

Why Open Access and IRs?

A Somewhat Scientific Motivation

*Hussein Suleman
hussein@cs.uct.ac.za*

*University of Cape Town
Department of Computer Science*

July 2008



sivulile



open access
south africa

Somewhat Scientific Motivation

Theory/Observation

Methodology

Evaluation

Related Work



Open Access



Theory



Research Discovery 1/3



Web

Results 1 - 10 of about 73 for **suleman fox madalli**. (0.59 seconds)

[UCT CS Research Document Archive - Digital Libraries](#)

Digital Libraries. **Fox**, Edward A., Hussein **Suleman**, Devika **Madalli** and Lillian Cassel (2003) Digital Libraries, in Practical Handbook of Internet Computing. ...
pubs.cs.uct.ac.za/archive/00000016/ - 6k - [Cached](#) - [Similar pages](#)

[UCT CS Research Document Archive - Design and Implementation of ...](#)

Suleman, Hussein, Edward A. **Fox** and Devika **Madalli** (2003) Design and Implementation of Networked Digital Libraries: Best Practices. ...
pubs.cs.uct.ac.za/archive/00000017/ - 6k - [Cached](#) - [Similar pages](#)
 [[More results from pubs.cs.uct.ac.za](#)]

[Documentation Research and Training Centre: Items for Author ...](#)

... 2003, Design and Implementation of Networked Digital Libraries: Best Practices, Hussein, **Suleman**; **Fox**, Edward A.; **Madalli**, Devika P. ...
<https://drtc.isibang.ac.in/items-by-author?author=Madalli%2C+Devika+P.> - 10k - [Cached](#) - [Similar pages](#)

[Documentation Research and Training Centre: Items for Author Fox ...](#)

... Date of Issue, Title, Authors. 2003, Design and Implementation of Networked Digital Libraries: Best Practices, Hussein, **Suleman**; **Fox**, Edward A.; **Madalli**, Devika



Research Discovery 2/3

UCT CS Research Document Archive

[Home](#) || [About](#) || [Browse](#) || [Search](#) || [Register](#) || [User Area](#) || [Help](#)

Digital Libraries

Fox, Edward A., Hussein Suleman, Devika Madalli and Lillian Cassel (2003) *Digital Libraries*, in *Practical Handbook of Internet Computing*. CRC Press.

Full text available as:

[PDF](#) - Requires [Adobe Acrobat Reader](#) or other PDF viewer.

EPrint Type: Book Chapter

[H Information Systems: H.0 GENERAL](#)

[H Information Systems: H.1 MODELS AND PRINCIPLES](#)

Subjects: [H Information Systems: H.3 INFORMATION STORAGE AND RETRIEVAL](#)

[A General Literature: A.1 INTRODUCTORY AND SURVEY](#)

[H Information Systems: H.4 INFORMATION SYSTEMS APPLICATIONS](#)

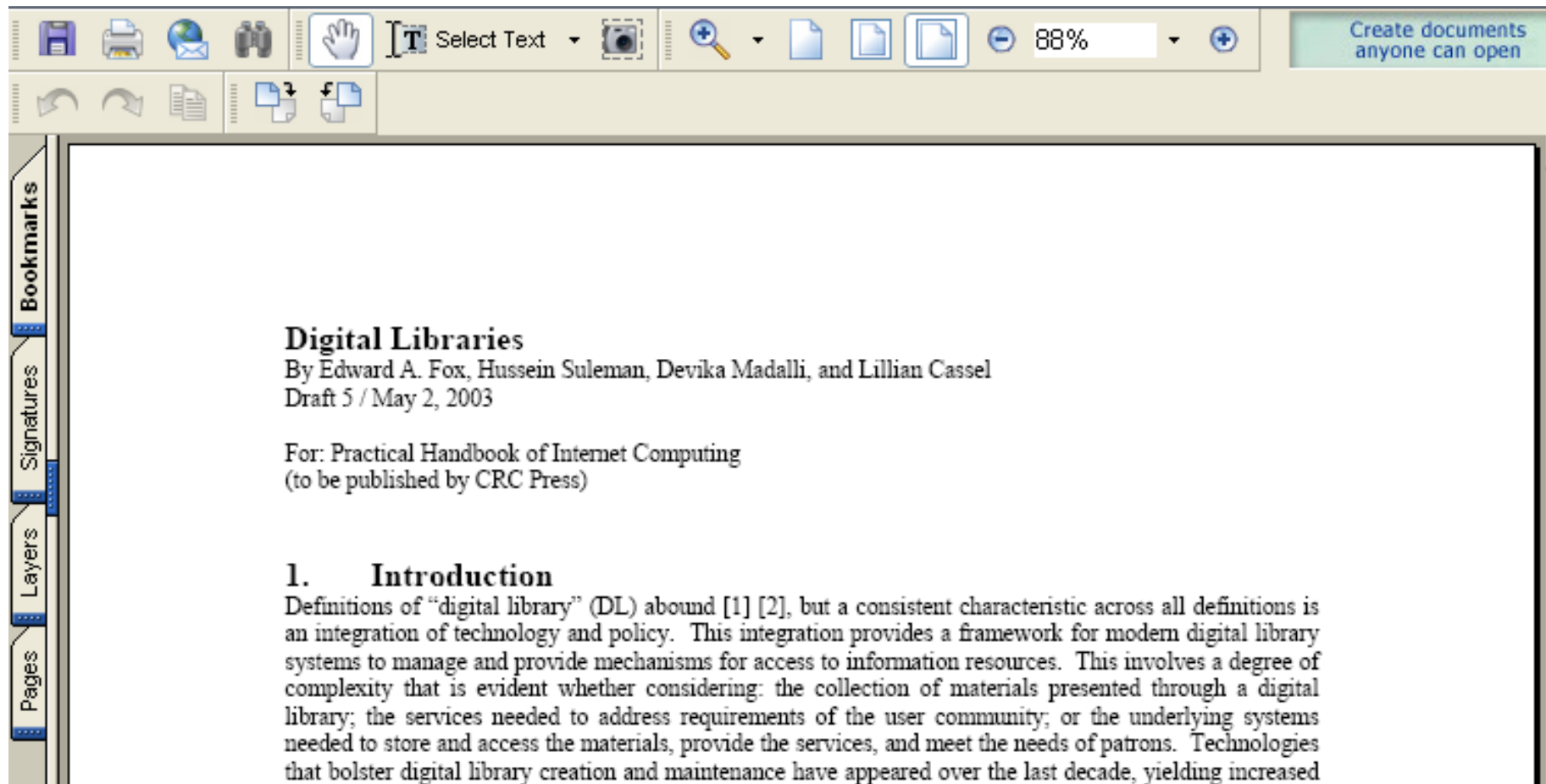
ID Code: 16

Deposited By: [Suleman, Hussein](#)

Deposited On: 11 July 2003



Research Discovery 3/3



The screenshot shows a PDF viewer interface. The top toolbar includes icons for save, print, email, zoom, and navigation. The document content is as follows:

Digital Libraries
By Edward A. Fox, Hussein Suleman, Devika Madalli, and Lillian Cassel
Draft 5 / May 2, 2003

For: Practical Handbook of Internet Computing
(to be published by CRC Press)

1. Introduction
Definitions of “digital library” (DL) abound [1] [2], but a consistent characteristic across all definitions is an integration of technology and policy. This integration provides a framework for modern digital library systems to manage and provide mechanisms for access to information resources. This involves a degree of complexity that is evident whether considering: the collection of materials presented through a digital library; the services needed to address requirements of the user community; or the underlying systems needed to store and access the materials, provide the services, and meet the needs of patrons. Technologies that bolster digital library creation and maintenance have appeared over the last decade, yielding increased



Open Access?

- 1734 hits directly from Google in March 2007.
- Example:
 - <http://www.google.com/search?q=questionnaire+system+UML>
 - Kritzinger, Pieter, Marshini Chetty, Jesse Landman, Michael Marconi and Oksana Ryndina (2003)
ChattaBox: A Case Study in Using UML and SDL for Engineering Concurrent Communicating Software Systems. In Proceedings Southern African Telecommunications Networks and Applications Conference, George, South Africa.



Open Access in General

- Principle of unrestricted access to information.
- Two common approaches:
 - Open Access Journals
 - Institutional Repositories
- What content?
 - Research output, theses, learning material, etc.
- Beneficial for:
 - Authors
 - Researchers seeking information
 - Institutions and funders



Methodology



UCT CS Research Repository

UCT CS Research
Document Archive

[Home](#) || [About](#) || [Browse](#) ||
[Search](#) || [Register](#) || [User Area](#)
|| [Help](#)

Welcome to UCT CS Research
Document Archive



Welcome to the [UCT Computer Science](#) Research Document Archive, which archives and makes accessible documents that are products and by-products of research in the department.

Search the Titles, Authors, Abstracts and Keywords :

[Browse](#)

Browse the archive by [Subject](#), [Year](#), [Lab](#) or [Type](#).

[Latest Additions](#)

View items added to the archive in the past week.

[Simple Search](#)

Search the archive using the most common fields.

- ☑ Author self-submission
- ☑ Checking of submissions
- ☑ Archive-everything!
- ☑ UCT-CS-specific metadata and classification systems
- ☑ Hierarchical browsing
- ☑ Simple and fielded searching
- ☑ OAI-PMH compliance



Why we have a repository

- ❑ It was faster than simply waiting for UCT!
- ❑ CS departments internationally archive technical reports (NCSTRL).
- ❑ Research websites don't last long (enough).
- ❑ UCT doesn't have an electronic thesis project yet.
- ❑ We need to improve ACCESS to our work.
- ❑ We need to preserve our research output.
- ❑ Bureaucracy (UCT, NRF, DoE, etc.) requires tracking publications.
- ❑ We (think we) know what we are doing.



What we archive

- ❑ Books and Book Chapters
- ❑ Conference Paper and Posters
- ❑ Journals (online and paginated)
- ❑ Newspaper and Magazine Articles
- ❑ Preprints
- ❑ Presentation Slides
- ❑ Conference Proceedings
- ❑ Departmental Technical Reports
- ❑ Electronic Theses and Dissertations
- ❑ Other Stuff ...



Infrastructure Requirements

- Software: EPrints v2.2.1
 - plus a few changes here and there.
- Server:
 - Tacked onto an existing machine at first!
 - 3GHz Pentium/512MB/160GB
- Operating System:
 - FreeBSD 5.0
- Web server:
 - Apache v1.3.7
- Administrator: shared with other systems ...



Community Building

- Filling the archive:
 - Get official support.
 - Twist arms of staff.
 - Fill archive with own publications to make others look bad.
 - Twist arms of staff even harder.
 - Get (student) researchers to twist student arms.
 - “The domino effect”.



Copyright

For work being deposited by its own author: In self-archiving this collection of files and associated bibliographic metadata, I grant UCT CS Research Document Archive the right to store them and to make them permanently available publicly on-line. I declare that this material is my own intellectual property and I have the right to archive and disseminate it through a departmental collection. I understand that UCT CS Research Document Archive does not assume any responsibility if there is any breach of copyright in distributing these files or metadata.

For work being deposited by someone other than its author: I hereby declare that the collection of files and associated bibliographic metadata that I am archiving at UCT CS Research Document Archive is in the public domain or is being archived with the permission of and on behalf of the author(s). If this is not the case, I accept full responsibility for any breach of copyright that distributing these files or metadata may entail.

Clicking on the deposit button indicates your agreement to these terms.

[< Back](#)[Deposit EPrint Later](#)[Deposit EPrint Now](#)

Metadata/Citation Rendering

Conference Paper

1. [Arnab, Alapan and Andrew CM Hutchison \(2004\) Digital Rights Management - An Overview of current challenges and solutions. In Venter, HS, JHP Eloff, L Labuschagne and MM Eloff, Eds. *Proceedings Information Security South Africa \(ISSA\)*, Gallagher Estate, Midrand, South Africa.](#)



Departmental Research Overview

2004

Conference Paper

1. [Arnab, Alapan and Andrew CM Hutchison \(2004\) Digital Rights Management - An Overview of current challenges and solutions. In Venter, HS, JHP Eloff, L Labuschagne and MM Eloff, Eds. *Proceedings Information Security South Africa \(ISSA\)*, Gallagher Estate, Midrand, South Africa.](#)

Conference Poster

1. [Landman, Jesse and Pieter Kritzinger \(2004\) DS-CDMA Fading Channels with Bursty IP Traffic Arrivals. In *Proceedings WiOpt 2004*, Cambridge, England.](#)

Departmental Technical Report

1. [Arnab, Alapan and Andrew Hutchison \(2004\) Digital Rights Management - A current review. Technical Report CS04-04-00, Department of Computer Science, University of Cape Town.](#)
2. [de Wet, Nico Dirk and Pieter S. Kritzinger \(2004\) Model-Based EIS Performability Analysis. Technical Report CS04-01-00, Department of Computer Science, University of Cape Town.](#)
3. [Perumal, Sameshan and Pieter Kritzinger \(2004\) A Tutorial on RAID Storage Systems. Technical Report CS04-05-00, Department of Computer Science, University of Cape Town.](#)
4. [Ryndina, Oksana and Pieter Kritzinger \(2004\) Improving Requirements Specification: Verification of Use Case Models with Susan. Technical Report CS04-06-00, Department of Computer Science, University of Cape Town.](#)

Electronic Thesis or Dissertation



Subgroup Research Overview

UCT CS Research Document Archive

[Home](#) || [About](#) || [Browse](#) || [Search](#) || [Register](#) || [User Area](#) || [Help](#)

Laboratory: Collaborative Visual Computing

2007

Conference Paper

1. [Blake, Edwin, David Nunez and Bertus Labuschagne \(2007\) Longitudinal Effects on Presence Suspension of Disbelief or Distrust of Naive Belief\(2007\)](#). In Moreno, Laura, Eds. *Proceedings PRESENCE 2007 The 10th Annual International Workshop on Presence*, pages 291-295, Barcelona, Spain.
2. [Labuschagne, Bertus, David Nunez and Edwin Blake \(2007\) Presence predicts false memories of virtual environment content](#). In Moreno, Laura, Eds. *Proceedings PRESENCE 2007 The 10th Annual International Workshop on Presence*, pages 297-301, Barcelona, Spain.

Journal (Paginated)

1. [Marais, Patrick and James Gain \(2007\) High fidelity compression of irregularly sampled height-fields](#). *South African Computer Journal* 38:40-50.
2. [Marais, Patrick, James Gain and Dave Shreiner \(2007\) Distance-Ranked Connectivity Compression of Triangle Meshes Computer](#). *Computer Graphics Forum (earlyOnline)*.

Other

1. [Maunder, Andrew, Gary Marsden and Richard Harper \(2007\) Shoot and Carry](#).



NRF / DoE Credit

- We already have a departmental listing of all research output.
- Where copyright does not allow, we include just a citation – no files – for completeness.

NDLTD Union Catalog Project

Suleman, Hussein (2004) *NDLTD Union Catalog Project*, in Fox, Edward A., Shahrooz Feizabadi, Joseph M. Moxley and Christian R. Weisser, Eds. *Electronic Theses and Dissertations: A Sourcebook for Educators, Students, and Librarians*, chapter 6, pages 73-77. Marcel Dekker, Inc..

Full text available as:

EPrint Type: Book Chapter

Subjects: [H Information Systems: H.4 INFORMATION SYSTEMS APPLICATIONS](#)

ID Code: 14

Deposited By: [Suleman, Hussein](#)

Deposited On: 11 July 2003



Interoperability

- ▣ Our archive is compliant with Open Archives Initiative's Protocol for Metadata Harvesting (OAI-PMH) v2.0.
- ▣ Metadata can be freely harvested by any service provider.
- ▣ baseURL: <http://pubs.cs.uct.ac.za/perl/oai2>



Communities and Metadata

- Participate in OAI:
 - Metadata can be in Dublin Core.
- Participate in NDLTD:
 - Metadata can be in ETDMS.
 - Set for theses and dissertations only.
- Participate in NCSTRL:
 - Metadata can be in RFC1807.
 - Set for technical reports only.
 - OAI-PMH Request:
 - http://pubs.cs.uct.ac.za/perl/oai2?verb=ListRecords&metadataPrefix=oai_rfc1807&set=747970653D746563687265706F7274



Migration

- <http://pubs.cs.uct.ac.za> is the “public view”
- <http://pubs.cs.uct.ac.za:1081> is the actual server.
- Apache rewriting rules are used to proxy to the actual server.
- Advantages:
 - Migration is trivial - we can move the server and nobody will know.
 - All resources have the most generic URL possible.
 - The repository can be co-located with other projects.



Research

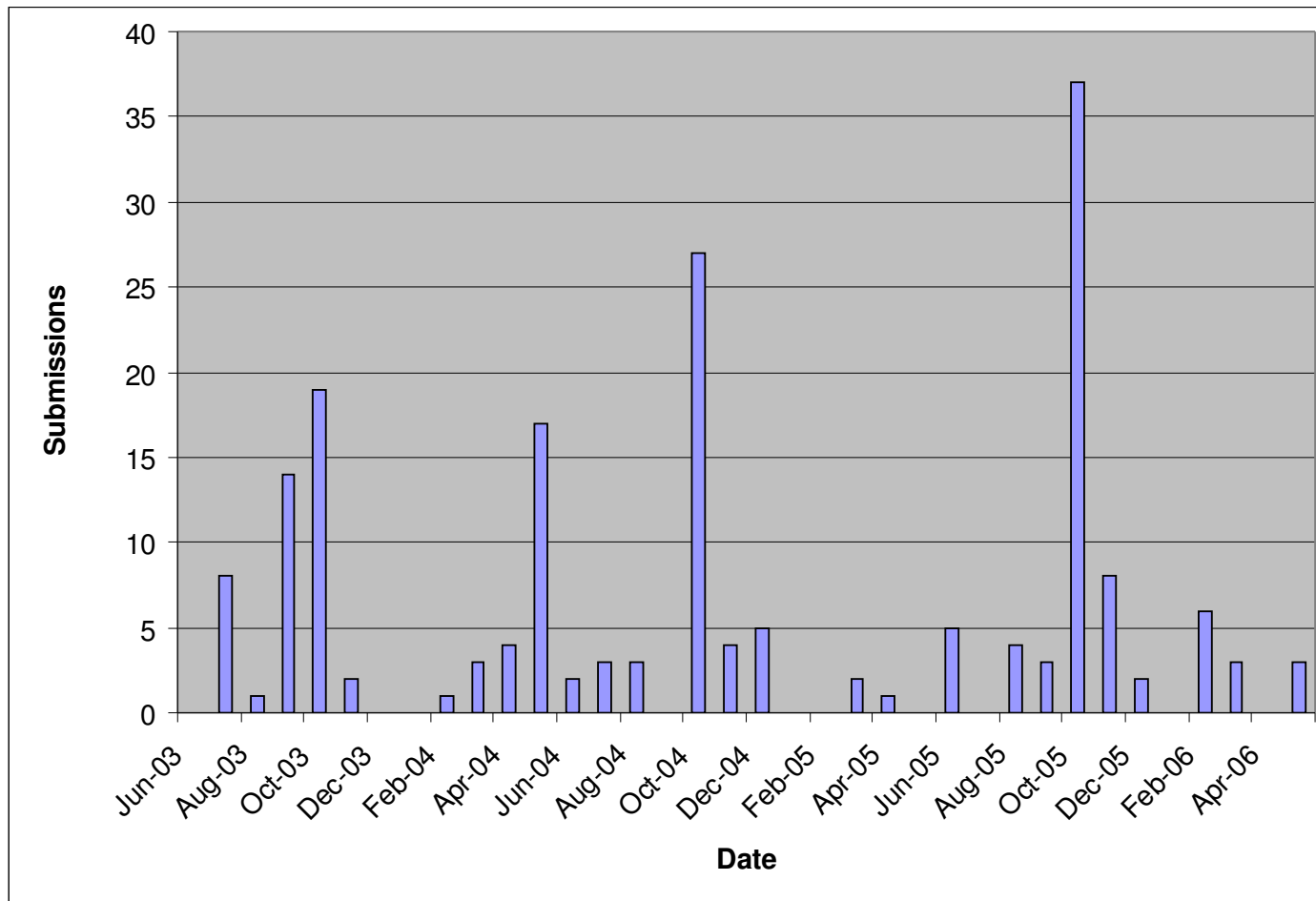
- Import metadata/files into DSpace
 - Student assignment to migrate metadata/content.
 - Based completely on OAI-PMH interface.
 - All 15 groups replicated basic EPrints functionality in DSpace with same data set.
- Higher-level services to enhance basic services provided by repositories:
 - Ongoing work into component-based systems ...
 - Ongoing work into scalability of systems ...
 - Ongoing work on interfaces to manage repositories more easily ...



Evaluation



Submissions to Archive over Time



Analysis of Access

- 187 items, 478520 log file entries
- 798 unique user agent types
 - 65% from crawlers and bots
 - 34 from metadata harvesters
- Data extraction from log file:
 - Unique IP addresses (visitors)
 - Source of access
 - Which resources are accessed
 - Access patterns of popular/old resources
 - Effect of submission sequence
 - Distribution of accesses
 - Analysis of partial responses

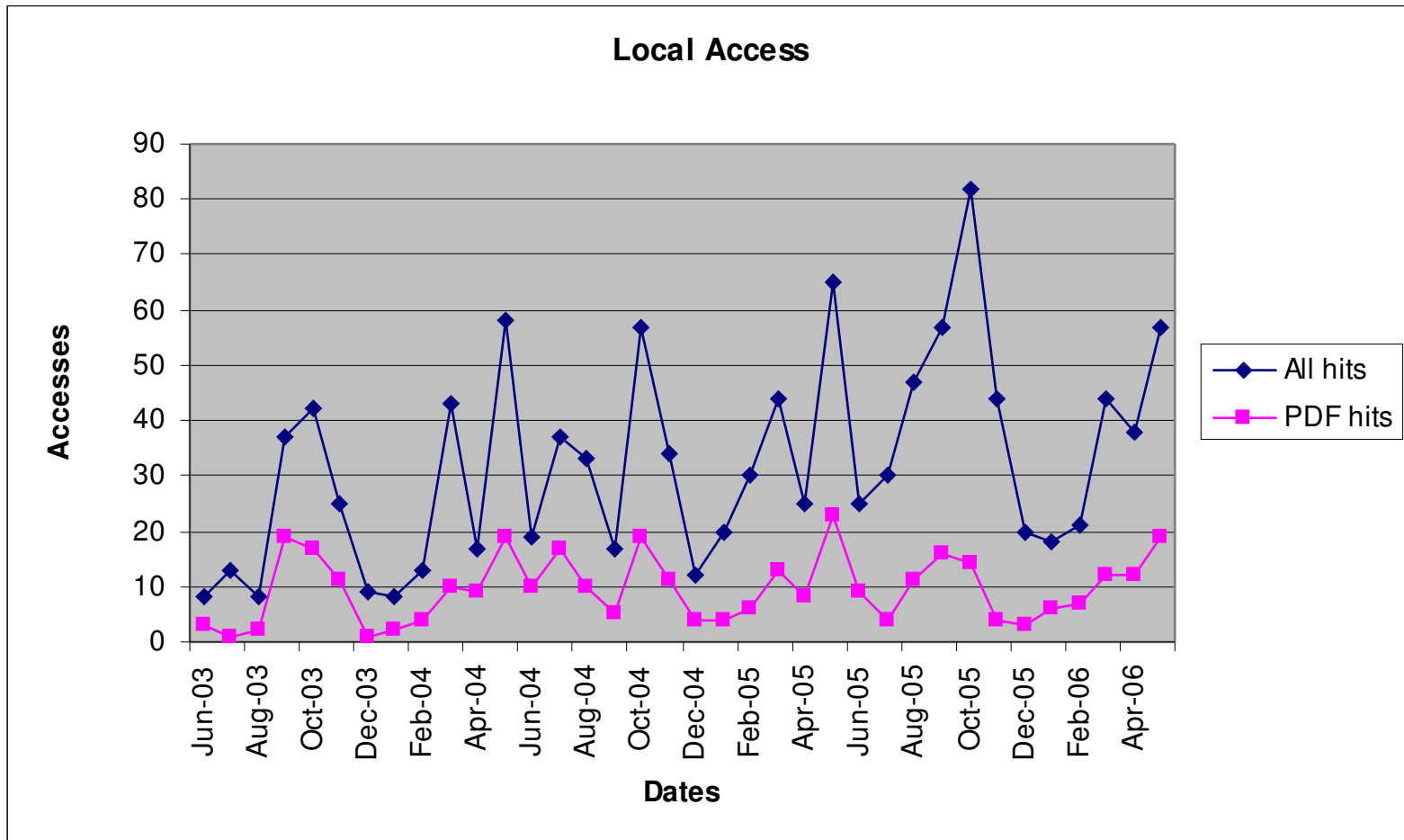


Disclaimer

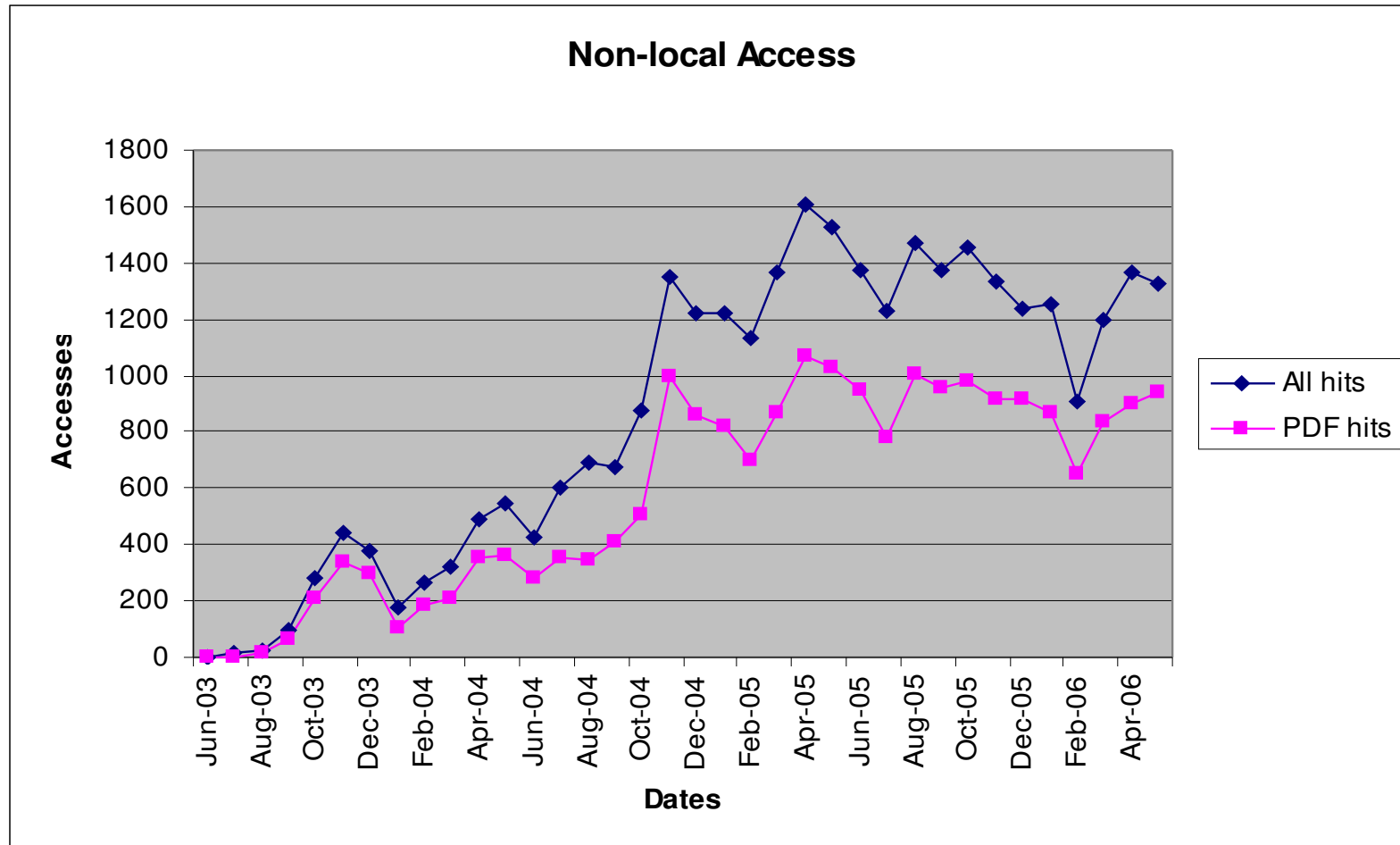
- Web log analysis is at best an approximation!
- Assumptions:
 - download = access
 - if useragent != bot, then real user
 - no caching
 - every user has a single IP address
 - ...



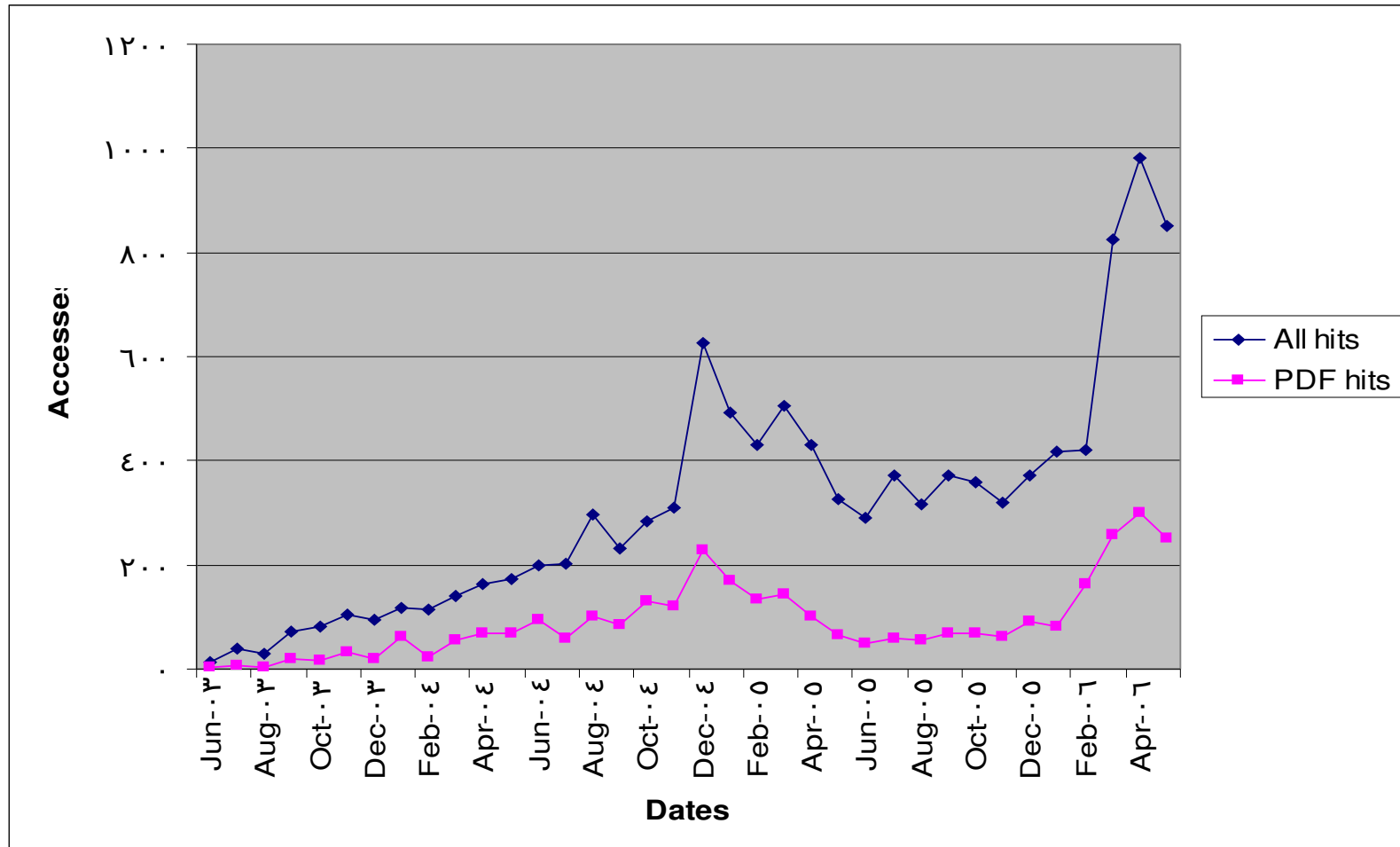
Unique Local IP Addresses



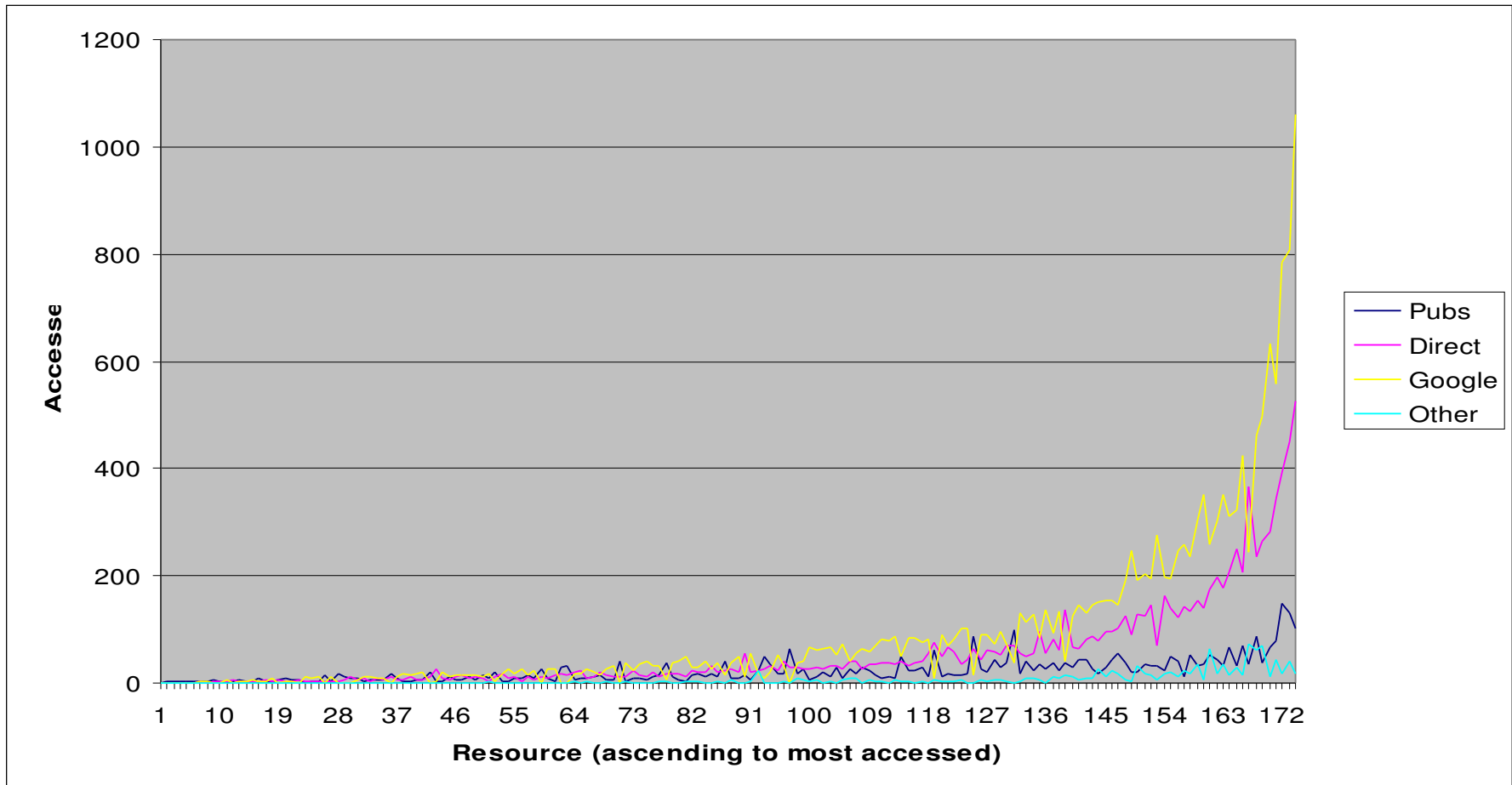
Unique Non-local IP Addresses



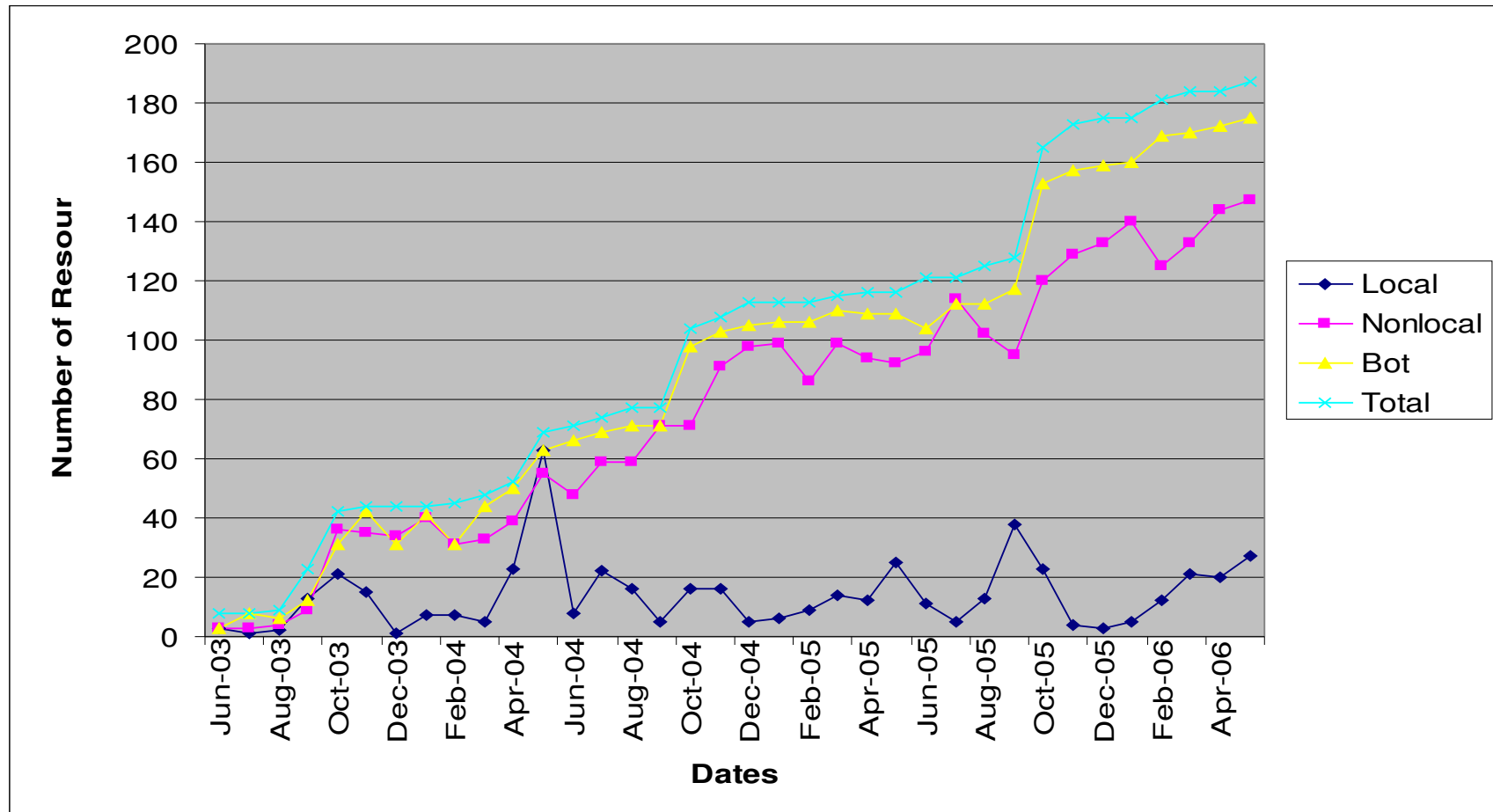
Unique Crawler IP Addresses



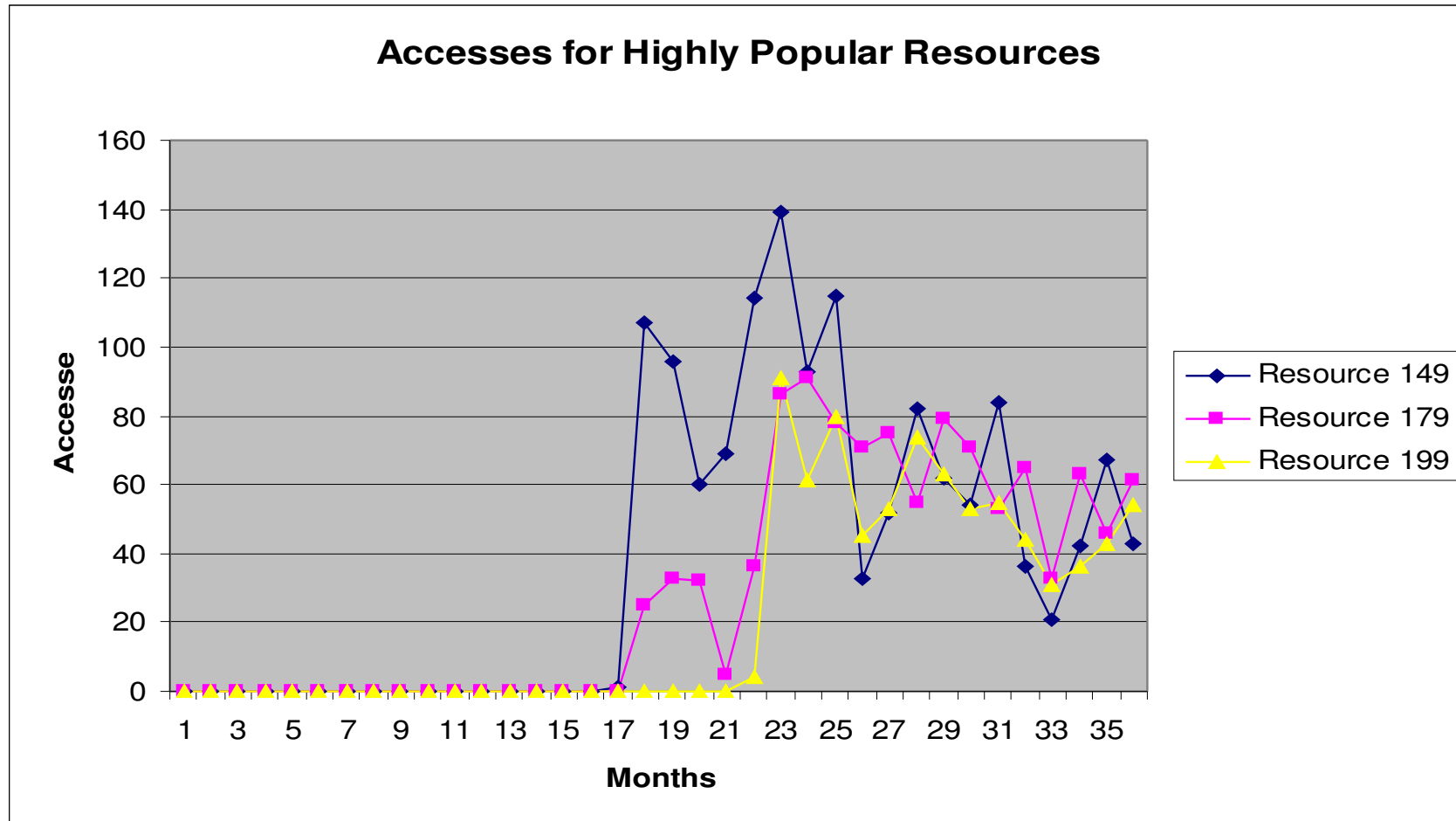
Source of Access



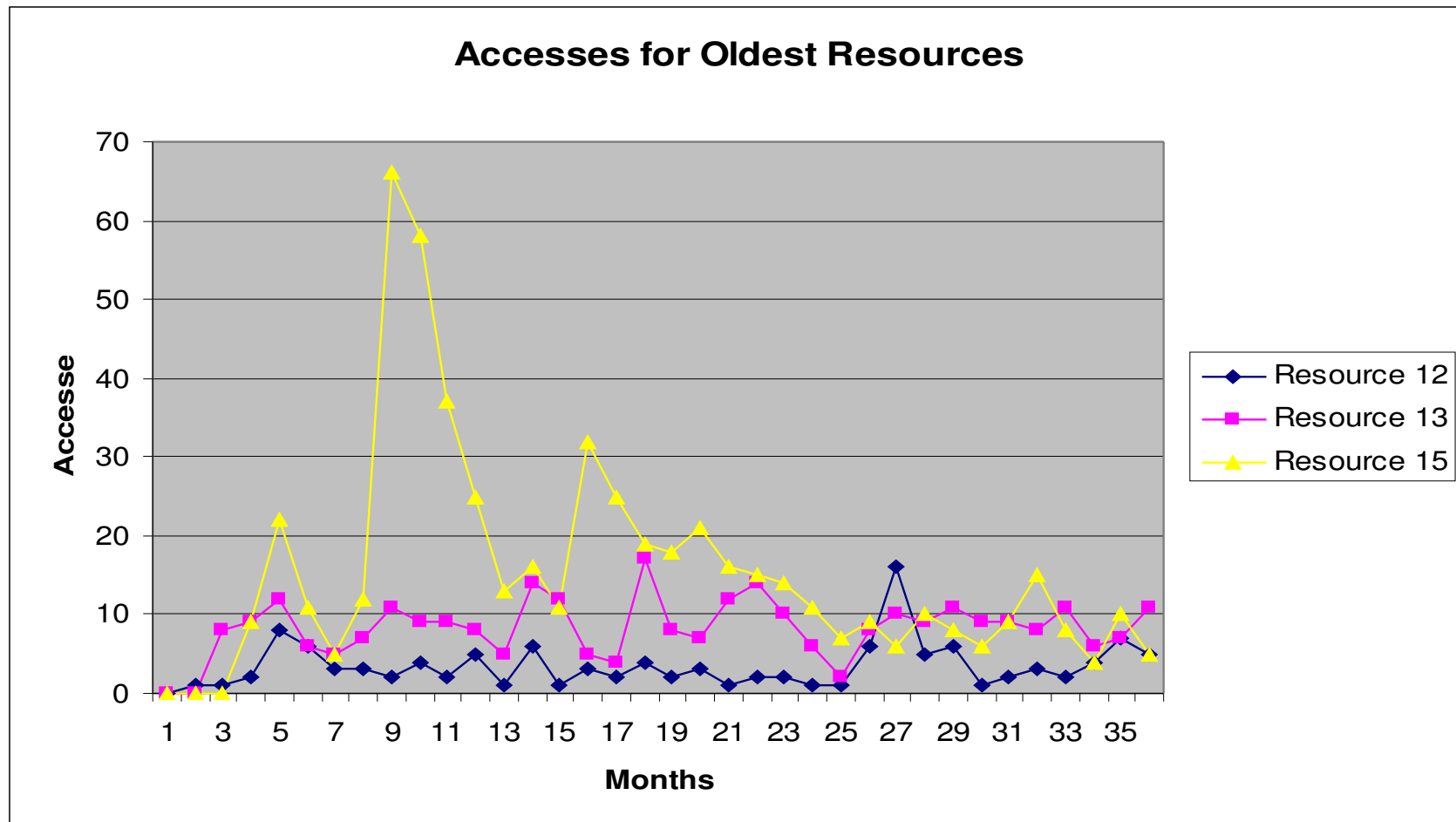
Which Resources are Accessed



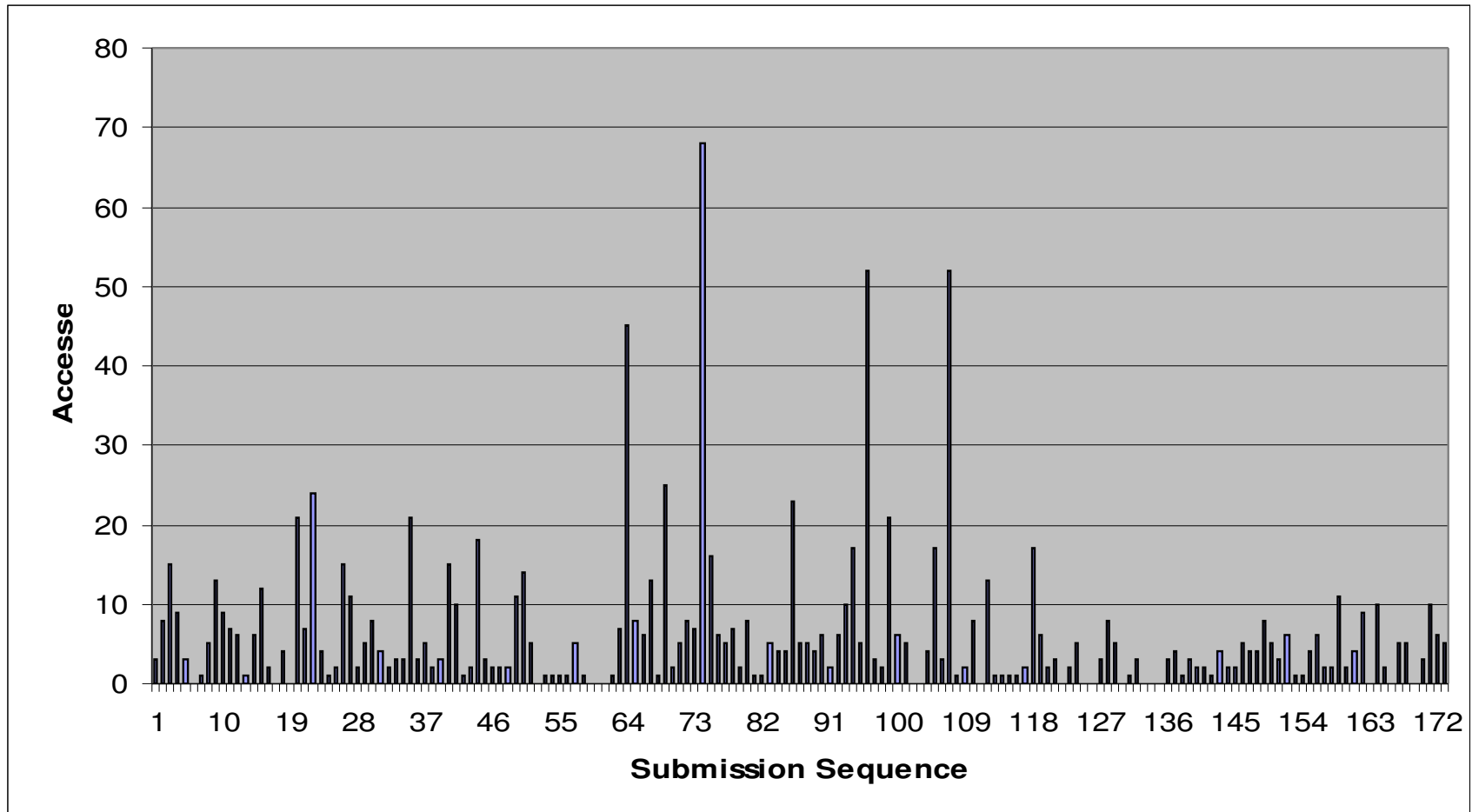
Access Patterns – Most Popular



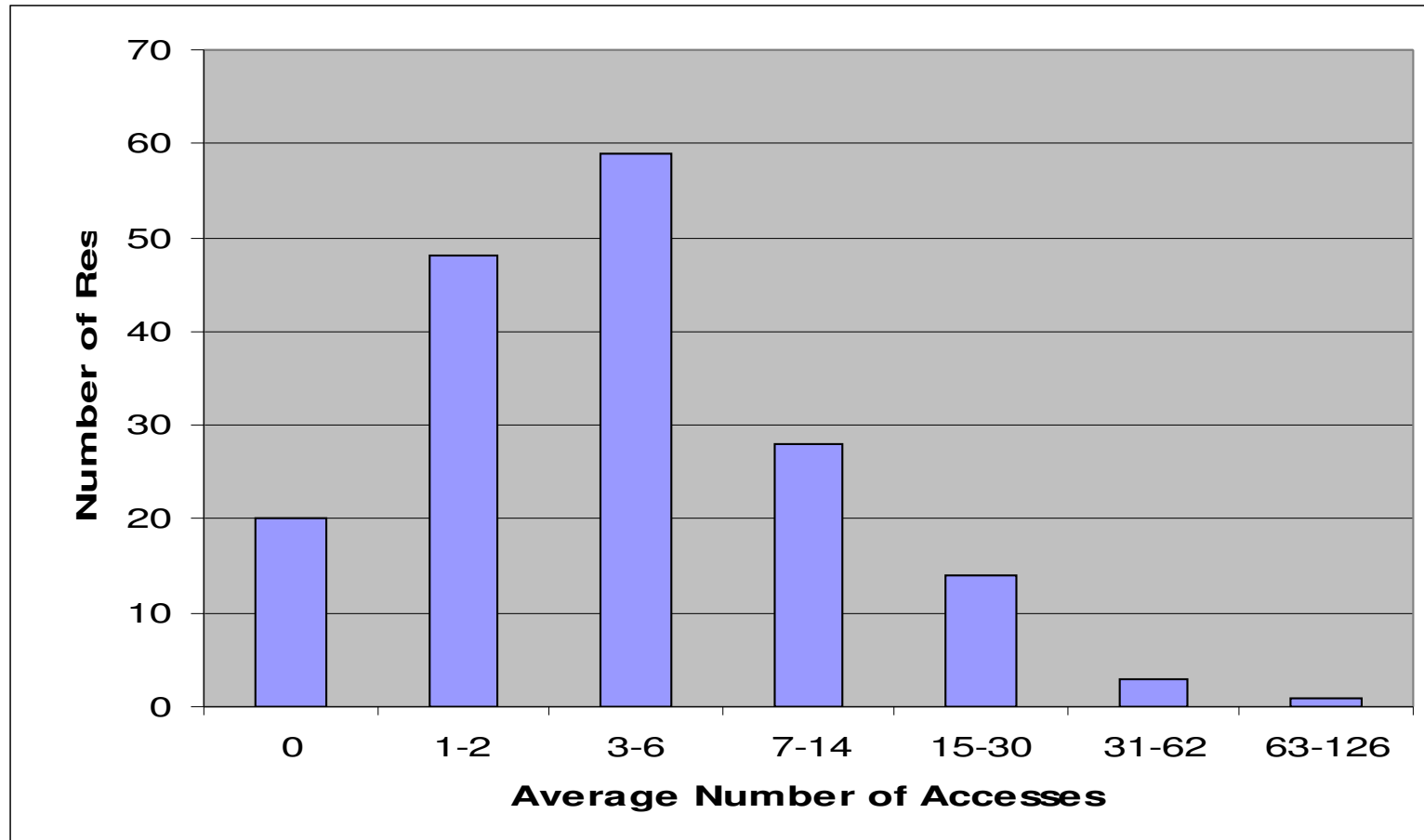
Access Patterns - Oldest Resources



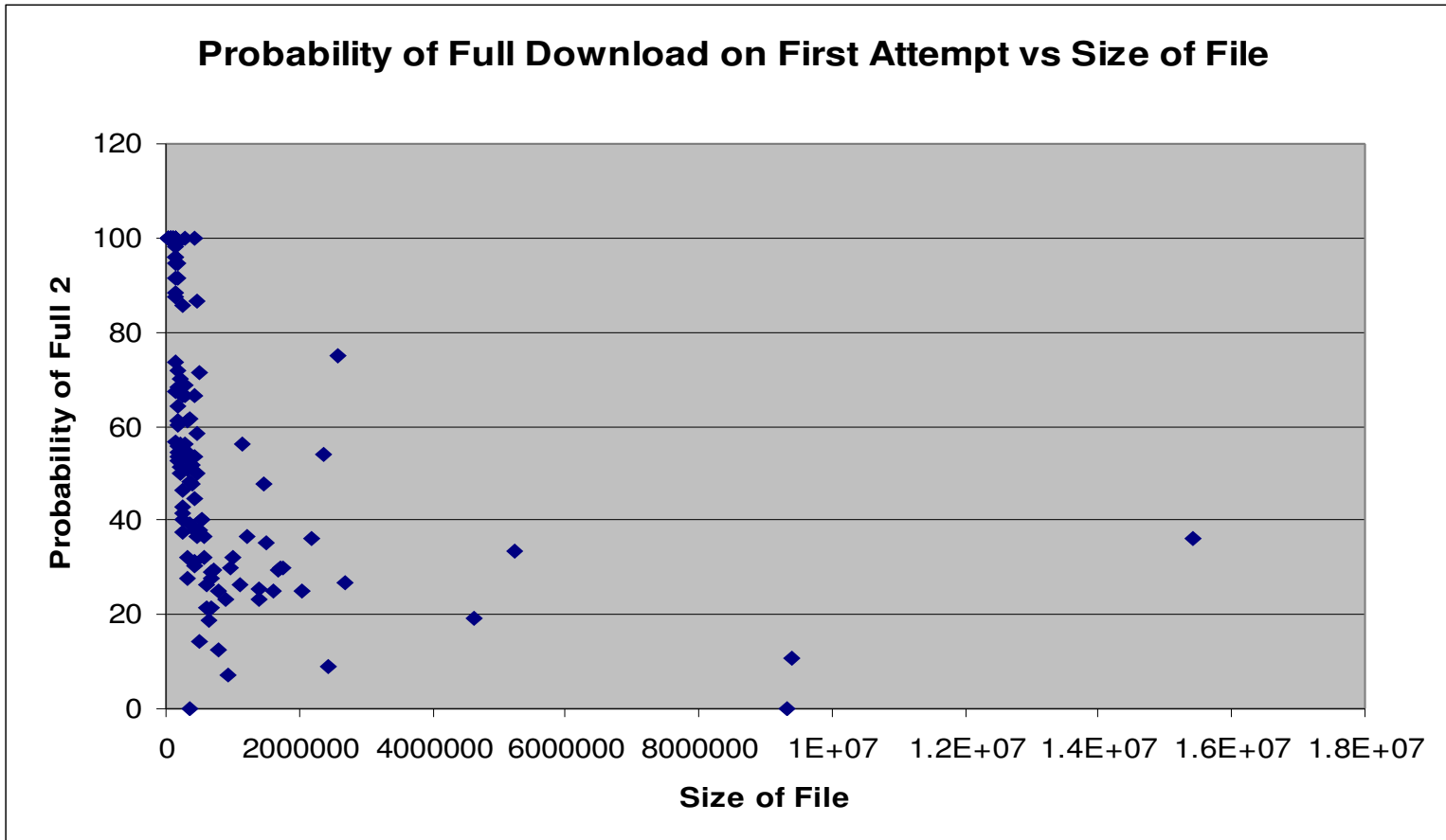
Effect of Submission Sequence



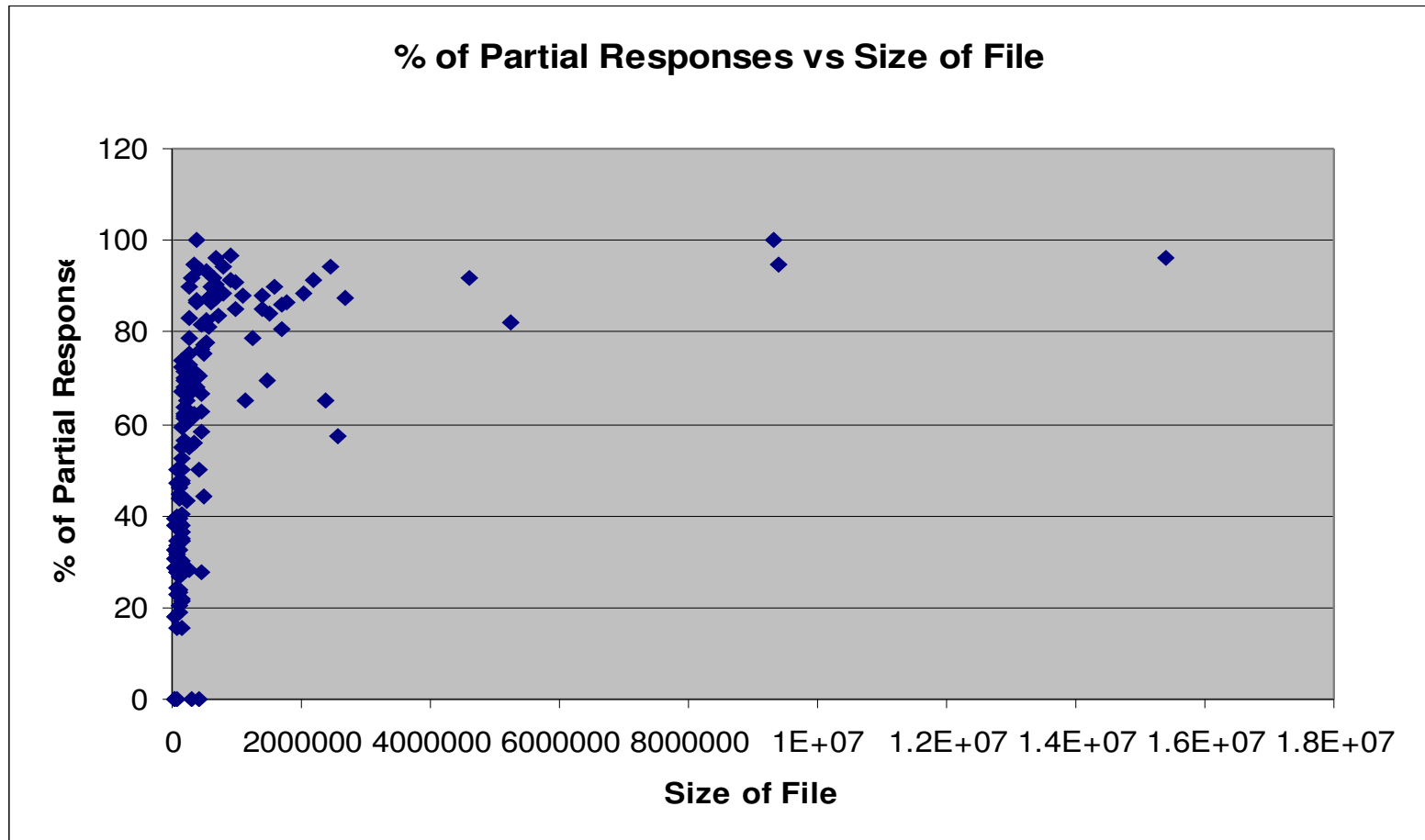
Distribution of Accesses



Probability of Success on First Download



Percentage of Partial Responses



Summary of Statistics

- ❑ The archive is regularly and aggressively indexed by search services.
- ❑ The vast majority of access to resources comes from outside UCT.
- ❑ Most users find their way to resources through Google and other search engines.
- ❑ Most resources are used in most months – by both crawlers/harvesters and end users.
- ❑ Access appears to be dictated by content, and not age of resources.
- ❑ While many downloads are successful immediately, larger files result in more partial downloads.



Related Work



Serials Crisis

- Library budgets are lower than in the past.
- Journals cost more.
- Journals are being bundled so there is little or no benefit in going electronic.
 - Electronic versions sometimes cost more!
- No long-term access – move by publishers to subscription model.



Failure of SUBJECT repositories

- In the 90s subject archives were popular e.g., arXiv, RePEc, NCSTRL.
- Problems:
 - Sustainability – repositories run by organisations with limited funding.
 - Development skill – staff needed to develop and maintain all software needed to run the archive.
 - Should the archive be centralised or distributed?
 - Owner of data was not best provider of services.
- Solution = Institutional Repositories



Change of COPYRIGHT in our favour

- ❑ Most society publishers will allow archiving on a website or IR e.g., ACM
- ❑ Most commercial publishers allow archiving on a website or IR after some time (typically 12-24 months), if not immediately.
- ❑ Newer commercial publisher agreements make greater allowance for IRs.
- ❑ You can always negotiate with a publisher!



International ETD Movement (700k+)

OCLC ONLINE COMPUTER LIBRARY CENTER

A Project of OCLC Research

XTCat NDLTD
NDLTD Union Catalog



[Identify](#) | [GetRecord](#) | [ListIdentifiers \(Resumption\)](#) | [ListMetadataFormats](#) | [ListRecords \(Resumption\)](#) | [ListSets](#)

```
responseDate 2007-11-04T13:20:01Z
request      http://alcme.oclc.org/ndltd/servlet/OAIHandler?verb=ListSets
```

ListSets

setSpec	setName	DC record count
ADTP	Australasian Digital Theses Program	30853
AUCKLAND	University of Auckland	375
BGMJU	Brigham Young University Theses	910
BICBF	Bibliothèque interuniversitaire de la Communauté française de Belgique	880
CALTECH	California Institute of Technology	3776
CCSD	CCSD theses-EN-ligne, France	7936
CRANFIELD	Cranfield University	109



National ETD Project

- Collaboration between CHELSA and NRF
- Aims to get every tertiary institution to archive theses in near future.
- National portal for discovery of ETDs.
- Approximately half of the institutions already have repositories:
 - Rhodes, Wits, Pretoria, SUN, UWC, ...
- NRF is supporting some institutions without capacity in the short term.



Open Access in Southern Africa

- Much interest since early late 90s.
- Many conferences and workshops to reskill archivists.
- Local support network: www.sivulile.org
- Possibility of future links with government/NRF systems (RIMS).
- Emerging repositories:
 - Rhodes, Pretoria, **DUT**, ...



Final Thoughts

- ❑ A document repository is a simple and effective way to increase visibility to the research of a department or institution.
- ❑ But it is imperative that documents are archived as soon as possible, to maximise visibility!
- ❑ In the context of national and international developments, open access to research is now more relevant than ever.
- ❑ **DUT already has the infrastructure - now all that is left is for researchers to make use of it!**



Links

- Sivulile
 - <http://www.sivulile.org/>
- UCT CS Research Archive
 - <http://pubs.cs.uct.ac.za/>
- Dspace
 - <http://www.dspace.org/>
- EPrints
 - <http://www.eprints.org/>
- Open Archives Initiative
 - <http://www.openarchives.org/>



Discussion...



*to find me, search on Google or
Facebook for "hussein suleman"*