi storici conterranno nell'URI, dopo l'opportuno prefisso, il nome della categoria o del periodo così come indicato nei dati iniziali. Dato che, per definizione, gli URI devono essere univoci si è resa necessaria un'operazione di disambiguazione dei nomi dei vari musei presenti nel catalogo. In alcuni casi è infatti possibile che alcuni di essi abbiano lo stesso nome (i musei civici di due città diverse hanno, ad esempio, entrambi la denominazione 'Museo Civico'). Durante la lettura delle informazioni è stata perciò implementata un'operazione di controllo di eventuali oggetti con nomi identici: in tal caso alla denominazione proposta nel catalogo è aggiunto il nome della città (il museo civico della città di Correggio sarà denominato, ad esempio, 'Museo Civico di Correggio', e allo stesso modo quelli di altre città); ciò assicura che la creazione degli URI relativi ai diversi musei avvenga in

modo corretto e che non esistano URI identici assegnati a risorse diverse.
La conversione di determinate tipologie di informazioni richiede l'attuazione di operazioni aggiuntive, oltre a quelle necessarie per modellare i dati secondo i modelli descritti. Ogni volta che incontra informazioni relative agli indirizzi, ad esempio, il trasformatore si occupa anche di reperirne le coordinate di geolocalizzazione, che sono utilizzate poi come identificativo 'preferito' per la risorsa Place assegnata all'oggetto analizzato (come descritto nella sezione precedente e indicato in figura 4.5). Tali informazioni sono ottenute utilizzando i servizi web di georeferenziazione messi a disposizione da Google Maps: in base all'indirizzo (fornito come stringa) è infatti possibile richiedere all'apposito servizio di Google, tramite invio di una richiesta HTTP, un documento XML che indica, tra le altre cose, anche le coordinate di latitudine e longitudine riferite a quello specifico luogo. Queste vengono poi assegnate, come letterali, alla risorsa Appellation che identifica la posizione del museo. Come si è più volte sottolineato talune risorse devono presentare dei collegamenti con dataset del Linked Data. A tal fine il trasformatore, durante la conversione di dati che richiedono la creazione di queste tipologie di risorse, si occupa anche di effettuare delle query SPARQL sui dataset esterni per ottenere gli URI di queste risorse esterne. Nello specifico questo accade per informazioni che riguardano città, provincie, categorie, sottocategorie, periodi storici e autori. In base ai nominativi relativi ad ognuna di esse è effettuata una query SPARQL al dataset di DBPedia, che consente di individuare l'eventuale risorsa che presenta una proprietà *rdfs:label* associata ad un letterale uguale al nominativo della risorsa considerata. Una volta individuata tale risorsa, di essa è ottenuto sia il suo stesso URI, che quelli di tutte le altre risorse che compaiono come oggetti del predicato owl:sameAs (le risorse di DBpedia presentano infatti tale collegamento con altre risorse, se esistenti, di FreeBase, GeoNames e LinkedGeoData). Tutti gli URI reperiti tramite questa interrogazione vengono utilizzati per creare link in uscita, sempre di tipo owl:sameAs, anche dal dataset di OpenERCH. Nel caso delle categorie, come già spiegato in sezione 4.3.1, viene effettuato un

controllo aggiuntivo per individuare anche risorse di DBpedia che indicano vere e proprie categorie; tale controllo è effettuato sul namespace dell'URI della risorsa individuata, in quanto tali risorse di categorizzazione presentano un namespace specifico. A seconda della tipologia viene poi creato il collegamento tramite il predicato rdfs:seeAlso piuttosto che owl:sameAs. Tutte le informazioni, relative sia alle risorse esterne che ai dati di geolocalizzazione, vengono salvate in file XML locali in modo che in eventuali successive trasformazioni possano essere recuperate localmente evitando un eccesso di richieste ai server esterni.

Per rendere più semplice e leggibile l'attività di trasformazione, il trasformatore si occupa di convertire ogni singola informazione tramite il modello presentato nella sezione 4.3.1, producendo quindi solo le proprietà di base necessarie per collegare le diverse risorse. Grazie alla descrizione in OWL dell'ontologia utilizzata, è possibile inferire nuova informazione (in particolare rappresentata dalla creazione di proprietà aggiuntive) in base alle regole e alle caratteristiche assegnate ad ogni proprietà e classe nel modello ontologico. CIDOC-CRM prevede, ad esempio, che ad ogni proprietà che collega due risorse, ne corrisponda una inversa (cioè in cui le classi di dominio e codominio sono invertite) che presenta un nome diverso dall'originale. Tutte queste informazioni aggiuntive vengono asserite grazie all'utilizzo di un ragionatore: tale componente, a cui è fornita sia l'ontologia che il modello RDF di base prodotto dal trasformatore, si occupa di aggiungere tutte le triple necessarie per esprimere completamente l'informazione in base a quanto indicato dall'ontologia e già fornito dagli statement esistenti. Se un'entità, ad esempio, presenta la proprietà *has_category* con una determinata categoria, la risorsa che rappresenta quest'ultima, grazie all'attività del ragionatore, in automatico presenterà anche la proprietà *is_category_of* (che è appunto la proprietà inversa di has_category) con tutte le entità che hanno tale categoria. Se, inoltre, una risorsa viene definita come istanza di una determinata classe ontologica, essa sarà istanza anche di tutte le altre classi gerarchicamente superiori a quella considerata. Ciò significa che tale entità presenterà

il predicato *rdf:type* con tutte queste classi. Per rendere le informazioni più facilmente leggibili (ed individuare più velocemente la classe specifica a cui una risorsa fa riferimento), tra l'entità considerata e la classe diretta di cui è istanza è aggiunta la proprietà *direct_type*. Il ragionatore utilizzato, messo a disposizione dalle librerie Jena, si occupa anche di asserire nuova informazione facendo riferimento a delle regole di inferenza, scritte secondo un modello di riferimento ben preciso all'interno di un file esterno. Tali regole consentono di verificare l'esistenza di determinati statement e, in caso affermativo, creare nuove triple. Tale funzionalità è utilizzata, ad esempio, per creare alcune delle proprietà 'scorciatoia' prima descritte.
Il risultato dell'attività di trasformazione è rappresentato da diversi file RDF, che vengono poi caricati sul repository di dati fornito da Sesame per essere facilmente acceduti.

4.4 Accesso ai dati e deferenziazione degli URI

I principi del Linked Data richiedono che gli URI utilizzati per identificare le risorse siano conformi al protocollo HTTP, cioé che sia possibile deferenziarli. Il meccanismo utilizzato in OpenERCH per ottenere tale caratteristica, come spiegato nella sezione 3.3.2, consiste nell'utilizzare degli URI differenti per fare riferimento ad una risorsa e a delle sue specifiche rappresentazioni. A partire dall'URI della risorsa è infatti possibile costruire un indirizzo che permette di ottenerne una sua rappresentazione nel formato desiderato (HTML oppure varie serializzazioni RDF).
Per ottenere tale risultato è stata creata un'applicazione web, basata sull'application server Tomcat, che ha essenzialmente il compito di fornire, a fronte di richieste HTTP effettuate agli indirizzi delle varie rappresentazioni delle risorse, i documenti che le descrivono. Le risorse presenti in OpenERCH ricadono infatti nella categoria di 'risorse non informative' (*non-information resource*), in quanto fanno riferimento a concetti astratti o veri e propri oggetti fisici che ovviamente non possono essere restituiti a fronte di una richiesta se

non tramite una loro rappresentazione documentale. Il Linked Data propone due diversi approcci per la deferenziazione di URI che si riferiscono a risorse di questo genere, quello utilizzato in OpenERCH si basa sui meccanismi di *negoziazione dei contenuti* e *redirezione* previsti dal protocollo HTTP. L'idea di base consiste nel fatto che il richiedente specifica, nella propria richiesta, il tipo di documento che preferisce e il Server può pertanto selezionare la risposta appropriata. Tale specifica viene fatta tramite un apposito *header* (chiamato *accept*) inserito nel pacchetto di richiesta HTTP.
L'applicazione è composta perciò da una *servlet* che gestisce le richieste effettuate sugli URI delle risorse e fornisce i risultati appropriati. Se l'indirizzo fornito si riferisce ad una generica risorsa (e non ad una sua specifica rappresentazione), la servlet analizza il contenuto presente nell'header e risponte tramite un pacchetto HTTP particolare con codice '*303 See Other*' che contiene un nuovo URI che si riferisce ad una nuova risorsa che descrive la risorsa non informativa iniziale. A questo punto il client potrà effettuare una nuova richiesta direttamente al nuovo indirizzo, ottenendo questa volta il documento che la rappresenta nel formato richiesto.
Per ottenere la descrizione di una risorsa l'applicazione, una volta ottenuto l'URI che la identifica, effettua una query SPARQL che le permette di ottenere tutte le triple RDF in cui tale risorsa compare come soggetto o oggetto. La costruzione dei documenti che hanno il compito di descrivere la risorsa avviene in modi diversi a seconda che il formato richiesto sia una specifica serializzazione RDF piuttosto che un documento HTML. Nel primo caso, infatti, le librerie Jena permettono di costruire direttamente un documento RDF (nel formato desiderato) a partire dai risultati dell'interrogazione, pertanto in tal caso l'applicazione torna semplicemente questo documento. Per ottenere invece una rappresentazione in HTML (come quella mostrata nel capitolo precedente in figura 3.2), viene prima ricavato un file RDF (serializzato in XML) contenente tutte le triple risultanti dalla query e su esso viene poi applicata una trasformazione XSLT che permette di ottenerne una rappresentazione in HTML. Questa presenta tutti i predicati e relativi oggetti

associati alla risorsa iniziale, rappresentati all'interno di una tabella. Come già spiegato, una risorsa può avere diversi tipi (ossia essere istanza di diverse classi ontologiche), per consentire una migliore visualizzazione è stato creato un piccolo controllo *javascript* che nasconde la lista di tutti i tipi e, a fronte del predicato rdf:type, mostra solamente il tipo diretto a cui la risorsa è associata (identificato anche dalla proprietà direct_type). Cliccando sull'apposito link è possibile comunque visualizzare tutti i tipi da cui discende la risorsa. In figura 4.7 è mostrato un esempio dei passaggi effettuati per ottenere la visualizzazione HTML di una risorsa.

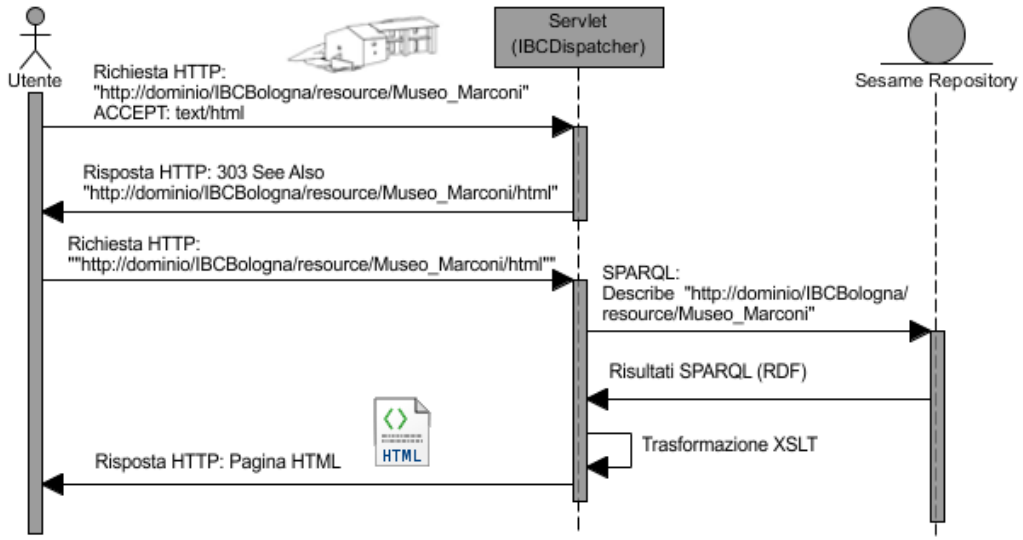


Figura 4.7: *Diagramma di sequenza che descrive l'attività di deferenziazione degli URI in OpenERCH tramite il meccanismo di negoziazione dei contenuti e redirezione.*

Un'apposita pagina web permette di effettuare interrogazioni SPARQL al dataset di OpenERCH tramite una semplice interfaccia, come spiegato in sezione 3.3.2 e mostrato in figura 3.3. Alcuni controlli javascript si occupano di cambiare i vari formati relativi ai risultati a seconda della tipologia di query inserita dall'utente, ricavabile dalla prima parola inserita nell'interrogazione (che ne indica appunto il tipo). La richiesta viene inoltrata ad un'altra

servlet che si occupa, a sua volta, di effettuare l'interrogazione sul repository messo a disposizione dallo store Sesame, all'interno del quale sono stati caricati i dati ottenuti dalla trasformazione. I risultati sono poi convertiti dalla servlet a seconda del formato richiesto. Anche in questo caso le API di Sesame permettono di ottenere i risultati in diverse serializzazioni RDF, che vengono restituite direttamente dalla servlet. Nel caso in cui, però, l'utente abbia richiesto i risultati in HTML, essa ottiene i dati in formato XML e si occupa di convertirli tramite una trasformazione XSLT. I fogli di stile creati sono diversi a seconda della tipologia di query, per le quali cambiano anche i formati dei risultati: nello specifico per le query di selezione i risultati sono rappresentati tramite una tabella che contiene gli URI ottenuti per ogni variabile richiesta nell'interrogazione, mentre per le query di costruzione o descrizione le triple ottenute sono rappresentate in una tabella che mostra i valori di soggetti, predicati e oggetti.

4.5 Applicazione di mash-up: aspetti tecnici e dettagli implementativi

Come spiegato nella sezione 3.4, del progetto OpenERCH fa parte anche l'applicazione ERCH-Mashup creata allo scopo di testare i meccanismi di accesso al dataset ottenuto dalla conversione e per dare un esempio di come questi possono essere utilizzati e combinati con quelli del Linked Data con i quali sono intrecciati. Nello specifico si tratta di un'applicazione web, anch'essa eseguibile in ambiente Tomcat, basata sull'utilizzo di servlet Java. L'interfaccia web proposta dall'applicazione permette di specificare dei criteri di ricerca per l'individuazione degli oggetti del catalogo, presentando prevalentemente una serie di *combo box* popolate dinamicamente tramite delle chiamate *ajax*. Alcuni controlli javascript si occupano infatti di effettuare delle chiamate dinamiche ad un'apposita servlet che, a sua volta, ha il compito di restituire, tramite delle interrogazioni SPARQL al repository di OpenERCH, i valori di scelta che saranno visualizzati nei menù. Tale servlet

si occupa, ad esempio, di creare interrogazioni che permettono di conoscere tutte le città di una determinata provincia, basandosi sulle risorse che rappresentano le città e le provincie assegnate ad ogni oggetto secondo il modello proposto nella sezione 4.3.1.
Compito principale dell'applicazione, oltre all'attività di filtro, è quello di fornire informazioni aggiuntive relative a diverse tipologie di informazione. Tale attività è svolta ancora una volta tramite la creazione dinamica di query SPARQL effettuate, però, non solo sui dati interni di OpenERCH ma anche sui dataset presenti nel Linked Data e descritti in sezione 4.1.2. Le informazioni aggiuntive vengono infatti reperite tramite delle richieste di tipo *GET* effettuate ad apposite servlet che, in base ai parametri indicati nella richiesta, sono in grado di comprendere la tipologia di dato rispetto al quale si intende effettuare la ricerca di informazione aggiuntiva. Le servlet si occupano di individuare eventuali risorse esterne collegate a quella di interesse e, tramite la creazione di interrogazioni, reperiscono informazioni relative alla loro descrizione presente nei dataset esterni. Tutti i risultati reperiti vengono infine utilizzati per creare apposite pagine web contenenti tutte le notizie aggiuntive reperite riguardo al concetto indicato.
Le informazioni relative a categorie e sottocategorie sono recuperate, ad esempio, tramite interrogazioni dirette sia al dataset di *DBpedia* che ad un dataset messo a disposizione dal progetto *Linked Open Data Italia*[4] che contiene una lista geo-referenziata dei musei italiani divisi in categorie.
Interrogando il dataset interno di OpenERCH viene innanzitutto recuperato l'URI della risorsa di DBpedia corrispondente alla categoria di interesse (il predicato di riferimento, come spiegato in sezione 4.3.1, è rdfs:seeAlso), utilizzato per la creazione delle interrogazioni successive. Di seguito si descrivono i meccanismi utilizzati per recuperare le varie tipologie di informazioni.

[4]Linked Open Data Italia è un progetto avviato da diverse aziende nell'ambito dell'*Open Government* per incentivare la pubblicazione di dati di pubblica utilità posseduti dalle pubbliche amministrazione. Propone alcuni dataset di vario tipo, ottenibili sia tramite *dump* in RDF che tramite interrogazioni a degli endpoint SPARQL. `http://www.linkedopendata.it`.

- La descrizione generale della risorsa e la relativa immagine vengono ottenute ricercando, nel dataset di DBpedia, gli oggetti con i quali la risorsa che indica la categoria è collegata tramite i predicati *dbpedia-owl:abstract* e *dbpedia-owl:thumbnail* (il namespace '*dbpedia-owl*' fa riferimento alle risorse identificate dall'ontologia di DBpedia): il primo fa riferimento ad un letterale (contenente la descrizione generale) e il secondo all'URL dell'immagine. La query costruita utilizza inoltre alcuni filtri per privilegiare, se presente, la descrizione in italiano, considerando quella in inglese solo nel caso in cui la prima non fosse presente.

- La pagina di Wikipedia e di Freebase relative alla categorie sono individuate in DBpedia tramite i predicati *foaf:page* e owl:sameAs; vengono quindi recuperati tramite delle interrogazioni che fanno riferimento a tali proprietà.

- Le persone collegate alla categoria considerata sono ottenute ricercando in DBpedia tutte le risorse che presentano, con la risorsa considerata, almeno uno tra i seguenti predicati: *dbpedia-owl:field*, *dbpprop:field* o *dbpedia-owl:mainInterest*. Per ognuna di esse viene inoltre recuperata la label, che conterrà il nome della persona.

- I musei internazionali relativi alla categoria considerata sono recuperati dal dataset di DBpedia considerando tutte le risorse istanze della classe ontologica *dbpedia-owl:Museum* che presentano la proprietà *dbpedia-owl:type* con la risorsa relativa alla categoria. Anche in questo caso per ognuno di essi è recuperata la label.

- I musei italiani della stessa categoria sono infine ottenuti dal dataset di Linked Open Data Italia, all'interno del quale le categorie dei musei sono identificate da risorse descritte in DBpedia. Vengono pertanto ricercati tutti i musei che presentano come categoria esattamente la risorsa individuata inizialmente.

Il procedimento appena descritto costituisce il meccanismo generale con cui vengono recuperate le informazioni aggiuntive relative alle diverse tipologie di dati, creando delle query che fanno riferimento di volta in volta a proprietà specifiche che consentono di ottenere le informazioni di interesse desiderate.
In alcuni casi è possibile, tuttavia, che venga richiesta la ricerca di notizie aggiuntive relative a risorse non presenti nel dataset di OpenERCH (la ricerca di informazioni addizionali sulla categoria arte, ad esempio, propone, come si è appena descritto, anche l'elenco di persone collegate a tale categoria, che possono non essere presenti nel catalogo di OpenERCH). Tale situazione viene resa nota alla servlet tramite un'apposito parametro. In questo caso la risorsa di partenza di DBpedia non è individuata tramite i collegamenti owl:sameAs o rdfs:seeAlso presenti nel dataset interno, ma ricercando direttamente in DBpedia la risorsa che presenta una label che coincide con il nome dell'oggetto di interesse.
Nel caso in cui si effettuino ricerce su luoghi geografici, vengono coinvolti anche i dataset di GeoNames e LinkedGeoData. Il primo è utilizzato per individuare eventuali città o luoghi geografici prossimi ad una specifica città della quale si stanno cercando informazioni aggiuntive. Anche in questo caso viene individuato (direttamente dai collegamenti presenti in OpenERCH, oppure passando da DBpedia in caso si tratti di una città non presente nel catalogo) l'URI della risorsa che rappresenta tale città nel dataset di GeoNames. Dato che quest'ultimo non fornisce un endpoint per le interrogazioni SPARQL, in questo caso è richiesto l'intero documento RDF di feature che, come descritto in sezione 4.1.2, contiene i dati dei luoghi vicini alla risorsa. Tramite le librerie Jena tale documento è analizzato e da esso sono recuperate le informazioni di interesse. Le notizie rilevate in questo modo sono unite a quelle recuperate da DBpedia tramite i meccanismi prima descritti.
Il dataset di LinkedGeoData è utilizzato per individuare i luoghi di interesse vicini ad un determinato museo. Un'apposita servlet, alla quale sono passate come parametro le coordinate di geolocalizzazione del museo considerato, si occupa, in questo caso, di effettuare una particolare interrogazione all'end-

point SPARQL fornito da LinkedGeoData. Dei particolari predicati proposti in tale dataset permettono infatti di recuperare tutte le risorse vicine ad un determinato punto spaziale indicato tramite coordinate di latitudine e longitudine. Le risorse così individuate vengono filtrate in base alla tipologia (bar, ristorante, chiesa, ecc.), recuperabile tramite la classe ontologica istanziata da ognuna di esse. Per ognuna vengono inoltre recuperati i dati relativi alle coordinate di locazione, utilizzati sia per creare una mappa che ne indica la posizione, che per fornirli ai servizi di Google che permettono di ottenere (se richiesto dall'utente) le informazioni stradali per raggiungere tale posizione a partire dall'indirizzo del museo considerato (anche per esso, come specificato prima, si possengono le coordinate di latitudine e longitudine).
Tutti i meccanismi fino ad ora descritti e utilizzati nella realizzazione di ERCH-Mashup, mettono in evidenza come sia possibile, tramite la creazione dinamica di query SPARQL ai vari endpoint, riuscire sia a filtrare facilmente le informazioni, riuscendo ad individuare esattamente i dati desiderati, che a correlare tra loro dati recuperati da diverse sorgenti creando facilmente valore aggiunto a livello informativo.

Capitolo 5

Valutazione

Nel capitolo precedente è stata presentata una descrizione tenica relativa alle principali scelte implementative adottate per la conversione dei dati e le altre operazioni e funzionalità realizzate tramite il progetto OpenERCH. Di seguito si vuole tentare una valutazione dei risultati ottenuti, relativamente agli obiettivi preposti, analizzando prima aspetti quantitativi quali tempistiche delle operazioni e tempi di risposta delle applicazioni e successivamente aspetti qualitativi relativi all'aderenza agli standard e principi di fondo previsti per le tecnologie adottate.

5.1 Analisi quantitativa

La breve analisi quantitativa che si propone di seguito si riferisce prevalentemente alle tempistiche relative alle principali operazioni svolte all'interno di OpenERCH. Come illustrato nelle sezioni precedenti, il progetto si articola in diverse componenti operative in qualche modo indipendenti tra loro, ognuna delle quali risponde a specifiche esigenze; di seguito si propongono alcune osservazioni relative ai tempi di risposta richiesti per le varie operazioni effettuate da ognuna di queste componenti.

- Trasformatore: valutazioni quantitative relative a tale componente possono essere effetuate valutando i tempi richiesti per la vera e propria

attività di conversione da dati grezzi a quelli aderenti al modello ontologico. Come descritto nella sezione 4.3.2, tale componente, nello svolgere la vera e propria trasformazione, si occupa anche di interrogare i dataset esterni per la creazione dei link in uscita. Il tempo totale della trasformazione risente pertanto dei tempi di risposta relativi a tali richieste esterne, che possono risultare variabili e dipendono dai server messi a disposizione da tali dataset. Nonostante tale variabilità, queste tempistiche risultano spesso molto ridotte e le risposte alle query effettuate vengono ottenute in genere nel giro di pochi secondi o frazioni di secondo. L'attività di trasformazione (che, si ricorda, include anche l'operazione di inferenza da parte del ragionatore) è stata lanciata diverse decine di volte durante lo sviluppo e non ha mai richiesto tempi superiori ai due minuti (tenendo conto anche dell'alta quantità di stampe prodotte per monitorare l'attività), a fronte di un input relativo ai metadati di tutti i siti presenti nel catalogo Samira e corrispondenti a oltre un migliaio di oggetti, dai quali è ricavato un modello finale contenente circa 250 mila triple RDF.

- Web application per l'accesso ai dati e la deferenziazione degli URI: come descritto nella sezione 4.4, le principali attività svolte in quest'ambito sono il recupero dei dati di interesse tramite interrogazioni al repository di Sesame e eventuali trasformazioni per la loro visualizzazione. Operazione preliminare a quelle appena citate è il caricamento del dataset ottenuto dalla conversione sul gestore dei dati fornito da Sesame (tramite un'interfaccia chiamata *Workbench*). I dati vengono caricati a partire dai file RDF prodotti dal trasformatore e la loro memorizzazione in Sesame richiede tempi che a volte arrivano a qualche minuto. Dato che si tratta di un'attività svolta in rare occasioni (solo se si ha la necessità di convertire nuovamente i dati, ad esempio a fronte di qualche aggiornamento al dataset) tale tempistiche possono essere considerate molto buone.
  I tempi di risposta relativi invece alle operazioni di deferenziazione degli

URI richieste all'apposita applicazione web risultano ottimi, in quanto si ottengono i relativi documenti nel giro di alcune frazioni o al massimo pochi secondi, anche nel caso di documenti HTML (che, come descritto nella sezione 4.4, richiedono trasformazioni XSLT). Ovviamente tali tempistiche dipendono dai tempi di risposta forniti da Sesame che quindi si comporta in modo molto positivo a fronte del dataset gestito.
Le interrogazioni SPARQL effettuate tramite interfaccia richiedono essenzialmente le stesse tempistiche ottenibili effettuando le stesse query direttamente dalle API fornite da Sesame (ciò significa, quindi, che i tempi richiesti dalla servlet per la gestione dell'interrogazione e le eventuali trasformazioni dei risultati sono irrisori). È impossibile fornire una stima precisa dei tempi richiesti per ottenere i risutati di un'interrogazione, in quanto questi sono fortemente dipendenti dalla tipologia stessa dell'interrogazione effettuata. A fronte di un numero di risultati non particolarmente elevato le risposte sono ottenibili entro alcuni secondi, tuttavia per query particolarmente sofisticate, che indicano numerosi pattern e forniscono elevati numeri di risultati, tali tempistiche possono allungarsi fino ad arrivare anche a qualche minuto.

- Applicazione di mash-up: in questo caso possono essere valutati i tempi di attesa richiesti dall'applicazione sia relativamente alle attività di ricerca/filtraggio che di recupero di informazioni aggiuntive. Come spiegato nelle sezioni 3.4 e 4.5, tale applicazione permette di effettuare ricerche sugli oggetti del catalogo in base a diversi parametri (provincia o città di appartenenza, categorie, sottocategorie, artisti e periodi storici associati), che vengono tradotte in query SPARQL effettuate sul dataset di OpenERCH. Tali interrogazioni possono pertanto essere abbastanza sofisticate e soprattutto fare riferimento ad un numero molto elevato di triple richiedendo dei tempi di risposta a volte troppo lunghi (fino a decine di secondi o qualche minuto). Tale problematica risiede probabilmente nel fatto che il dataset propone un numero molto elevato di triple, derivanti principalmente dall'attività di inferenza, e

questo impatta sicuramente sul meccanismo di *matching* effettuato dal motore che effettua le interrogazioni. Dato che molte delle informazioni descritte da tali triple inferite sono sostanzialmente inutili per il tipo di ricerche proposte, è stato creato un secondo repository contenente solo gli statement prodotti dal trasformatore senza l'utilizzo del ragionatore (le query di ricerca utilizzate, in effetti, non fanno uso di eventuali proprietà inferite). Le interrogazioni per le ricerche vengono effettuate su tale nuovo insieme di dati che risulta molto più compatto e permette di ottenere dei tempi di risposta molto più ridotti che non superano mai qualche secondo.

Per quanto riguarda il reperimento di informazioni aggiuntive, in tal caso le tempistiche dipendono prevalentemente da quelle fornite dagli endpoint dei dataset sui quali vengono effettuate le interrogazioni per il reperimento dei dati. Nel complesso, considerando sia il reperimento dei dati che la loro elaborazione effettuata dalle servlet per la creazione delle pagine web finali visualizzate, entro pochi secondi è possibile ottenere i risultati richiesti. Alcune problematiche si incontrano solo nel caso di informazioni aggiuntive su città e derivano essenzialmente dalle richieste fatte al dataset di GeoNames che risulta spesso non raggiungibile o fornisce errori di risposta: in tal caso le richieste vengono ripetute fino ad un massimo di 5 volte (in genere dopo qualche tentativo il servizio risponde bene), superate le quali queste informazioni vengono scartate nella pagina web finale prodotta.

In base ai risultati appena presentati e descritti, si può giungere alla conclusione che le tempistiche prodotte da OpenERCH risultano accettabili per un utilizzo sereno e senza particolari problematiche dei dati da esso messi a disposizione e di eventuali applicazioni che su di essi possono essere costruite, anche se fanno riferimento a informazioni esterne, come dimostrato dagli ottimi tempi di risposta forniti dall'esempio dell'applicazione di mash-up creata.

5.2 Aderenza ai principi del Linked Data e scoperta informativa

Come più volte sottolineato, l'obiettivo principale dell'attività svolta nell'ambito del progetto OpenERCH consiste nel rendere facilmente accessibili e utilizzabili i dati sui beni culturali della regione Emilia Romagna utilizzando le tecniche proposte dal Linked Data che, come si è analizzato, risultano tra le più adatte per rendere tali informazioni raggiungibili e consumabili anche in un'ottica di relazione con altri dati correlati. Valutare qualitativamente l'attività svolta potrebbe risultare complicato, in quanto non esistono dei parametri oggettivi e facilmente determinabili. Al fine di misurare in qualche modo la bontà del lavoro svolto si può fare, tuttavia, riferimento ai principi su cui l'intero concetto di Linked Data è basato, valutando se ognuno di essi è stato rispettato e in che modalità.
Una prima valutazione può essere effettuata in base ai principi generali proposti in [BER07] (e presentati nella sezione 2.4.2).

- *Utilizzare gli URI per identificare le risorse.* In OpenERCH, come ampiamente descritto nelle sezioni precedenti, ogni risorsa è identificata da specifici URI costruiti secondo delle regole ben precise e tenendo conto dei principi generali per la creazione di indirizzi facilmente accessibili e memorizzabili proposti in [SCA08].

- *Utilizzare gli URI in conformità al protocollo HTTP, in modo che possano essere consultati.* Anche tale principio ha rappresentato un punto saldo del lavoro svolto in OpenERCH, richiedendo la creazione di tutto il meccanismo relativo alla deferenziazione degli URI descritto in sezione 4.4. Tramite il meccanismo della negoziazione dei contenuti è possibile ottenere una rappresentazione documentale di qualsiasi risorsa presente in OpenERCH: ogni URI è quindi accessibile e consultabile. È importante specificare, in quest'ambito, che il dataset non è ancora stato formalmente caricato su un dominio web nel momento in cui si

scrive, principalmente a causa di alcune problematiche relative alle licenze d'uso con cui i dati sono proposti nel portale Samira da cui sono stati recuperati. L'attività di eventuale modifica o adattamento di tali licenze richiede precise procedure burocratiche talvolta molto lunghe, al termine delle quali i dati saranno probabilmente resi disponibili via web. Tale operazione, tuttavia, non incide sulle attività svolte e fino ad ora descritte: tutti i meccanismi e le applicazioni create non richiedono modifiche o adattamenti particolari per essere pubblicate e funzionare correttamente su un dominio web.

- *Utilizzare gli standard (RDF e SPARQL) come modello per la rappresentazione dei dati.* Il dataset costruito in OpenERCH è completamente basato sul modello a triple di RDF e facilmente accessibile, sia tramite interfaccia grafica che tramite apposite librerie, tramite interrogazioni gestite da un endpoint SPARQL. Tali standard sono pertanto stati utilizzati pienamente per la rappresentazione dei dati.

- *Includere link agli URI che identificano altre risorse, in modo che possano essere scoperte nuove informazioni.* L'attività di collegamento con risorse esterne ha rappresentato uno dei punti di maggiore attenzione durante tutto il lavoro, si è descritto ampiamente come molte risorse presentino proprietà che costituiscono dei collegamenti con quelle di altre risorse di dataset esterni permettendo la creazione di percorsi di scoperta informativa palesati anche tramite gli esempi proposti dall'applicazione di mash-up.

Si può pertanto affermare con facilità e franchezza che tutti i principi sono stati rispettati e adempiuti pienamente. Oltre a quelli appena descritti si può fare riferimento anche all'interpretazione di tali principi proposta dal Linking Open Data Project, che fornisce una lista di regole da rispettare per richiedere la pubblicazione di un proprio dataset all'interno del progetto. Nello specifico, oltre a regole che fanno riferimento essenzialmente a quanto

già richiesto dai principi precedenti, sono richieste le seguenti direttive, di cui si valuta l'aderenza:

- *Il dataset deve contenere almeno mille triple.* Come già asserito precedentemente il dataset di OpenERCH contiene 250 mila triple che descrivono oltre mille oggetti che a loro volta fanno riferimento a diversi altri concetti tra città, persone, attività e così via.

- *Il dataset deve essere connesso con link RDF ad altri dataset già presenti nel Linking Open Data Project, sono richiesti arbitrariamente almeno 50 collegamenti.* Tutti i dataset con cui sono stati effettuati i collegamenti fanno parte del Linking Open Data Project; le triple che fanno riferimento a collegamenti di questo genere sono oltre 2100 (si pensi a tutte le risorse che si riferiscono a città, provincie, categorie, attori che presentano collegamenti con tre o quattro risorse diverse).

- *L'accesso all'intero dataset deve essere possibile tramite crawling RDF, dump RDF o tramite un endpoint SPARQL.* Il *crawling* si riferisce alla possibilità di navigare tra le risorse visualizzando per ognuna di esse una rappresentazione sia accedendo ad uno specifico indirizzo che utilizzando eventuali browser RDF. In OpenERCH tale attività è possibile grazie ai meccanismi di negoziazione dei contenuti realizzati per la deferenziazione degli URI delle risorse. Il *dump RDF* è ottenibile facendo riferimento ai documenti RDF creati dall'attività di trasformazione. Si è infine più volte spiegato come OpenERCH disponga di un endpoint per le interrogazioni SPARQL. Tutti questi meccanismi di accesso sono pertanto stati implementati.

La modellazione dei dati di OpenERCH rispetta quindi tutti i principi proposti dal Linked Data e ciò permette di rispondere alle principali esigenze per le quali tale attività è stata svolta. I dati, come si è visto, possono essere facilmente filtrati permettendo di effettuare ricerche molto specifiche tramite la creazione di query SPARQL anche relativamente semplici. Il modello ontologico a cui si è fatto riferimento ha permesso inoltre di avere una

rappresentazione delle informazioni molto formale e precisa, consentendo di arricchire ogni singolo concetto considerato e rappresentando una concreta alternativa ai modelli tipici di rappresentazione incentrati sugli oggetti, da cui derivano tutte le principali limitazioni e problematiche descritte nella sezione 1.1. A tutto questo può essere aggiunto l'ulteriore, importante vantaggio di avere collegamenti con dati esterni, che permettono di arricchire qualli già posseduti correlandoli in modalità nuove lasciando agli utilizzatori la possibilità di creare valore di interesse. L'applicazione di mash-up creata a scopi esemplificativi ha fornito un esempio concreto di come sia facilmente possibile ottenere importante valore aggiunto direttamente utilizzabile e utile per migliorare l'accesso e la vivibilità delle informazioni. Tutto questo, in aggiunta alla perfetta aderenza ai principi del Linked Data, porta, in definitiva, a valutare la modellazione di OpenERCH come un'attività che apporta un importante miglioramento qualitativo all'accesso e all'utilizzo dei dati messi a disposizione dalla regione Emilia Romagna, aprendo nuovi scenari di utilizzo e ampie possibilità di creazioni applicative su essi basate.

Conclusioni

L'obiettivo dominante che ha portato alla realizzazione del lavoro presentato in questo documento di tesi si basa, come più volte sottolineato, sulla convinzione che le tecnologie legate al *Semantic Web* rappresentino lo strumento ideale per la rappresentazione della conoscenza in ambito artistico e culturale. Il modello a grafo tipico dell'RDF, l'utilizzo di modelli ontologici per la definizione formale delle informazioni e la possibilità di associare e arricchire le proprie sorgenti dati sfruttando carichi informativi esterni rappresentano sicuramente le caratteristiche principali che consentono di ottenere una modellazione dei dati relativi a beni artistici e culturali in grado di esprimere a pieno il valore informativo degli oggetti. Il dataset prodotto tramite il progetto *OpenERCH*, e tutte le applicazioni ralizzate per consentirne un accesso e un utilizzo conforme agli standard e ai principi del *Linked Data*, costituisce un esempio pratico e diretto di come una rappresentazione precisa e formale delle informazioni permetta di ottenere numerosi vantaggi legati principalmente alle attività di accesso diretto e specifico ai dati e al loro arricchimento tramite l'integrazione con quelli di domini informativi diversi e complementari.

L'attività svolta ha lo scopo, in qualche modo, di mostrare come sia possibile, sfruttando queste tecnologie, rendere disponibili i dati in un formato aperto e aderente agli standard, consentendone un utilizzo il più possibile generico e non legato esclusivamente alla sola attività di consultazione. *ERCH-Mashup*, applicazione di mash-up realizzata come esempio, ha in effetti messo in evidenza come, facendo riferimento ad un modello formale e ad una rete di dati effettivamente esistente e utilizzabile, si riesca facilmente ad individua-

re con precisione l'informazione di interesse abbattendo le frontiere imposte dai singoli domini informativi e reperendo, tramite modalità standard, i dati indifferentemente dalla sorgente che li contiene. Tutto ciò inoltre non ha richiesto la creazione di nessun modello dati o meccanismo realizzato ad hoc, ma anzi l'utilizzo di tecniche riutilizzabili e di un insieme di dati che risultano effettivamente reimpiegabili per scopi e realizzazioni software libere e attualmente non pensate. L'aderenza ad un modello ontologico standard e ormai diffuso porta infine l'ulteriore vantaggio di permettere una comprensione più facile dello schema con cui i dati sono espressi, invogliando e semplificando le attività di correlazione con altri eventuali dataset simili che potranno in futuro essere inseriti all'interno della rete di dati del Linked Data.
Esistono diversi aspetti che possono essere migliorati e potenziati all'interno di OpenERCH. Come spiegato nella descrizione delle caratteristiche del sistema realizzato proposta in questo documento, oltre alla conversione vera e propria l'attività svolta ha rigurdato anche la creazione di tutti i meccanismi necessari e previsti dal Semantic Web per rendere il dataset creato accessibile sia tramite interrogazioni che tramite la deferenziazione degli URI relativi alle diverse risorse in esso rappresentate. Una prima, importante, attività che potrà essere svolta al di là del presente progetto di tesi consiste nel rendere effettivamente reperibile ed utilizzabile il dataset prodotto all'interno del *Linking Open Data Project*. Le tempistiche relative alla regolarizzazione di diverse procedure burocratiche inerenti le licenze d'uso dei dati utilizzati nell'attività di tesi non ha infatti permesso, al momento della pubblicazione di questo documento, di rendere effettivamente reperibili in un dominio web i dati ottenuti dal lavoro svolto. Come illustrato nella valutazione dei risultati conseguiti (presentata nel quinto capitolo), tuttavia, tutte le applicazioni realizzate possono essere pubblicate all'interno di un dominio web senza necessità di particolari modifiche e il loro utilizzo permette di aderire completamente a tutti i requisiti richiesti per entrare a far parte del *Web of Data*. Un auspicabile e imminente sviluppo futuro del progetto consiste pertanto nel rendere raggiungibile il dataset creato all'interno di un dominio

web per chiederne la formale pubblicazione e ammissione all'interno della rete di dati formata dal Linking Open Data Project. In questo modo le informazioni prodotte potranno non solo essere utilizzate molto più facilmente per i più disparati scopi applicativi (in relazione con le altre già presenti con le quali presentano collegamenti e associazioni) ma anche essere oggetto di future correlazioni da parte di nuovi dataset inseriti nella rete.
Come spiegato nel terzo capitolo, i dati oggetto di lavoro riguardano prevalentemente musei, teatri storici e altre tipologie di edifici e collezioni di interesse storico, artistico e culturale dislocati nella regione Emilia Romagna. Il dataset di OpenERCH può tuttavia essere arricchito ulteriormente tramite l'inclusione di nuovi dati relativi agli oggetti veri e propri che sono contenuti all'interno dei contenitori presi in esame. Informazioni riguardanti reperti storici o archeologici piuttosto che opere ed oggetti d'arte di vario tipo sono infatti presenti nel portale *Samira* dal quale i dati oggetto del lavoro sono stati recuperati; le questioni relative alle licenze d'uso di cui si accennava prima non hanno però consentito di poterli includere nel dataset realizzato. Il modello ontologico di riferimento risulta abbastanza vasto e completo da fornire tutte le classi e le relazioni necessarie per la rappresentazione della conoscenza relativa a tali entità. Un ulteriore, possibile, attività futura può consistere quindi nella creazione di un modello aderente all'ontologia CIDOC-CRM che permetta la conversione di tutte le informazioni relative anche a tali oggetti, ottenendo quindi una completa trasformazione semantica dell'intero catalogo.
Gli sviluppi appena proposti fanno riferimento alle attività inerenti la conversione vera e propria dei dati. Come si è più volte sottolineato, però, il vero obiettivo dell'attività svolta consiste nel rendere tali informazioni accessibili e disponibili per diversi scopi a chi ne abbia interesse, permettendo la creazione di applicativi come quello di mash-up proposto. Numerose e in qualche modo non prevedibili risultano perciò le attività future che potranno essere svolte tramite l'utilizzo di questi dati, creando delle applicazioni che possano rispondere a diverse esigenze proprio secondo i principi tipici e che stanno

alla base del web semantico. I dati prodotti con OpenERCH, presentando anche diverse informazioni relative alla geolocalizzazione degli oggetti, potrebbero, per fare un esempio, risultare molto interessanti e di grande utilità per applicazioni (adesso molto diffuse ed utilizzate) che si occupano di creare percorsi geografici o turistici di vario tipo, fornendo servizi simili a quelli delle mappe di Google. La stessa applicazione di mash-up (il cui sviluppo, si ricorda, ha essenzialmente uno scopo di esempio pratico e di valutazione su come i risultati ottenuti possono essere utilizzati) potrebbe essere oggetto, per esempio, di ulteriori miglioramenti tramite la ricerca di nuove tipologie di informazione o tramite la realizzazione di una nuova interfaccia più sofisticata e accessibile (magari pensata per i dispositivi mobili).
In aggiunta alle attività appena proposte, è importante, per concludere, osservare come nuovi interessanti sviluppi possono delinearsi indipendentemente da quanto è stato realizzato nell'ambito di OpenERCH e descritto in questo documento di tesi. Si è visto infatti come risulti importante avere dei dati correlati con quelli che rientrano in discipline e domini diversi e come questa correlazione porti alla creazione di valore aggiunto. È stato inoltre raccontato come il Web of Data sia in qualche modo una rete in continua evoluzione, e come esistano diversi progetti che puntano ad utilizzare ed espandere sempre di più la rete informativa da esso messa a disposizione. Gli sviluppi di cui si parla si rifersicono pertanto alla possibilità, in futuro, di correlare i dati di OpenERCH con nuove sorgenti informative di interesse, incrementando la quantità di collegamenti in uscita e dando la possibilità, di conseguenza, di creare arricchimenti ancora più utili e interessanti. Si può pensare, ad esempio, ad un futuro in cui nella rete di dati semantici entrino a far parte numerosi altri dataset che raccolgono altre informazioni su musei o altri beni culturali di altre regioni, che risulterebbero sicuramente di fondamentale interesse per arricchire il valore di quelli proposti in OpenERCH. Più la rete di dati si espanderà, specie se in riferimento al campo artistico e culturale, maggiori saranno le prospettive future di utilizzo di sorgenti dati come quella realizzata in questo lavoro di tesi.

Bibliografia

[ABV11] Aliaga D. G., Bertino E., Valtolina S.: "DECHO-A Framework for the Digital Exploration of Cultural Heritage Objects". *ACM Journal on Computing and Cultural Heritage*, 3(3), Friedlander A. (Ed.), Marzo 2011 (`http://portal.acm.org/citation.cfm?id=1921619`).

[ACG98] Auer P., Cavallini F., Giffi E., Lattanzi M.: *Strutturazione dei dati delle schede di catalogo - Normativa per la strutturazione e il trasferimento dei dati*, ICCD, 1998 (`http://www.iccd.beniculturali.it/getFile.php?id=160`).

[AH08] Allemang D., Hendler J.: *Semantic Web for the Working Ontologist - Modeling in RDF, RDFS and OWL*, Morgan Kaufmann Publ, 2008.

[AMD10] Antoine I., Meghini C., Dekkers M., Gradmann S. et al: *Europeana Data Model Primer v1.0*, 05 Agosto 2010 (`http://www.europeana-libraries.eu/web/europeana-project/technicaldocuments/`).

[AS06] Alesso H. P., Smith C. F.: *Thinking on the Web: Berners-Lee, Gödel and Turing*, Wiley-Interscience, 2006.

[BAK00] Baker T.: "A grammar of Dublin Core". *D-Lib Magazine*, 6(10), Laurence Lannom (Ed.) , 2000 (`http://www.dlib.org/dlib/october00/baker/10baker.html`).

[BCC09] Baruzzo A., Casoto P., Challapalli P., Dattolo A.: "An Intelligent Service Oriented Approach for Improving Information Access in Cultural

Heritage". *Journal of Digital Information*, 10(6), John Leggett (Ed.), 2009 (http://sole.dimi.uniud.it/~paolo.casoto/papers/7.pdf).

[BCE11] Belisario E., Cogo G., Epifani S., Forghieri C.: *Come si fa Open Data? Istruzioni per l'uso per Enti e Amministrazioni Pubbliche*. Convegno Amministrazioni Intelligenti che cambiano il paese, Roma, 14 aprile 2011, E-GOV - Cultura e Tecnologie per l'Innovazione in collaborazione con l'Associazione Italiana per l'Open Government, 2011.

[BCH07] Bizer C., Cyganiak R., Heath T.: *How to Publish Linked Data on the Web*, 2007 (http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/), ultima visita: 31 Maggio 2011.

[BER07] Berners-Lee T.: *Linked Data - Design Issues*, 27 Luglio 2006 (ultima modifica: 18 Giugno 2009), (http://www.w3.org/DesignIssues/LinkedData.html), ultima visita: 31 Maggio 2011.

[BF00] Berners-Lee T., Fischetti M.: *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*, HyperCollins, 2000.

[BHB09] Bizer C., Heath T., Berners-Lee T.: "Linked Data - The Story So Far". *International Journal on Semantic Web and Information Systems*, 3(5), Sheth A. (Ed.), 1-22, 2009 (http://dblp.uni-trier.de/db/journals/ijswis/ijswis5.html#BizerHB09).

[BLK09] Bizer C., Lehmann J., Kobilarov G., Auer S. et al: "DBpedia - A crystallization point for the Web of Data". *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3), Finn T. (Ed.), 154-165, 2009 (http://portal.acm.org/citation.cfm?id=1640848).

[CAN05] Cantone F.: "Shared Technologies in archeologia: nuove prospettive di gestione e condivisione di dati in rete". *Archeologia e Calcolatori*, 16, Paola Moscati (Ed.), 271-290, 2005 (http://soi.cnr.it/archcalc/indice/PDF16/CANTONE271-290.pdf).

[CDG10] Crofts N., Doerr M., Gill T., Stead S., Stiff M.: *Definition of the CIDOC Conceptual Reference Model*, ICOM/CIDOC Documentation Standards Group e CIDOC CRM Special Interest Group, Gennaio 2010 (`http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.2.pdf`).

[COY07] Coyle K.: "Managing Technology. The Library Catalog in a 2.0 World". *The Journal of Academic Librarianship*, 33(2), David Kohl (Ed.), 289-291, 2007 (`http://www.kcoyle.net/jal_33_2.html`).

[DHL03] Doerr M., Hunter J., Lagoze C.: "Towards a Core Ontology for Information Integration". *Journal of Digital Information*, 4(1), John Leggett (Ed.), 2003 (`http://jodi.ecs.soton.ac.uk/Articles/v04/i01/Doerr/`).

[DOS03] Daconta M. C., Obrst L. J., Smith K. T.: *The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management*, Wiley, 2003.

[GAR01] Gartner Consulting, *The Emergence of Distributed Content Management and Peer-to- Peer Networks*. The Gartner Group, Gennaio 2001.

[HFB09] Hebeler J., Fisher M., Blace R., Perez-Lopez A.: *Semantic Web Programming*, Whiley, 2009.

[LPS05] Lagoze C., Payette S., Shin E., Wilper C.: "Fedora: An architecture for complex objects and their relationships". *International Journal on Digital Libraries*, 6(2), Springer (Ed.), 124-138, 29 Dicembre 2005 (`http://www.springerlink.com/content/x7224797g8703g30/fulltext.pdf`).

[MJV09] Marijn K., Jaap K., V. de Keizer: "Information Retrieval in Cultural Heritage". *Interdisciplinary Science Reviews*, 34(2-3), Willard McCarty (Ed.), 2009, 268-284 (`http://staff.science.uva.nl/~mhakoole/publications/2009/kool:info09.pdf`).

[MMS10] Masci M. E., Merlitti D., Scarselli T.: *Mapping tra le Schede di catalogo ICCD e VRA Core 4.0. Indicazione dei livelli minimi di catalogazione - Versione 2.0*, MiBAC, 08 Luglio 2010 (`www.culturaitalia.it/pico/documenti/museiditalia/Allegato_4.doc`).

[PDA05] Paolini P., Di Blas N., Alonzo F.: "ICT per i beni culturali : esempi di applicazione". *Mondo digitale*, 4(5), AICA (Ed.), 44-61, Settembre 2005 (`www.mondodigitale.net/Rivista/05_numero_quattro/Paolini_p._44-61.pdf`).

[SAA08] Schreibera G., Aminb A., Aroyoa L., van Assema M. et al: "Semantic annotation and search of cultural heritage collections: The Multimedia E-Culture demonstrator". *Web Semantics: Science, Services and Agents on the World Wide Web*, 6, Finn T. (Ed.), 243-249, 2008 (`http://oai.cwi.nl/oai/asset/13260/13260A.pdf`).

[SCA08] Sauermann L., Cyganiak R., Ayers D., Völkel M.: *Cool URIs for the Semantic Web*. W3C Interest Group Note, 03 Dicembre 2008 (`http://www.w3.org/TR/cooluris/`).

[SET09] Segaran T., Evans C., Taylor J.: *Programming the Semantic Web*, O'Reilly, 2009.

[SIG09] Signore O.: "Representing Knoledge in Archaeology: From Cataloguing Cards to Semantic Web". *Archeologia e Calcolatori*, 20, Paola Moscati (Ed.), 118-128, 2009 (`http://soi.cnr.it/archcalc/indice/PDF20/10_Signore.pdf`).

[SMM05] Signore O., Missikoff O., Moscati P.: "La Gestione della Conoscenza in Archeologia: Modelli, Linguaggi e Strumenti di Modellazione Concettuale Dall'XML al Semantic Web". *Archeologia e Calcolatori*, 16, Paola Moscati (Ed.), 291-319, 2005 (`http://soi.cnr.it/archcalc/indice/PDF16/SIGNORE291-319.pdf`).

[TBS03] Tansley R., Bass M., Stuve D., Branschofsky M., Chudnov D.: "The DSpace institutional digital repository system: current functionality". *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, IEEE Computer Society, 87-97, 2003 (`http://hdl.handle.net/1721.1/26705`).

Ringraziamenti

Grazie ...

Alla mia famiglia che mi ha permesso di arrivare fino a quì.

Ad Anna per tutto il sostegno e la pazienza con cui mi ha sopportato.

A NSI e, in particolare, al dott. Giuseppe Frangiamone per la possibilità che mi hanno dato e per la fiducia riposta.

Ad Andrea, Giuseppe, Emanuele, Patricia per il supporto e la sopportazione.

A Catia, Cristian, Antonio, Mattia e tutti i colleghi che hanno reso unica questa esperienza universitaria.

A tutti i miei amici che, in un modo o nell'altro, mi hanno sostenuto.

A chiunque abbia avuto la pazienza e l'interesse di leggere questo documento di tesi.