

If You Harvest arXiv.org, Will They Come?

Michael L. Nelson, Johan Bollen

Old Dominion University
Department of Computer Science
Norfolk VA 23529 USA

{mln,jbollen}@cs.odu.edu

ABSTRACT

We examine which NASA Technical Report Server (NTRS) repositories have received the most downloads during 15 months of operation. In particular, we explore the collection development policy of including non-NASA scientific, technology and medicine (STM) repositories. We found that three of the four non-NASA repositories included in NTRS contributed little to the overall download totals.

Categories and Subject Descriptors

H.3.7 [Information Storage and Retrieval]: Digital Libraries.

1. INTRODUCTION

The NASA Technical Report Server (NTRS) is an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) compliant aggregator [1]. Using primarily OAI-PMH, it harvests metadata records from 17 repositories. When NTRS was created, there were few scientific, technology and medicine (STM) OAI-PMH repositories, so non-NASA STM repositories were included: arXiv.org physics eprint server, BioMed Central, Energy Citation Database (Department of Energy), and the Aeronautical Research Council (the UK equivalent of NASA's predecessor, NACA).

2. ANALYSIS

NTRS offers two searching modes: simple and advanced. In the simple mode, fielded searches are not supported and only NASA repositories are searched. In advanced searches, fielded searching is supported and users have the option of including non-NASA repositories in their search. Thus users never receive non-NASA results unless they explicitly request them. We examined NTRS logs from April 28, 2003 until June 30, 2004. NTRS is instrumented to record when a user requests a download for the full-text content. Table 1 and Figure 1 describe the repositories and their monthly use. The Energy Citation Database, BioMed Central and arXiv.org contributed little to the download totals, while contributing significantly to the total number of holdings in NTRS. ARC represents a significant number of downloads. This indicates users are willing to select non-NASA repositories from the advanced search interface (logs show the advanced search is used 2X as the simple), and the prominence of both NACA and ARC suggests an interest in historical aeronautical publications.

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, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

JCDL '05, June 7–11, 2005, Denver, Colorado, USA

Copyright 2005 ACM 1-58113-876-8/05/0006...\$5.00.

3. CONCLUSIONS

ARC was accessed frequently but the other non-NASA repositories were not despite their large holdings. The subject matter of ARC is similar to the NASA repositories, suggesting that NTRS remains aerospace-focused and the presence of other STM materials has yet to expand its user base. arXiv.org is perhaps the most well-known OAI-PMH repository and is harvested by at least 10 different OAI-PMH service providers, but its presence did not guarantee its use in NTRS.

Table 1. Repository Profiles

Total Downloads	Repository	Current Metadata Records	Estimated Full-Content %
151524	NASA LaRC	4903	100%
72122	NACA	7640	100%
65508	NASA JPL	19121	100%
10184	ARC UK	2647	100%
4493	NASA MSFC	546	100%
2413	NASA JSC	129	80%
1584	ECD DOE	20738	70%
1269	NASA CASI	494645	5%
1181	arXiv.org	272266	100%
809	NASA GISS	1771	40%
403	NASA Genesis	33	100%
390	NASA RIACS	61	100%
166	BioMed Central	18454	100%
52	NASA ARC	354	0%
14	NASA SSC	39	100%
2	NASA KSC	82	100%
1	NASA GSFC	11	100%

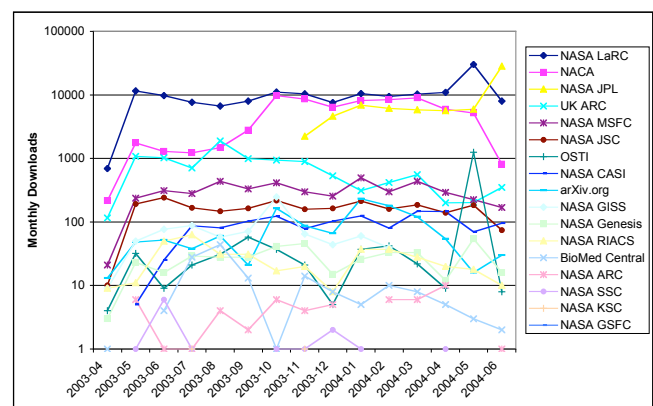


Figure 1. Monthly Downloads Per Repository

4. REFERENCES

- [1] Nelson, M.L., Rocker, J., Harrison, T.L. OAI and NASA Scientific and Technical Information, *Library Hi-Tech*, 21, 2 (2003), 140-150.