시맨틱 웹 기반의 연구성과물 검색시스템에 관한 연구

A Study on the Semantic Web based Research Results Information Retrieval System

박동진*

목 차

Key Words : metadata, semantic web, RDF/RDFS, RQL, research results information

Abstract

It has lately been recognized that the sharing and exchanging of the research results information is the critical factor to improve the research productivity. So many institutions are planning or developing the information systems which provide the research information services for researcher. But it has very difficulty in integrating the research resources information due to the dispersion and heterogeneity in data sources, and semantic and structural difference in describing data. We propose the semantic web based methodology and conceptual framework for raising the interoperability of metadata about research results information, which will support the integration of the distributed research data for information services in the end. Finally we proposed the conceptual architecture of research information service system which shows the main components, the functional requirements, and the principal and design direction at implementing the system.

Ⅰ. 서론

최근 국가적으로 그리고 각 연구기관에서 R&D 투자의 확대와 아울러 연구생산성의 향상을 위하여 많은 노력을 기울이고 있다. 특히 연구성과정보의 유통을 활성화하기 위하여 국가 R&D 종합관리체계구축의 일환으로서 국가연구성과정보관리시스템을 개발중에 있으며, 각 연구기관에서도 연구개발성과정보들을 수집, 교환 및 평가하는 정보시스템을 계획하거나 운영 중에 있다. 그러나 현재 국가는 물론 각 기관에서도 연구성과정보를 각자의 방법으로 정의하고 또한 상이한 체제로 관리하고 있음으로 해서 구조적인(structural) 그리고 의미적인

* 공주대학교 산업시스템공학과

(semantical) 면에서 전 연구기관들의 정보들을 통합하여 서비스하기가 매우 어려운 실정이다. 즉 연구성과정보의 호환성에 문제가 있는 것이다.

현재 연구성과정보의 서비스는 첫째, 국가적으로 각 연구기관의 데이터베이스를 통합한 후 포탈을 통하여 정보를 서비스하는 방법과, 둘째, 각 연구정보서비스기관에서 수집한 연구정보를 구매하는 방법과, 셋째, 각 연구기관의 웹 서버에 접속하여 해당 연구기관에서 제공하는 연구정보를 서비스 받는 방법과, 마지막으로 사용자가 웹에서 범용검색엔진을 통하여 연구정보를 수집하는 방법 등이 있다. 본 연구는 국가수준에서 연구성과정보를 서비스 해주는 첫 번째 방법의 보완차원에서 고려될 수 있다. 분산 데이터베이스 기술을 이용한 각 연구기관의 데이터베이스 통합시 제도적으로 그리고 기술적으로 다음과 같은 상당한 제약이 있다. 먼저 데이터베이스의 보안차원에서 각 기관은 소극적으로 대처할 수 있으며, 다음으로 공개하는 정보제공의 적시성 및 다양성에도 많은 제약이 있다. 이는 각 기관자체적으로 연구와 관련된 다양한 정보들을 체계적 관리하고 있지 않기 때문에 정보 제공의 어려움이 있는 것이다. 즉 많은 연구성과정보가 비구조적인 문서의 형태로 존재하기 때문에 데이터베이스화하기가 어렵다. 따라서 현재 추진 중이 연구성과정보 시스템이 완료되더라도 정보서비스의 양과 질적인 면에서 연구자의 정보서비스 욕구를 충분히 만족시킬 수 없을 것이다.

본 연구는 시맨틱 웹 기술을 도입하여 연구성과정보의 메타데이터에 대한 상호운용성을 제고하는 방안을 다음과 같이 제시함으로써 궁극적으로 연구성과정보의 통합을 지원하고자 한다. 특히 시맨틱 웹 기반의 검색시스템 개발에 있어서 필요한 주요기능, 구성요소, 적용기술 및 상세 설계의 방향을 제시한다. 제2장에서는 연구의 배경을, 제3장에서 시스템의 개념적 구조를 제시하고 제4장에서 결론을 맺는다.

II. 연구의 배경

2.1 연구성과정보 상호운용성 수준

정보의 상호운용성(interoperability)이란 사용자가 두개 또는 그 이상의 기술적인(technical) 시스템에서 만족할 수 있는 방법으로 정보를 직접 교환할 수 있는 상태를 말한다(Mooney 2001). 즉 상호운용성은 이기종간 분산환경의 정보시스템을 서로 연동하여 단일의 통합 환경으로 서비스를 제공하는 개념이며 동시에 기능적 요구사항이다(이수상 2004). 통합의 기능적인 면은 3가지 면이 있다. 첫째, 연합(federation) 수준으로 상호간에 엄격한 프로토콜을 준수하여 통합을 하는 것이다. 둘째, 다소 느슨한 연합형태인 수확(harvesting) 수준으로서 각 시스템별로 메타데이터를 교환함으로써 상호연동이 가능한 수준의 통합이다. 마지막으로 수집(gathering)수준으로 검색엔진으로 수집하여 서비스하는 수준의 통합이다(Maly et al. 2001)

현재 국내의 각 연구기관 연구정보시스템의 데이터베이스는 이질적(heterogeneous)이고 분산된(distributed) 환경이다. 또한 한 조직 내에서도 연구정보들이 데이터베이스화 되어 있지 않고 다양한 형태의 문서로 흩어져서 저장 관리되고 있다. 정보의 통합에 있어서 첫 번째 방법인 연합에 의한 통합은 연구정보의 접근에 있어서 매우 효율적인 방법이나, 국내의 연구성과정보가 이질적으로 분산된 데이터베이스에 그리고 일부는 웹 상에 흩어져 있기 때문에 이를 논리적 그리고 물리적으로 통합하기가 매우 어렵다. 또한 세 번째 통합 방법인 일반 검색엔진을 이용한 수집 수준으로는 소위 정보과다(overload)의 문제와 정보결핍

(deficiency)의 문제가 동시에 발생되므로 전문 연구자들에게 있어서 정보검색 만족도는 매우 낮을 수 밖에 없다. 따라서 현실적으로 연구성과정보서비스는 상호운용성의 기능적인 면에서 보면 수확 수준에서 고려되어야 한다. 즉 각 기관의 연구성과정보관리의 다양성을 인정하고 단지 메타데이터의 상호운용성을 제고하면서 연구성과정보의 통합을 촉진하는 방안이 필요한 것이다.

2.2 연구성과정보 서비스의 요구사항

연구성과정보는 R&D 전주기에 걸쳐서 다양한 문서 및 매체로 발생되며 그 범위나 대상에 있어서도 통일된 의견은 없다. 연구성과정보서비스 사용자인 전문 연구자의 관점에서 보면 정밀한 검색을 통한 구체적이고 정확한 정보를 요구한다. 즉 검색 목적 및 내용이 다양하고 검색의 대상도 폭도 넓어질 필요가 있다. 특히 연구성과물 측면에서 보면 공식적인 채널을 통하여 발표된 연구보고서나 논문 뿐 아니라, 공식적인 출판경로를 통하여 입수하기 힘든 회색문헌(grey literature)에 대한 정보까지 요구한다(Jeffery 1999). 즉 R&D 관점에서의 회색문헌인 출판전 배포문, 연구계획 제안서, 컨퍼런스 발표논문, 기술보고서, 학위논문, 회의/세미너/워크샵 자료, 기술계약에 관한 정보, 그리고 특허정보, 연구원 및 연구기관 등에 관한 정보까지 포함된다(Almeida 1999; Auger 1996). 국내 연구소 연구원들에 대한 설문조사의 결과에 따르면 회색문헌은 최신의 연구결과이므로 과학기술의 발전을 위해서 회색문헌의 학술적 가치 및 활용 상의 가치가 매우 크며, 각 기관은 물론 국가적 차원에서 이를 체계적으로 관리해야 할 필요가 있음을 주장하였다(남영준 2002).

그러나 현재 국가적으로 제공되는 연구정보서비스인 과학기술통합검색(www.yeskisti.net)와 국가연구개발종합시스템(www.kordi.go.kr/index.jsp)를 분석하면 다음과 같은 문제점이 있다. 첫째, 정보검색 측면에서 연구성과물 중 연구보고서와 같이 공식적으로 발표된 것들에 대한 정보만을 검색 가능하고 각종 회의 및 세미나 자료, 논문, 단행본 등을 포함하는 기타의 연구성과물에 대한 정보의 검색이 불가능하다. 둘째, 연구성과측정 측면에서 특정 과제와 관련하여 산출된 모든 성과물에 대한 추적 및 관리가 시스템적으로 불가능하다. 셋째, 연구기관별 연구성과물 유통과 정보교환이 불가능하다.

또한 웹의 활용으로 인하여 많은 연구정보자원이 웹 상에 존재한다. 이는 연구자가 연구정보시스템에 연구정보를 입력하기보다는 연구자 자신 혹은 연구과제의 웹페이지에 연구성과물과 및 연구에 관한 정보를 저장하기 때문이다. 따라서 웹상에 넓게 분포되어 있는 연구정보의 접근도 중요하게 다루어져야 한다.

2.3 연구성과정보 메타데이터의 표현

메타데이터는 데이터에 관한 구조화된 데이터로서 자원과는 독립적으로 존재하면서 자원에 대한 다양한 접근점과 네트워크 주소를 포함하는 데이터이다(Miller 1998). 메타데이터는 다음과 같이 활용된다. 첫째, 자원의 의미를 요약해 준다. 둘째, 자원의 검색을 가능하게 한다. 셋째, 자원의 필요성에 대한 판단을 가능하게 해 준다. 넷째, 다른 자원들과의 연관성을 말해준다(Steinacker et al. 2001).

대상이 되는 자원을 메타데이터로 기술하기 위해서는 먼저 적절한 메타데이터 요소의 집합(element set)을 선택하고, 각 데이터 요소에 대하여 어휘를 정의하고 그리고 표현 방법인 스킴(scheme)을 결정한다. 많은 경우 메타데이터 스킴은 기존의 표준이

나 온톨로지(ontology)들로부터 도입한다. 대표적인 웹자원을 기술하는 표준 메타데이터가 Dublin Core 이다(Dublin Core Metadata Initiative 2005). 특히 과학기술 분야의 연구개발 성과물을 위해서는 최근에 CRIS(Current Research Information System)에서 CERIF(Common European Reaearch Information Format)-2004 메타데이터를 제시하였다.

2.4 시맨틱 웹을 활용한 연구성과 정보서비스

시맨틱 웹은 기존의 웹 서비스의 한계를 극복하기 위하여 1998년 Berners-Lee 가 주창한 차세대 웹 기술이다. 시맨텍 웹은 웹이 제공하고 있는 정보를 잘 정의된 온톨로지를 기반으로 하여 표현함으로써 물리적 논리적으로 분산되어 있는 애플리케이션 간의 상호운용성을 제공하기 위한 것이다(이승희 외 2003). 즉 시맨틱 웹에서는 다른 데이터 구조를 갖고 있는 애플리케이션도 온톨로지를 통해 상대방의 정보를 이해하고 처리할 수 있다. 이때 사람이 인식하고 이해하는 것이 아니라 기계, 즉 컴퓨터가 이해하고 정보를 처리하는 것을 말한다. 이와 같이 지능형 처리 즉, 컴퓨터가 이해할 수 있는 정보를 만든 다는 것은 메타데이터에 대한 접근법에서 출발한 것이다. 시맨틱 웹은 메타데이터를 기반으로 한 온톨로지와 지식표현(knowledge representation)이 핵심이다(이재호 2002)

시맨틱 웹은 계층화된 구조(layered structure)이다. 제일 하단에는 구문(syntax)을 전달할 수 있도록 XML이 기초를 이루고 있다. RDF는 정보표현 프레임워크를 제공하고 있으며, RDFS는 데이터 모델링의 구조로서 클래스와 속성을 정의하고 그들 사이의 관계를 정의한다. 그 위의 온톨로지는 RDF/RDFS를 포함하며, 지능적 처리가 가능하도록 정보의 논리관계도 표현하는 것이다.(Berners_Lee et al. 2001)

온톨로지란 특정 도메인에 있어서 어휘들의 의미를 기술하고 어휘들 간의 상호관계들에 대한 표현으로 어떤 영역을 기술하기 위해 사용하는 개념과 그리고 유사한 어휘들을 모아 놓은 어휘 집에 관한 규격이다(W3C 2005). Lopatenko와 그의 동료들에 따르면 연구정보검색시 다음과 같은 문제가 있을 때 시맨틱 웹 기술이 활용되어 질 수 있다고 주장한다(Lopatenko et al. 2002). 첫째, 데이터 소스가 분산되어 있으며 데이터가 구조적으로 그리고 의미적으로 서로 차이가 나며, 정보 검색 시 이를 고려해야 하는 경우, 둘째, 정교한(sophisticated) 정보 검색이 요구될 때, 셋째, 기존의 데이터 구조와 새로운 데이터의 구조가 서로 호환될 필요가 있을 때, 넷째, 기존의 시맨틱 웹 데이터를 활용하여야 할 때, 다섯째, 데이터베이스에 직접적인 접근이 불가능 할 때, 마지막으로 다른 시맨틱 웹 기반의 구조와 호환될 필요가 있을 때이다.

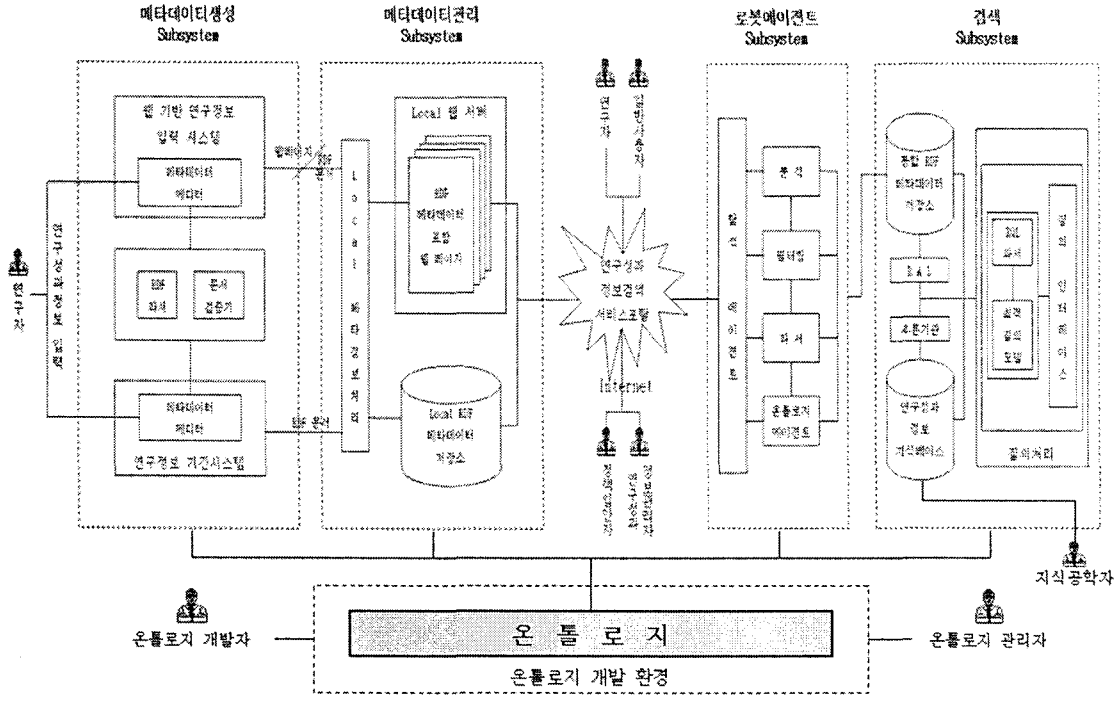
Lopatenko와 동료들에 의해서 유럽의 연구기관간의 연구정보교환을 위해 연구성과 정보시스템 프로토타입인 AURIS-MM(Austrian Research Information System - Multimedia Enhanced)이 개발되었다 (http://derpi.tuwien.ac.at/~andrei/AURIS-MM-plan.html). 이것은 기존의 오스트리아연구정보시스템인 AURIS에 시맨틱 웹 기술을 채택하여 기능적으로 확장한 시스템이다. 이 시스템은 유럽에서 표준 연구성과 정보 메타데이터로 인정되는 CERIF-2000을 기반으로 온톨로지를 개발하고 이를 RDF/RDFS 로 표현하였다. 본 연구는 AURIS-MM과 같은 취지에서 우리나라의 특수성을 반영한 시맨틱 웹 기반의 연구성과정보서비스 체계를 제안하고자 한다.

III. 메타데이터 기반 시스템의 개념적 구조

본 장에서는 시맨틱 웹 기술의 적용관점에서 메타데이터 기반 연구성과정보시스템의 개념적 구조를 제시한다. <그림 1>에서처럼 시스템은 온톨로지를 기반인프라로 하며 기능적으로 크게 4개의 서브시스템으로 구성된다. 메타데이터생성 서브시스템과 메타데이터관리 서브시스템은 각 연구기관에서 관리되며, 로봇에이전트 서버시스템과 검색 서브시스템은 국가수준에서 연구성과정보를 서비스하는 기관에서 관리한다. 각 서브시스템의 구성요소 및 설계특성을 보면 다음과 같다.

3.1 메타데이터생성 서브시스템

메타데이터생성 서브시스템은 각 연구기관의 웹 환경 및 기간시스템 환경에서 구동되는데 핵심은 메타데이터 편집기와 RDF 파서 및 문서검정기이다. 메타데이터 편집기는 RDF 메타데이터 작성을 위한 환경을 제공한다. 즉 연구성과정보 온톨로지를 로드시켜 비쥬얼화 함으로서 사용자가 메타데이터용 엔터티를 선택하고, 세부사항인 속성



<그림 1> 연구성과정보시스템 개념적 체계도

둘을 입력하게 한다. 또한 생성된 엔터티와 기존의 엔터티간의 관계도 설정한다. 또한 메타데이터 편집기는 웹 주석(annotation) 기능이 있어 웹 페이지에 작성된 메타데이터를 삽입하는 기능도 있다. 마지막으로 생성된 메타데이터는 메타데이터 저장소로 보내고, 메타데이터를 포함하고 있는 웹 페이지는 웹 서버로 보낸다. RDF 파서는 연구성과정보 메타데이터가 올바르게 작성되었는지를 확인하기 위해서 RDF 문서에 대하여 구문검사를 하는 실시하는 기능이 있으며, 텍스트, 트리, 그래프에서 저작중인 RDF 및

XML 문서에 대한 파싱을 실시한다. RDF 문서 유효성 검증기(validator)는 RDF 문서를 생성하기 바로 전 단계에게 문서의 유효성을 검사한다. 유효성검사는 메타데이터에 포함된 의미(semantic)를 가진 여러 어휘들의 DTD와 스키마를 검증함으로서 표현의 정확성을 확보하는 것이다.

3.2 메타데이터관리 서브시스템

메타데이터관리 서브시스템은 생성된 메타데이터를 관리하는 기능들을 포함한다. 로칼메타정보처리 기능은 각 연구자의 홈페이지, 각 연구과제의 홈페이지, 그리고 연구정보시스템 등으로부터 작성된 RDF 문서를 주기적으로 수집하여 웹 서버 혹은 메타데이터 저장소에 RDF 메타데이터 문서를 저장하고 기존의 저장된 메타데이터를 관리하는 기능을 한다. 그리고 RDF 문서가 포함된 웹 페이지는 웹 서버에 등록함으로써 연구성과정보를 연구기관에서 직접 외부에 공개할 수도 있다. 연구기관 내에서 작성된 모든 RDF 메타데이터를 저장하는 저장소(repository)는 다양한 유형의 데이터베이스가 사용될 수 있지만 관계-객체혼합형데이터베이스가 RDF 스키마의 구조를 모델화하는데 이점이 있다.

3.3 로봇에이전트 서브시스템

로봇에이전트는 RDF 메타데이터를 수확(harvesting)하는 기능을 하는 지능형 에이전트이다. 에이전트는 웹을 자동적으로 순화하는 프로그램으로서 온톨로지를 지식 기반으로 각 연구기관에서 새로 등록한 RDF 메타데이터를 수집하고, 필터링하여 검색시스템의 통합 RDF 저장소로 가져오는 역할한다. 로봇에이전트는 RDF 문서의 분석을 위해서 RDF 파서가 필요하다. 또한 로봇 에이전트는 기능적으로 보면 도메인 온톨로지와 RDF 구조를 조회하는 온톨로지 에이전트(Ontology Agent)와 RDF 문서를 탐색하여 수집하는 탐색 에이전트(Search Agent)로 구분된다.

3.4 검색 서브시스템

전체 시스템에서 가장 핵심적인 역할을 하는 것이 검색 서브시스템이며 기술적으로 가장 많은 검토가 필요한 부분이다. 검색 서브시스템은 크게 질의처리모듈, 지식베이스 및 추론기관, 통합메타데이터 저장소 등으로 나누어진다. RQL을 기반으로 하는 질의 모듈은 사용자 검색 폼으로부터 입수된 정보검색 요구사항을 RQL 질의 문장으로 변환시키고 RQL 파서를 이용하여 최적의 질의 모델을 생성시킨 후 검색을 요청한다. 지식베이스와 추론기관은 지식베이스에 포함된 규칙(rule)과 사실(fact)을 활용하여 새로운 결론을 추론(reasoning abilities)할 수 있는 능력과 온톨로지의 스키마 탐색(schema exploration) 능력이 검색 엔진의 정확도를 더욱 향상시킨다. 통합 RDF 메타데이터 저장소는 모든 연구기관으로부터 입수된 메타데이터를 저장하는 곳이다. 로칼 메타데이터 저장소와는 같은 구조지만 대용량의 RDF 데이터를 처리해야 하므로 검색시스템의 성능에 매우 중요한 역할을 하나 아직 기술적으로 완전히 해결되지 않고 있는 실정이다. RAL(Repository Abstraction Layer)은 관계형데이터베이스의 RDBMS의 역할을 하는 것으로 RDF 데이터와 시스템간의 독립성을 유지하고, 각종 RQL 질의어를 처리하고, RDF 관리(administration) 기능을 포함한다.

IV. 결론

연구성과정보시스템에 있어서 시맨틱 웹 기술의 적용은 기존의 문제점을 해결할 뿐 아니라 정보서비스 및 상호운용성의 제고 차원에서 많은 가능성을 보인다. 본 연구에서는 기술적 관점에서 구체적인 시스템 구현 가능성을 검토하고 개념적 구조를 제시하였다.

구체적으로 본 연구의 결과는 연구성과정보의 통합 면에서 다음과 같은 기대효과가 있다. 즉 시맨틱 웹 기술의 채택은 이질적이고 분산된 데이터 소스로부터 정보를 수확(harvesting)방법으로 시스템을 통합하게 한다. 또한 국가차원이나 각 연구기관 차원에서 야기되는 연구성과 데이터의 구조 및 의미의 비일치성(discrepancies)에 따른 호환성 및 상호운용성 문제를 해결할 수 있는 온톨로지 기반의 어프리케이션 개발을 가능하게 한다. 본 연구에서 제시한 시스템은 연구성과정보시스템과 연동되어 기존의 문제점들을 해결할 수 있다.

다음으로 정보 검색면에서는 온톨로지 기반의 지능적 정보 검색(intelligent information retrieval)을 가능하게 한다. 기존의 검색시스템에서는 어휘나 용어가 시스템에 따라서 다르게 표현되며, 정보 수요자나 제공자가 서로 다른 관점으로 정보를 이해한다. 그러나 온톨로지 기반의 검색은 정보 검색자로 하여금 도메인 지식을 적용하게 하여 더욱 정교하고 강력한 검색을 가능하게 한다. 그러나 본 연구에서는 개념적 방향과 핵심적인 기술사항 만을 검토 하였을 뿐이며 구체적인 구현에 있어서 발생될 문제점들에 대한 검토는 없다. 아직 시맨틱 웹에 대한 기술체계가 완성되지 않았으며 이를 지원하는 툴도 충분하지 않다. 따라서 본 연구에서 언급된 사항들을 구현하기 위한 기술적인 세밀한 검토가 추후에 필요할 것이다.

참 고 문 헌

1. 이승희, 신문수, 정무영. 2003. RDFS+OWL을 이용한 생물학적 데이터의 지식표현과 추출, 『한국경영과학회/대한산업공학회 2003 춘계공동학술대회』, 1136-1141.
2. 이재호. 2002. 시맨틱 웹 기술을 적용한 전자상거래 표준 운용체계연구, 『전자상거래표준화통합포럼』.
3. Almeida, Mario do G.G. 1999. "Control Access for Grey Literature in Brazil: A Proposal." *Proceeding on the fourth International Conference on Grey Literature.*
4. Berners_Lee T., J Hendle, and O. Lassila. 2001. *The Semantic Web*, Scientific American.
5. Jeffery. K. 1999. "An Architecture for Grey Literature in a R&D." *Proceeding on the fourth International Conference on Grey Literature.*
6. Lopatenko, A., A. Asserson, G. Keith, and K. Jeffery. 2002. "CERIF - Information Retrieval of Research Information in a Distributed Heterogeneous Environment." *Proceeding on the 6th International Conference on Current Research Information Systems*
7. Maly, K., M. Zubair, and X. Liu. 2001. "Kepler - An OAI Data/Service Provider for the Individual." *D-Lib Magazine*, Vol. 7. No. 1. 2005.[cited 2006. 5.10]. <*www.dlib.org/dlib/april01/maly/04maly.html*>.
8. Mooney, S. 2001. "Interoperability." *D-Lib Magazine*, Vol. 7, No. 1,2005.[cited 2006. 5.10]. <*www.dlib.org/dlib/january01/mooney/01mooney.html*>.
9. Steinacker A., A. Ghavam and R. Steinmetzm. 2001. "Metadata Standards for Web-Based Resource", *IEEE Multimedia*, Vol. 8, Issue 1, pp. 70-76.
10. W3C, "OWL Web Ontology Language Overview."[cited 2006. 5.10]. <*www.w3.org/TR/owl-features/*>.