OAI-PMH: Open Archives Initiative Protocol for Metadata Harvesting

T.B. Rajashekar
National Centre for Science Information (NCSI)
Indian Institute of Science, Bangalore 560 012
(E-Mail: raja@ncsi.iisc.ernet.in)

Prepared for presentation in the Workshop on Open Access, MSSRF, Chennai, 2-4 May 2004

NCSI, IISc

Acknowledgements

- In preparing this presentation, I have used material from several presentations on OAI-PMH by other authors
- I gratefully acknowledge these sources

Digital Repositories: Current Situation

- Mushrooming number and variety of distributed digital repositories (~ archives, digital libraries)
- Use variety of hardware, software, database solutions
- Different search and retrieval interfaces
- Most of the content not indexed by web search engines
- Content resides in backend databases – not picked up by web search engines

T.B. Rajashekar
NCSI, IISc

Problems faced by Users

- How users identify and retrieve relevant information from different repositories?
- Visiting and searching individual repositories is very expensive
- Key Requirement: How do we support cross searching?

T.B. Rajashekar
NCSI, IISc

Current Solutions

- Federated/ distributed searching
  - Z39.50 IR protocol
- Metadata harvesting
  - OAI-PMH protocol

What is a protocol?

A protocol is a set of rules defining communication between systems. FTP (File Transfer Protocol) and HTTP (Hypertext Transport Protocol) are examples of protocols used for communication between systems across the Internet.

T.B. Rajashekar
NCSI, IISc

Federated/ distributed searching

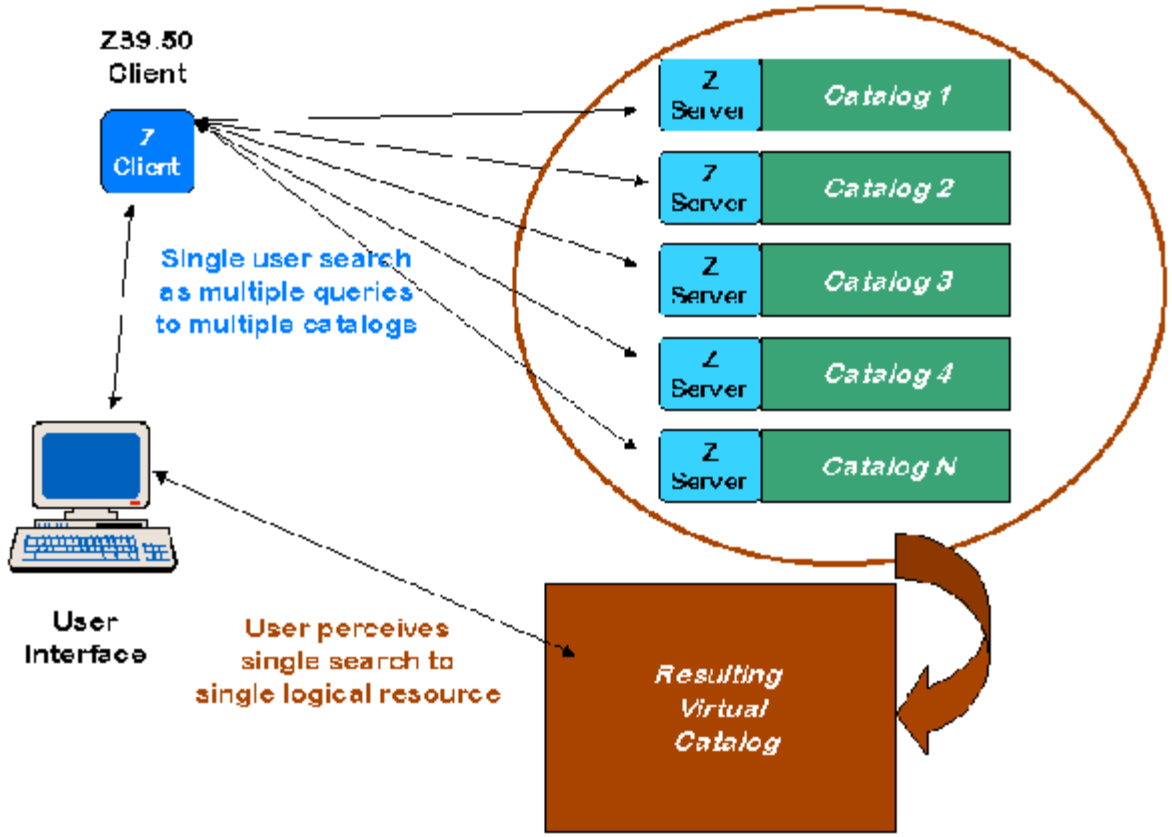
- Protocol: "Information Retrieval (Z39.50): Application Service Definition and Protocol Specification", (ISO/ ANSI standard) (v1-1991, v2-1992, v3-1995)
- Client-Server model (TCP/IP Service)
- Process:
  - Client (‘Origin’) sends queries, formatted according to Z39.50, to repository Server (“Target”).
  - Server translates this to local query format, searches the database, sends the results to the client, formatted according to Z39.50
  - Client translates the results and presents it to the user
- Client can send queries to as many related z39.50 compliant servers as possible

T.B. Rajashekar
NCSI, IISc

Z39.50 protocol …

- Example implementation: Distributed searching of library catalogues/ bibliographic databases
- Problem - performance
  - Implementation not easy
  - Does not scale well (if nodes > 100)
  - Network bandwidth
  - Z39.50 implementation at client (“Origin’) end
- Z30.50 resources: http://lcweb.loc.gov/z3950/agency/ (Z39.50 International Maintenance Agency, Library of Congress)

T.B. Rajashekar
NCSI, IISc



T.B. Rajashekar
NCSI, IISc

Metadata Harvesting Protocol

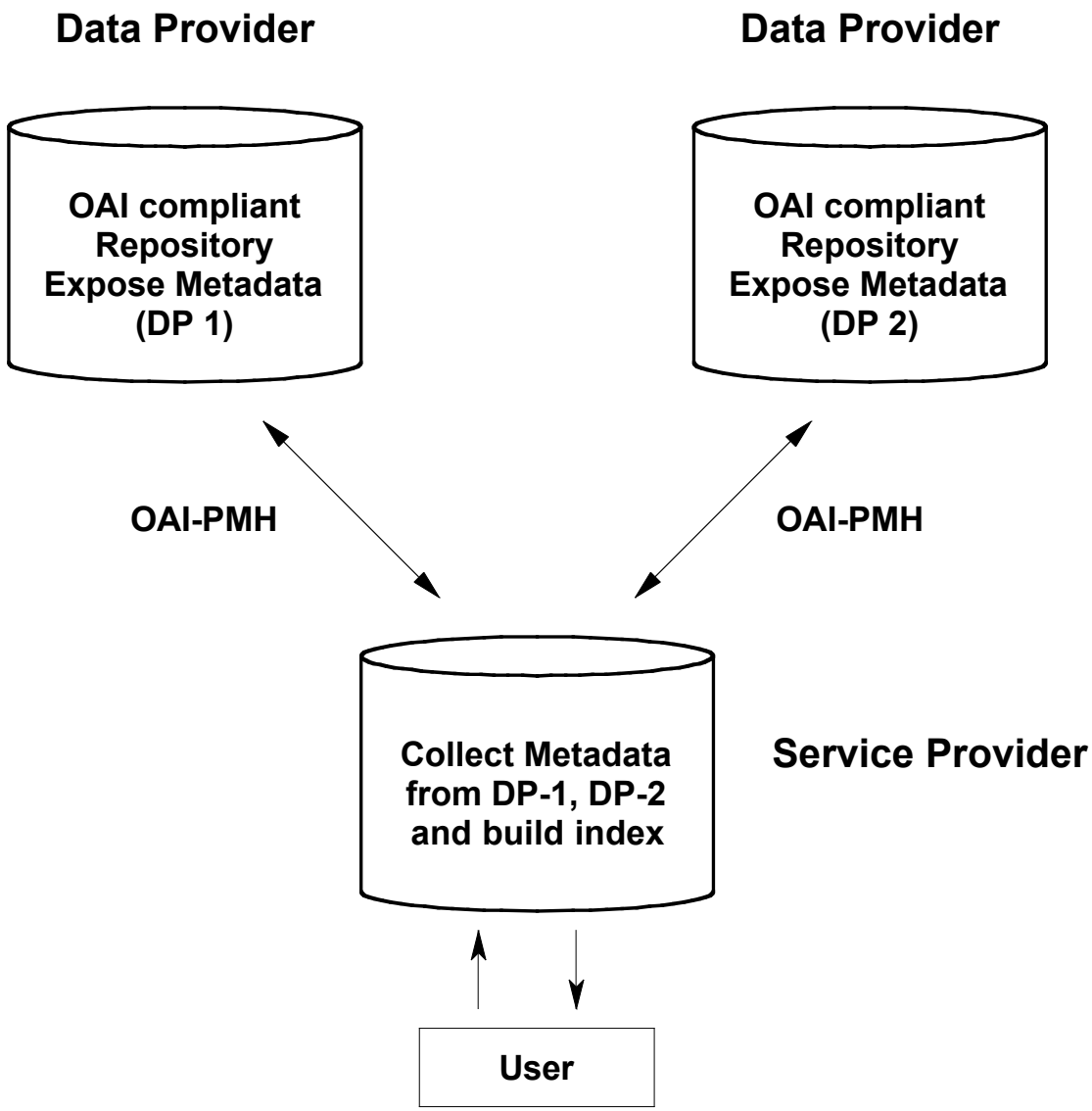
- Protocol: OAI-PMH: Open Archives Initiative Protocol for Metadata Harvesting
- OAI (Open Archives Initiative)
  - OAI is an initiative to develop and promote interoperability standards that aim to facilitate the efficient dissemination of content. (http://www.openarchives.org/)
- Lightweight harvesting protocol for sharing metadata between services
- Defines a mechanism for harvesting XML-formatted metadata from repositories
- Two key players: Data Providers and Service Providers

T.B. Rajashekar
NCSI, IISc

OAI-PMH Protocol…

- Data Provider
  - maintains one or more repositories (web servers) that support the OAI-PMH as a means of exposing metadata.
  - respond to OAI-PMH queries over HTTP, and deliver metadata in XML format
  - OAI-PMH compliance
- Service Provider
  - issues OAI-PMH requests over HTTP to data providers and uses the metadata as a basis for building value-added services (e.g. central indexing and searching)
- Users
  - Search the central metadata index at the service provider, browse metadata and obtain full document from individual repository
  - No need to install any software

Repository interoperability through OAI protocol



T.B. Rajashekar
NCSI, IISc

OAI-PMH Protocol…

- Harvesting
  - in the OAI context, harvesting refers specifically to the gathering together of metadata from a number of distributed repositories (e.g. eprint archives) into a combined data store

T.B. Rajashekar
NCSI, IISc

OAI-PMH: Brief History

- Santa Fe convention – July 1999 – call for single search interface to different archives (Ginsparg, Luce and Sompel)
- Creation of UPS [Universal Preprint Service] – October 1999 – metadata harvesting
- UPS name changed to OAI
- OAI-PMH V. 1.0 [01/2001]
- OAI-PMH V. 2.0 [06/2002]

T.B. Rajashekar
NCSI, IISc

T.B. Rajashekar
NCSI, IISc

What's in the Name

Open

The protocol is openly documented, and is compliant with open Standards – HTTP, DC and XML

Archives

Archive/Repository - contains collection of document-like objects

Initiative

OAI is happening at break-neck speed

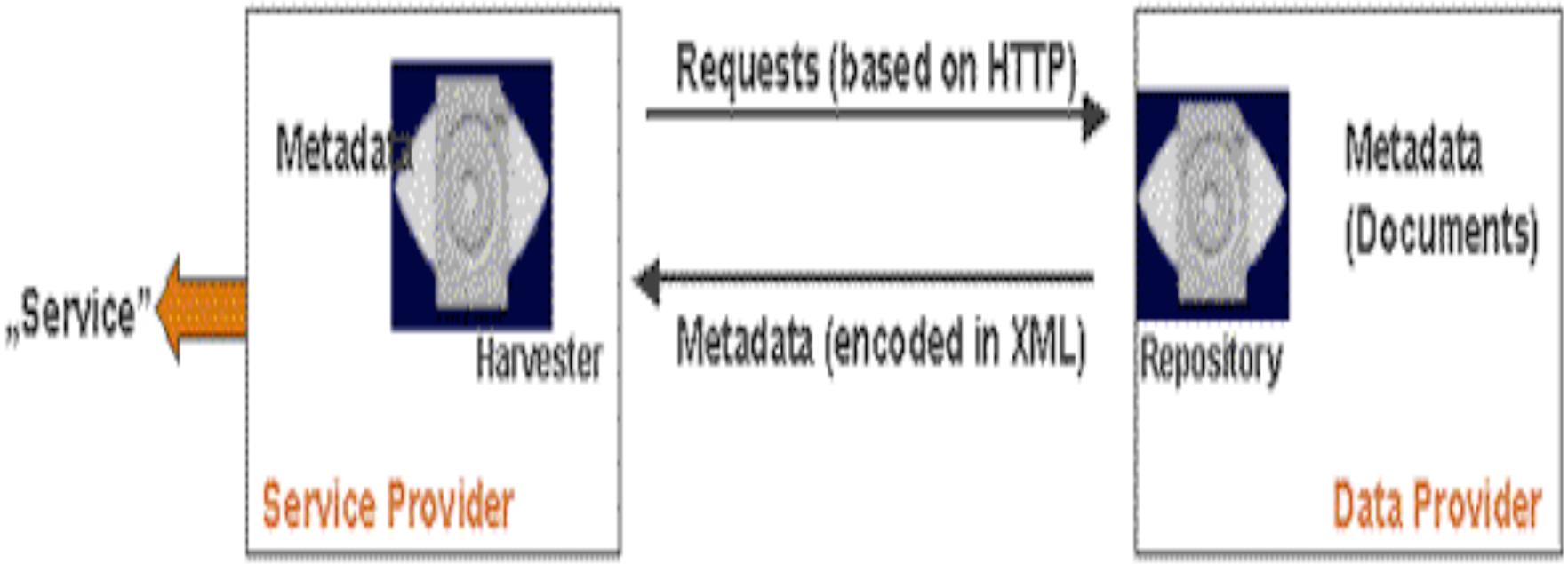
T.B. Rajashekar
NCSI, IISc

OAI-PMH v.2.0 [06/2002]

- Low-barrier interoperability specification
- Metadata harvesting model: data provider / service provider
- Metadata about resources
- HTTP based
- XML responses
- Unqualified Dublin Core
- Stable; No backward compatibility
- Future releases will be backward compatible

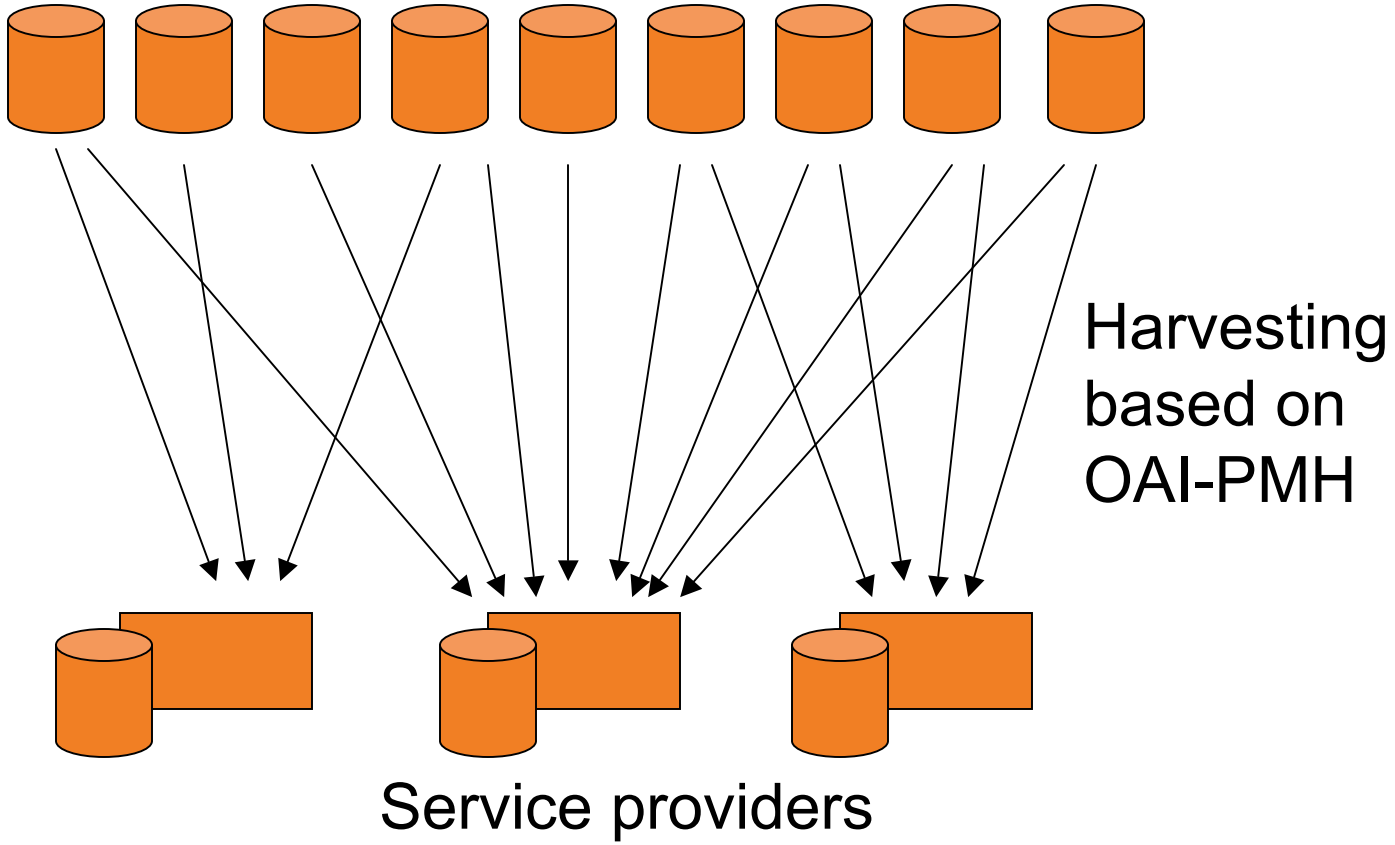
T.B. Rajashekar
NCSI, IISc

Basic Functioning of OAI-PMH



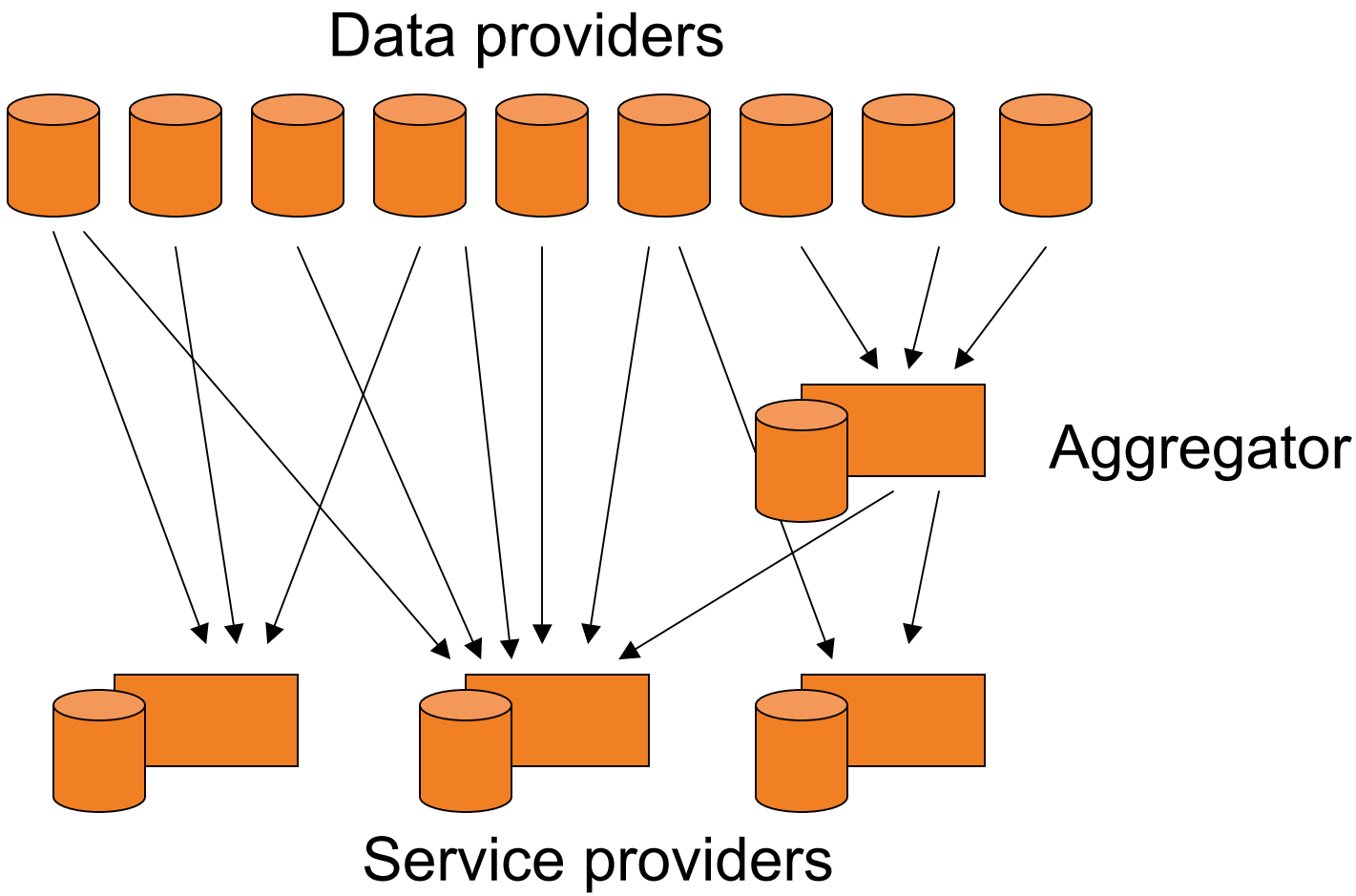
T.B. Rajashekar
NCSI, IISc

Multiple data and service providers



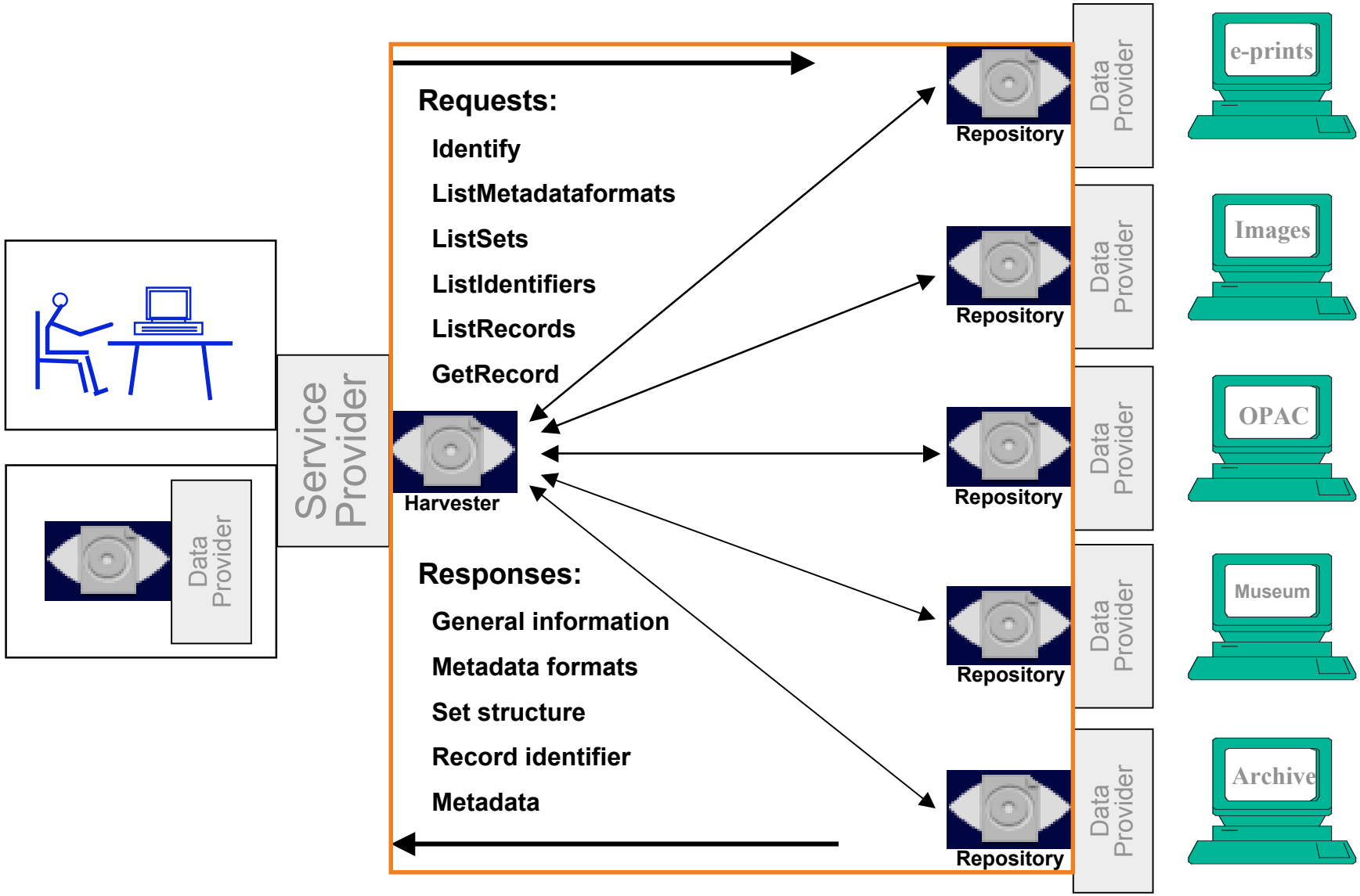
T.B. Rajashekar
NCSI, IISc

Aggregators



T.B. Rajashekar
NCSI, IISc

OAI-PMH: Structure Model



T.B. Rajashekar
NCSI, IISc

OAI-PMH: Protocol Overview

- Protocol is based on HTTP
- Request arguments are issued as GET or POST methods
- Responses are encoded in XML syntax
- Supports any metadata format (at least: Dublin Core)

T.B. Rajashekar
NCSI, IISc

OAI-PMH: Protocol Overview…

- Data providers may support granularity for service providers for selective harvesting:
  - Define a logical set hierarchy
  - Date stamps (last change of metadata set)
- Error messages are http based
- Supports flow control
- Supports six request types (known as ‘verbs’)
  - e.g. **http://archive.org?verb=ListRecords&metadataformat=oai_dc&from=2002-11-01**

T.B. Rajashekar
NCSI, IISc

Protocol Details: Definitions

- Harvester
  - client application issuing OAI-PMH requests
- Repository
  - network accessible server, able to process OAI-PMH requests correctly
- Resource
  - object the metadata is “about”, nature of resources is not defined in the OAI-PMH
- Item
  - component of a repository from which metadata about a resource can be disseminated
  - has a unique identifier

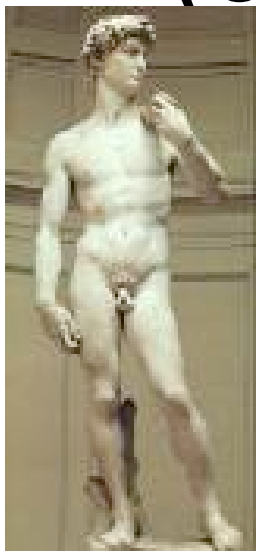
T.B. Rajashekar
NCSI, IISc

Protocol Details: Definitions (2)

- **Record**
  - metadata in a specific metadata format
- **Identifier**
  - unique key for an item in a repository
- **Set**
  - optional construct for grouping items in a repository

T.B. Rajashekar
NCSI, IISc

Protocol Details: Definitions (3)



resource

item = identifier

Metadata about *David*

item

Dublin Core metadata

MARC metadata

SPECTRUM metadata

record

T.B. Rajashekar
NCSI, IISc

Uniqueness and Persistence

- Each record must be uniquely addressable by a distinct identifier
  - (identifier + metadataPrefix)
- Each metadata entity should ideally be persistent to guarantee that service providers can always refer back to the source.

T.B. Rajashekar
NCSI, IISc

OAI Verbs (Request Types)

- Six different request types
  - Identify
  - ListSets
  - ListMetadataFormats
  - ListIdentifiers
  - GetRecord
  - ListRecords

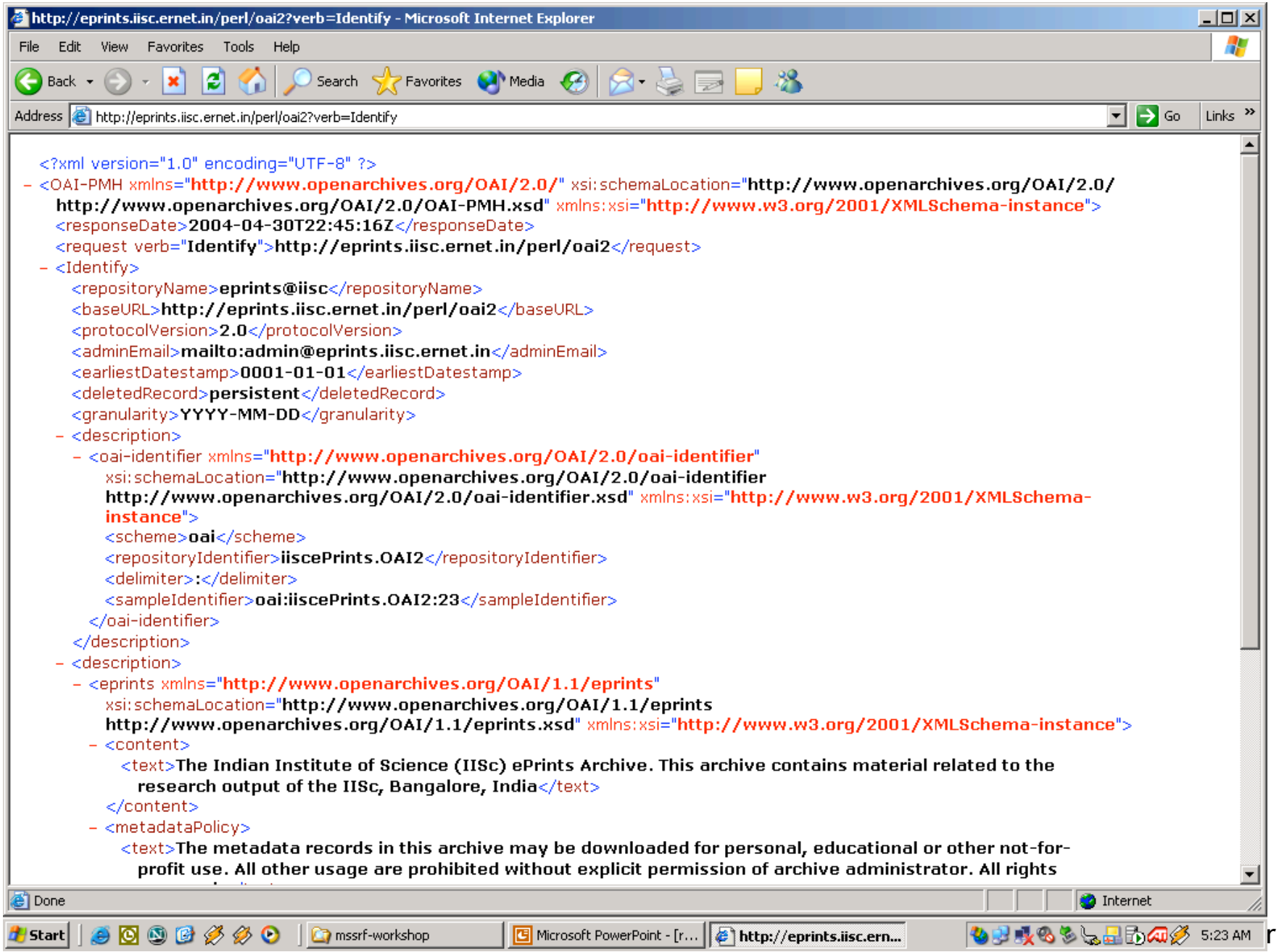
T.B. Rajashekar
NCSI, IISc

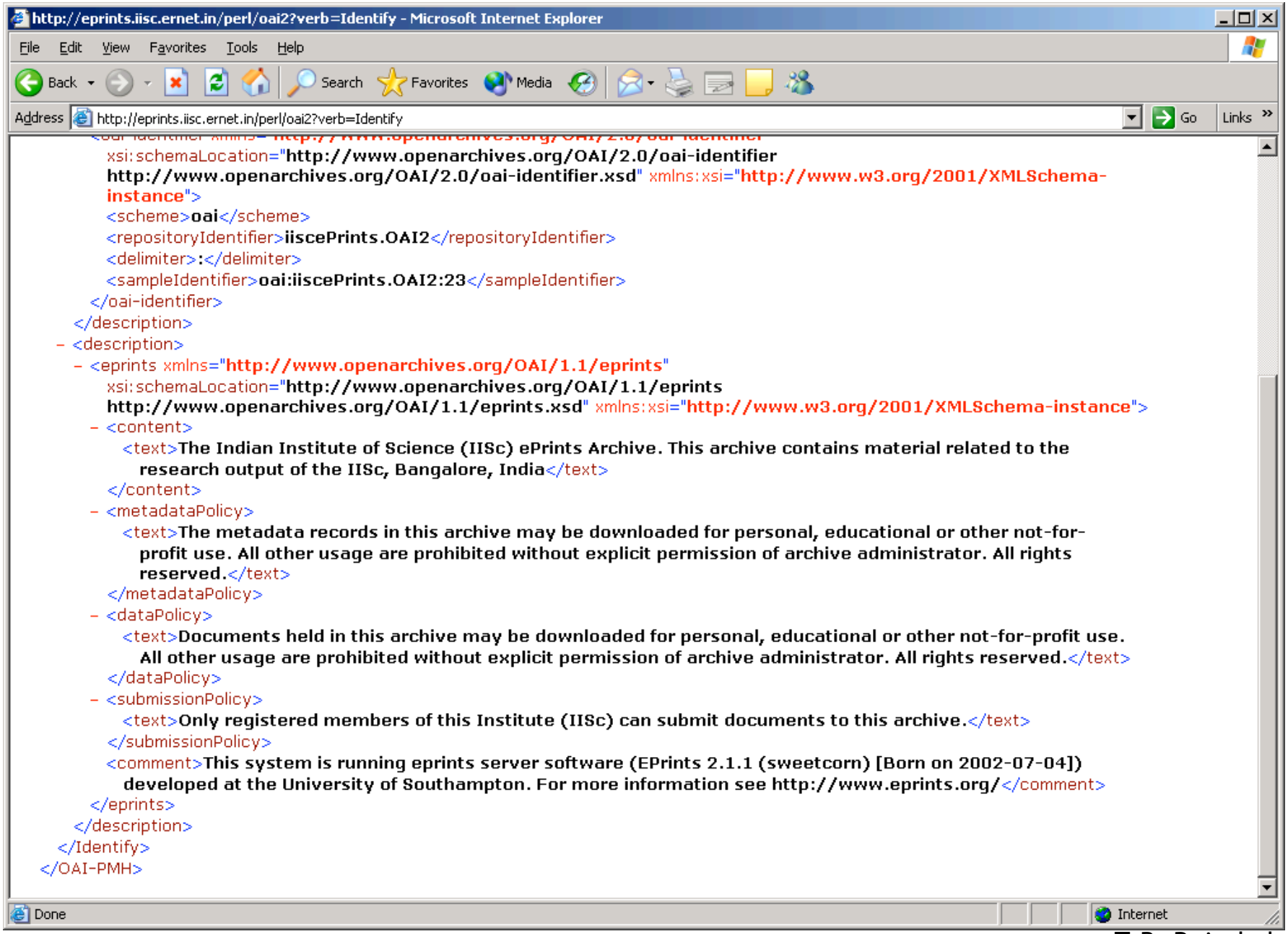
OAI Verbs - Identify

- Purpose
  - Return general information about the archive and its policies (e.g., date stamp granularity)
- Parameters
  - None
- Sample URL
  - http://eprints.iisc.ernet.in/perl/oai2?verb=Identify

T.B. Rajashekar
NCSI, IISc

Identify Request





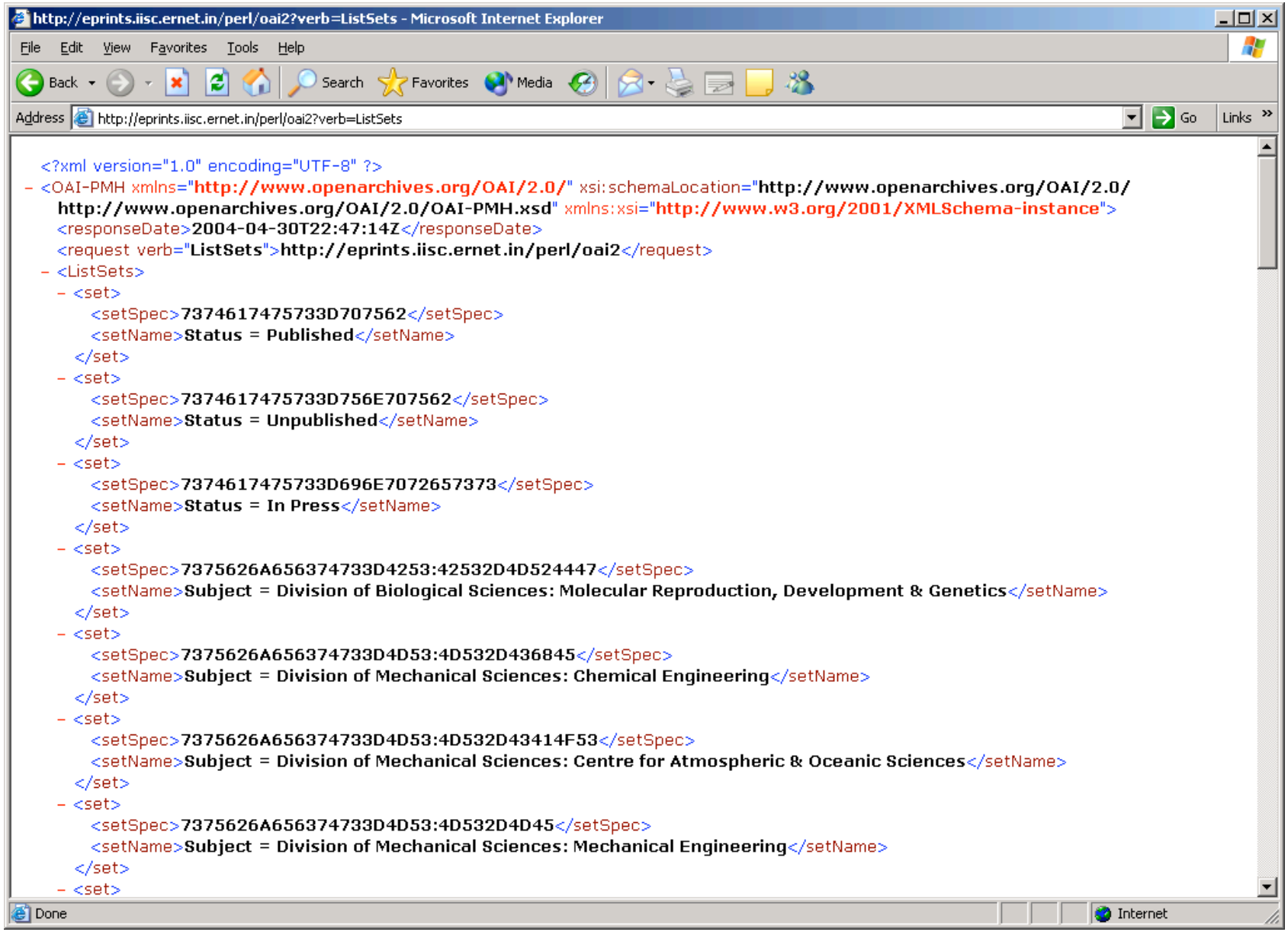
T.B. Rajashekar
NCSI, IISc

OAI Verbs - ListSets

- Purpose
  - Provide a listing of sets in which records may be organized
- Parameters
  - None
- Sample URL
  - http://eprints.iisc.ernet.in/perl/oai2?verb=ListSets

T.B. Rajashekar
NCSI, IISc

ListSets Request



T.B. Rajashekar
NCSI, IISc

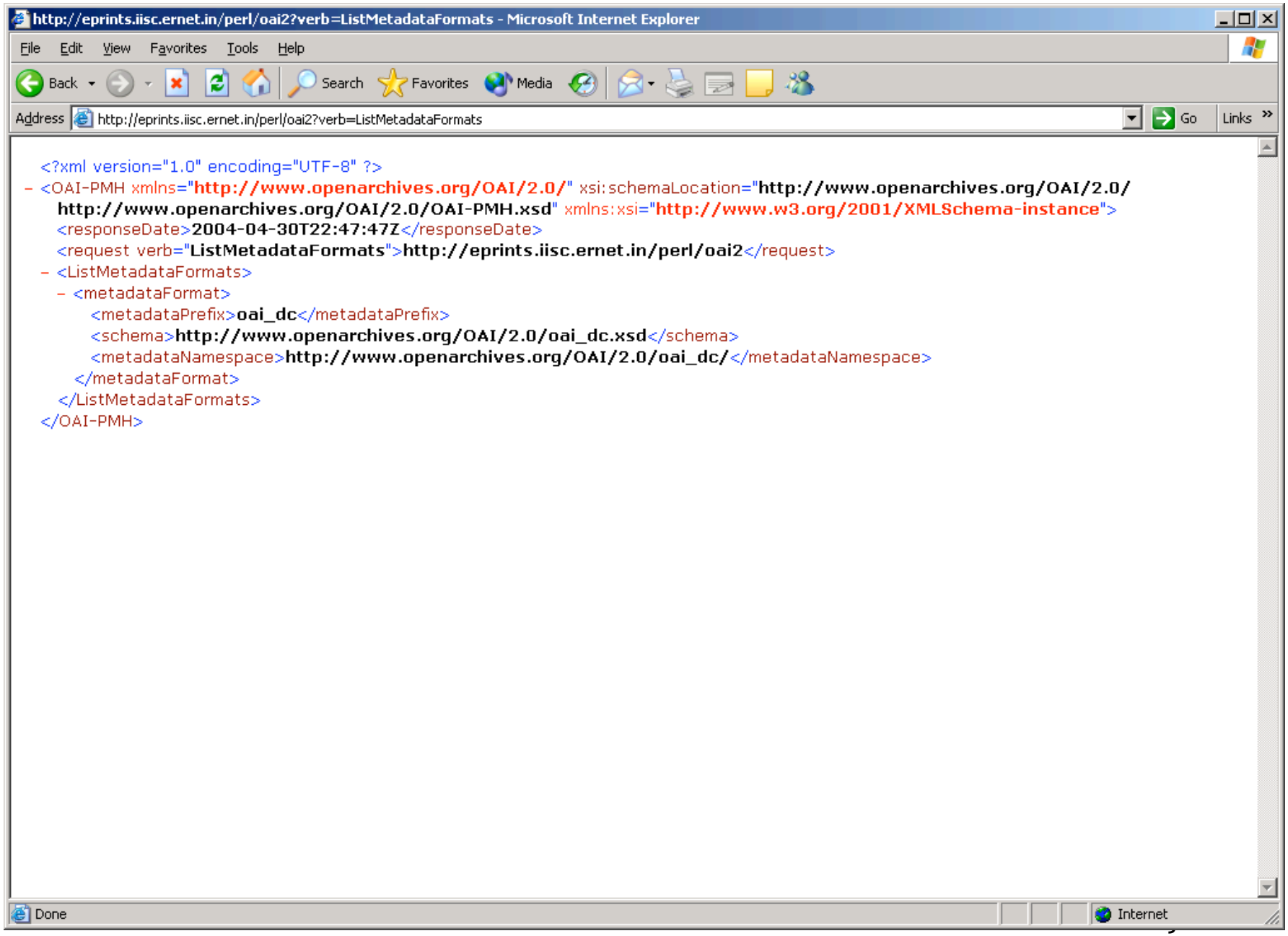
OAI Verbs - ListMetadataFormats

- Purpose
  - List metadata formats supported by the archive as well as their schema locations and namespaces
- Parameters
  - identifier – for a specific record (O)
- Sample URL

  http://eprints.iisc.ernet.in/perl/oai2?verb=ListMetadataFormats

T.B. Rajashekar
NCSI, IISc

ListMetadataFormats Request

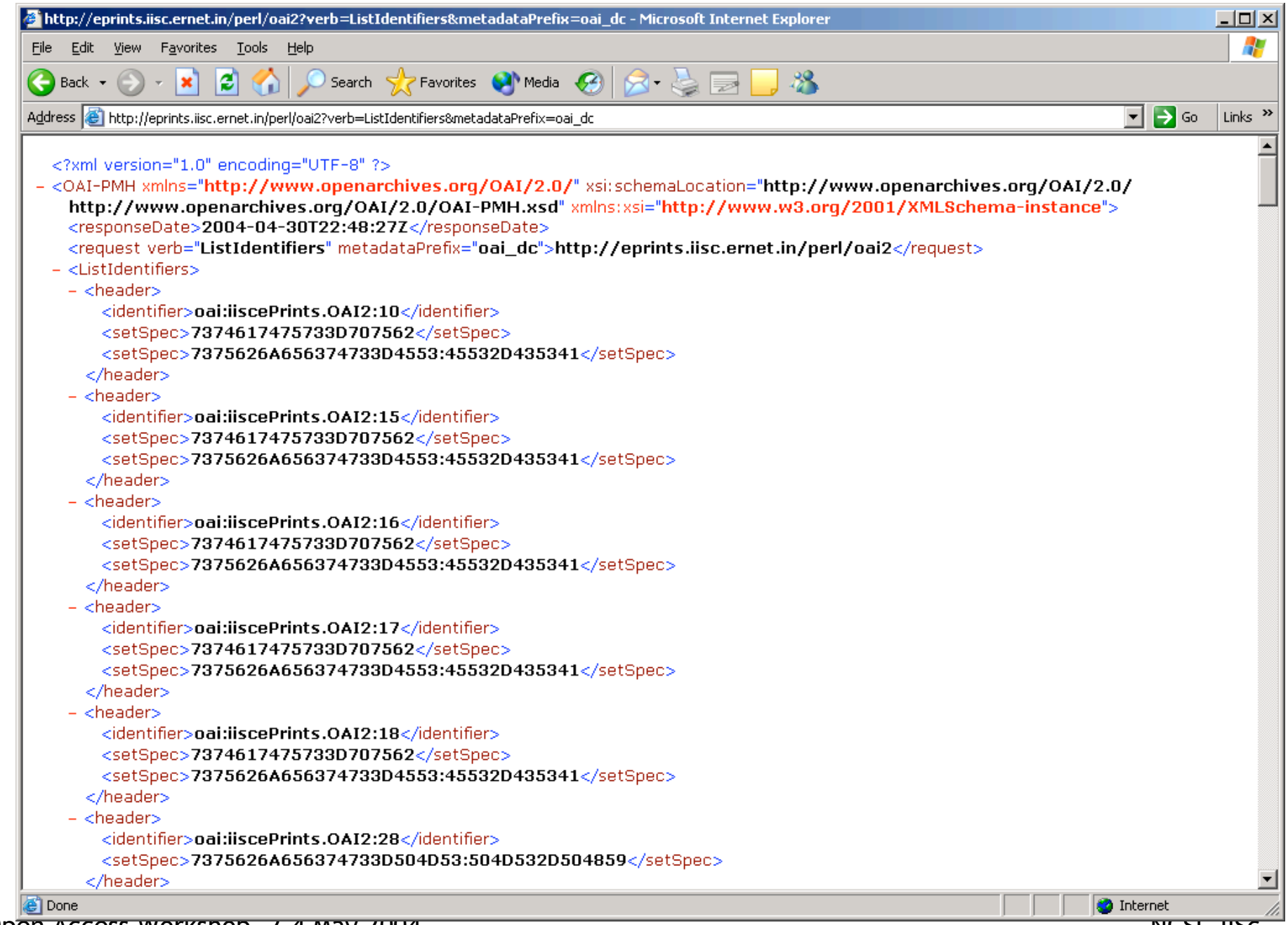


OAI Verbs - ListIdentifiers

- Purpose
  - List headers for all items corresponding to the specified parameters
- Parameters
  - from – start date (O)
  - until – end date (O)
  - set – set to harvest from (O)
  - metadataPrefix – metadata format to list identifiers for (R)
  - resumptionToken – flow control mechanism (X)
- Sample URL
  - http://eprints.iisc.ernet.in/perl/oai2?verb=ListIdentifiers&metadataPrefix=oai_dc

T.B. Rajashekar
NCSI, IISc

ListIdentifiers Request

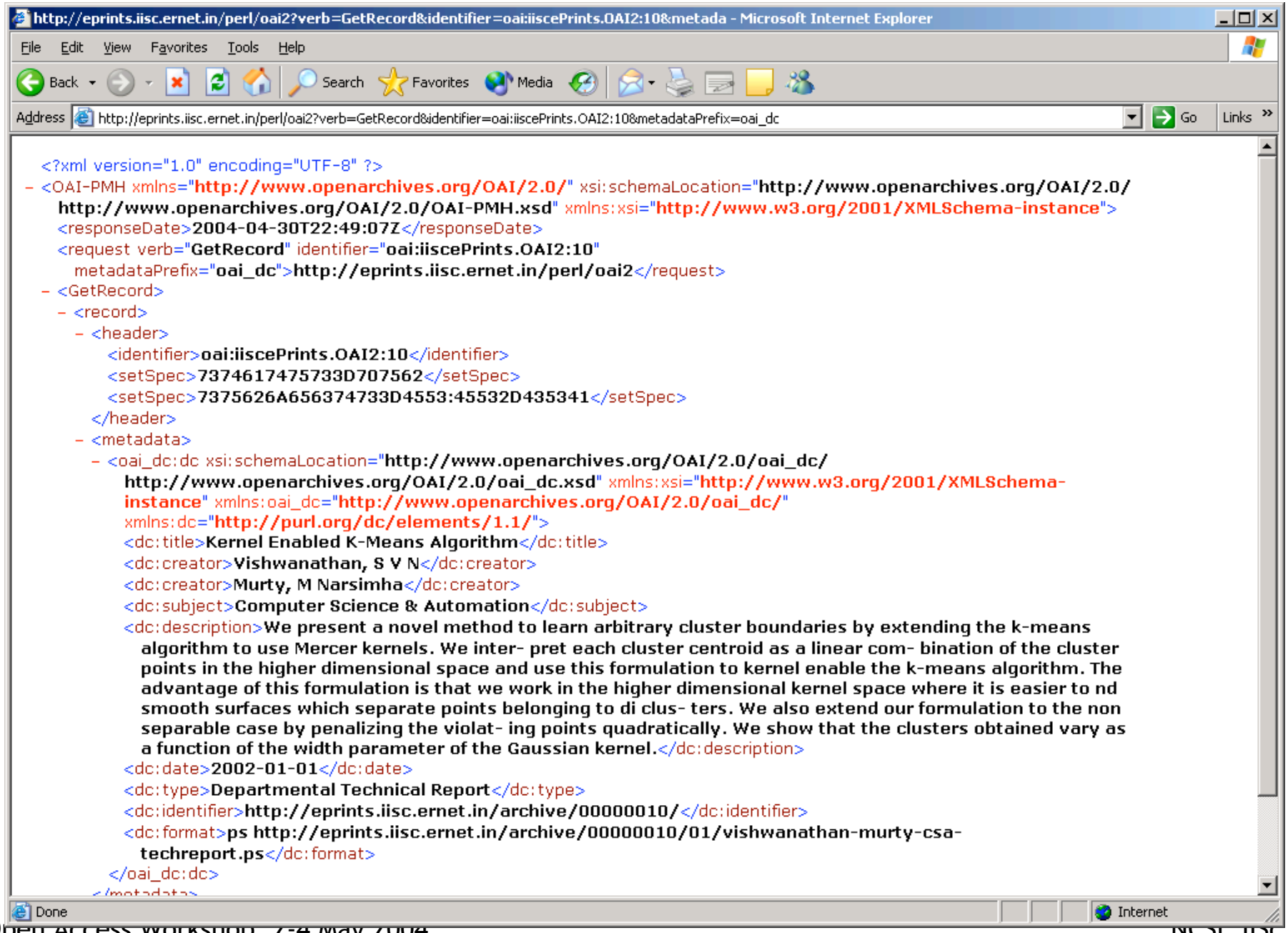


OAI Verbs - GetRecord

- Purpose
  - Returns the metadata for a single item in the form of an OAI record
- Parameters
  - identifier – unique id for item (R)
  - metadataPrefix – metadata format for the record (R)
- Sample URL
  - http://eprints.iisc.ernet.in/perl/oai2?verb=GetRecord&identifier=oai:iiscePrints.OAI2:10&metadataPrefix=oai_dc

T.B. Rajashekar
NCSI, IISc

GetRecord Request

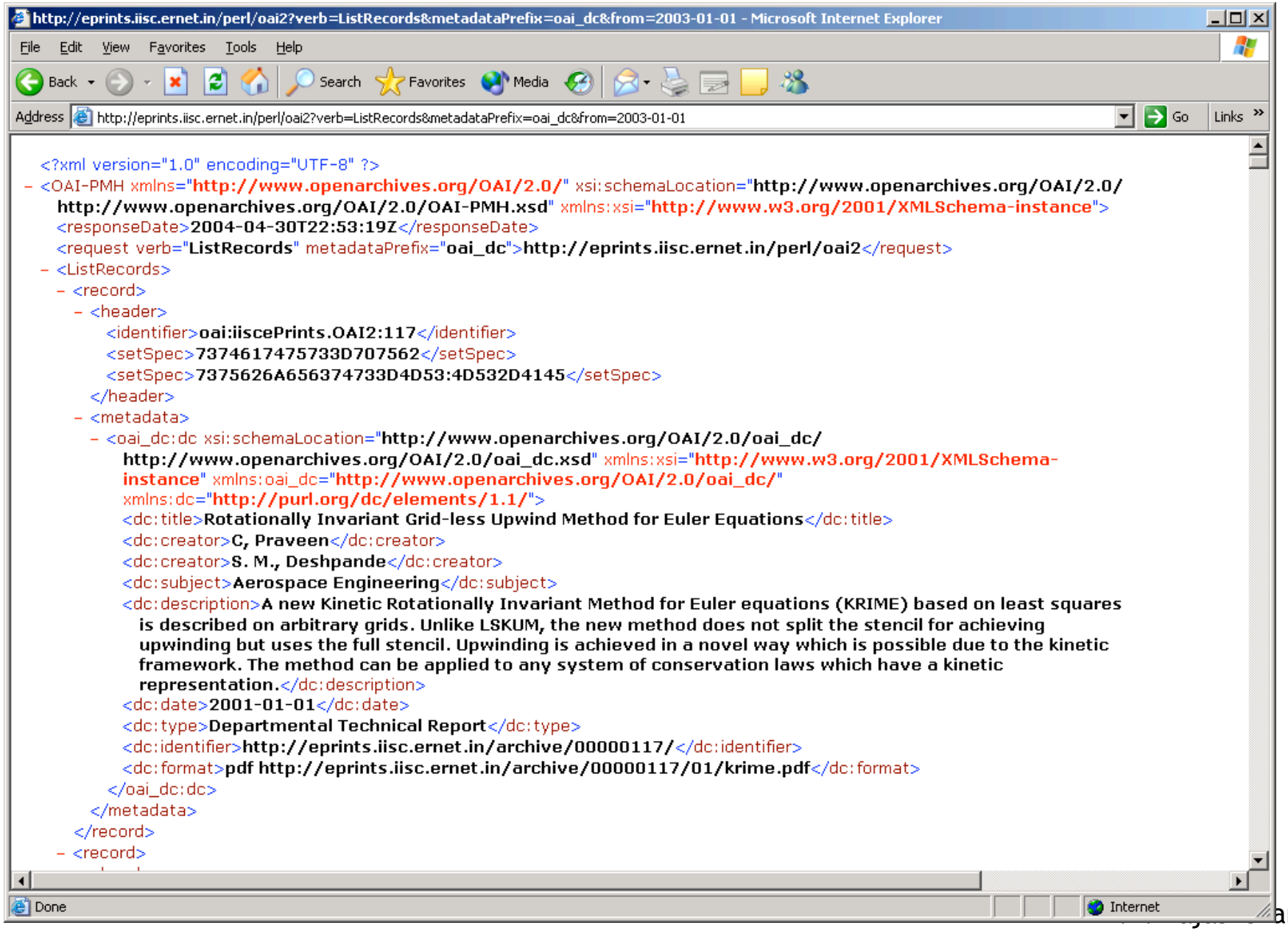


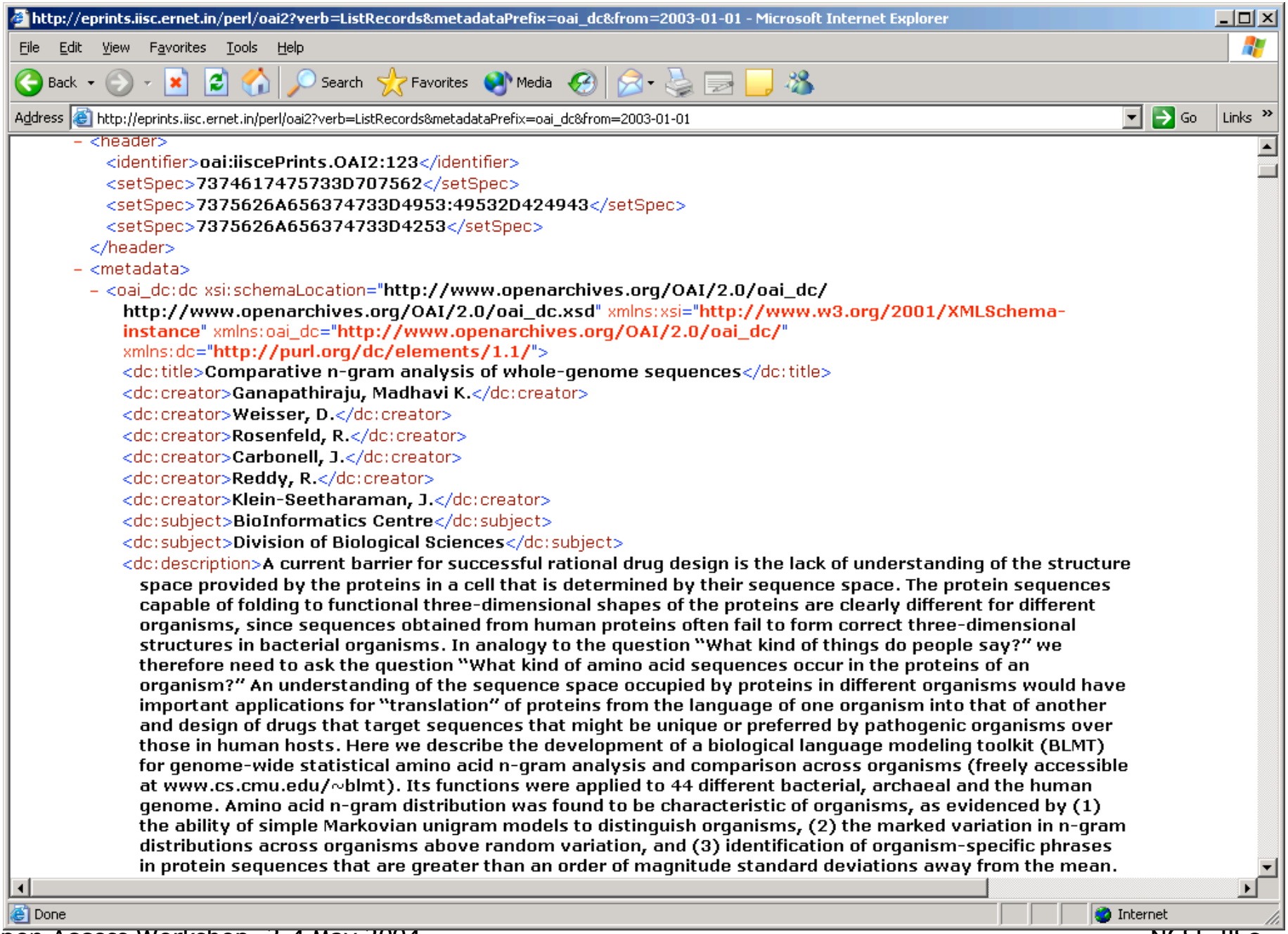
OAI Verbs - ListRecords

- Purpose
  - Retrieves metadata records for multiple items
- Parameters
  - from – start date (O)
  - until – end date (O)
  - set – set to harvest from (O)
  - resumptionToken – flow control mechanism (X)
  - metadataPrefix – metadata format (R)
- Sample URL
  - http://www.anarchive.org/cgi-bin/OAI?verb=ListRecords&metadataPrefix=oai_dc&from=2003-01-01

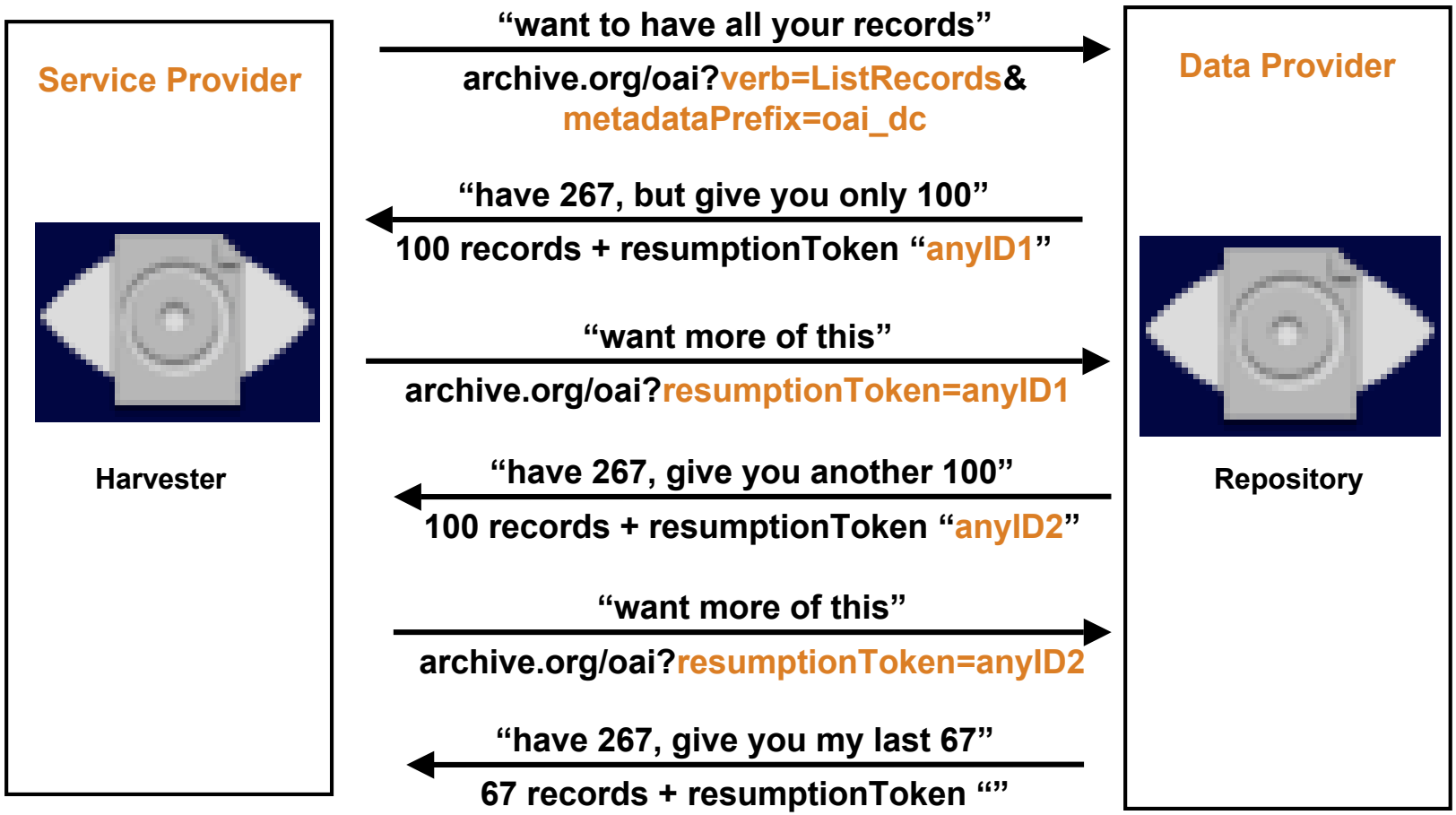
T.B. Rajashekar
NCSI, IISc

ListRecords Request




http://eprints.iisc.ernet.in/perl/oai2?verb=ListRecords&metadataPrefix=oai_dc&from=2003-01-01 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://eprints.iisc.ernet.in/perl/oai2?verb=ListRecords&metadataPrefix=oai_dc&from=2003-01-01
Go Links
- <header>
<identifier>oai:iiscePrints.OAI2:123</identifier>
<setSpec>7374617475733D707562</setSpec>
<setSpec>7375626A656374733D4953:49532D424943</setSpec>
<setSpec>7375626A656374733D4253</setSpec>
</header>
- <metadata>
- <oai_dc:dc xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
xmlns:dc="http://purl.org/dc/elements/1.1/">
<dc:title>Comparative n-gram analysis of whole-genome sequences</dc:title>
<dc:creator>Ganapathiraju, Madhavi K.</dc:creator>
<dc:creator>Weisser, D.</dc:creator>
<dc:creator>Rosenfeld, R.</dc:creator>
<dc:creator>Carbonell, J.</dc:creator>
<dc:creator>Reddy, R.</dc:creator>
<dc:creator>Klein-Seetharaman, J.</dc:creator>
<dc:subject>BioInformatics Centre</dc:subject>
<dc:subject>Division of Biological Sciences</dc:subject>
<dc:description>A current barrier for successful rational drug design is the lack of understanding of the structure space provided by the proteins in a cell that is determined by their sequence space. The protein sequences capable of folding to functional three-dimensional shapes of the proteins are clearly different for different organisms, since sequences obtained from human proteins often fail to form correct three-dimensional structures in bacterial organisms. In analogy to the question "What kind of things do people say?" we therefore need to ask the question "What kind of amino acid sequences occur in the proteins of an organism?" An understanding of the sequence space occupied by proteins in different organisms would have important applications for "translation" of proteins from the language of one organism into that of another and design of drugs that target sequences that might be unique or preferred by pathogenic organisms over those in human hosts. Here we describe the development of a biological language modeling toolkit (BLMT) for genome-wide statistical amino acid n-gram analysis and comparison across organisms (freely accessible at www.cs.cmu.edu/~blmt). Its functions were applied to 44 different bacterial, archaeal and the human genome. Amino acid n-gram distribution was found to be characteristic of organisms, as evidenced by (1) the ability of simple Markovian unigram models to distinguish organisms, (2) the marked variation in n-gram distributions across organisms above random variation, and (3) identification of organism-specific phrases in protein sequences that are greater than an order of magnitude standard deviations away from the mean.
Done
Internet

Protocol Details: Flow Control



T.B. Rajashekar
NCSI, IISc

OAI Compliant Tools

- eprints.org (http://www.eprints.org)
- Dspace (http://dspace.org)
- CDSware (http://cdsware.cern.ch)
- Kepler (http://kepler.cs.odu.edu/)

A guide to Institutional Repository Software. 2nd edition. Open Society Institute. January 2004. Contains summary information about each repository software and a very detailed feature and functionality table. http://www.soros.org/openaccess/software

T.B. Rajashekar
NCSI, IISc

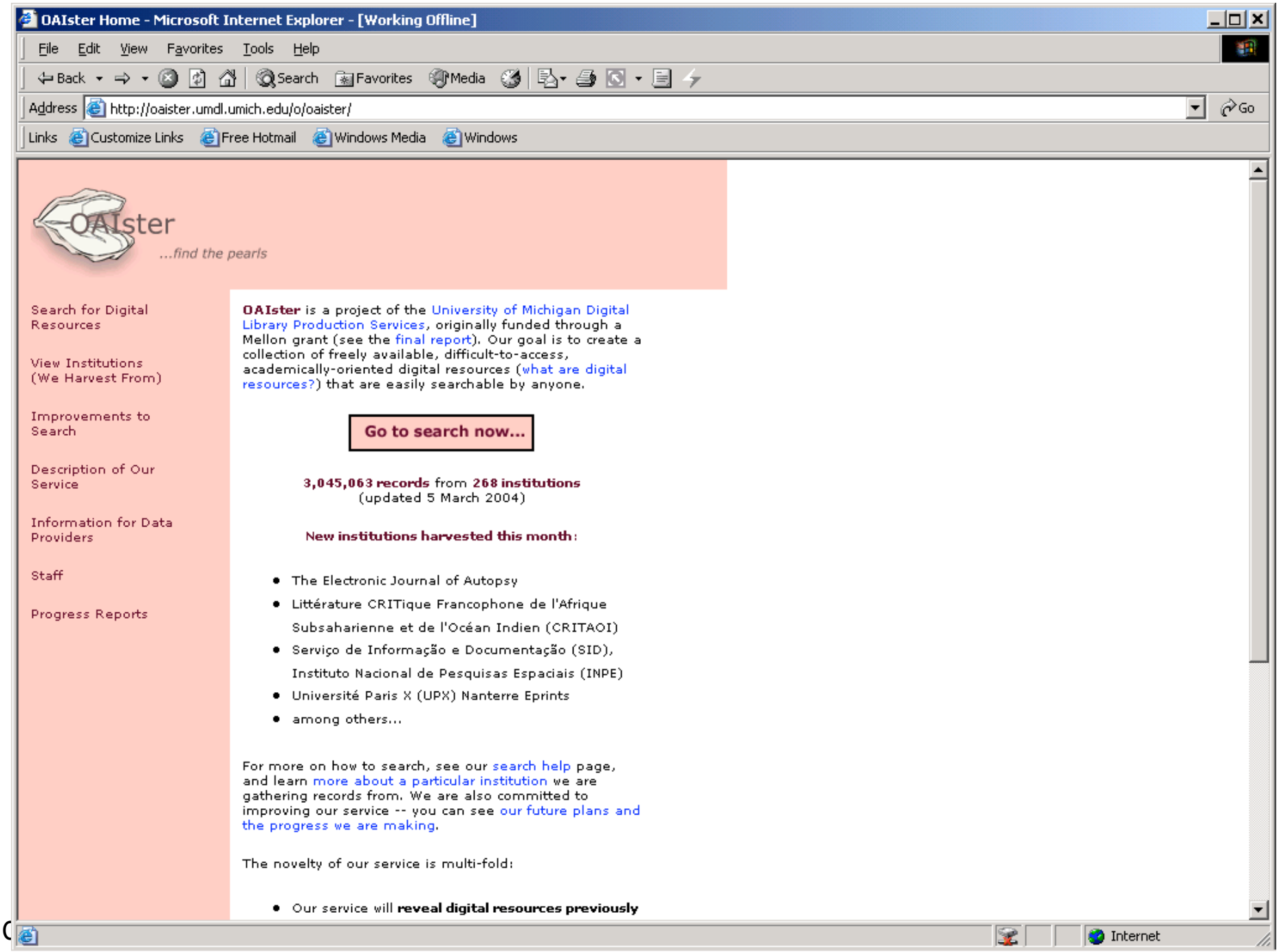
OAI-PMH Based Services

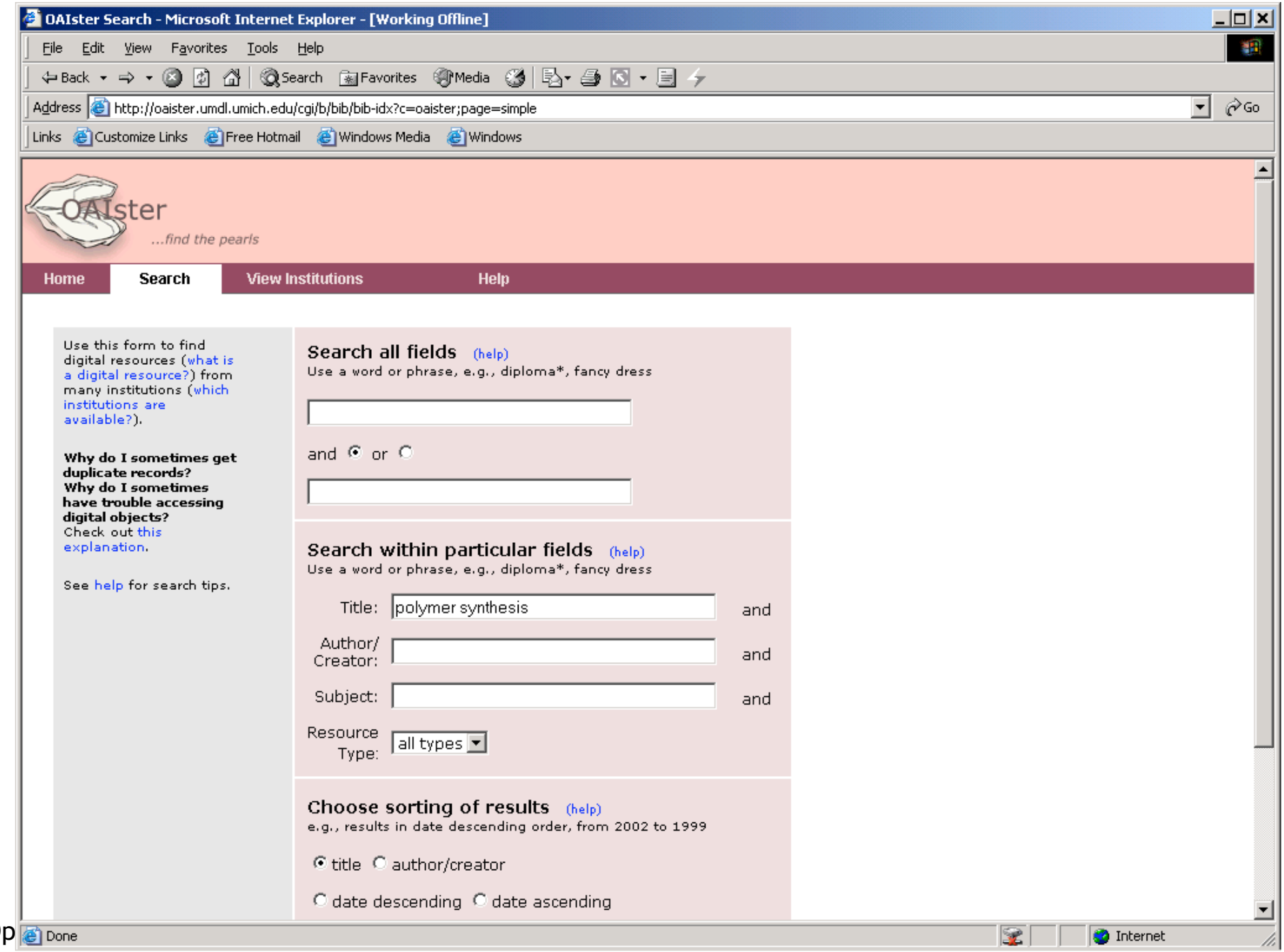
- Repository Explorer:
  - http://oai.dlib.vt.edu/cgi-bin/Explorer/oai2.0/testoai/
- Serach engines
  - Arc: http://arc.cs.odu.edu/
  - MyOAI: http://www.myoai.org/
  - Physnet: (subset of arXive, IOP…)
    http://physnet.uni-oldenburg.de/oai/query.php
  - OAIster: http://oaister.umdl.umich.edu/o/oaister/

T.B. Rajashekar
NCSI, IISc

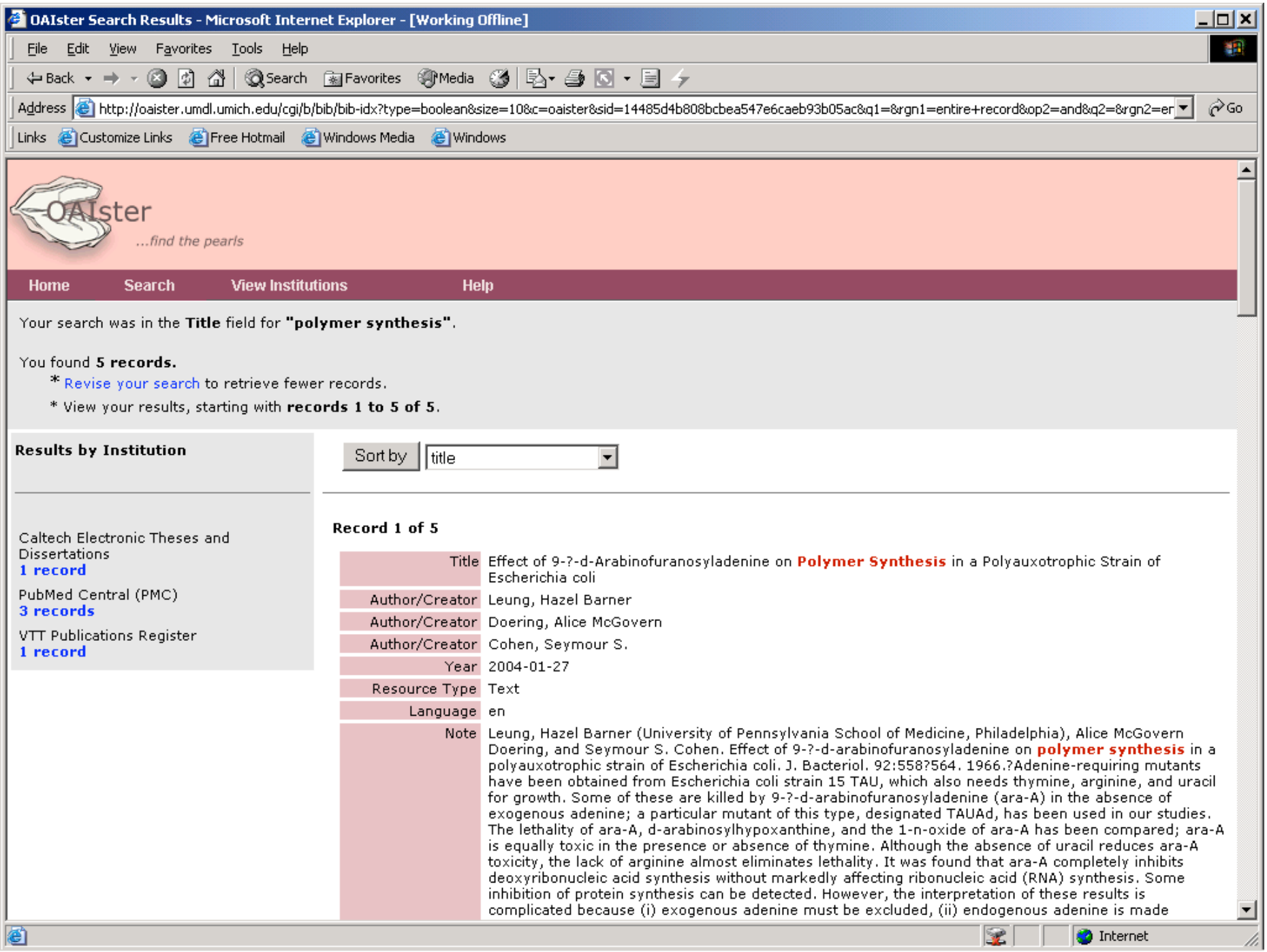
OAI Cross-Archive search Example

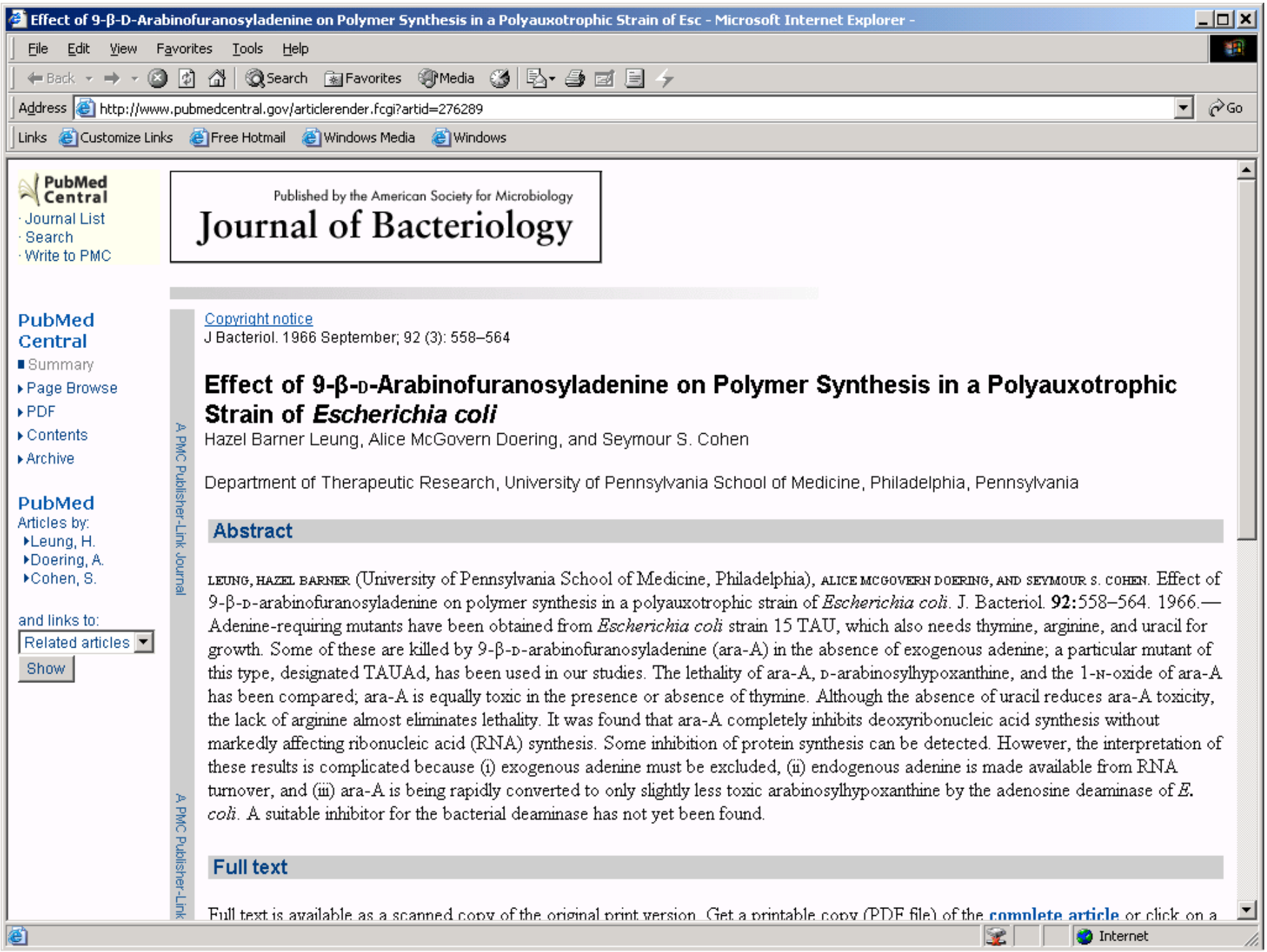
T.B. Rajashekar
NCSI, IISc


OAIster Home - Microsoft Internet Explorer - [Working Offline]
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://oaister.umdl.umich.edu/o/oaister/
Go
Links Customize Links Free Hotmail Windows Media Windows
OAIster
...find the pearls
Search for Digital Resources
View Institutions (We Harvest From)
Improvements to Search
Description of Our Service
Information for Data Providers
Staff
Progress Reports
OAIster is a project of the University of Michigan Digital Library Production Services, originally funded through a Mellon grant (see the final report). Our goal is to create a collection of freely available, difficult-to-access, academically-oriented digital resources (what are digital resources?) that are easily searchable by anyone.
Go to search now...
3,045,063 records from 268 institutions
(updated 5 March 2004)
New institutions harvested this month:
The Electronic Journal of Autopsy
Littérature CRITique Francophone de l'Afrique Subsaharienne et de l'Océan Indien (CRITAOI)
Serviço de Informação e Documentação (SID), Instituto Nacional de Pesquisas Espaciais (INPE)
Université Paris X (UPX) Nanterre Eprints
among others...
For more on how to search, see our search help page, and learn more about a particular institution we are gathering records from. We are also committed to improving our service -- you can see our future plans and the progress we are making.
The novelty of our service is multi-fold:
Our service will reveal digital resources previously
Internet


OAIster Search - Microsoft Internet Explorer - [Working Offline]
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://oaister.umdl.umich.edu/cgi/b/bib/bib-idx?c=oaister;page=simple
Go
Links Customize Links Free Hotmail Windows Media Windows
OAIster
...find the pearls
Home
Search
View Institutions
Help
Use this form to find digital resources (what is a digital resource?) from many institutions (which institutions are available?).
Why do I sometimes get duplicate records?
Why do I sometimes have trouble accessing digital objects?
Check out this explanation.
See help for search tips.
Search all fields (help)
Use a word or phrase, e.g., diploma*, fancy dress
and or
Search within particular fields (help)
Use a word or phrase, e.g., diploma*, fancy dress
Title: polymer synthesis and
Author/ Creator: and
Subject: and
Resource Type: all types
Choose sorting of results (help)
e.g., results in date descending order, from 2002 to 1999
title author/creator
date descending date ascending
Done
Internet

Op


OAIster Search Results - Microsoft Internet Explorer - [Working Offline]
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://oaister.umdl.umich.edu/cgi/b/bib/bib-idx?type=boolean&size=10&c=oaister&sid=14485d4b808bcbea547e6caeb93b05ac&q1=&rgn1=entire+record&op2=and&q2=&rgn2=er
Go
Links Customize Links Free Hotmail Windows Media Windows
OAIster
...find the pearls
Home Search View Institutions Help
Your search was in the Title field for "polymer synthesis".
You found 5 records.
* Revise your search to retrieve fewer records.
* View your results, starting with records 1 to 5 of 5.
Results by Institution
Caltech Electronic Theses and Dissertations
1 record
PubMed Central (PMC)
3 records
VTT Publications Register
1 record
Sort by title
Record 1 of 5
Title Effect of 9-?-d-Arabinofuranosyladenine on Polymer Synthesis in a Polyauxotrophic Strain of Escherichia coli
Author/Creator Leung, Hazel Barner
Author/Creator Doering, Alice McGovern
Author/Creator Cohen, Seymour S.
Year 2004-01-27
Resource Type Text
Language en
Note Leung, Hazel Barner (University of Pennsylvania School of Medicine, Philadelphia), Alice McGovern Doering, and Seymour S. Cohen. Effect of 9-?-d-arabinofuranosyladenine on polymer synthesis in a polyauxotrophic strain of Escherichia coli. J. Bacteriol. 92:558?564. 1966.?Adenine-requiring mutants have been obtained from Escherichia coli strain 15 TAU, which also needs thymine, arginine, and uracil for growth. Some of these are killed by 9-?-d-arabinofuranosyladenine (ara-A) in the absence of exogenous adenine; a particular mutant of this type, designated TAUAd, has been used in our studies. The lethality of ara-A, d-arabinosylhypoxanthine, and the 1-n-oxide of ara-A has been compared; ara-A is equally toxic in the presence or absence of thymine. Although the absence of uracil reduces ara-A toxicity, the lack of arginine almost eliminates lethality. It was found that ara-A completely inhibits deoxyribonucleic acid synthesis without markedly affecting ribonucleic acid (RNA) synthesis. Some inhibition of protein synthesis can be detected. However, the interpretation of these results is complicated because (i) exogenous adenine must be excluded, (ii) endogenous adenine is made
Internet


Effect of 9-β-D-Arabinofuranosyladenine on Polymer Synthesis in a Polyauxotrophic Strain of Esc - Microsoft Internet Explorer -
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.pubmedcentral.gov/articlerender.fcgi?artid=276289
Go
Links Customize Links Free Hotmail Windows Media Windows
PubMed Central
· Journal List
· Search
· Write to PMC
Published by the American Society for Microbiology
Journal of Bacteriology
PubMed Central
Summary
Page Browse
PDF
Contents
Archive
PubMed
Articles by:
Leung, H.
Doering, A.
Cohen, S.
and links to:
Related articles
Show
A PMC Publisher-Link Journal
Copyright notice
J Bacteriol. 1966 September; 92 (3): 558–564
Effect of 9-β-D-Arabinofuranosyladenine on Polymer Synthesis in a Polyauxotrophic Strain of Escherichia coli
Hazel Barner Leung, Alice McGovern Doering, and Seymour S. Cohen
Department of Therapeutic Research, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania
Abstract
LEUNG, HAZEL BARNER (University of Pennsylvania School of Medicine, Philadelphia), ALICE MCGOVERN DOERING, AND SEYMOUR S. COHEN. Effect of 9-β-D-arabinofuranosyladenine on polymer synthesis in a polyauxotrophic strain of Escherichia coli. J. Bacteriol. 92:558–564. 1966.—Adenine-requiring mutants have been obtained from Escherichia coli strain 15 TAU, which also needs thymine, arginine, and uracil for growth. Some of these are killed by 9-β-D-arabinofuranosyladenine (ara-A) in the absence of exogenous adenine; a particular mutant of this type, designated TAUAd, has been used in our studies. The lethality of ara-A, D-arabinosylhypoxanthine, and the 1-N-oxide of ara-A has been compared; ara-A is equally toxic in the presence or absence of thymine. Although the absence of uracil reduces ara-A toxicity, the lack of arginine almost eliminates lethality. It was found that ara-A completely inhibits deoxyribonucleic acid synthesis without markedly affecting ribonucleic acid (RNA) synthesis. Some inhibition of protein synthesis can be detected. However, the interpretation of these results is complicated because (i) exogenous adenine must be excluded, (ii) endogenous adenine is made available from RNA turnover, and (iii) ara-A is being rapidly converted to only slightly less toxic arabinosylhypoxanthine by the adenosine deaminase of E. coli. A suitable inhibitor for the bacterial deaminase has not yet been found.
Full text
Full text is available as a scanned copy of the original print version. Get a printable copy (PDF file) of the complete article or click on a
A PMC Publisher-Link
Internet

Effect of 9-β-D-Arabinofuranosyladenine on Polymer Synthesis in a Polyauxotrophic Strain of Esc - Microsoft Internet Explorer -

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.pubmedcentral.gov/articlerender.fcgi?artid=276289 Go

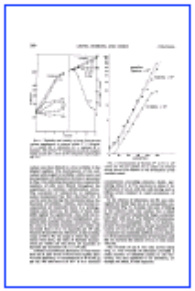
Links Customize Links Free Hotmail Windows Media Windows

Full text

Full text is available as a scanned copy of the original print version. Get a printable copy (PDF file) of the **complete article** or click on a page below to browse page by page.

558

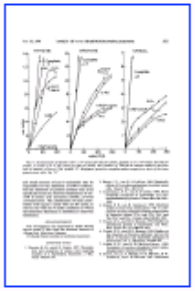
559



560

561

562



563

564

Internet

Summary

- Low-cost mechanism for harvesting metadata records from one system to another
- Based on HTTP and XML – Web-friendly
- Development over last 2-3 years has seen move from specific (discovery of e-prints) to generic (sharing descriptions of any resource)

T.B. Rajashekar
NCSI, IISc

Summary…

- Recommends simple DC as record format but extensible to any format encoded in XML
- OAI-PMH is **not** a search protocol
- Metadata and full-text typically made freely available – but not a requirement
  - OAI-PMH can be used between closed groups

T.B. Rajashekar
NCSI, IISc

Related Resources

- OAI Web site:
  - http://www.openarchives.org/
- Open Archives Forum
  - http://www.oaforum.org/tutorial/index.php

T.B. Rajashekar
NCSI, IISc