Fu, Y. , Mostafa, J. (2004). Integration of Biomedical Text and Sequence OAI Repositories. Proceedings of IEEE/ACM Joint Conference on Digital Libraries 2004, Tuscon, Arizona

Integration of Biomedical Text and Sequence OAI Repositories

Yueyu Fu
Laboratory of Applied Informatics Research
Indiana University, Bloomington
IN, 47405-3907
(812)856-4182, 01
yufu@indiana.edu

Javed Mostafa
Laboratory of Applied Informatics Research
Indiana University, Bloomington
IN, 47405-3907
(812)856-4182, 01
jm@indiana.edu

ABSTRACT
Archived biomedical literature and sequence data are growing rapidly. The Open Archives Initiative's Protocol for Metadata Harvesting (OAI-PMH) [1] provides a convenient way for data sharing. But it has not been tested in the biomedical domain, especially in dealing with different types of data, such as protein, and gene sequences. We built four individual OAI-PMH repositories based on different biomedical resources. Using the harvested data from the four repositories we created an integrated OAI-PMH repository, which hosts the linked literature and sequence data in a single place.

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *collection, dissemination, standards, system issues, user issues.*

General Terms
Design, Experimentation, Standardization

Keywords
Open Archive Initiative, Metadata, Interoperability

1. INTRODUCTION
Literature and sequence data are both very important in the biomedical domain. Archiving this information in OAI-PMH repositories and also integrating literature and sequence information will greatly help biomedical researchers in developing new research hypotheses and experiments. To achieve this, it's necessary to extend OAI-PMH [3]. We created our own metadata schemas for the gene and protein sequence data. As data providers, four individual OAI-PMH repositories were built using existing tools, such as the OAICAT. As a service provider, using the harvested data from the above repositories we created an integrated OAI-PMH repository. This integrated repository has

the metadata of the literature and the sequence data and links the two data sets in a uniform OAI-PMH compliant format.

OAI-PMH has not been tested in the biomedical domain especially in dealing with different types of data, such as protein and gene sequences. Our main goal for this project was to create and integrate different biomedical resources using OAI-PMH. We concentrated on developing repositories for four different resources: Medline for biomedical literature, Refseq for gene DNA sequence, Refseqp for protein sequence and Swissprot for protein sequence. Another goal was to develop an integrated repository that harvests the four repositories using OAI-PMH and incorporates a search interface that permits searching of the integrated repository.

The paper will briefly outline the architecture of the four repositories and the integrated repository. It will describe how the repositories can be harvested and it will describe the search function of the integrated repository. The paper will conclude with examples of searches that can be conducted on the integrated repository.

2. ARCHITECTURE
The overall architecture of the repositories is shown below (see Figure 1).

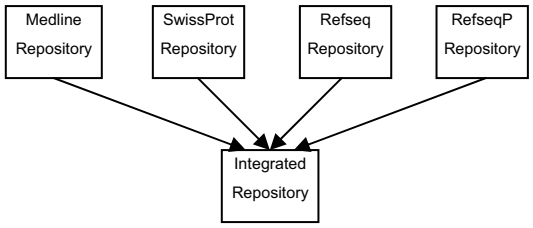

Figure 1. Architecture of the repositories

To build an OAI repository for MEDLINE records, 50557 MEDLINE XML records were downloaded from the NLM site. The MEDLINE fields were mapped to unqualified Dublin Core format and subsequently stored in an OAI-PMH repository. Each OAI item represents a MEDLINE record and its OAI identifier is the corresponding MEDLINE id. Only the default OAI metadata format, oai_dc, is available for each OAI item.

From the NCBI site, 4032 RefSeq records linked from our MEDLINE subset and that contain gene sequences were downloaded. The standard Dublin Core format is not suitable for RefSeq sequence data. Hence, we created a simple RefSeq XML

schema for the RefSeq OAI repository [2]. Two OAI metadata formats are provided for each OAI item:

- refseq: contains the refseq records in our refseq XML format.
- oai_dc: contains only the accession id in the title field to satisfy the mandatory requirement of OAI [1].

Also, 2072 Refseq records linked from our MEDLINE subset and that contain protein sequences were downloaded. A simple RefseqP XML schema was created for the RefSeqP OAI repository. Two OAI metadata formats are provided for each OAI item:

- refseqp: contains the refseq records in our refseqp XML format.
- oai_dc: contains only the accession id in the title field to satisfy the mandatory requirement of OAI.

Our SwissProt dataset includes 1503 SwissProt records linked from our MEDLINE subset and that contain protein sequences. We created a simple SwissProt XML schema for the SwissProt OAI repository. Two OAI metadata formats are provided for each OAI item:

- swissprot: contains the swissprot records in our swissprot XML format.
- oai_dc: contains only the accession id in the title field to satisfy the mandatory requirement of OAI.

3. REPOSITORY INTEGRATION

All the above four OAI repositories can be harvested using OAI-PMH protocol. To build the integrated OAI repository, those four repositories were harvested separately. The harvested results were automatically merged into a single OAI repository by an application we implemented using Java. An XML schema was created for the integrated repository. Two OAI metadata formats are available for each OAI item:

- enable_oai: contains the integrated records in our enriched XML format (see Figure 2).
- oai_dc: contains the MEDLINE fields to satisfy the mandatory requirement of OAI.

```
<record>
<title>Cytoglobin: a novel globin type … in vertebrate tissues.</title>
<creator>Thorsten Burmester</creator>
<subject>Amino Acid Sequence</subject>
<description>Vertebrates possess multiple … needs of muscle cells.</description>
<publisher> Mol Biol Evol</publisher>
<date>2000-Apr</date>
<type>Journal Article</type>
<format>XML</format>
<identifier>MedlineID: 21918419</identifier>
<identifier>PMID: 11919282</identifier>
<language>eng</language>
<swissprotid>Q8UUR3</swissprot>
<refseqid>NM_152952</refseqid>
<refseqpid>NP_694484</refseqpid>
</record>
```

Figure 2. A enable_oai record from the integrated repository.

4. INTERACTIONS

The integrated repository can be accessed in the following three ways:

- Interaction by OAI repository administrators via a web browser interface. Any service providers interested in harvesting this repository can explore it at its OAI base URL http://tara.slis.indiana.edu:8080/oaicat/.
- Harvested by OAI harvesters based on OAI-PMH protocol.
- Interaction by users via a web search interface (see Figure 3).

The search interface of the integrated repository provides basic search functions for title, creator, subject, publication date, MEDLINE id, SwissProt id, RefSeq id, and RefSeqP id. Users can search on any combination of the fields. The search results include the full record retrieved.

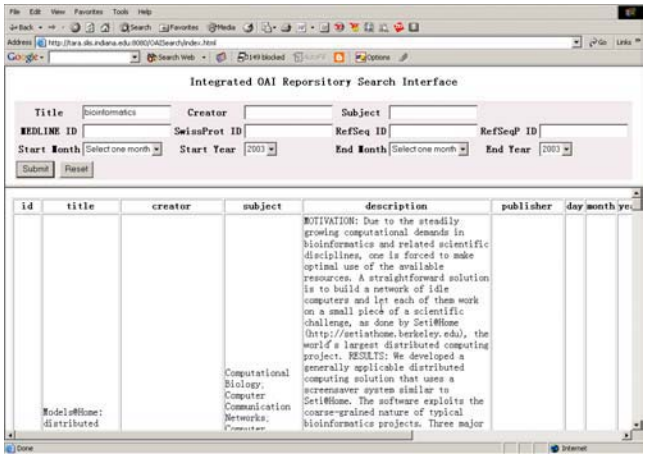

Figure 3. Search functions for the integrated repository

5. CONCLUSION

Our system demonstrates that OAI-PMH can be successfully applied in the biomedical domain and different data resources can be integrated into a uniform format. In addition, search functions can be supported for accessing the integrated repository.

6. ACKNOWLEDGMENTS

This work was partially supported through a grant from the National Science Foundation Award#:0333623.

7. REFERENCES

[1] Lagoze, C., Sompel, H. V., Nelson, M., and Warner, S. *The Open Achieves Initiative Protocol for Metadata Harvesting-Version 2*. 2002.

< http://www.openarchives.org/OAI/openarchivesprotocol.html>

[2] Sompel, H. V., Lagoze, C., Nelson, M., and Warner, S. *Implementation Guidelines for the Open Achieves Initiative Protocol for Metadata Harvesting*. 2002.

[3] Sompel, H. V., Young, J. A., and Hickey, T. B. Using the OAI-PMH Differently. *D-Lib Magazine*, 9(7/8), 2003.