

OpenArXiv = arXiv + RDBMS + Web Services*

Justin Fisher
Penn State University
jmf390@psu.edu

Hyunyoung Kil
Penn State University
hkil@psu.edu

Dongwon Lee
Penn State University
dongwon@psu.edu

<http://openarxiv.ist.psu.edu/>

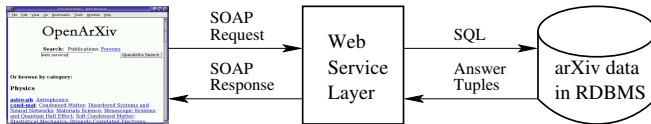


Figure 1: Interactions in OpenArXiv.

Extended Abstract

The *arXiv* (<http://arxiv.org>) is one of the popular scientific digital libraries. Since 1991, it has been the major forum for disseminating scientific results in Physics, Mathematics, Nonlinear Sciences, Computer Science, and Quantitative Biology. Although the number of publications in arXiv is smaller than other scientific digital libraries, since arXiv is self-archived by authors and often contains publications in 3–4 various formats (e.g., ps, pdf, doc, tex), the quality of extracted meta-data is excellent and the amount of the required storage for data and meta-data is substantial.

The *OpenArXiv* project aims to significantly improve the arXiv digital library in two ways: (1) by managing digital documents with an RDBMS and exploiting state-of-the-art database techniques, we add more sophisticated and flexible services, e.g., contents-based search, advanced query processing and triggers technology; and (2) by utilizing the standard XML-based web services framework, we build a programmable interface to arXiv so that not only human users but also software agents can freely access the contents of arXiv in many applications. Note that our API set is a superset of OAI-PMH¹, and thus there is a straightforward lossless mapping from our API set to OAI-PMH.

Note that while arXiv affords flexibility in allowing authors to provide their own metadata for the publication, this can result in less uniformity in terms of naming or punctuation conventions. As a result, when the raw data of arXiv

*The project is partially supported by Microsoft Scientific Data Intensive Computing (SciData) Award, 2005.

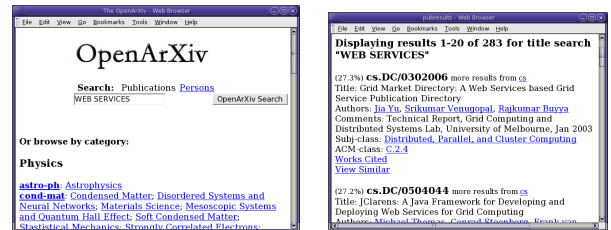
¹<http://www.openarchives.org/OAI/2.0/>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...\$5.00.

API signature	Description
<code>int HowManyACMClasses (string ClassType)</code>	Return the number of publications with the <code>ClassType</code>
<code>ArrayList getAllAuthors (string pubid)</code>	Return the names of all authors under whom a publication identified by <code>pubid</code> was submitted

Table 1: OpenArXiv web service API examples.



(a) Search

(b) Result

Figure 2: OpenArXiv client example.

were gathered, some basic cleaning was necessary.

Figure 1 illustrates the interactions between components of OpenArXiv. The underlying RDBMS (MS SQL Server 2000) contains tables for publications, persons, document, subject and field-specific classification (e.g., ACM and MSC) classes. The web service layer contains a collection of APIs, initially mined by tools like [1], encapsulates every functionality provided by the arXiv web site – notably *browse* and *search*. When a software agent invokes an API function as a SOAP request, the web service layer issues a SQL command correspondent to the function and gives back the result which the DBMS returned. Table 1 shows examples of these APIs. Figure 2 shows the screen-shots of “search” and its results on a client program that mimics the look and feel of the arXiv site.

Furthermore, by exploiting advanced functionalities of an RDBMS, the OpenArXiv easily supports several useful features such as full-text approximate matching with relevance ranking, as shown below:

```

SELECT pub.*, ft.*
FROM pub INNER JOIN FREETEXTTABLE(pub, title, 'X') ft
ON pub.pubid = ft.[Key]
ORDER BY ft.rank DESC
  
```

1. REFERENCES

- [1] Y.-H. Lu, Y. Hong, J. Varia, and D. Lee. “Pollock: Automatic Generation of Virtual Web Services from Web Sites”. In *ACM Symp. on Applied Computing (SAC)*, Santa Fe, NM, Mar. 2005.