

# ECODIGIT

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

## **D3.4 Proof-of-Concept**

Acronimo Progetto:

**EcoDigit**

Titolo Progetto:

**Ecosistema digitale per la fruizione  
e la valorizzazione dei beni e delle  
attività culturali della regione Lazio**

---

## D3.4

<b>Work Package:</b>	WP3 T3.3	
<b>Data di Sottomissione:</b>	2 Ottobre 2019	
<b>Inizio Progetto:</b>	2 Ottobre 2018	
<b>Durata Progetto:</b>	15 Mesi	
<b>Reponsabile Deliverable:</b>	Massimo Mecella (RM1)	
<b>Versione:</b>	v1.0	
<b>Stato:</b>	Finale	
<b>Autore:</b>	Luigi Asprino	ISTC-CNR
	Ludovica Marinucci	ISTC-CNR
	Andrea Giovanni Nuzzolese	ISTC-CNR
	Valentina Presutti	ISTC-CNR
	Massimo Mecella	RM1
	Miguel Ceriani	RM1
<b>Altri contribuenti al lavoro riportato nel deliverable:</b>	Marialuisa Mongelli	ENEA
	Massimo Mecella	RM1
	Marco Canciani	RM3
	Federico Meschini	UNITUS
<b>Reviewer:</b>	Reviewer	Affiliazione

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

Luigi Asprino, Ludovica Marinucci, Andrea Giovanni Nuzzolese, Valentina Presutti, Massimo Mecella e Miguel Ceriani. *D3.4 Proof-of-Concept*. Deliverable Progetto EcoDigit. 2019

## Revisioni

Versione	Data	Modificata da	Commento
0.1	1/9/2019	Luigi Asprino	Creazione Documento
0.2	6/9/2019	Luigi Asprino	Prima Bozza
0.3	14/9/2019	Ludovica Marinucci, Andrea Giovanni Nuzzolese e Valentina Presutti	Prima Revisione
0.4	25/9/2019	Massimo Mecella, Miguel Ceriani	Seconda Revisione

## Executive Summary

Le cinque università statali del Lazio in rete con CNR, ENEA e INFN si candidano a costituire il Centro di Eccellenza del Distretto Tecnologico per i beni e le attività Culturali (DTC) del Lazio. La mission del Centro è costituire un centro di aggregazione ed integrazione di competenze nel settore delle tecnologie per i beni e le attività culturali. In questo contesto, il progetto **EcoDigit-Ecosistema digitale per la fruizione e la valorizzazione dei beni e delle attività culturali del Lazio** ha l'obiettivo di arricchire il sistema Anagrafe delle Competenze del DTC con una piattaforma middleware che faciliti l'integrazione di nuove sorgenti di dati e consenta la pubblicazione e il riuso di servizi per la valorizzazione e la fruizione del patrimonio culturale del Lazio.

Il Work Package 3 del progetto si occupa: *(i)* di identificare le potenziali sorgenti (T3.1); *(ii)* di definire un modello di ingresso che le sorgenti dovranno rispettare per poter essere integrate e di selezionare degli strumenti metodologici e tecnologici che una sorgente può adottare per rendere i propri dati compatibili con il modello di ingresso (T3.2); *(iii)* di dimostrare la fattibilità dell'approccio di integrazione proposto (T3.3).

Il presente deliverable, frutto del lavoro del task T3.3, ha l'obiettivo di fornire una prova di concetto al fine di dimostrare come una sorgente, censita nel Task 3.1 possa essere resa compatibile con al modello di ingresso e integrata nel sistema usando gli strumenti definiti nel Task T3.2.

## Indice

<b>1</b>	<b>Introduzione</b>	<b>5</b>
1.1	Obiettivi Work Package . . . . .	5
1.2	Obiettivo del deliverable . . . . .	5
1.3	Relazione con le altre attività del progetto . . . . .	5
1.4	Outline documento . . . . .	5
<b>2</b>	<b>ArCo: Knowledge Graph del Patrimonio Culturale Italiano</b>	<b>7</b>
2.1	Integrazione di ArCo nella Digital Library . . . . .	7
<b>3</b>	<b>S&amp;TDL - Science and Technology Digital Library</b>	<b>13</b>
<b>4</b>	<b>Dati Inseriti provenienti da Applicativo Software</b>	<b>14</b>
<b>5</b>	<b>Conclusioni</b>	<b>18</b>

## 1 Introduzione

### 1.1 Obiettivi Work Package

Il Work Package 3 del progetto si occupa: *(i)* di identificare le potenziali sorgenti (T3.1); *(ii)* di definire un modello di ingresso che le sorgenti dovranno rispettare per poter essere integrate e di selezionare degli strumenti metodologici e tecnologici che una sorgente può adottare per rendere i propri dati compatibili con il modello di ingresso (T3.2); *(iii)* di dimostrare la fattibilità dell'approccio di integrazione proposto (T3.3).

### 1.2 Obiettivo del deliverable

Il presente deliverable, frutto di parte del lavoro del task T3.3, ha l'obiettivo di fornire una prova di concetto di come una sorgente, censita nel Task 3.1 possa essere resa compatibile con al modello di ingresso e integrata nel sistema usando gli strumenti definiti nel Task T3.2.

### 1.3 Relazione con le altre attività del progetto

**Influenza degli altri task del progetto sul lavoro descritto in questo documento.** Il lavoro descritto in questo documento si è avvalso dei risultati dei task: "T3.1 Censimento ed individuazione delle sorgenti potenziali", le sorgenti integrate nel sistema sono state identificate attraverso l'attività di censimento; "T3.2 Definizione del modello di integrazione di una sorgente", in quanto le sorgenti vengono rese conformi al modello di ingresso definito nel deliverable D3.2 [1] attraverso la metodologia e gli strumenti censiti nel deliverable D3.3 [2].

**Influenza del lavoro descritto in questo documento sugli altri task del progetto.** Grazie al lavoro descritto in questo documento, il "Task 4.4 Prototipo di servizio avanzato per la fruizione dei beni culturali nel dominio della formazione" beneficerà di una serie di contenuti raccolti e integrati seguendo il modello di integrazione definito nel task T3.2.

### 1.4 Outline documento

Le successive sezioni sono suddivise per sorgente, cioè dedicheremo una sezione differente per ogni sorgente integrata. Nella sezione verrà descritto come gli strumenti descritti nel deliverable D3.3 [2] sono stati implementati al fine di integrare la sorgente in oggetto. In particolare verrà mostrato il processo che ha portato all'integrazione di tre sorgenti eterogenee. La prima sorgente, ArCo (cf. Sezione 2) è un Knowledge Graph del patrimonio culturale

italiano, tramite uno SPARQL endpoint<sup>1</sup>. La seconda sorgente integrata, S&TDL (cf. Sezione 3) è un sistema informativo integrato in grado di gestire grandi volumi di documenti, dati e informazioni, volto allo sviluppo di iniziative di digital library, repository, sistemi di digital preservation, ecc. Le informazioni integrate da S&TDL sono disponibili in formato XML. La terza sorgente integrata è un applicativo software (cf. Sezione 4) che permette di raccogliere dati inseriti manualmente attraverso Google Form.

---

<sup>1</sup><http://dati.beniculturali.it/arco/>

## 2 ArCo: Knowledge Graph del Patrimonio Culturale Italiano

L'Istituto Centrale per il catalogo e la Documentazione (ICCD) del Ministero dei Beni e Delle Attività Culturali e del Turismo (MiBACT) in collaborazione con l'Istituto di Scienze e Tecnologie della Cognizione del Consiglio Nazionale delle Ricerche (ISTC-CNR) ha avviato nel 2017 il progetto ArCo (Architettura della Conoscenza). Il progetto, concluso nel 2019, aveva l'obiettivo di progettare una rete di ontologie per la strutturazione della conoscenza per i beni culturali e rendere disponibili i dati del Catalogo generale dei beni culturali secondo i paradigmi dei Linked Open Data (LOD).

Il progetto ha valorizzato il contesto dei beni culturali rendendo i dati accessibili, rintracciabili e riusabili da utenti e applicazioni. I dati sono stati strutturati sulla base di modelli ontologici che:

- rispecchiano l'analiticità strutturale delle schede ministeriali ICCD utilizzate per la descrizione dei beni culturali;
- restituiscono tutta la complessità dell'oggetto "bene culturale" descritto;
- valorizzano e rafforzano la ricchezza semantica del Catalogo e delle relazioni esplicite e implicite presenti nelle schede;
- assicurano interoperabilità attraverso operazioni di "allineamento semantico" tramite relazioni di equivalenza con altre ontologie specifiche del dominio beni culturali (tra le quali CIDOC-CRM ed EDM) ed altre ontologie e modelli concettuali consolidate o sviluppate anche in domini più specifici (es. Bibframe, Ric-O).

Il progetto ha prodotto un Knowledge Graph (accessibile tramite interfaccia SPARQL al link<sup>2</sup>) che il risultato della trasformazione in Linked Open Data di più di 800 mila schede catalografiche. Il Knowledge Graph risultate da questa trasformazione è contiene più di 169 milioni di triple. La struttura semantica del Knowledge Graph è tramite una rete ontologica contenente 7 moduli.

### 2.1 Integrazione di ArCo nella Digital Library

La rete ontologica a supporto di ArCo offre una caratterizzazione semantica molto ricca degli oggetti del patrimonio culturale. Il modello di ingresso del sistema (vedi Deliverable D3.2 [1]), invece, tratta gli oggetti in maniera molto più generica rispetto al modello ontologico di ArCo. Questo è dovuto al fatto che il sistema (middleware EcoDigit insieme con la Digital Library) deve poter fornire una vista comune di una più ampia varietà di tipologie di oggetti.

<sup>2</sup><http://dati.beniculturali.it/sparql>



Quindi, dal punto di vista concettuale la procedura di integrazione di ArCo nella Digital Library prevede i seguenti passi.

1. *Selezione.* Il primo passo consisterà nella selezione degli oggetti del patrimonio culturale di interesse per il caso d'uso del sistema. Dato che il sistema ha un focus sulle attività della regione Lazio verranno selezionati da ArCo tutti i beni storici o artistici presenti sul territorio della regione Lazio.
2. *Estrazione.* Per ogni oggetto selezionato bisognerà estrarre le informazioni di interesse per il sistema (e.g. nome, descrizione, scheda catalografica del bene etc.).
3. *Trasformazione.* I dati/metadati estratti da ArCo dovranno essere strutturati secondo la sintassi e semantica di ingresso attesa dal sistema (vedi Deliverable D3.2 [1] e D3.3 [2]).

Questa procedura è stata implementata ed eseguita tramite LOD-RH (vedi Deliverable D3.3 [2]), lo strumento sviluppato al fine da permettere l'estrazione di dati RDF accessibili tramite uno SPARQL endpoint. Nel resto della sezione documenteremo LOD-RH è stato configurato per integrare ArCo nella Digital Library.

**Selezione.** L'attività di selezione si occupa di selezionare dallo SPARQL endpoint le URI che identificano le risorse da importare. LOD-RH permette di selezionare delle risorse di interesse da uno SPARQL endpoint definendo una SELECT query (in linguaggio SPARQL). Questa query è mostrata nel riquadro 1. La query sfrutta la geolocalizzazione degli oggetti fornita da ArCo. I dati relativi alla geolocalizzazione degli oggetti sono modellati secondo la Location Ontology della rete ontologica di ArCo<sup>3</sup> e la Core Location Vocabulary di OntoPia<sup>4</sup>. La query quindi selezionerà tutte le risorse (i.e. ?resource) aventi un indirizzo (a-loc:hasCulturalPropertyAddress) identificato con ?cultAdd, tale che ?cultAdd è riferito alla regione Lazio, identificata con la URI regione:lazio. Questa query, eseguita sullo SPARQL endpoint di ArCo, seleziona 56214 risorse<sup>5</sup>.

```

PREFIX a-loc: <https://w3id.org/arco/ontology/location/>
PREFIX CLV: <https://w3id.org/italia/onto/CLV/>
PREFIX regione: <https://w3id.org/arco/resource/Region/>

SELECT DISTINCT ?resource WHERE {
  ?resource a-loc:hasCulturalPropertyAddress ?cultAdd .
  ?cultAdd CLV:hasRegion regione:lazio
}
    
```

**Riquadro 1:** Query usata da LOD-RH per selezionare da ArCo le risorse di interesse.

<sup>3</sup><https://w3id.org/arco/ontology/location>

<sup>4</sup><https://w3id.org/italia/onto/CLV>

<sup>5</sup>Dati relativi all'ultimo accesso eseguito in data 1 Ottobre 2019.

**Estrazione.** Il passo di estrazione si occupa di estrarre una serie di informazione di interesse per ognuna delle risorse selezionate nel passo di selezione. LOD-RH consente di estrarre dallo SPARQL endpoint delle informazioni di interesse definendo una CONSTRUCT query in linguaggio SPARQL. La query viene mostrata nel riquadro 2. La query permette di estrarre le informazioni relative alla catalogazione (e.g. nome, descrizione, identificativi) e alla geolocalizzazione del bene. Queste informazioni sono modellate in ArCo secondo le seguenti ontologie: ArCo Core<sup>6</sup>, ArCo Location<sup>7</sup>, ArCo Catalogue<sup>8</sup>, Core Location Vocabulary<sup>9</sup> e l0<sup>10</sup> di OntoPiA. LOD-RH eseguirà questa CONSTRUCT query una volta per ogni possibile ?resource restituita dalla query mostrata nel riquadro 1. Ad ogni esecuzione LOD-RH sostituirà (i.e. *binding* della variabile) le occorrenze della variabile ?base con i valori associati a ?resource. Quindi nel caso di ArCo sono eseguite 56214 CONSTRUCT (una per ogni risorsa restituita dalla SELECT). I modelli restituiti dalle CONSTRUCT sono salvati in file differenti serializzati secondo il formato RDF/XML.

```

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX arco: <https://w3id.org/arco/ontology/arco/>
PREFIX a-loc: <https://w3id.org/arco/ontology/location/>
PREFIX a-cat: <https://w3id.org/arco/ontology/catalogue/>
PREFIX CLV: <https://w3id.org/italia/onto/CLV/>
PREFIX l0: <https://w3id.org/italia/onto/l0/>

CONSTRUCT {
  ?base rdfs:label ?l .
  ?base arco:description ?d .
  ?base arco:uniqueIdentifier ?ud .
  ?base a-loc:hasCulturalPropertyAddress ?loc .
  ?base a-cat:isDescribedByCatalogueRecord ?cr .
  ?cr a-cat:systemRecordCode ?rc .
  ?loc CLV:hasCity ?city .
  ?city l0:name ?name .
  ?base a-loc:hasTimeIndexedTypedLocation ?tl .
  ?tl a-loc:atSite ?s .
  ?s l0:name ?tlname .
} WHERE {
  ?base rdfs:label ?l .
  OPTIONAL {
    ?base arco:description ?d .
  }
}

```

<sup>6</sup><https://w3id.org/arco/ontology/arco>

<sup>7</sup><https://w3id.org/arco/ontology/location>

<sup>8</sup><https://w3id.org/arco/ontology/catalogue>

<sup>9</sup><https://w3id.org/italia/onto/CLV>

<sup>10</sup><https://w3id.org/italia/onto/l0>

```
?base arco:uniqueIdentifier ?ud .
?base a-cat:isDescribedByCatalogueRecord ?cr .
?cr a-cat:systemRecordCode ?rc .
OPTIONAL{
  ?base a-loc:hasCulturalPropertyAddress ?loc .
  ?loc CLV:hasCity ?city .
  ?city l0:name ?name .
}
OPTIONAL{
  ?base a-loc:hasTimeIndexedTypedLocation ?tl .
  ?tl a-loc:atSite ?s .
  ?s l0:name ?tlname .
}
}
```

**Riquadro 2:** CONSTRUCT query usata da LOD-RH per estrarre da ArCo le informazioni di interesse.

**Trasformazione.** Il passo di trasformazione opera una ristrutturazione dei dati, trasformandoli secondo il modello di ingresso del sistema. Avendo usato LOD-RH come strumento per l'estrazione esistono due possibilità per effettuare questa trasformazione.

La prima soluzione sfrutta il meccanismo di estrazione di LOD-RH. Infatti, modificando la clausola CONSTRUCT della query mostrata nel Riquadro 2 è possibile attuare la ristrutturazione dei dati della sorgente contestualmente con l'estrazione. Nel Riquadro 3 viene mostrata la query CONSTRUCT modificata per effettuare la trasformazione contestualmente con all'estrazione dei dati. Ai fini di esempio viene mostrata la ristrutturazione solo di alcuni attributi della risorsa di interesse, cioè titolo, descrizione e identificativo.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX arco: <https://w3id.org/arco/ontology/arco/>
PREFIX a-loc: <https://w3id.org/arco/ontology/location/>
PREFIX a-cat: <https://w3id.org/arco/ontology/catalogue/>
PREFIX CLV: <https://w3id.org/italia/onto/CLV/>
PREFIX l0: <https://w3id.org/italia/onto/l0/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>

CONSTRUCT {
  ?base dc:title ?l .
  ?base dc:description ?d .
  ?base dc:identifier ?ud .
} WHERE {
  ?base rdfs:label ?l .
OPTIONAL {
```

```
?base arco:description ?d .
}
?base arco:uniqueIdentifier ?ud .
}
```

**Riquadro 3:** CONSTRUCT query usata da LOD-RH per estrarre da ArCo e trasformare secondo il modello di ingresso della Digital Library le informazioni di interesse.

La seconda soluzione per ristrutturare i dati richiede la definizione di una trasformazione XSLT. Questa seconda modalità permette di definire trasformazioni più modulari che potrebbero essere molto utili nel caso in cui la trasformazione da implementare risulta essere consistente sia nel numero di attributi da trasformare che nel tipo di operazioni di trasformazione da implementare (e.g. concatenazione di due o più attributi). Nel riquadro 4 viene mostrato un estratto della trasformazione applicata sui dati provenienti dallo SPARQL endpoint di ArCo. La trasformazione completa è disponibile online al seguente indirizzo<sup>11</sup>. La trasformazione XSLT mostrata nel riquadro 4 definisce un template (i.e. `xsl:template`) per le entità ottenute da ArCo, per ognuna di queste entità definisce un individuo del tipo `DUL:Object`, mantenendo la URI originale dell'oggetto che viene salvata nella variabile `{uriObject}`. Inoltre, per ogni entità, il template verifica che abbia una etichetta (espressa tramite il predicato `rdfs:label`) e in caso positivo la label viene trasformata nel `dc:title` dell'oggetto generato. Nel fare ciò il testo della label in input viene normalizzato.

```
<xsl:stylesheet version="1.0"
...
xmlns:xalan="http://xml.apache.org/xalan">
...
<xsl:template match="...">
  <xsl:variable name="uriObject" select="@rdf:about" />
  <rdf:RDF>
    <DUL:Object rdf:about="{uriObject}">
      ...
      <xsl:if test="rdfs:label">
        <dc:title>
          <xsl:text disable-output-escaping="yes">&lt;![CDATA[</xsl:text>
          <xsl:value-of select="normalize-space(rdfs:label/text())" />
          <xsl:text disable-output-escaping="yes">]]&gt;</xsl:text>
        </dc:title>
      </xsl:if>
      ...
    </DUL:Object>
  </rdf:RDF>
</xsl:template>
```

<sup>11</sup>[https://raw.githubusercontent.com/ecodigit/workspace/master/trasformazioni/arco\\_culturalproperty/transform.xml](https://raw.githubusercontent.com/ecodigit/workspace/master/trasformazioni/arco_culturalproperty/transform.xml)

```
...  
</xsl:stylesheet>
```

**Riquadro 4:** Estratto della trasformazione applicata sui dati provenienti dallo SPARQL endpoint di ArCo.

### 3 S&TDL - Science and Technology Digital Library

Al fine di valutare la capacità del modello di ingresso presentato nel Deliverable D3.2 [1] di rappresentare informazioni contenute in una delle sorgenti del D3.1 [3] si è proceduto alla trasformazione e integrazione dei dati contenuti *Science & Technology Digital Library* (S&TDL)<sup>12</sup>. S&TDL consiste in un sistema informativo integrato in grado di gestire grandi volumi di documenti, dati e informazioni, volto allo sviluppo di iniziative di digital library, repository, sistemi di digital preservation, ecc. Attualmente la S&TDL gestisce centinaia di migliaia di risorse informative, corrispondenti a milioni di oggetti digitali, suddivisibili in: (i) *prodotti della ricerca*, ovvero dataset e contenuti digitalizzati di interesse storico e culturale; (ii) *attività di ricerca*, ovvero progetti, istituzioni e attori operanti nei settori della ricerca e sviluppo (R&S) e dei beni archivistici e culturali. L'analisi di questo dataset, scelto per la sua varietà di contenuti, è stata possibile grazie alla collaborazione con il team del progetto *Anagrafe delle Competenze*<sup>13</sup> che ha reso disponibile il dettaglio degli schemi XML dei dati nella Repository della S&TDL. Le schede XML sono relative in particolare alle sezioni della S&TDL: Persone, Prodotti, Progetti, Organizzazioni e Dati della ricerca.

**Trasformazione.** Come detto i dati di S&TDL sono originariamente in formato XML formattati secondo uno schema interno alla Digital Library. Per integrarli nel sistema è stato necessario implementare delle trasformazioni XSLT per strutturare i dati di S&TDL secondo lo schema definito nel Deliverable D3.2 [1]. Queste trasformazioni sono state implementate dal team di sviluppo del progetto Anagrafe sotto le indicazioni del team del progetto EcoDigit. In particolare, l'interazione tra i due team per la definizione delle trasformazioni è stata la seguente. Un'analisi congiunta dei due team agli schemi della S&TDL ha portato: (i) alla definizione di una serie di requisiti che sono stati implementati dal team di EcoDigit nella prima versione modello di ingresso (documentato nella prima versione del Deliverable D3.2 [1]). (ii) alla mappatura (definita dal team EcoDigit) dei campi degli schemi XML della S&TDL nello modello di ingresso (anche questa attività è documentata nel Deliverable D3.2 [1]); In seguito il team di Anagrafe sulla base delle indicazioni derivanti dalla precedente analisi ha provveduto alla realizzazione della mappatura tra i campi degli schema XML della S&TDL nello modello di ingresso. Questa mappatura è stata implementata in linguaggio XSLT.

---

<sup>12</sup><https://dl.cnr.it/>

<sup>13</sup><https://www.dtc lazio.it/anagrafe-delle-competenze>

## 4 Dati Inseriti provenienti da Applicativo Software

Al fine di permettere l'inserimento manuale di dati non disponibili in formato strutturato sono stati sviluppati una serie di form usando l'applicativo Google Form. Questi form permettono l'arricchimento dei dati del sistema con dati forniti manualmente da utenti. In particolare, sono stati sviluppati i seguenti form:

- (i) *Form Persone*. Questo form, disponibile online all'indirizzo<sup>14</sup>, permette l'inserimento di dati relativi a persone non ancora censite dal Centro di Eccellenza. Oltre ai dati anagrafici della persona, il form permette di inserire: l'organizzazione a cui la persona afferisce, attività e interessi di ricerca, pubblicazioni della persona (che possono essere fornite secondo il formato definito dal sistema IRIS), e i progetti ai quali la persona partecipa o ha partecipato.
- (ii) *Form Organizzazioni*. Il form Organizzazioni, disponibile online all'indirizzo<sup>15</sup>, permette l'inserimento di dati relativi ad organizzazioni non ancora censite dal Centro di Eccellenza. Il form permette di inserire i dati relativi all'organizzazione quali: nome, acronimo, indirizzo, descrizione, numero di telefono, fax, email, sito web e i membri dell'organizzazione.
- (iii) *Form Oggetti*. Il form Oggetti, disponibile online all'indirizzo<sup>16</sup>, permette di inserire (meta)dati relativi ad oggetti del dominio dei beni culturali. Il form permette di inserire le seguenti informazioni relative all'oggetto: nome, descrizione, localizzazione, identificativo, autori dell'oggetto, le pubblicazioni relative all'oggetto, l'organizzazione che detiene i diritti dell'oggetto. Il form, inoltre, permette di fornire una classificazione dell'oggetto che censito rispetto alle categorie per il prototipo presentate nel Deliverable D3.2 [1].

**Trasformazione dei Risultati Secondo il Modello di Ingresso.** Un applicativo software è stato sviluppato al fine di garantire che i dati raccolti tramite i form sopraelencati vengano esportati da Google Form, strutturati secondo il modello di ingresso definito nel Deliverable D3.2 [1], e integrati nel sistema. Questo software è stato sviluppato in linguaggio Java ed è disponibile online sul repository GitHub del progetto<sup>17</sup>.

Google Form fornisce sotto forma di Spreadsheet i dati raccolti tramite i form. Quindi, per ottenere i dati, l'applicazione usa le Sheets API di Google<sup>18</sup>. Dato l'identificativo di uno Spreadsheet Google le API permettono di accedere al contenuto delle sue righe.

<sup>14</sup>[https://docs.google.com/forms/d/e/1FAIpQLScg8\\_fUzv9qEKYL5Mewa0CIIdNE0QvI-zCfMgBm0zNyJgPuJow/viewform](https://docs.google.com/forms/d/e/1FAIpQLScg8_fUzv9qEKYL5Mewa0CIIdNE0QvI-zCfMgBm0zNyJgPuJow/viewform)

<sup>15</sup>[https://docs.google.com/forms/d/e/1FAIpQLSf3hRxcwjzMkmmS4gRlzmNUFc5CjGmtLg3LuLHcLo\\_TM0Gspw/viewform](https://docs.google.com/forms/d/e/1FAIpQLSf3hRxcwjzMkmmS4gRlzmNUFc5CjGmtLg3LuLHcLo_TM0Gspw/viewform)

<sup>16</sup>[https://docs.google.com/forms/d/e/1FAIpQLSfzrFn2AVmhSBpQV06w5-RU5WIw0vbbq\\_HM3-DWRo1RubsL5A/viewform](https://docs.google.com/forms/d/e/1FAIpQLSfzrFn2AVmhSBpQV06w5-RU5WIw0vbbq_HM3-DWRo1RubsL5A/viewform)

<sup>17</sup><https://github.com/ecodigit/workspace>

<sup>18</sup><https://developers.google.com/sheets/api>

Per implementare la trasformazione dei dati è stata usata la tecnica del templating (cf. Deliverable D3.3 [2]). Il *templating* è una tecnica che permette di popolare un file (nel nostro caso in formato RDF/XML) con dati provenienti da un applicativo software. Il funzionamento di questi sistemi può essere descritto come segue.

Viene realizzato un documento di template avente la struttura sintattica e semantica desiderata (nel nostro caso quella attesa dalla Digital Library). Per definire i template dell'applicazione è stato usato il linguaggio FTL. Nel documento di template nei punti che dovrebbero contenere i valori reali legati alla risorsa descritta dal file si inseriscono delle variabili. Nel repository GitHub<sup>19</sup> sono disponibili i template usati per realizzare la trasformazione dei dati provenienti dai form. Il riquadro 5 mostra, a scopo illustrativo, un estratto del template usato per trasformare i dati raccolti dal form Persone.

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
...
xmlns:geo_wgs84="http://www.w3.org/2003/01/geo/wgs84_pos#">
  <foaf:Person rdf:about="{person.URI}">
    <foaf:givenname>{person.givenName}</foaf:givenname>
    <foaf:family_name>{person.familyName}</foaf:family_name>
    <#if person.img??>
    <foaf:img rdf:resource="{person.img}" />
    </#if>
    ...
  </foaf:Person>
</rdf:RDF>
```

**Riquadro 5:** Estratto del template usato per trasformare secondo il modello di ingresso i dati raccolti dal form Persone.

Nel riquadro viene definito il template per la specifica di una entità di tipo persona (i.e. foaf:Person) identificata dalla URI corrispondente alla variabile `{person.URI}`. Nome e Cognome della persona sono forniti dalle variabili `{person.givenName}` e `{person.familyName}`, la URL all'immagine profilo della persona è rappresentata dalla variabile `{person.img}`. Mentre nome e cognome della persona sono attributi obbligatori, l'immagine profilo è un attributo opzionale. Per gli attributi opzionali, il linguaggio FTL permette di definire una clausola `#if`, ad esempio, nel caso di dell'immagine profilo, la persona sarà associata ad una immagine (specificata tramite foaf:img) solo se la variabile `{person.img}` è associata ad un qualsiasi valore.

Il nome delle variabili del template richiama il nome delle variabili del modello usato dall'applicativo software (e.g. nel caso in cui l'applicazione sia realizzata in Java le variabili vengono nominate secondo gli attributi delle classi Java che contengono i dati di interesse). Nel riquadro 6 viene mostrato un estratto della classe Java che si occupa di popolare il template mostrato nel riquadro 5.

<sup>19</sup><https://github.com/ecodigit/workspace/tree/master/templates>



```
public class Person extends Agent {

    private String givenName, familyName, img;

    public String getGivenName() {
        return givenName;
    }

    public void setGivenName(String givenName) {
        this.givenName = givenName;
    }

    public String getFamilyName() {
        return familyName;
    }

    public void setFamilyName(String familyName) {
        this.familyName = familyName;
    }

    public String getImg() {
        return img;
    }

    public void setImg(String img) {
        this.img = img;
    }
    ...
}
```

**Riquadro 6:** Estratto della classe Java usata per memorizzare i dati della Persona.

Template e oggetto software contenente i dati vengono passati ad un template engine che si occupa di popolare il template con i dati mantenuti nell'oggetto software. Come template engine è stato usato Apache FreeMarker<sup>20</sup>. Un esempio di codice che implementa questa trasformazione è mostrato nel riquadro 7. Nello snippet di codice viene creato un nuovo oggetto `p` di tipo `Persona`. Di questa persona viene fornito nome, cognome e immagine profilo dalle variabili `spreadSheetGivenName`, `spreadSheetFamilyName` e `spreadSheetImg`, popolate con dati ottenuti dal Google Form. A questo punto l'oggetto `p` viene passato al template engine di Apache FreeMarker (i.e. oggetto `temp`) che si occupa di processarlo sostituendo alle variabili del template `person.ftlh` i valori delle variabili dell'oggetto `p`. Il risultato ottenuto viene a questo punto scritto dal file writer `out`.

<sup>20</sup><https://freemarker.apache.org/>

```
...  
Person p = new Person();  
p.setGivenName(spreadSheetGivenName);  
p.setFamilyName(spreadSheetFamilyName);  
p.setImg(spreadSheetImg);  
Map<String, Object> root = new HashMap<>();  
root.put("person", p);  
Template temp = TransformerConfiguration.getInstance()  
    .getFreemarkerCfg().getTemplate("person.ftlh");  
FileWriter out = new FileWriter(new File(...));  
temp.process(root, out);  
...
```

**Riquadro 7:** Esempio di codice Java che effettua la trasformazione dei dati di una Persona censita tramite il form.

## 5 Conclusioni

In questo documento è stata fornita una prova di concetto al fine di dimostrare come delle sorgenti dati, anche se non originariamente definite secondo il modello di ingresso definito nel Deliverable D3.2 [1], possano diventarlo usando gli strumenti elencati dal Deliverable D3.3[2].

In particolare è stato descritto il processo che ha portato all'integrazione di tre sorgenti eterogenee. La prima sorgente, ArCo (cf. Sezione 2) è un Knowledge Graph del patrimonio culturale italiano, tramite uno SPARQL endpoint<sup>21</sup>. La seconda sorgente integrata, S&TDL (cf. Sezione 3) è un sistema informativo integrato in grado di gestire grandi volumi di documenti, dati e informazioni, volto allo sviluppo di iniziative di digital library, repository, sistemi di digital preservation, ecc. Le informazioni integrate da S&TDL sono disponibili in formato XML. La terza sorgente integrata è un applicativo software (cf. Sezione 4) che permette di raccogliere dati inseriti manualmente attraverso Google Form.

---

<sup>21</sup><http://dati.beniculturali.it/arco/>

## Riferimenti bibliografici

- [1] Luigi Asprino, Ludovica Marinucci, Andrea Giovanni Nuzzolese e Valentina Presutti. *D3.2 Modello di ingresso*. Deliverable Progetto EcoDigit. 2019.
- [2] Luigi Asprino, Ludovica Marinucci, Andrea Giovanni Nuzzolese e Valentina Presutti. *D3.3 Studio sugli strumenti di supporto*. Deliverable Progetto EcoDigit. 2019.
- [3] Miguel Ceriani e Massimo Mecella. *D3.1 Report sul Censimento*. Deliverable Progetto EcoDigit. 2019.