Un catalogo per la descrizione di risorse archeologiche

Nicola Aloia, Franca Debole, Carlo Meghini

Istituto di Scienza e Tecnologie dell'Informazione del CNR, Pisa, Italia
{nicola.aloia,franca.debole,carlo.meghino}@isti.cnr.it

Abstract. This paper discusses the registry developed by the ARIADNE project for describing the archaeological resources that are made available by the partners of the project for the purposes of discovery, access and integration on a research infrastructure. These resources include: data, services and language resources, such as metadata formats, vocabularies and mappings. The registry is addressed to cultural institutions, private or public, which wish to describe their assets in order to make them known to e-infrastructures.

Keywords: catalogue, registry, archaeological resources

1 Introduzione

Con il rapido progredire delle tecnologie informatiche e della comunicazione, si assiste a una diffusione sempre più capillare di strumenti automatici a supporto del lavoro del ricercatore. Le comunità scientifiche hanno tutte a disposizione un'ampia gamma di software che, in applicazioni isolate o in servizi globalmente offerti, si fanno carico di acquisire, gestire, elaborare e visualizzare dati di tutti i tipi per una varietà di scopi. Questa ricchezza di offerta, potenzialmente positiva, rischia però di diventare un ostacolo alla creazione di uno spazio comune della conoscenza: se colta in contesti non comunicanti, essa può creare l'effetto opposto, cioè frammentazione della conoscenza. Per evitare un esito tanto negativo, la comunità informatica propone da tempo il ricorso alle infrastrutture informatiche, come tecnologia in grado di dare supporto al lavoro dello scienziato mettendolo al tempo stesso al centro di una comunità che condivide conoscenza, dati e strumenti, con ciò mettendo riparo alla frammentazione esistente e prevenendo che se ne generi di nuova. Il progetto europeo ARIADNE nasce in questo contesto, e si propone di creare un'infrastruttura a supporto del lavoro del ricercatore in archeologia, rendendo disponibili i dati e i servizi fin qui sviluppati nelle singole comunità nazionali o settoriali, e integrando tali dati e servizi laddove possibile per creare una base di conoscenza condivisa e universalmente accessibile.

Nel contesto di ARIADNE, come pure in quello dei molti progetti paralleli finanziati dalla Commissione Europea nel settore delle infrastrutture, assume particolare rilevanza il catalogo (*catalog*), o registro (*registry*), delle risorse

dell'infrastruttura. Tale catalogo ha lo scopo di descrivere i dati e i servizi che formano lo spazio informativo dell'infrastruttura, al fine di dare supporto a operazioni fondamentali per l'accesso alla conoscenza, quali la scoperta (*discovery*) dei dati e dei servizi, la loro visualizzazione in esplorazione navigazionale (*browsing*) e il loro accesso.

Il presente lavoro dà conto della struttura informativa del catalogo di ARIADNE, presentando il modello dei dati sottostante, e di questo indicando le principali classi e proprietà, e le rispettive tassonomie. Viene fatto anche un breve cenno all'implementazione del modello che il progetto ARIADNE sta attualmente realizzando, con il contributo di partner scientifici e tecnologici europei.

2 Il modello dei dati del catalogo di ARIADNE

Come accennato nell'introduzione, il progetto ARIADNE ha l'obiettivo di integrare le varie infrastrutture di dati di ricerca archeologica esistenti, per fornire agli studiosi, nella loro pratica di ricerca, potenti strumenti, tecnologicamente avanzati, ai fini della fruizione dei dati e dei servizi che la comunità rende disponibili. Integrare la grande quantità di dati e di strumenti tecnologici esistenti esige di scoprirne le caratteristiche: a questo scopo nasce il catalogo di ARIADNE, che censisce e descrive quanto è disponibile tra i partner del progetto, e più in generale dell'intera comunità degli archeologi, per individuare, tramite raffinati meccanismi di ricerca, le risorse candidate per l'integrazione. In questo paragrafo presentiamo il modello dei dati su cui si basa il catalogo di ARIADNE. Tale modello, che noi chiamiamo ACDM (ARIADNE Catalogue Data Model), descrive le risorse disponibili tra i vari partner del progetto. In Fig. 1 è mostrato un diagramma UML semplificato di ACDM, che comprende le sue classi e associazioni più rilevanti. La definizione del modello e conseguentemente l'implementazione degli strumenti per il suo utilizzo sono state oggetto di frequenti aggiornamenti e revisioni, in seguito ai nuovi requisiti e alle conoscenze acquisite durante lo sviluppo del progetto stesso. Per una descrizione dettagliata e aggiornata di ACDM si rimanda alla documentazione ufficiale presente sul sito del progetto (www.ariadne-infrastructure.eu).

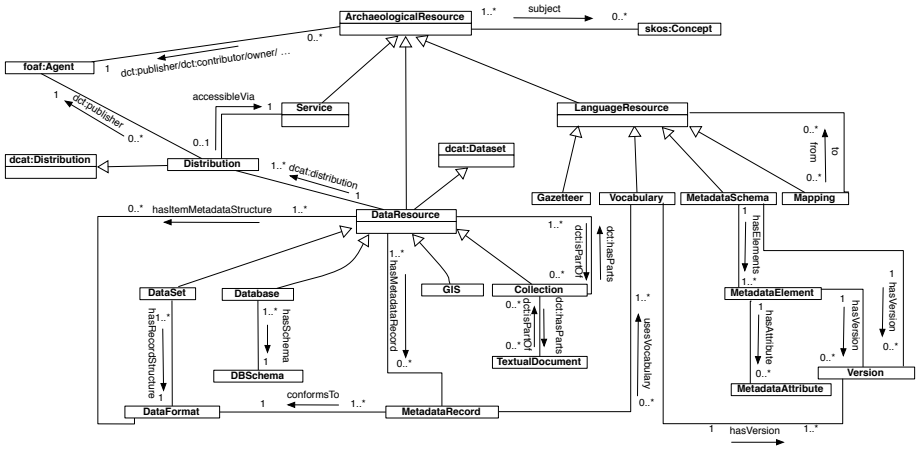

Fig. 1: Diagramma UML di ACDM

Come illustrato in Fig. 1, le risorse archeologiche in ARIADNE, modellate come sottoclassi di *ArchaeologicalResource,* sono di tre tipi differenti:

> **Risorse dati**, modellate con *DataResource*: rappresentano contenitori di dati come dataset, databases, GIS e collezioni.
>
> **Risorse linguistiche**, modellate con *LanguageResource*: comprendono vocabolari, schemi di metadati, gazzeteer e definizioni di mapping tra schemi di metadati.
>
> **Servizi**, modellati con *Service*: le sue istanze rappresentano i vari servizi disponibili presso i partners di ARIADNE.

DataResource, LanguageResource e *Service* sono classi astratte utili a definire proprietà e associazioni comuni alle loro sottoclassi ed ereditano le associazioni definite in *ArchaeologicalResource.*

Modellare in maniera esaustiva questi tipi di risorse è un compito complesso, per cui abbiamo focalizzato la nostra attenzione sugli obiettivi del progetto, riutilizzando, quanto più possibile, standard esistenti, consapevoli dei requisiti d'interoperabilità con altre realtà presenti nell'area in cui s'inquadra il progetto ARIADNE. In particolare per descrivere *DataResource* abbiamo esteso il vocabolario DCAT (Data Catalog Vocabulary) [1], per *LanguageResource* abbiamo usato la definizione ISO/IEC (AISO AiISI) 11179 "*Specification and Standardization of Data Elements*" [2], per *Service* abbiamo esteso il vocabolario con cui *Dbpedia* descrive il software.

2.1 Il vocabolario DCAT

DCAT è un vocabolario RDF, pubblicato dal Government Linked Data Working Group del W3C come raccomandazione per descrivere datasets e cataloghi sul Web, al fine di consentire la loro reperibilità e utilizzo. Nelle dichiarazioni degli autori il modello DCAT "è particolarmente adatto a rappresentare cataloghi di dati di varie amministrazioni, come ad esempio Data.gov e data.gov.uk" ed è stato proposto come uno strumento per la pubblicazione di datasets in modalità Open Data. Attualmente vari datasets sono stati pubblicati secondo le specifiche DCAT e vari progetti europei ne raccomandano ufficialmente l'adozione. L'adozione di DCAT in ARIADNE perciò ci pone nella situazione ideale per pubblicare i dati del progetto anche come Open Data.

DCAT utilizza un numero di classi e di relazioni provenienti da altri vocabolari ben noti come *foaf:Agent*, *skos:Concept*, *Dublin Core*. Le principali classi del modello sono *dcat:Catalog* che rappresenta una raccolta curata di metadati relativi ai dataset, *dcat:Dataset* che rappresenta una collezione curata di dati e *dcat:Distribution* che rappresenta la disponibilità dei vari dataset in differenti formati. In Fig. 2 è mostrato il diagramma delle classi di DCAT.

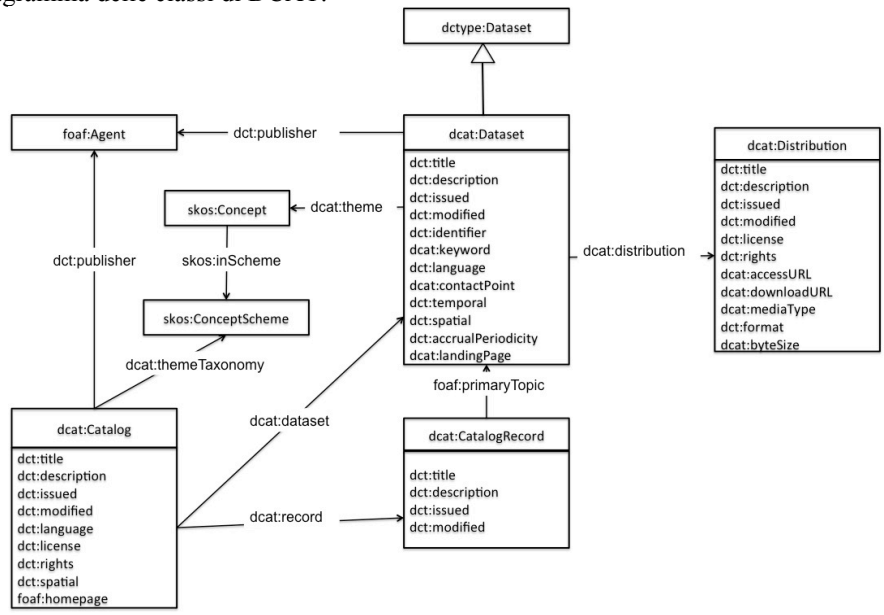

Fig. 2: Il modello DCAT

3 Le classi principali di ACDM

In questo paragrafo descriviamo brevemente alcuni dettagli delle principali classi di ACDM.

3.1 ArchaeologicalResource

Questa classe definisce le proprietà comuni alle sue sottoclassi, in gran parte utilizzando i termini del vocabolario DCAT cui si aggiungono proprietà per esprimere le politiche di accesso e gli identificatori originali delle risorse. *ArchaeologicalResource* ha come sottoclassi:

- *DataResource*
- *LanguageResource*
- *Service*

Le principali associazioni definite sono:

- *dct:publisher:* associa un'istanza di *ArchaeologicalResource* con l'istanza della classe *foaf:Agent* relativa all'ente che ha reso pubblicamente disponibile la risorsa;
- *dct:creator:* associa un'istanza di *ArchaeologicalResource* con l'istanza della classe *foaf:Agent* relativa al principale responsabile della creazione della risorsa;
- *owner:* associa un'istanza di *ArchaeologicalResource* con l'istanza della classe *foaf:Agent* relativa al proprietario legale della risorsa;
- *legalResponsible:* associa un'istanza di *ArchaeologicalResource* con l'istanza della classe *foaf:Agent* relativa al responsabile legale della risorsa;
- *scientificResponsible:* associa un'istanza di *ArchaeologicalResource* con l'istanza della classe *foaf:Agent* relativa al responsabile scientifico della risorsa;
- *dct:subject* associa un'istanza di *ArchaeologicalResource* con un soggetto presente in un determinato vocabolario, modellato con la classe *skos:Concept*.

3.2 DataResource

Questa classe specializza *ArchaeologicalResource* e descrive le risorse archeologiche che sono contenitori di dati. È una classe astratta che definisce le proprietà e le associazioni comuni alle sue sottoclassi (*DataSet*, *Databases*, *GIS*, *Collection*, vedi Fig. 1). Tra le varie proprietà di *DataResource*, che principalmente usa i termini del vocabolario DCAT, segnaliamo *dct:temporal* e *dct:spatial*, che forniscono informazioni spazio temporali sulla risorsa. Le associazioni principali che hanno questa classe come dominio sono:

- *dct:isPartOf:* nel caso la risorsa faccia parte di una collezione, questa proprietà associa la risorsa all'istanza di collezione cui appartiene;
- *dcat:distribution*: associa la risorsa con una o più istanze della classe *Distribution* (cioè con i vari formati accessibili della risorsa);
- *hasItemMetadataStructure*: associa la risorsa col formato dei metadati dei membri della risorsa (es. i metadati di ogni record di un dataset, o i metadati di ogni elemento di una collezione);
- *hasMetadataRecord*: associa la risorsa con i suoi metadati (es. i metadati che descrivono una collezione).

DataResource ha le seguenti sottoclassi:
Collection: definiamo una collezione archeologica come un'aggregazione di risorse, dette *elementi* della collezione. Gli elementi di una collezione sono singoli oggetti (ad esempio, immagini, testi, video, etc.) o istanze di *DataResource*, (cioè istanze delle sue sottoclassi *DataSet*, *Databases*, *GIS* o *Collection*); per esempio, una collezione può includere un documento di testo, un insieme di immagini, uno o più dataset e altre collezioni. Per ragioni d'interoperabilità *Collection* è una sottoclasse di *dcmitype:Collection*. La principale associazione che ha questa classe come dominio è *dct:hasParts*, che associa un'istanza di collezione con le sue componenti.

Database: le istanze di questa classe sono un insieme di record omogeneamente strutturati, gestiti tramite un sistema di gestione di database, come ad esempio MySQL. La principale associazione di questa classe è *hasSchema*, che associa un'istanza di database con la sua definizione strutturale rappresentata da un'istanza della classe *DBSchema*.

Dataset: un dataset archeologico è definito come un insieme di record omogeneamente strutturati, costituiti da campi che contengono valori. L'associazione principale che ha questa classe come dominio è *hasRecordStructure*, che associa un dataset con la sua definizione strutturale rappresentata da un'istanza della classe *DataFormat*.

GIS: questa classe ha come istanze record di dati gestiti da un Geographical Information Systems (GISs).

3.3 LanguageResource

Questa è la classe di tutte le risorse linguistiche descritte nel catalogo a fini di riutilizzo o integrazione all'interno della comunità di ARIADNE. Un'istanza di *LanguageResource* è una risorsa di natura linguistica, sia in linguaggio naturale (ad esempio un *gazzetter*) o in un linguaggio formale (ad esempio un vocabolario o uno schema di metadati). *LanguageResource* comprende anche definizioni di *mapping*, intese come associazioni tra due espressioni di risorse linguistiche, definizioni che possono essere di tipo formale o informale.

3.4 Service

Mentre per la descrizione dei *Dataset* è stato possibile adottare un vocabolario standard (DCAT), per quanto riguarda la descrizione dei servizi da censire in ARIADNE, la situazione è più complessa, poiché esistono diversi vocabolari, nessuno dei quali si è affermato come standard. Sulla base di evidenze raccolte tramite i rapporti del progetto, abbiamo classificato i servizi nelle seguenti categorie, che riflettono il modo in cui un servizio è accessibile (Fig. 3):

- ***StandAloneService:*** servizi che possono essere scaricati e installati su una macchina.
- ***WebService***: servizi accessibili sul Web tramite un'API.
- ***ServiceForHuman:*** servizi accessibili sul Web solo tramite una GUI.
- ***InstitutionalService***: servizi offerti da istituzioni, il cui accesso deve essere negoziato attraverso un'interazione personale con i rappresentanti di questa istituzione.

In ACDM abbiamo introdotto la classe astratta *Service* che descrive le proprietà e le associazioni comuni a tutti i servizi. *Service* è una sottoclasse di *ArchaeologicalResource*, per cui eredita tutte le proprietà e le associazioni di questa classe (Fig. 1).

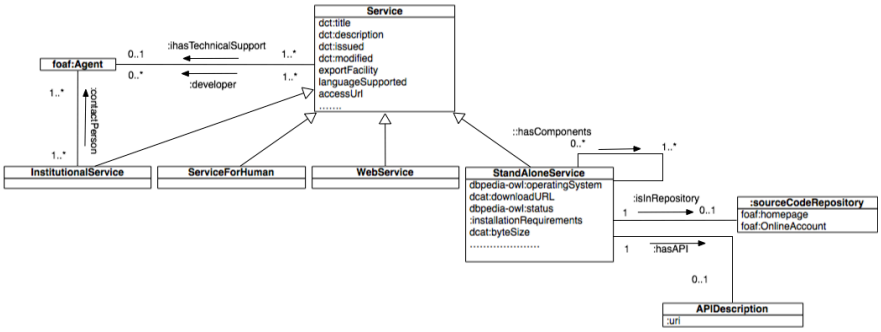

Fig. 3 Digramma UML dei Servizi

4 Strumenti per la gestione del catalogo

Il modello dei dati descritti nei paragrafi precedenti è la base su cui è stato implementato il servizio di *Registro* del progetto ARIADNE (da ora in avanti ci riferiremo al servizio solamente col nome Registro). Il Registro è il componente software, che utilizza una base di dati SQL, in cui sono memorizzati e resi disponibili tutte le risorse del catalogo, consentendo agli utenti di creare/modificare record attraverso un'API REST o attraverso una GUI Web. In Fig. 4 è mostrato il workflow

per l'acquisizione dei dati, realizzato tramite l'infrastruttura MoRe (Metadata & Object Repository) fornita dai partner di ARIADNE presso il Digital Curation Unit di Athena Research Centre (http://www.dcu.gr/).

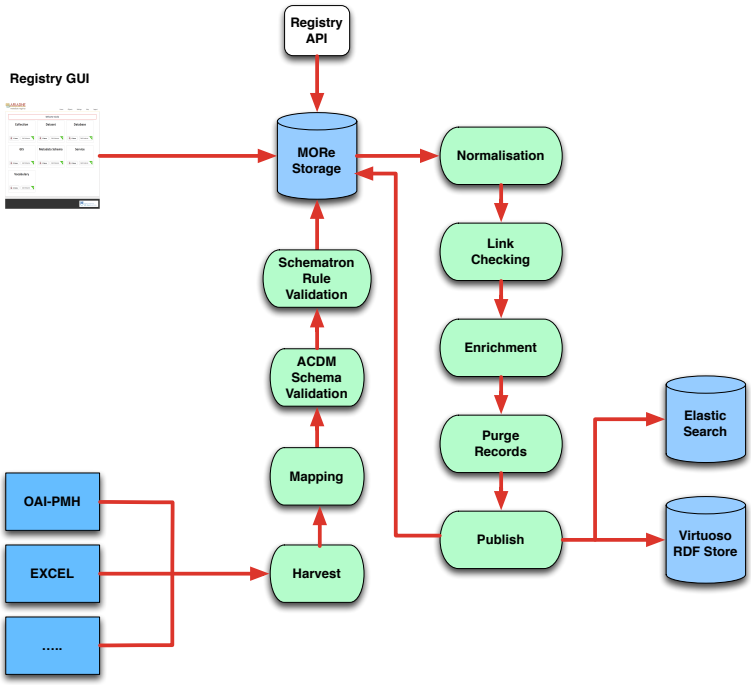

Fig. 4 Workflow per l'acquisizione dei dati del registro.

Come mostrato in Fig. 4, i record sono memorizzati nell'infrastruttura MoRe, tramite un'interfaccia grafica basata sul Web (Fig. 5), tramite chiamate alle API Rest di MoRe o importati direttamente da file esterni e da server OAI-PMH. I record importati in maniera batch sono sottoposti ad una fase di mapping dal modello esterno e validati tramite l'XML Schema di ACDM. In una fase successiva, i record consolidati nello storage di MoRe sono trasformati in RDF e memorizzati in un gestore di triple (Virtuoso nel nostro caso); un'altra trasformazione consente di memorizzare i record del catalogo in un sistema di Information Retrieval (Elastic Search nel nostro caso). Prima di essere pubblicati in Virtuoso o in Elastic Search i record ACDM passano attraverso le fasi di normalizzazione, link checking e enrichment.

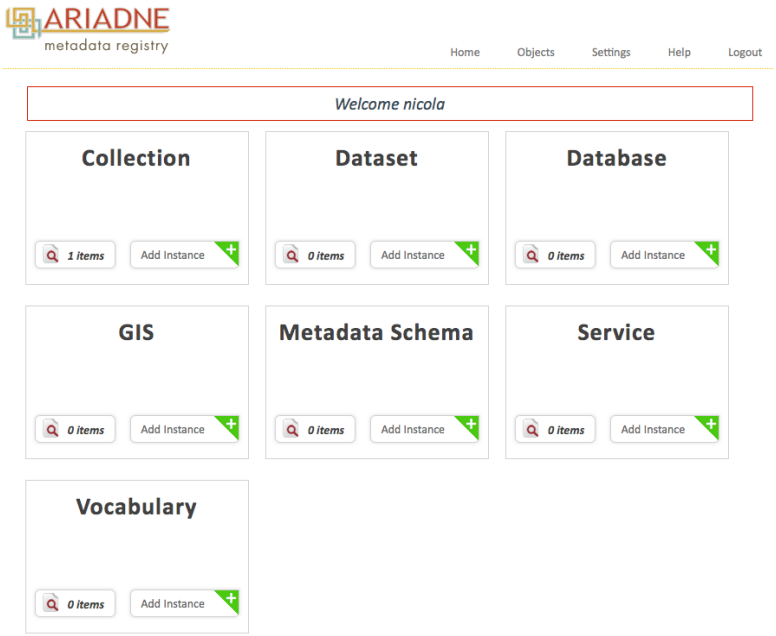

Fig. 5 Interfaccia grafica basata sul Web

5 Conclusioni

Il catalogo di ARIADNE è in fase avanzata di sviluppo ed è attualmente popolato con le descrizioni di decine di migliaia di risorse dati, fornite dalle più importanti istituzioni archeologiche europee, che si sono offerte di collaborare alla fase di messa a punto del catalogo stesso. Per la fine del progetto, il catalogo offrirà una base informativa stabile, accessibile via web, e fornita delle operazioni di scoperta, navigazione e accesso cui si è accennato sopra.

Da quel momento, il catalogo sarà messo a disposizione dell'intera comunità scientifica dell'archeologia, che potrà utilizzarlo come punto unico di accesso al patrimonio delle conoscenze del settore.

Ringraziamenti

Ringraziamo il progetto ARIADNE, finanziato dalla Commissione Europea nell'ambito del Settimo Programma Quadro, contratto n. FP7-INFRASTRUCTURES-2012-1-313193.

Riferimenti

1. DCAT http://www.w3.org/TR/vocab-dcat/
2. ISO 11179 Part1 Framework for the Specification and Standardization of Data Elements (2004)
3. MoRe (http://more.dcu.gr/).