Enabling Interoperability For Autonomous Digital Libraries: An API To CiteSeer Services

Yves Petinot[1,2], C. Lee Giles[1,2,3], Vivek Bhatnagar[2,3], Pradeep B. Teregowda[1], Hui Han[1,3]

[1]Department of Computer Science and Engineering
The Pennsylvania State University
111 IST Building
University Park, PA 16802
{petinot,hhan}@cse.psu.edu

[2]eBusiness Research Center
The Pennsylvania State University
401 Business Administration Building
University Park, PA 16802
{vivekb,pbt105}@psu.edu

[3]School of Information Sciences and Technology
The Pennsylvania State University
332 IST Building
University Park, PA 16802
{giles}@ist.psu.edu

Abstract
We introduce CiteSeer-API, a public API to CiteSeer-like services. CiteSeer-API is SOAP/WSDL based and allows for easy programatical access to all the specific functionalities offered by CiteSeer services, including full text search of documents and citations and citation-based document discovery. CiteSeer-API is currently showcased on SMEALSearch [10], a digital library search engine for business academic publications.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: retrieval models. H.3.7 [Digital Libraries]: dissemination, standards, system issues.

General Terms
Design, Standardization.

Keywords
CiteSeer, API, digital libraries, SOAP, WSDL.

1. Introduction
Digital Libraries (DL) systems remain strongly proprietary in the way they collect, index, store and present their document collections. Efforts for access normalization, such as that of the OAI-PMH [5] community address the issue of presenting collections in a standard format that allows, if not interoperation of those systems, at least the creation of meta-systems able to virtually aggregate many heterogenous collections. The interoperation of DL systems is however not addressed by those efforts. In this paper we consider the case of CiteSeer-like servers [3,4,8] and how they can be made more interoperable. We introduce CiteSeer-API, a SOAP/WSDL based API to CiteSeer-like servers which we envision will serve as the corner stone for seamless interoperability with and between CiteSeer services. We provide an overview of the current functionalities supported by CiteSeer-API and outline what we feel are the next necessary steps for enabling CiteSeer services on the Semantic Web. Registration to CiteSeer-API is available at [1].

JCDL'04, June 7–11, 2004, Tucson, Arizona, USA.

2. Motivations
CiteSeer servers have been brought to OAI-PMH compliance so that their metadata collection can be accessed by metadata harvesters [8]. Still, CiteSeer servers feature many functionalities that cannot be accommodated by OAI-PMH, including full text document and citation search and citation-based document discovery. Our motivations for CiteSeer-API are therefore (1) to provide programatical access for all the functionalities supported by CiteSeer-like systems; (2) to enable interoperability of CiteSeer services with distributed and heterogeneous DL systems.

3. CiteSeer Object URIs
CiteSeer servers manipulate three main concepts (object classes): Document, Citation and Group [1]. In order to enable interoperability with distributed DL systems (e.g. interlinking), CiteSeer-API assigns a Unique Resource Identifier (URI) to each of these objects. The URI format associated with each type of resource is presented in Table 1.

Table 1: CiteSeer Object URIs Formats

Resource Type	*URI Format*
Document	http://<server>/document/<enc>/<doc-id>
Citation	http://<server>/citation/<enc>/<cite-id>
Group	http://<server>/group/<enc>/<group-id>

These URIs are typically returned by the search methods and can then be used with the object access or bibliography-oriented methods to access extended information on the correspondig Document, Citation or Group resource.

4. CiteSeer-API Methods
Following is a succinct description of the methods supported by CiteSeer-API. Comprehensive reference is available at [1].

4.1. Search Methods
- **findDocumentsByText**: document full text search
- **findCitationsByText**: citation text search

4.2. Object Access Methods
- **getDocument**: retrieve a Document object
- **getCitation**: retrieve a Citation object
- **getGroup**: retrieve a Group object

4.3. *Bibliography-Oriented Methods*

The following methods are all relative to a specific Document *D* in the collection

- **getCitations**: get Citations made by *D*
- **getCitedBy**: get Documents citing *D*
- **getCoCitation**: get *D*'s co-citation set
- **getActiveBibliography**: get D's active bibliography set

4.4. *Miscellaneous Methods*

- **getNewDocumentAdditions**: list most recent additions
- **getDocumentText**: get full ASCII text of a document
- **getDocumentAsDC**: get document Dublin Core record

4.5. *Registration and Administrative Methods*

In the perspective of enabling access to CiteSeer-like services on the Semantic Web, the action of registering with the API service is also part of the API.

- **register**: register to CiteSeer-API, e-mails personal key
- **getUserProperty**: get user property
- **setUserProperty**: set user property

5. Accessing CiteSeer-API

As illustrated in Figure 1, CiteSeer-API proposes a new interface to CiteSeer servers which is complementary to the regular web-interface and the OAI-PMH interface. The CiteSeer-API service, which is also HTTP based, is advertised through its WSDL description. The WSDL schema was intentionaly kept simple to ensure compatibility with most WSDL toolkits and users are expected to generate access stubs based on the current WSDL description.

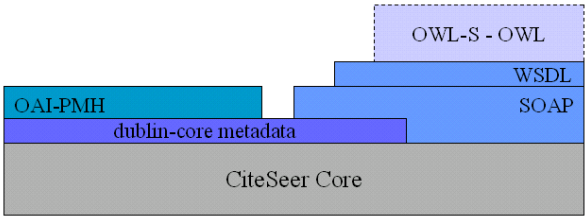

Figure 1: Protocols Stack for CiteSeer servers

6. Interoperability Through CiteSeer-API

CiteSeer-API currently uses CiteSeer's internal – arbitrary - identifiers as values for doc-id, cite-id and group-id, which corresponds to a *no-encoding* value for *enc* (cf. Table 1). Such identifiers, however, do not allow for interactions with other DL systems, including other CiteSeer servers themselves. [2] advocates that resource identifiers should be features that can be computed directly from the original digital resources and should therefore be implemented as digital signatures (e.g. checksums) of these resources. The CiteSeer software libraries makes use of the SHA [9] algorithm in order to identify exact file duplicates and therefore we are extending CiteSeer-API to support Document URIs that use these checksums as doc-id (enc now takes the value *SHA*). Document URIs created from the document checksums will allow heterogenous DL to locate CiteSeer resources while having limited knowledge of the system itself.

7. Usage Scenarios

We envision the following applications for CiteSeer-API : (1) creation of alternative user interfaces to CiteSeer servers; (2) use of CiteSeer corpuses for training and testing purposes; (3) seamless interlinking and cooperation with heterogenous DL systems; (4) Document version control applications; (5) DL mirroring.

8. Related And Future Work

Since its release, many research projects have created alternative interfaces or wrappers to the CiteSeer.org web-site in order to provide alternative visualization of citation-based relationships. We believe that CiteSeer-API will simplify such taks by facilitating the integration of CiteSeer services in third party applications.

To fully leverage the framework presented here, it is desirable to bring it into the context of the Semantic Web. CiteSeer-API is described using WSDL. Although this allows for the automatic generation of code stubs to programmatically access the CiteSeer services, the WSDL description does not carry the semantics of the underlying service. Our next effort will therefore be to apply semantics on CiteSeer-API using OWL [6] and OWL-S [7] ontologogies. It is worth noting that the task performed by CiteSeer-like services - autonomous citation indexing - will require that the OWL-S description of the service be combined with an OWL-encoded syntactic document ontology.

9. Conclusions

We introduced CiteSeer-API, a SOAP/WSDL-based API to CiteSeer-like services. CiteSeer-API was design not only to allow interaction between CiteSeer-like services but also with other DL systems. In this regard, the choice of resource identifiers that stem from the resources themselves is fundamental to ensure the interoperability of CiteSeer services with heterogenous DL systems. While CiteSeer-API turns CiteSeer-like niche search engines into actual web-services, it still requires developers to have an understanding of the service in order to make use of it. The addition of semantics concepts to CiteSeer-API using OWL-S will enable automated agents to discover, register and seamlessly exploit CiteSeer-like services. We encourage research groups to take advantage in their own projects of the functionalities and data available through CiteSeer-API.

10. References

[1]: homepage of CiteSeer-API, http://smealsearch.psu.edu/api

[2] : Crespo, A.; Garcia-Molina, H.. Archival Storage for Digital Libraries, Third ACM Conference on Digital Libraries. Pittsburgh, PA, USA, June 23-26, 1998

[3]: C.L. Giles, K. Bollacker, S. Lawrence, "CiteSeer: An Automatic Citation Indexing System", In *Proceedings of the 3rd ACM Conference on Digital Libraries (DL'98)*, pp 89-98, 1998.

[4]: S. Lawrence, K. Bollacker, C.L. Giles, "Distributed Error Correction", In Proceedings of the 4th ACM Conference on Digital Libraries, p. 232, 1999.

[5]: "The Open Archives Initiative Protocol for Metadata Harvesting", http://www.openarchives.org/OAI/openarchivesprotocol.htm.

[6]: OWL Web Ontology Language Reference, http://www.w3.org/TR/2004/REC-owl-ref-20040210/

[7]: OWL-S , http://www.daml.org/services/owl-s/1.0/

[8]: Y. Petinot, P.B. Teregowda, H. Han, C.L. Giles, S. Lawrence, A. Rangaswamy and N. Pal, "eBizSearch: an OAI-Compliant Digital Library for eBusiness", In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL 2003)*, pp 199-209, Houston (TX), May 2003.

[9]: FIPS 180-1, "Secure Hash Standard", NIST, US Department of Commerce, Washington D.C., Apr. 1995.

[10]: homepage of SMEALSearch, http://smealsearch.psu.edu