한국 문화유산 정보의 시맨틱 구조 조성을 위한 연구: CIDOC CRM 적용 목적으로 기존 포탈 정보의 기반 조성을 위하여

Constructing a Foundation for Semantic Structure of Korean Heritage Information: A Study on Creating a Substructure of Korean Heritage Portal by Implementing CIDOC CRM

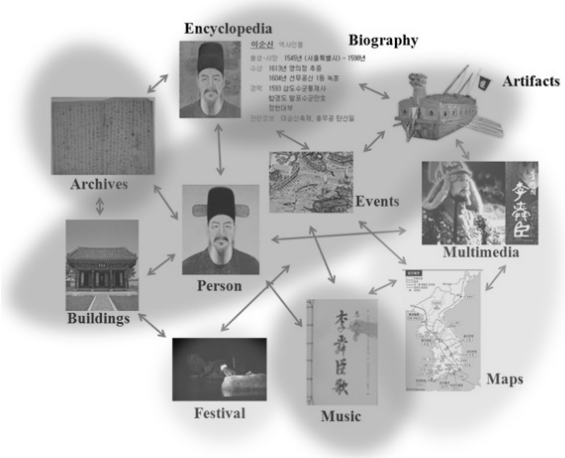
차소영, 한국과학기술원 문화기술대학원, astro1307@kaist.ac.kr
김정화, 한국과학기술원 문화기술대학원, jwkim21@kaist.ac.kr
Soyoung CHA, Graduate School of Culture Technology, KAIST
Jung Wha KIM, Graduate School of Culture Technology, KAIST

한국의 문화유산정보 포탈을 통합하여 시맨틱 웹을 조성하기 위한 기반 구축을 목표로 한다. 세계박물관협의회가 문화유산정보의 시맨틱 활용을 위해 개발한 CIDOC CRM 온톨로지를 적용, 시맨틱 포탈구축에 필요한 사전 조건 및 환경 조사를 내용으로 한다.

1. 서론

근래에 있어 웹은 없어서는 안 될 도구이다. 웹상에는 방대한 양의 정보가 있고 이들은 구조적으로 연결이 되어 있다 [1]. 하지만 내용면으로 보았을 때, 이러한 연결고리는 구조적인 것만큼 탄탄하지 못하다. 이러한 정보의 의미연결은 문화유산 분야를 대할 때 더 의미를 갖게 된다. 문화유산 정보는 단순히 정보로만 존재하는 것이 아니라 정보의 내용이 서로 복잡하게 얽혀있기 때문에 이 연구의 당위성을 제공한다.

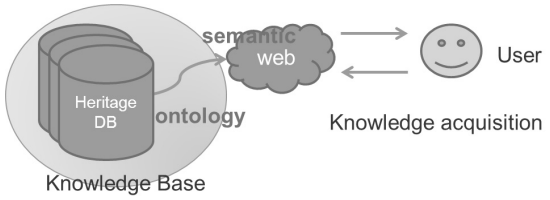
시맨틱 웹(Semantic Web)은 이러한 정보를 처리하기에 훌륭한 방법 중 하나이다. 시맨틱 웹은 서로 다른(heterogeneous) 정보들을 연결시키기 때문이다. 기존의 웹은 인간에 의해 정보가 처리되었지만, 시맨틱 웹은 컴퓨터 프로그램으로 웹상의 정보를 처리할 수 있다 [2]. 이러한 시맨틱 웹은 온톨로지로 구성된



<그림 1> 여러 다른 종류들로 이루어진, 그러나 의미적으로 서로 연관된 문화유산 정보구조 예시 - 임진왜란

다. CIDOC CRM은 문화유산 분야를 기술하기 위해 특별히 고안된 온톨로지로, 문화유산 시맨틱 웹을 구성하기에 적절하다.

이 연구에서는 한국의 문화유산 시맨틱 포



<그림 2> 문화유산 DB로부터 사용자에 이르는 도식화

털을 구축하기 위한 초석을 마련하는데 그 목적을 둔다. 기존의 문화유산 포털을 구성하고 있는 문화유산 데이터베이스를 온톨로지 구조화하여, 정보제공자들은 보다 쉽게 정보를 제공·관리하며 정보 이용자들은 정보를 좀 더 효율적이고 의미 있게 구성하여 사용할 수 있는 발판을 마련한다.

2. 배경

2.1 시맨틱 웹

시맨틱 웹(Semantic Web)은 World Wide Web Consortium (W3C)에서 Tim Bemers-Lee에 의해 만들어졌다 [2]. 시맨틱 웹은 웹상에서 머신이 정보의 뜻 - 혹은 “semantics”-을 이해할 수 있는 방법 혹은 기술을 말한다 [4]. 시맨틱 웹 기술은 웹상의 데이터 간에 관계나 인덱스를 만들어서 머신이 정보를 해석할 수 있게 한다.

이러한 시맨틱 웹으로 구현된 포탈은 정보제공자들이 정보를 쉽게 추가하거나 변형할 수 있고, 정보 수용자들은 같은 데이터에서 서로 다른 지식을 재-구조화 할 수 있다.

2.2 CIDOC Conceptual Reference Model

CIDOC CRM(Conceptual Reference Model)은 다양한 문화유산 정보를 통합, 교환, 연결하기 적합하게 만든 온톨로지이다 [5]. 이는 국제박물관협회(International Council of Museums)의 국제 도큐멘테이션 위원회 (International Committee for Documentation)에 의해 컴퓨터 공학자, 고고학자, 뮤지엄 큐레이터, 예술사학자, 서지학자, 물리학자, 철학자들의 협력 하에 개발되었다 [6].

3. 해외 관련 연구

문화유산 정보는 그 본질적인 특성 상 시맨틱 웹상에서 표현하기 적절하다. 이러한 특성 때문에 유럽 국가에서는 국가 문화유산 시맨틱 포털 구현을 위해 1990년대부터 이에 대한 연구들이 시작되었다.

<표 1> 해외 사례 분석

국가	Finland	United Kingdom	Netherlands
프로젝트	FinnONTO 2003-2010	STAR project since 2004	MultimediaN E-Culture project
목표	표준 메타데이터 구축 / 통합 핀란드 온톨로지 구축 / 온톨로지 서비스 제공	서로 다른 데이터베이스를 새로운 시스템에 적용가능한 모델 구현	새로운 문화유산 시맨틱 웹 구현
문제점	메타데이터 스키마 간의 시맨틱 상호운용성	현재 혹은 예전의 데이터를 새로운 시스템에 적용	국제 시소러스와 국내 시소러스와의 불일치
해결책	통합 온톨로지 개발	CRM 리모델링 CRM 매핑	시맨틱 어노테이션

핀란드에서는 국가의 각 분야에 대한 온톨로지는 이미 마련된 상태에서, 하나의 통합된

국가 포털을 구축하기 위한 통합 온톨로지 개발이 주된 목적이었다. [7]

영국의 경우 문화유산 분야에서 과거와 현재, 그리고 미래의 데이터들 간에 의미 관계를 나타낼 수 있는 온톨로지 개발이 주되니 목적이었다. 영국의 경우, CIDOC CRM을 바탕으로 개발하였지만, 영국의 헤리티지에 맞도록 특화한 CRM-EH라 불리는 온톨로지를 개발하고 있다. [8]

네덜란드의 경우, 시소러스와 관련하여 문제를 겪고 이었다. 국제적으로 통용되고 있는 시소러스(AAT 혹은 ULAN)와 네덜란드 자국 시소러스의 불일치를 해결하기 위해 시맨틱 어노테이션(annotation) 방법을 사용하였다. [9-10]

4. 한국의 문화유산 정보 현황

현재 한국에는 국가기록원, 문화재청, 국가문화유산포털, 유네스코, 각종 뮤지엄 사이트, 그리고 대학교 도서관 등 문화유산 정보를 다루는 포털들이 많이 있다. 하지만 대부분의 문화유산 관련 웹사이트들은 같은 데이터를 공유하고 있다. 문화유산포털과 이뮤지엄, 그리고 문화재청의 메타데이터는 같다. 유네스코 홈페이지의 데이터 내용 또한 문화재청에서 발견할 수 있다.

문화재청은 문화재의 관리·보호·지정 등의 사무를 관장하기 위해 설립한 기관으로 문화유산에 관련된 정보 뿐 아니라, 행정 전반에 관련된 정보를 포함하고 있다. 문화유산포털은 이러한 문화유산 관련 정보를 포함하고 있고, 유네스코 홈페이지에서는 이들 중 유네스코 문화유산으로 지정된 문화재에 한해 정보를 공유하고 있다. 이뮤지엄 또한 문화유산에 관련된 정보를 공유하고 있는데, 이곳에는 뮤지엄 네트워크도 형성되어 있다. 국가기록원은 헤리티지와 관련된 국가 기록유산들이 보관되어 있다.

한국 문화유산 관련 메타데이터

한국의 문화유산 정보를 구술하는 메타데이터는 크게 두 가지가 있다. 하나는 1996년 국립중앙박물관에서 고안된 '한국유물분류표준'이고, 다른 하나는 국가기록원에서 사용하고 있는 '전자기록물 보존 메타데이터'이다. '한국유물분류표준'은 133개의 요소로 구술되는데, 이를 분류하여 보면 설명, 관리, 보존에 관한 요소는 있으나, 기술이나 이용에 관련된 요소는 부족함을 알 수 있다. '전자기록물 보존 메타데이터'은 문화유산 뿐 아니라 국가 기관의 기록을 모두 다루기 때문에 문화유산에 특화된 메타데이터는 아니나, 문화유산 보존이나 관리에 관한 유용한 정보를 포함하고 있다.

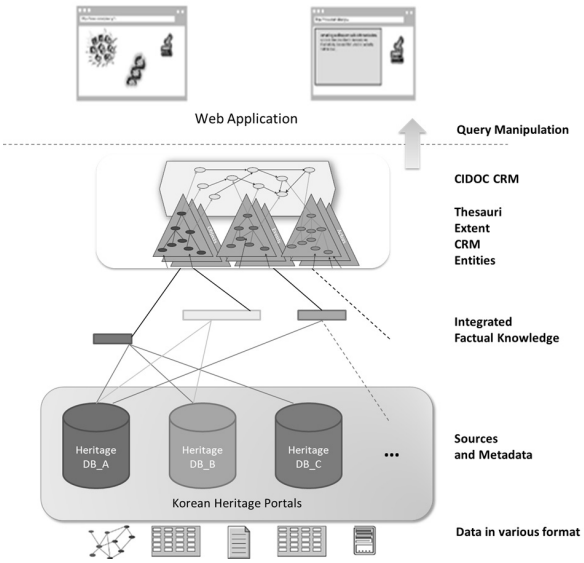
5. 한국 시맨틱 문화유산 포털

우리나라의 문화유산 정보 현황은 만족스러운 편이 아니다. 외국과 비교하였을 때, 메타데이터의 스키마 부족으로 다룰 수 있는 정보의 양이 부족하다. 또한 문화유산을 다루기에 적절한 시소러스도 부족하다.

<표 2> 국내외 문화유산 정보 현황 비교

	해외사례	한국
리소스	지식 구조를 만들 수 있을 만큼 충분	스키마 부족으로 부족
메타데이터 스키마	CIMI, Dublin Core	한국유물분류표준, 전자기록물 보존 메타데이터
시소러스	AAT, ULAN 국내 시소러스 시맨틱 어노테이션 방법	적절한 시소러스 부재
온톨로지	국내 온톨로지 CIDOC CRM	-
웹 구현	시맨틱 웹	-

시맨틱 문화유산 포털을 구현하기 위해서, 현재의 데이터로부터 온톨로지 구조를 세우는 것이다. 이 과정에서 CIDOC CRM을 도입하고자 한다. 이러한 온톨로지 구조를 바탕으로 쿼리 조작을 통해 시맨틱 문화유산 포털을 구현할 수 있다. 연구에서는 온톨로지 구조를 구현하는데 야기되는 문제점을 도출하고 이에 대한 해결책을 제안하려 한다.



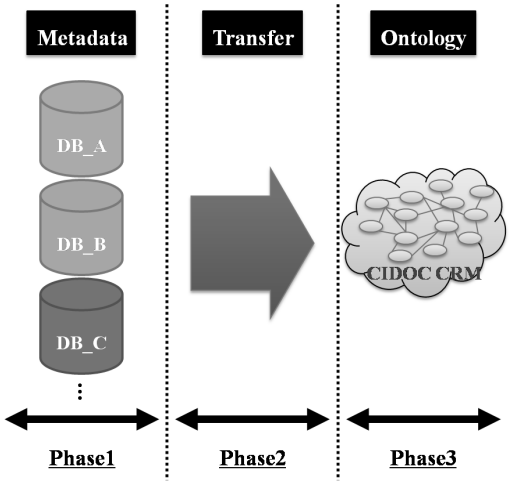
<그림 3> 시맨틱 문화유산 포털 구현 과정

온톨로지 구조 구현과정에서 야기되는 문제는 크게 3개의 카테고리로 구분될 수 있다. 하나는 메타데이터 단계에서 발생하는 문제이고, 다른 하나는 온톨로지 구조화 하는 과정에서 발생하는 문제, 마지막은 온톨로지 구조에서 발생하는 문제이다.

5.1 메타데이터 단계에서 발생하는 문제

한국의 헤리티지 메타데이터가 완벽하지 않기 때문에 야기되는 문제이다. 메타데이터가 외국에서 사용되는 것과 같지 않기 때문에 이를 적합하게 온톨로지 구조화 할 수 없다.

1) 시간을 구술하는 데이터의 부정확성



<그림 4> 온톨로지 구조 구현과정에서 야기되는 문제 분류

CIDOC CRM은 이벤트를 주축으로 하여 만들어진 온톨로지이다. 문화유산 역시 역사적 사건과 관련되어 있기 때문에 시간을 기술하는 요소는 매우 중요하다. CIDOC CRM을 구성하는 83개의 엔티티 중 35개가, 132개의 프로퍼티 중 26가지가 시간에 관련된 요소이다. 온톨로지가 “Domain(entity) → property → range(entity)”로 기술되기 때문에 31,850가지의 시간적 표현이 가능한 것이다. 한국 메타데이터의 시간 속성 부족은(<표 3> 참조) 문화유산 온톨로지 구조를 구축하는데 있어 많은 정보의 손실인 것이다.

<표 3> DC, CIMI, 한국유물분류표준의 시간 속성 비교1

Dublin Core	CIMI	한국유물 분류표준
Creator	creatorGeneral	작가, 제작처
	creatorRole	
	creatorDateOfBirth	
	creatorDateOfDeath	
	creatorNationalityCultureRace	
	creatorInfo	
Date	DateOfBrith	문화재지정일자,

	dateOfDeath	발굴일자, 이동일자
	DateOfOrigin	
	dateCollected	
Relation	relatedObjects	
	relatedObects-rendition	
	relatedTextualReferences	
	stylePeriod	시대, 국적
	periodName	
Indentifier	objectID	유물번호, 원판번호
Source	resource	입수처, 출토지
Language	objectLanguage	
-	contextHistorical	참고자료
	contextArchaelogical	

2) 기술·관리 관련 데이터의 부족

문화유산은 과거의 발자취이기도 하지만, 현재와 그리고 미래에도 연관된 것이다. 그렇기 때문에 기술적인 면이나 관리적인 면에 대한 정보도 중요하다. CIDOC CRM에도 이와 관련된 프로퍼티가 19가지가 존재한다. 하지만, 이와 관련된 한국유물분류표준의 스키마는 매우 부족하다. 이와 관련하여 필요한 요소들을 전자기록물 보존 메타데이터로부터 보충 받을 수 있다. (<표 4>, <표 5>)

<표 4> 전자기록물 보존 메타데이터 중 기술 및 관리 관련 요소

상위항목	하위항목
상세정보	유형
	요약정보
	포맷
	저장정보
	구조
	규모
	전자기록물 여부
	기술적 요건
	구기록물 정보
	언어
보존정보	보존종류
	보존장소
	보존기간
	보존일시
	보존행위자
	보존행위설명
	향후 보존

<표 5> DC, CIMI, 한국유물분류표준의 시간 속성 비교2

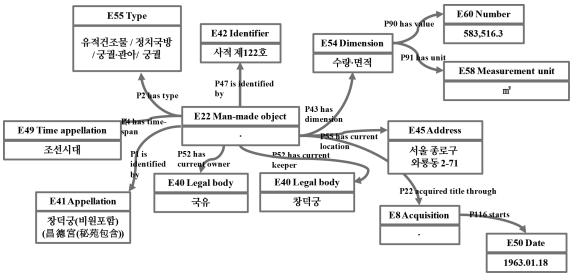
Dublin Core	CIMI	한국유물분류표준
Rights	copyrightRestriction	보험기록 / 현존여부
	provenance	
	repositoryName	
	repositoryPlace	
Contributor	association	
	associationGeneral	
	association-name	
	owner	
Format	quantity	크기, 형태 / 보존처리 내용 / 유물상태
	dimension	
	processTechnique	
	materialMedium	
	condition	
	exhibitionHistory	
Publisher Coverage	nationalityCultureRace	
	creditLine	
	wallTextlabel	
	protectionStatus	
	filedCollector	
	administrativeEvent	
	administrator	

1),2)를 종합하여 온톨로지 구조를 구현하는데 추가로 필요한 스키마를 찾아보면 <표 6>과 같다. 기존의 한국유물분류표준 만으로 정보를 CIDOC CRM을 이용하여 표현하면 <그림 5>와 같다. 이를 <표 6>에서 제안한 스키마를 추가로 고려하여 표현하면 <그림 6>과 같다.

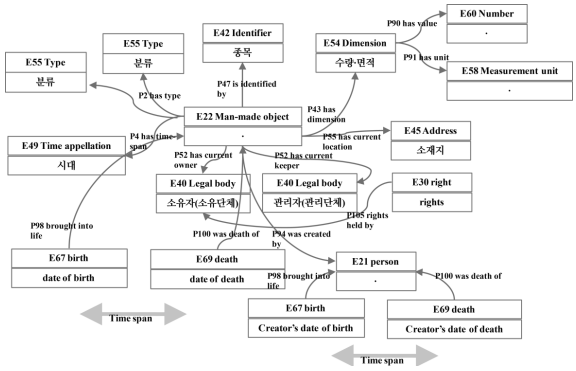
추가된 한국유물분류표준은 온톨로지 구축을 하였을 때 그 표현의 양을 깊게 만든다. 이는 정보 구조를 복잡하게 만들고, 복잡한 정보구조로 인하여 문화유산 정보들이 서로 연결되어 시맨틱 포털을 더 풍부하게 만들 것이다.

<표 6> 한국유물분류표준에 추가될 스키마

시간을 구술하는 데이터의 부정확성	기술·관리 관련 데이터의 부족
Date of birth	Date of preservation
Date of death	Duration of preservation
Creator's date of birth	Copyright
Creator's date of death	Technical aspects
변화 상태도 함께 구술되어야 함. (감쇄, 증가, 파괴, 변형 등)	



<그림 5> 한국유물분류표준으로만 표현한 CIDOC CRM 매핑

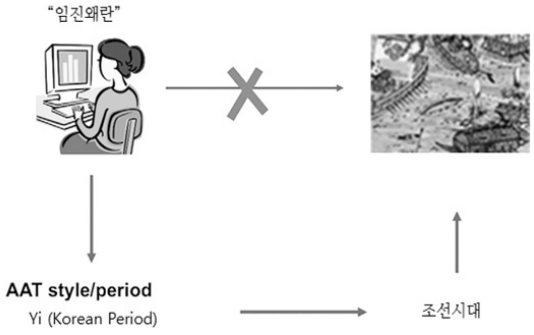


<그림 6> 추가된 한국유물분류표준으로 표현한 CIDOC CRM 매핑

5.2 온톨로지 구조화 하는 과정에서 발생하는 문제

1) 시소러스

시소러스는 도메인 언어를 구조화 할 때 필요한 요소이다. 이는 '개념상의 등가(conceptual equivalence)'를 나타내는 제한된 용어들의 리스트이다. 하지만 한국의 시소러스는 아직 불충분하다. 게다가 네덜란드의 상황과 마찬가지로, 국제적으로 사용되는 시소러스와 우리나라에서 사용되는 시소러스의 구조와 표현에 차이가 있다. 언어적 차이를 고려한다 하더라도, 이 불일치를 연결시켜줄 수 있는 시맨틱 어노테이션 작업이 요구된다.



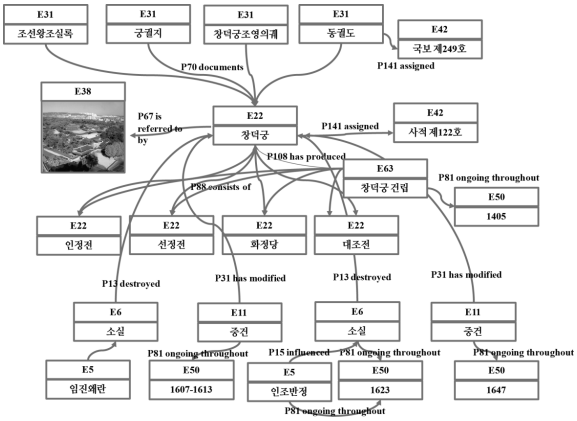
<그림 7> 국제적으로 사용되는 시소러스와 국내 시소러스 사이의 불일치

2) 설명문

한국의 메타데이터는 다른 나라와는 다르게 설명을 많이 포함하고 있다. 그리고 이 설명 안에 방대한 양의 정보를 가지고 있다. 이 정보를 CIDOC CRM으로 표현하면 <그림 8>과 같다. <그림 8>은 <그림 5>와 달리 문화유산 자체에 대한 정보뿐만 아니라 이를 둘러싼 역사적 사건 또한 보여준다.

설명 속에는 방대한 양의 정보가 포함되어 있지만, 이는 사람의 작업을 필요로 한다. 이를 다루기 위해서는 스키마 매핑을 한 후, 데이터 리프팅(혹은 트랜스퍼링)과정을 거쳐야 한다. 설명문에서 필요한 데이터를 뽑아내어 분류하기 위해서는 자연어처리(Natural Language Processing)이 필요한데, 외국의 경우 TEI [11]나 Grey literature [12]와 같은 방법이 사용되고 있다. 하지만 언어, 문법, 어순, 단어표현의 차이가 있기 때문에 이를 그대로 적용할 수 없고, 이를 따

로 처리할 수 있는 프로그램이 요구된다.



<그림 8> 설명문 내용을 추가하여 표현한 CIDOC CRM 매핑

3) 번역

온톨로지 구조 구현을 효율적으로 하기 위해서 CIDOC CRM 한국어판이 요구된다. 프랑스, 독일, 일본 등의 나라에서는 이미 자국어로 CIDOC CRM을 제공하고 있다. 이에 대한 지침은 “Guidelines for Translaing the CIDOC CRM"[1]에서 제공하고 있다.

5.3 온톨로지 구조에서 발생하는 문제

비록 CIDOC CRM이 문화유산을 기술하기에 적합하게 고안되었지만, 적용과정에 대한 지시사항이 없기 때문에 몇몇 문제가 있다. 다른 데이터가 같은 CIDOC CRM 표현을 가질 수 있고, 반대로 같은 데이터임에도 CIDOC CRM 표현은 서로 다를 수 있다. 이에 대한 문제는 다른 나라에서도 제기되었으나 해결책은 아직 제시되고 있지 않아, 이 연구에서는 문제점을 제기하기만 하겠다.

6. 결론 및 제언

본 연구에서 한국 시맨틱 문화유산 포털을 구현하기 위한 첫 번째 단계로 온톨로지 구조 구현과정에서 야기되는 문제점과 해결책을 제시하였다. 문제에 대한 실제적인 해결을 하지는 못하였지만 이 연구를 바탕으로 하여 해결책 마련을 위한 연구가 문화유산과 웹 전문가의 협업을 통해 이루어진다면 한국의 문화유산 정보 생산·관리·이용이 더 수월해질 것이다.

<표 7> (요약) 문제점과 해결방안

문제점		해결방안
메타데이터 단계에서 발생하는 문제	메타데이터 표현의 부족	메타데이터 스키마 추가
온톨로지 구조화 하는 과정에서 발생하는 문제	시소러스	시맨틱 어노테이션
	설명문	한국에 맞는 NLP
	번역	Guidelines for Translating the CIDOC CRM
온톨로지 구조에서 발생하는 문제	적용과정 지시사항 부재	-

1) http://www.cidoc-crm.org/translation_guidelines.html

참고문헌

M. Doerr and D. Iorizzo, "The dream of a global knowledge network--A new approach," Journal on Computing and Cultural Heritage (JOCCH), vol. 1, 2008, p.5.

T. Lee, et al., "The semantic web," Scientific American, vol. 284, 2001, pp.34-43.

T. Gruber, "It Is What It Does: The Pragmatics of Ontology. Invited presentation to the meeting of the CIDOC Conceptual Reference Model committee," Smithsonian Museum, Washington, DC, 2003.

W3C. Retrieved March 13, 2008., W3C Semantic Web Frequently Asked Questions.

N. Crofts, et al., "Definition of the CIDOC CRM conceptual reference model. Reference document," International Council of Museums, 2005.

M. Doerr. 2007-03-22, Past and Future of ISO21127:2006 or CIDOC CRM. (http://cidoc.mediahost.org/standard_crm(en)(E1).xml)

E. Hyvonen, "FinnONTO--Building the Basis for a National Semantic Web Infrastructure in Finland," 2006.

C. Binding, et al., "Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction via the CIDOC CRM," Research and Advanced Technology for Digital Libraries, 2008, pp.280-290.

G. Schreiber, et al., "Multimedian e-culture demonstrator," The Semantic Web-ISWC 2006, 2006, pp.951-958.

G. Schreiber, et al., "Semantic annotation and search of cultural-heritage collections: The MultimediaN E-Culture demonstrator," Web Semantics: Science, Services and Agents on the World Wide Web, vol. 6, 2008, pp.243-249.

C. Ore and O. Eide, "TEI and cultural heritage ontologies: Exchange of information?," Literary and Linguistic Computing, vol. 24, 2009, p. 161.

A. Vlachidis, et al., "Excavating grey literature: A case study on rich indexing of archaeological documents by the use of Natural Language Processing techniques and knowledge based resources," 2009