ECODIGIT

Ecosistema Digitale per la Fruizione e la Valorizzazione
dei Beni e delle Attività Culturali della Regione Lazio

D3.1 Report sul Censimento

Acronimo Progetto:	**EcoDigit**
Titolo Progetto:	**Ecosistema digitale per la fruizione e la valorizzazione dei beni e delle attività culturali della regione Lazio**

D3.1

Work Package:	WP3 T3.1	
Deliverable Dovuto il:		2 Gennaio 2019
Inizio Progetto:		2 Ottobre 2018
Durata Progetto:		15 mesi
Reponsabile Deliverable:		Massimo Mecella
Versione:		0.4
Stato:		Versione Finale
Autore:	Miguel Ceriani RM1 Massimo Mecella RM1	
Altri contribuenti al lavoro riportato nel deliverable:	Valentina Presuttti ISTC-CNR Marialuisa Mongelli ENEA Antionio Budano INFN Maria Prezioso RM2 Marco Canciani RM3 Giovanni Fiorentino UNITUS	
Reviewer:	Ludovica Marinucci ISTC-CNR Luigi Asprino ISTC-CNR	

Per citare questo documento si prega di utilizzare il seguente record bibliografico

Miguel Ceriani e Massimo Mecella. *D3.1 Report sul Censimento*. Deliverable Progetto EcoDigit. 2019

Revisioni			
Versione	**Data**	**Modificata da**	**Commento**
v 0.1	20/3/2019	Miguel Ceriani	Creazione Documento
v 0.2	29/03/2019	Ludovica Marinucci	Revisione documento
v 0.3	01/04/2019	Miguel Ceriani	Revisione documento
v 0.4	12/04/2019	Miguel Ceriani	Revisione documento

Executive Summary

Il presente deliverable D3.1 “Report sul Censimento” descrive l’indagine sulle fonti di dati di interesse per il progetto, presenti nel Lazio. Il questionario proposto è stato compilato finora 43 volte, per descrivere altrettante fonti di dati.

L’analisi delle informazioni inserite rivela l’eterogeneità delle fonti, sia per quanto riguarda la tipologia dei contenuti, che la granularità dei dati ed i formati utilizzati. In particolare, per quanto riguarda il livello di copertura delle fonti di dati di interesse, si rileva che al momento la maggior parte delle fonti censite sono quelle gestite direttamente dai partner, mentre poche fonti ”esterne” sono censite. Si rende necessario da questo punto di vista un ampliamento qualitativo della copertura, ad esempio diffondendo maggiormente questo questionario in forma mirata.

Gli inserimenti rappresentano i dataset che i partner considerano rilevanti e sono ricchi per quanto riguarda la descrizione delle tematiche e le tipologie di dati fornite, pur se spesso scarsi di informazioni tecniche su come accedere praticamente questi dati. I contenuti vanno da informazioni sulla ricerca accademica a modelli 3D di strutture architettoniche, da iscrizioni antiche a dati per la pianificazione territoriale sostenibile, a eterogenei dataset di beni culturali. Una considerevole fetta di questi contenuti sono localizzabili geograficamente.

Complessivamente le risposte forniscono un interessante quadro per iniziare ad esplorare l’ecosistema in cui si sta proponendo il sistema EcoDigit. La diversità dei dati disponibili obbliga a pensare ad un sistema flessibile che si possa adattare a diverse tipologie di dati e forme di organizzazione degli stessi.

Indice

Elenco delle figure

1 Introduzione

1.1 Obiettivi del Work Package

Il Work Package 3 del progetto EcoDigit ha l'obiettivo di analizzare, progettare e sviluppare metodologie e strumenti software per l'aggregazione e integrazione di molteplici tipologie di dati che riguardano i beni culturali nella Regione Lazio.

1.2 Obiettivo del deliverable

L'obiettivo di questo deliverable è descrivere l'attività di censimento delle fonti di dati nel Lazio che possano essere di interesse per il progetto.

1.3 Relazione con le altre attività del progetto

La ricognizione delle fonti di dati di interesse permetterà di precisare tecnologie e metodi necessari per effettuarne l'accesso e integrazione (WP2 e WP3), nonché il tipo di applicazione concreta che è possibile costruire per il dimostratore (WP4).

1.4 Outline documento

Nella Sezione 2 *Questionario* si descrive il contenuto e modalità di somministrazione del questionario. La Sezione 3 *Analisi dei Risultati* propone un'analisi delle risposte a questo questionario. Nella Sezione 4 *Conclusioni*, infine, si propone una riflessione sul processo e si tirano le conclusioni.

2 Questionario

Per effettuare il censimento, si è utilizzato un questionario online, fatto circolare tra il partner di progetto e all'esterno a partire dai contatti dei singoli partner. Ogni compilazione del questionario corrisponde alla descrizione di una singola fonte di dati.

2.1 Domande

Il questionario si compone di 26 domande, suddivise in 5 sezioni:

1. Informazioni generali
   (a) Indirizzo email
   (b) Referente (nome e cognome)
   (c) Referente (email)
   (d) Denominazione Sistema Informativo / Banca Dati / Repository / Repertorio
   (e) Ente responsabile
2. Tipologia ed informazioni tecniche sui dati
   (a) Composto da e/o integrato in/con Sistema/i
   (b) Identificativo dataset / sistema
   (c) Il sistema contiene dati primari o secondari ?
   (d) Dati georiferiti?
   (e) Copertura geografica
   (f) Scala della copertura geografica
   (g) Utilizzo della Banca Dati
   (h) Tipologia dei dati e breve descrizione
   (i) Tematica
   (j) Modalità e frequenza di aggiornamento
   (k) Consistenza: n. record e presenza di file allegati
3. Modalità di accesso ed utilizzo
   (a) Standard metadati utilizzati
   (b) Sistema informatico utilizzato
   (c) Formato di esportazione dei dati
   (d) Interoperabilità
   (e) Protocolli/standard di interoperabilità
   (f) Integrazione con altri dataset/banche dati/sistemi
4. Aspetti gestionali e legali
   (a) Presenza di linee guida, manuali, policy di gestione
   (b) Servizi connessi con l'utilizzo
   (c) Licenze d'uso per dati, documenti e metadati
5. Note
   (a) Inserire ulteriori aspetti rilevanti ai fini del DTC

Tutte le domande sono a testo libero. Unica eccezione è ”Modalità e frequenza di aggiornamento”, in cui si suggeriscono alcune risposte predefinite (*Giornaliera*, *Settimanale*, *Mensile*, *Trimestrale*, *Semestrale*, *Annuale*), lasciando comunque la possibilità di inserire testo libero anche in questo caso.

3 Analisi dei Risultati

Al momento della scrittura di questo deliverable, sono state ricevute 43 risposte. Il questionario rimane comunque online e ulteriori risposte verranno considerate nelle successive fasi di progetto.

3.1 Metodologia

Trattandosi per lo più di campi liberi, le riposte ricevute sono molto eterogenee, non solo nel contenuto ma anche nella forma. Data la necessità di uniformare e integrare le risposte per poterle confrontare, si è proceduto ad analizzare manualmente i dati. L'analisi quantitativa, che verrà descritta in dettaglio nei paragrafi successivi, è realizzata a valle di questa fase di normalizzazione e accorpamento tra sinonimi e concetti semanticamente contigui. A titolo di un esempio, RM2 è descritto nelle risposte come ente in sei modi diversi. Senza questa e altre normalizzazioni una qualunque analisi quantitativa avrebbe avuto poco senso.

3.2 Enti Partecipanti

Il questionario è stato compilato dai partner, i quali a loro volta lo hanno inviato ai loro contatti chiedendo di estenderlo ulteriormente a enti fornitori di dataset potenzialmenti interessanti. Analizzando i valori inseriti nel campo *Ente responsabile*, si può vedere in che modo i diversi enti sono rappresentati nel censimento.

La quasi totalità delle fonti di dati censite sono gestite da enti partner di progetto (vedi Figura 1). Questo risultato segnala che la circolazione tra enti esterni al progetto non ha dato i risultati sperati e rende necessario un ripensamento della strategia di diffusione del questionario per avere una copertura più completa. Questo non toglie comunque valore ai risultati attuali nell'ottica di un'iniziale esplorazione delle fonti esistenti nel Lazio.

Andando nel dettaglio dell'analisi dei singoli enti (vedi Figura 2), si nota una certa sproporzione relativa alla quantità di dataset censiti in relazione all'ente che ha compilato il questionario. Ad esempio, tra gli enti partner si rileva che mentre da RM3 sono stati censiti 16 dataset, da parte di ENEA ne risulta censito soltanto uno. Questa sproporzione deriva dall'eterogeneità delle immissioni per quanto riguarda la granularità considerata. Da un estremo ENEA ha censito con un unico inserimento un'intera piattaforma che permette l'accesso a molteplici dataset di tipo diverso, mentre RM3 ha censito singolarmente modelli 3d

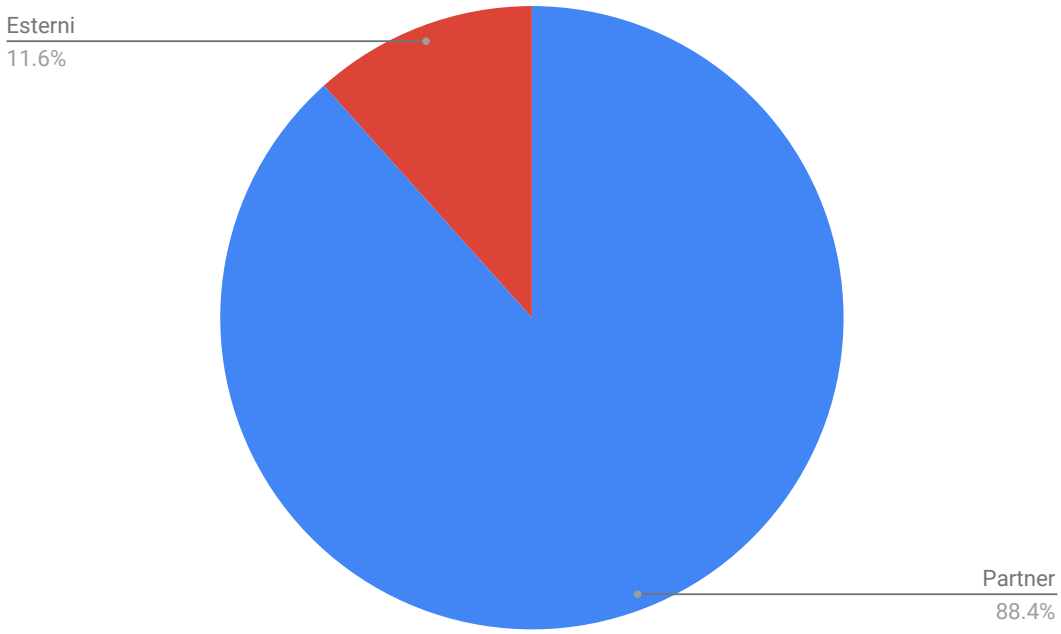

Figura 1: Distribuzione dei valori del campo *Ente responsabile*, distinti tra enti partner ed esterni al progetto

e dati di specifici monumenti, ognuno separatamente. Questo rappresenta un limite della forma aperta del presente questionario, oltre che un esempio della problematicità riscontrabile nel definire a che granularità i dataset debbano essere descritti e pubblicati. Nonostante lo schema del questionario sia stato il risultato di un gruppo di lavoro congiunto, al quale hanno partencipato esponenti degli enti partner sia del progetto EcoDigit sia di Anagrafe delle Competenze, che ha procedutto con più iterazioni a formlulare le domande ritenute più signiticative per cogliere la qualità dei dati, si sottolinea il carattere esplorativo anche dello schema del presente questionario, che si intende rimodulare a seguito dei risultati censiti procedendo ad ulteriori raffinazioni con criteri più precisi.

Riteniamo in ogni caso utile far notare, nel procedere alle successive analisi quantitative, che questa sproporzione nelle granularità incide ovviamente sui conteggi di occorrenze per tipologie di dati e simili classificazioni. Nonostante ciò, si è preferito evitare normalizzazioni sui conteggi che avrebbero avuto in ogni caso un certo livello di arbitrarietà. Per compensare questo problema e per il tipo di dati che stiamo considerando (anche una singola fonte dati atipica può essere particolarmente importante proprio per la sua specificità) l'analisi non è mai unicamente quantitativa, ma punta ad analizzare le diverse fonti nella loro eterogeneità.

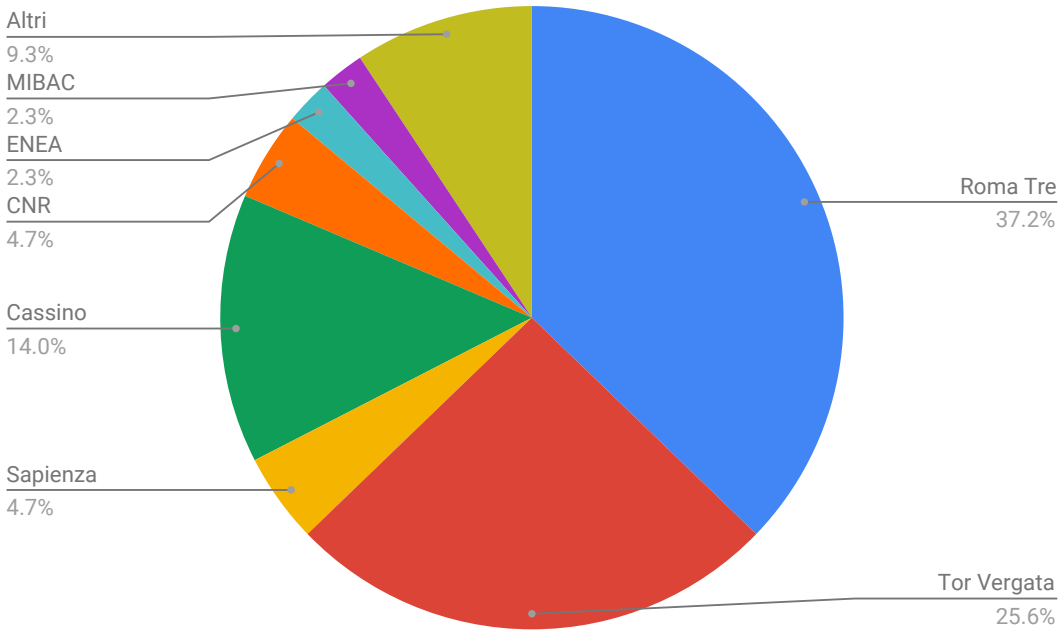

Figura 2: Distribuzione dei valori del campo *Ente responsabile*

3.3 Tematiche e tipologie dei dati

Attraverso una serie di campi, il questionario permette di descrivere la tematica dei dati e la loro tipologia. Analizzando il contenuto del campo *Tematica* si possono individuare e raggruppare gli ambiti coperti dalle fonti dati censite (Figura 3). Buona parte dei dataset (17 su 43) descrivono strutture architettoniche di interesse culturale. Si tratta, con un'unica eccezione, di fonti gestite dal Dipartimento Di Architettura di Roma Tre. Sei dataset riguardano iscrizioni di interesse paleografico. Due contengono informazioni su insiemi di beni culturali di tipo eterogeneo (in un caso beni culturali ecclesiastici gestiti dalle CEI, in un altro caso beni culturali analizzati da ENEA). Un dataset, STeMA-VAS, contiene dati legati alla pianificazione territoriale sostenibile. Ben sette dataset contengono informazioni su prodotti della ricerca accademica, non associabili direttamente (almeno in base alle informazioni fornite nel questionario) a uno specifico ambito dei beni culturali. Per otto dei 43 inserimenti, infine, non c'è nessuna informazione sulla tematica a cui i dati fanno riferimento.

Il campo *Tipologia* ha permesso di raccogliere informazioni sulle specifiche tipologie di dati disponibili (vedi Figura 4). Ogni fonte di dati offre in generale molteplici tipologie di dati. La maggior parte delle fonti (28) offrono qualche tipo di informazione testuale (ed esempio, articoli scientifici o annotazioni). Attorno alla metà offre immagini (che possono essere foto, ricostruzioni, ecc.) e tante (tutti i dataset su strutture architettoniche) offrono modelli 3D. Cinque contengono una descrizione di attività accademiche (pubblicazioni, strutture di afferenza, relazioni). Tre offrono altri contenuti multimediali (ad esempio, video) ed una dati geo-localizzati di tipo GIS. Per dieci fonti le tipologie offerte non sono descritte.

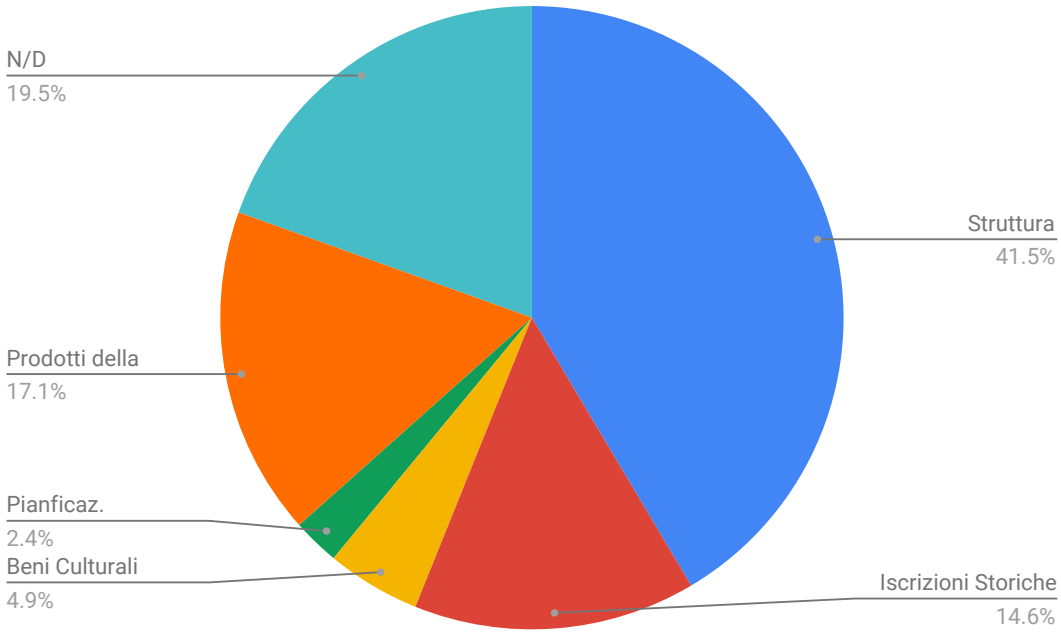

Figura 3: Distribuzione dei valori del campo *Tematica*

Ai partecipanti era inoltre richiesto di dichiarare se i dati forniti dalla fonte siano primari o secondari, ovvero derivati da un'altra fonte (vedi Figura 5). E' interessante notare che circa tre quarti delle fonti (33 su 43) contengono almeno in parte dati originali. Per più della metà (24 su 43) i dati sono soltanto o preponderantemente primari.

Come particolare aspetto della fonte censita, si chiedeva poi se i dati fossero georiferiti, ovvero se la descrizione di oggetti dotati di posizione nello spazio fosse associata alle relative coordinate (vedi Figura 6). Soltanto quattro fonti sono considerate georiferite in forma completa, mentre altre nove vengono descritte come georiferite parzialmente. Per le altre 30, venti sono dichiaratamente non georiferite mentre di dieci non vengono fornite informazioni.

Per capire il tipo di copertura geografica offerta dalle fonti di dati censite, è stato chiesto di descrivere la scala geografica di ogni fonte (vedi Figura 7). Le fonti di dato descritte si distribuiscono su diverse scale. Quelle a livello di singolo monumento sono 11, mentre quelle a scala urbana/territoriale sono 7. A scala provinciale/regionale sono 4, mentre 8 sono a livello nazionale. Una sola, infine, è a livello internazionale mentre di 11 non abbiamo informazioni o il concetto di scala geografica non è applicabile.

3.4 Interoperabilità

Nell'ottica di pianificare l'infrastruttura in grado di integrare queste fonti, il questionario prevedeva una serie di domande tecniche sugli standard utilizzati per offrire i dati. Purtroppo i corrispondenti campi sono stati compilati solo in una piccola parte dei casi. Senza un'ulte-

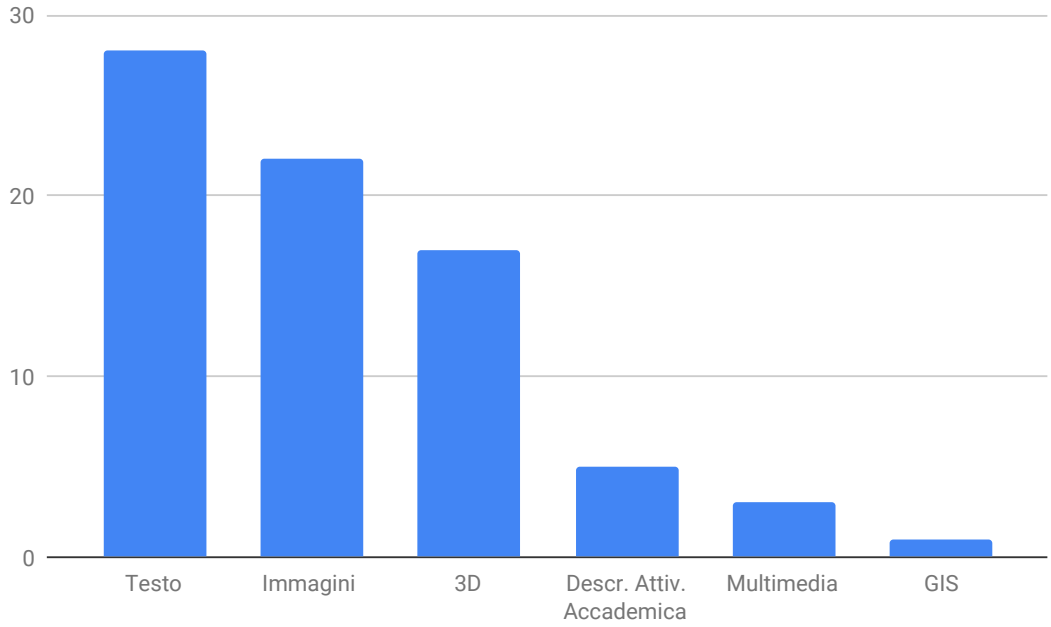

Figura 4: Distribuzione dei valori del campo *Tipologia*

riore analisi non è possibile assumere se chi ha effettuato l'inserimento ignorava la risposta a questi quesiti oppure se le fonti dati di cui non si hanno informazioni non supportano effettivamente nessun livello di interoperabilità. Nel seguito si procede in ogni caso ad analizzare i casi per cui abbiamo informazioni.

Per quanto riguarda il *Formato di esportazione dei dati* (vedi Figura 8), abbiamo l'informazione in 13 casi su 43.

- Dati strutturati di uso generale:
  - database relazionali descritti come DBF (un caso) o SQL (un altro caso, non è chiaro se si tratta di una connessione a un database o un dump),
  - strutture XML (un caso) e MODS (anche basato su XML, due casi),
  - tabelle CSV (quattro casi) ed Excel (tre casi).
- Formati strutturati dedicati ad applicazioni specifiche:
  - formati per citazioni bibliografiche (tre casi),
  - per modelli 3D (quattro casi),
  - per GIS (un caso),
  - per modelli biomeccanici di corpi nello spazio (un caso).
- Formati non strutturati:

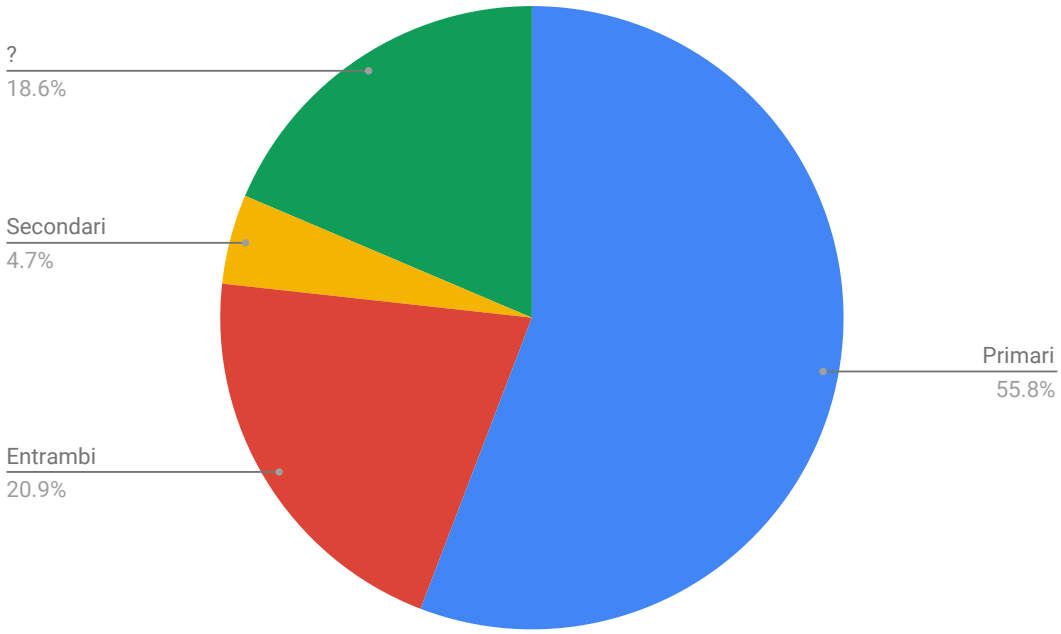

Figura 5: Distribuzione dei valori del campo *Il sistema contiene dati primari o secondari?*

- immagini JPEG (tre casi),
- documenti PDF (un caso),
- testo semplice.

Ancora meno informazioni abbiamo su eventuali *Standard metadati utilizzati* (vedi Figura 9). Il campo è stato riempito in nove casi e in sei di questi la risposta è che non è utilizzato nessuno standard. Le uniche fonti dati di cui sappiamo che si utilizza qualche standard per i metadati sono tre. Due utilizzano il vocabolario Dublin Core, mentre una usa MODS.

Il campo *Interoperabilità* (vedi Figura 10), infine, era probabilmente quello più importante per identificare in quale modo queste fonti dati possano essere integrate. Questo campo è stato compilato con informazioni significative soltanto in cinque casi. Da questi inserimenti sappiamo che due fonti di dati supportano OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting), una WebDAV, una possiede un API e una è in qualche modo integrabile con U-GOV.

4 Conclusioni

Il presente deliverable ha descritto l'indagine sulle fonti di dati di interesse per il progetto. E' stato proposto un questionario composto da 26 domande, che è stato compilato finora 43

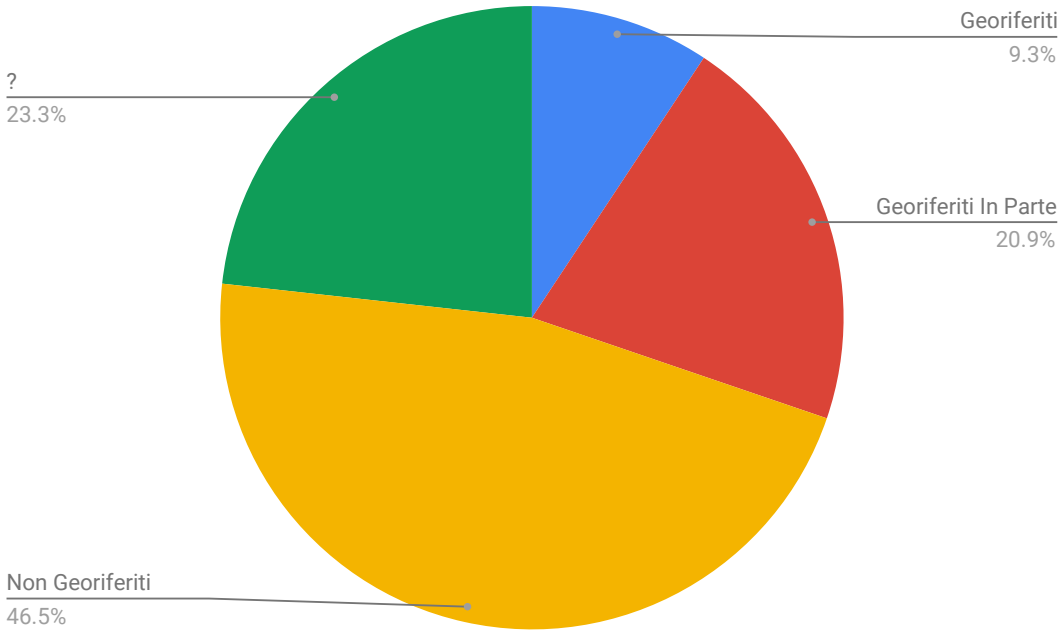

Figura 6: Distribuzione dei valori del campo *Georiferito*

volte, per descrivere altrettante fonti di dati. L'analisi dei risultati rivela alcuni limiti dell'indagine, ma anche molti elementi interessanti sulle tematiche e tipologie di dati da considerare nel progetto.

I principali limiti sono i seguenti:

- la scarsa risposta al censimento al di fuori dei partner di progetto (di fatto perdendo la possibilità di analizzare altre interessanti fonti di dati),
- l'eterogeneità nelle granularità considerate per descrivere i sistemi e
- la scarsa presenza di informazione tecniche su come interagire con le fonti dati descritte (forzando quindi ad analizzare ogni fonte separatamente per capire se e come è possibile accedervi).

Questi limiti non sono probabilmente dovuti a particolari errori nel processo, quanto al fatto che l'obiettivo di censire un insieme di fonti di dati in un ambito complesso ed eterogeneo non può essere realizzato in maniera "completa" solamente attraverso la condivisione di un questionario per un tempo limitato. Per completare le informazioni in funzione dei requisiti di progetto, questo censimento può essere considerato l'avvio di un processo che includerà necessariamente un'interazione dialettica tra i partner e con gli stakeholder esterni al progetto. Come parte di questo processo, si propone di mantenere il censimento aperto per tutta la durata del progetto.

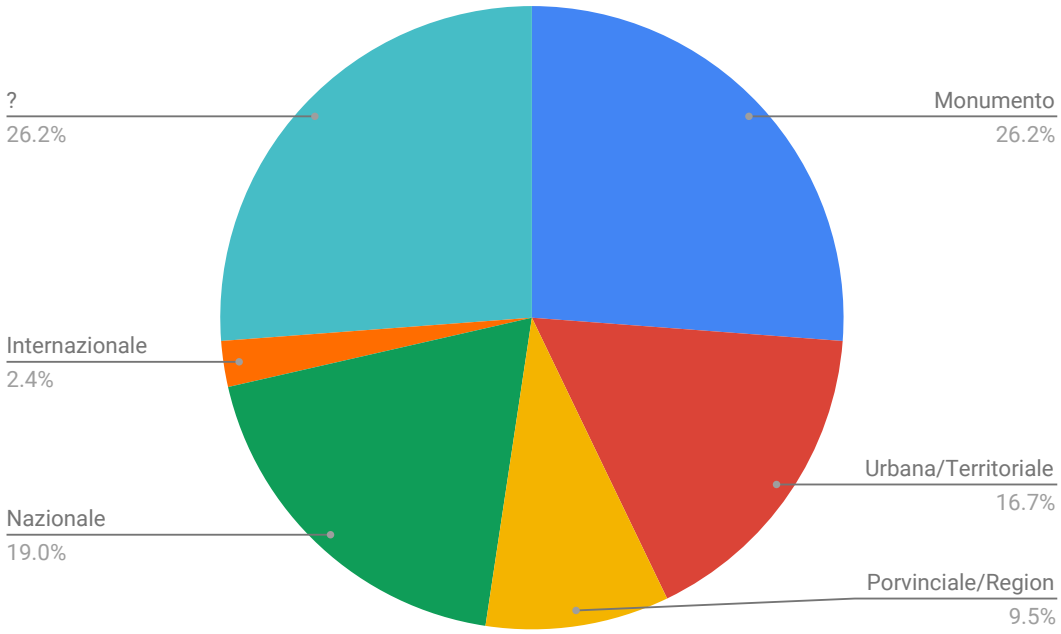

Figura 7: Distribuzione dei valori del campo *Scala geografica*

Quello che può essere detto al momento è che le principali tematiche descritte dai dati disponibili sono costituite da opere e dati architettonici di intesse culturale, iscrizioni di interesse paleografico, sistematizzazioni di prodotti di ricerca, insieme a qualche fonte di dati ad ampio spettro su diversi tipi di beni culturali. Si rileva inoltre che, per le informazioni fornite, sono disponibili diverse tipologie di dati (testo, multimedia, 3D, ecc.), ma pochi dati strutturati che utilizzino un qualche standard condiviso. Questo potrebbe complicare il progetto del sistema di integrazione, ma allo stesso rappresenta un'interessante sfida in un'ottica di ricerca.

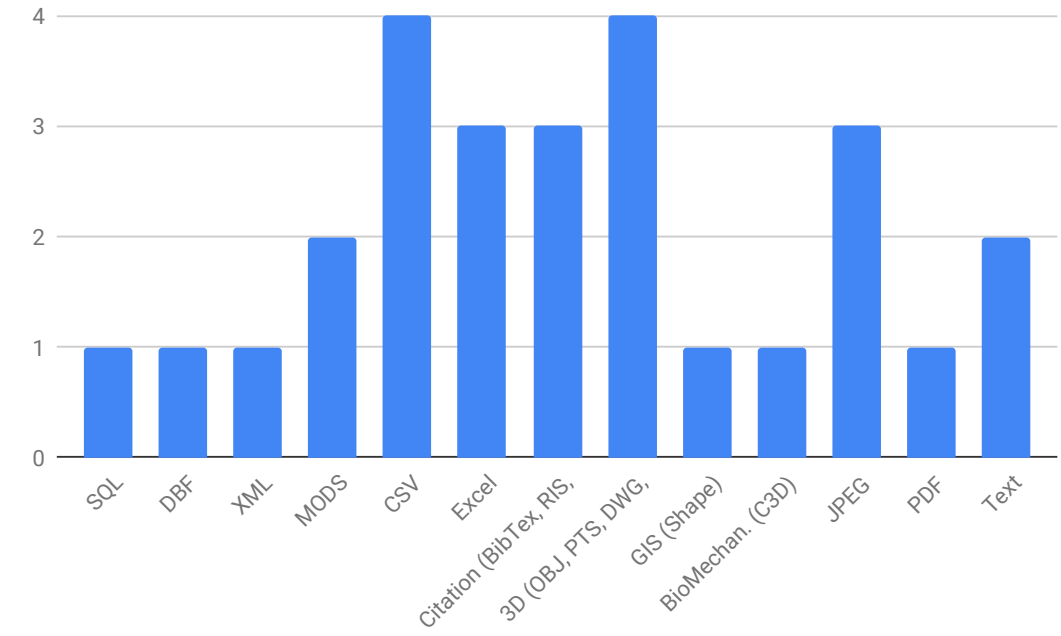

Figura 8: Distribuzione dei valori del campo *Formato di esportazione dei dati*

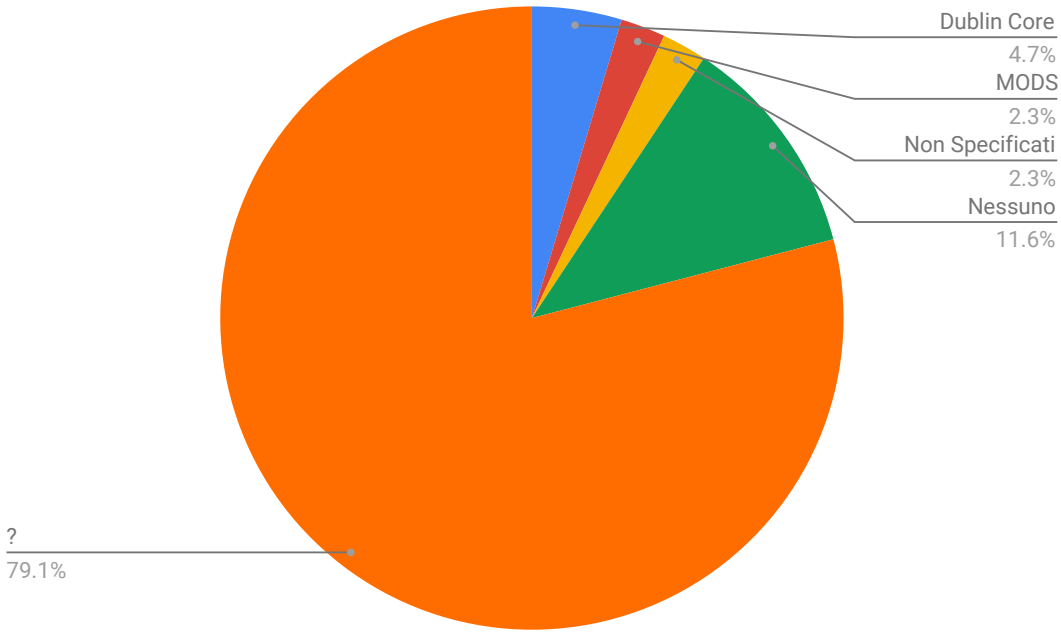

Figura 9: Distribuzione dei valori del campo *Standard metadati utilizzati*

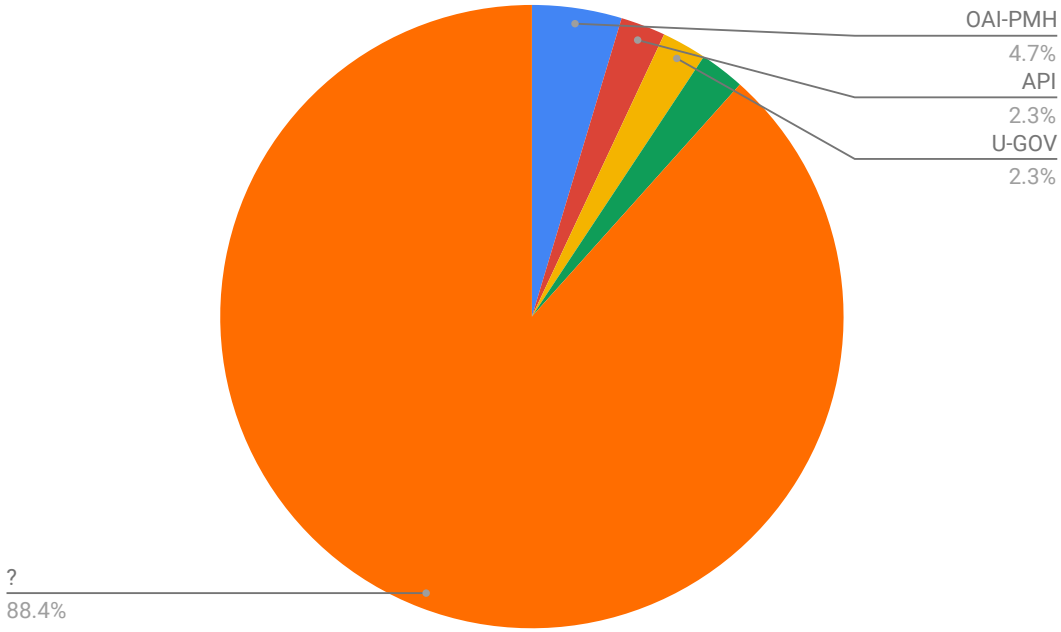

Figura 10: Distribuzione dei valori del campo *Interoperabilità*