

péter's picks & pans



Péter Jacsó
University of Hawaii

CiteBaseSearch, Institute of Physics Archive, and Google's Index to Scholarly Archive

The crown jewel of the
Open Citation Project, in
my eyes, is the CiteBase
Search service.

This month I dedicate my column to some physics databases. Physicists, particularly those who specialize in astrophysics and high-energy physics, have had the best digital information resources both in the traditional and the new Web-based information world. These range from the high-quality, fee-based classic INSPEC indexing/abstracting database to the millions of free indexing/abstracting records made available by the American Physical Society (APS), the American Institute of Physics (AIP), and the Institute of Physics (IoP), as well as free full-text articles offered by the NASA/ADS repository and Cornell's arXiv.org preprint archive. Physicists deserve to be spoiled—they have been one of the most innovative groups in developing open access digital archives and free subsets of those. Along with the Institute for Scientific Information, they made the most out of cited references through Web links.

My picks include the CiteBase Search information service, one of the many superb research projects of the University of Southampton, and the archive of the Institute of Physics (IoP), with its impressive search engine. My pan is Google's very poor implementation of the index of the full-text journal article archives of nine prominent scholarly publishers, which illustrates that not all is gold that glitters or goes by the name of Google.



the picks

CITEBASE SEARCH

The University of Southampton (also known as *soton* from its Web address), the U.K. host of the arXiv.org preprint archives, is also the home of many innovative, open access digital projects—the eprints software that makes self-archiving of research papers quick and easy, the cogprints.org archive of research papers in cognitive sciences, the Journal of Digital Information, and Psycology. Most of these tools and services have been developed in the framework of the Open Citation Project, which was funded by the enlightened Joint Information Systems Committee (JISC) of the U.K.

The crown jewel of the Open Citation Project, in my eyes, is the CiteBase Search service [<http://citebase.eprints.org>]. CiteBase contains the nearly 300,000 records in arXiv.org, cogprints.org, and BioMed Central. You can search



citebase Search

[Help and About](#) | [Impact Health-Warning](#)

Citebase is currently only an experimental demonstration. Users are cautioned not to use it for academic evaluation yet. Citation coverage and analysis is incomplete and hit coverage and analysis is both incomplete and noisy.

[Metadata](#) [Citation](#) [OAI Identifier](#)

Author(s) (explanation)

Title/Abstract Keywords "energy transfer"

Publication title

Creation Date from until

Rank matches by:

Showing 1 - 10 of 1491 found [1-100] Query took 3.568 seconds

Single and Double BFKL Pomeron
Hard Processes | Abstract/Citation
 167.00 Mueller, A H ; Patel, Bimal
 ... Onium-onium scattering at high en
 hard scattering in the large $\ln_c s$...
 expression is explicit and could be evaluated numerically. It has a ...

Dipole Picture of High Energy
 iv.org/hep-ph/9403256
 e a dipole picture of high energy
 zero while the other two have momentum transfer, t. This

CiteBase search results.

by article title, serials title, author, year, and abstract. This is useful but not unique. What makes CiteBase so attractive is that the results can be ranked by most cited author, most cited paper, and several other criteria related to how often authors and/or items are formally cited and accessed through the links provided by CiteBase to the full text. The number of look-ups are referred to as "hits" and "impact" in the parlance of CiteBase.

At the article level, the records include highly informative citation and impact statistics, enhanced by superb citation analysis/impact charts and tables, including co-citation analysis. You can even generate customized citation/impact charts (although the Java applet works a tad slowly) that provide an insight into the correlation between the citedness of the items and how often the items were looked up and how many hits they got.

The dominant part of the collection is related to physics. The chief developer of CiteBase, Tim Brody, provides ample warning about the trial nature of the citation analysis feature, which is based on autonomous extraction of references from full-text documents, but I did not find any glitches. The scope is currently limited to the U.K. mirror site of arXiv.org and two other archives. Even with these limitations, this project shows the perfect model for the ultimate advantages of not only self-archiving scholarly documents but also of linking to full text—and offering citation/impact analysis on the fly to help researchers make an informed decision in selecting the most-relevant papers on a topic from a combination of archives.

INSTITUTE OF PHYSICS ARCHIVE

In my July/August 2003 "Picks & Pans" column, I praised the American Institute of Physics' archive that recently was renamed to Scitation. Its British counterpart, the Institute of Physics (IoP), deserves as much praise for its outstanding archive that offers sophisticated features, many of them available free to anyone [<http://iop.org/EJ/search>]. All the IoP journals were digitized from their first issue, amounting to more than 100,000 articles. The full text is searchable, as well as the metadata (author, title, abstract, keywords). Searches can be limited to a

Full Text

<http://arXiv.org/abs/hep-ph/9403256>
 Nucl.Phys. B425 (1994) 471-488

Identifier (Harvest Date) [oai:arXiv.org/hep-ph/9403256](http://oai.arXiv.org/hep-ph/9403256) (2003-11-13)

dc:date 1994-03-10
 1994-03-09

dc:type text

- [Graph of this Article's Citation/Hit History](#)
- [All Articles Cited by this Article \(Reference List\)](#)
- [Top 5 Articles Citing this Article](#)
- [All Articles Citing this Article](#)
- [Top 5 Articles Co-cited with this Article](#)
- [All Articles Co-Cited with this Article](#)

This Article's Citation/Hits History ([explain?](#))

Use the [Correlation Generator](#) to explore the correlation between download impact ("hits") and citation impact.

Caution!	To this Article	To Authors (mean)
Citations Identified	167	95.5928
uk.arXiv.org Web Hits	10	11.1072

CiteBase citation and impact report options.

given time period and to 12 subdisciplines (plasma physics, applied physics, computer science, etc.) through an intuitive search template with pull-down menus and radio buttons.

The results can be sorted by date, author, affiliation, or relevance, and—very importantly—clustered by topics if there are more than 25 hits, activating Vivisimo's cluster engine by a simple click. My test search about "string theory" with the query phrase in the title yielded 158 hits, clustered into more than 20 topics, including the more specific topic of matrix string theory. Clicking on this topic narrowed the search down to 12 results. IoP's output options in terms of content, style, format, and destination are perfect.

The full text of the articles are available only for subscribers (except the ones from open access journals, such as the New Journal of Physics, and the articles chosen for the free IoP Select subset), but the list of cited and citing papers are available free—a feature not true at many other publishers' archives. There are links to preprint versions of cited and citing references hosted by the Stanford Linear Acceleration Center (SLAC) in the SPIRE archive, which in turn offers links to other archives, including CiteBase, with its impressive citation/impact analysis features. IoP's own HyperCite links take you to records of citing and cited articles in several archives, including IoP's, plus those of NASA/ADS and APS, as well as indexing/abstracting databases, such as PubMed and INSPEC (which requires subscription to display the I/A record, of course). IoP shows perfect integration of the features of the Verity software and Vivisimo, making the search process a delightful experience, highly efficient and flexible. Its linking to cited and citing references in other collections is exemplary and exudes synergy.



the pan

GOOGLE'S INDEX TO SCHOLARLY ARCHIVES

It is hard to believe, but not everything that is Googled is gold. Google

Search IOP Electronic Journals

Either: Search article headers and abstracts:

"string theory" in [Help](#)
 in
 in

Or: Search full text of articles:

[What is a cluster?](#)

Select year range:

Search all years [Help](#)
 Search from to

Select a journal, subject category or EJs Collection:

Search all journals [Help](#)
 Search specific journal(s)

[To select more than one journal, hold down the Control key (PC) or Option key (MAC)]

- Journal of Physics A: Mathematical and General
- Journal of Physics B: Atomic, Molecular and Optical Physics
[includes Journal of Physics B: Atomic and Molecular Physics]
- Journal of Physics: Condensed Matter
[includes Journal of Physics C: Solid State Physics]
- Journal of Physics F: Metal Physics
[includes Journal of Physics F: Metal Physics]
- Journal of Physics D: Applied Physics
[includes British Journal of Applied Physics]
- Journal of Physics G: Nuclear and Particle Physics
[includes Journal of Physics G: Nuclear Physics]

The search template of the native search engine of IoP

Search results clustered by subject

Additional search options

Export your search results, access your search history and save searches from the main search results page. Clustering is a new service. Please tell us what you think.



- * "string theory"
- [AdS](#)
- [Black hole](#)
- [Type 0 string theory](#)
- [Wave](#)
- [Matrix string theory](#)
- [Noncommutative](#)
- [Heterotic string theory](#)
- [Cosmological internal space](#)
- [Little string theory](#)
- [Sigma](#)
- [Duality cascade](#)
- [Non-commutative](#)
- [Acceleration, Universe](#)
- [Low-energy string](#)
- [Target space](#)
- [Two-dimensional string theory](#)
- [Inflation, Potential](#)

Category "string theory" contains 158 documents.

1. Chirality Change in String Theory

M. R. Douglas and C-G. Zhou, C-G. Zhou
Journal of High Energy Physics 2004 No 06 (June 2004) 014-014
[View article in new window](#)

It is known that string theory compactifications leading to low energy effective theories with different chiral matter content (e.g. different numbers of standard model generations) are connected through phase transitions, described by non-trivial quantum fixed point theories. We point out that such compactifications are also connected on a purely classical level, through transitions that can be described using standard effective field theory. We illustrate this with examples, including some in which the transition proceeds entirely through supersymmetric configurations.

2. The duality between IIB string theory on PP-wave and $\mathcal{N}=4$ SYM: a status report

Rodolfo Russo and Alessandro Tanzini
Classical and Quantum Gravity 21 No 10 (21 May 2004) S1265-S1295
[View article in new window](#)

The aim of this report is to give an overview of the duality between type IIB string theory on the maximally supersymmetric PP-wave and the BMN sector of the $\mathcal{N}=4$ super-Yang-Mills theory. The general features of the string and the field theory descriptions are reviewed, but the main focus of this report is on the comparison between the two sides of the duality. In particular, it is first explained how free IIB strings emerge on the gauge theory

Search results by Verity can be clustered by Vivisimo instantly.

was presented this spring with the full-text archive of nine scholarly publishers, material invisible to Google's spiders. It was a good idea to let users know about the presence of free bibliographic records with abstracts from thousands of high-quality journals (with links to the full-text versions for subscribers or pay-as-you go customers). However, the implementation of this project by Google is very poor. Google did a careless, rush job of indexing the full text of some of the articles in the archives, giving no consideration to its inherently rich set of metadata in a consistent pattern [www.google.com/cobrand?restrict=crossref&cof=AWPID%3AAbbd6d01e9a530922&q=].

In my test searches across five of the nine archives, each of which has good or excellent native search engines, Google's results kept fluctuating throughout the test period and showed inconsistency. Its only consistency was that it always yielded far fewer results than the native search engines. To simplify the test procedure, I created a polysearch engine for submitting the same query to the native and the Google search engines (limited to the specific publisher's domain) as a full-text and title-only search in one fell swoop for a series of tests.

Searching the IoP archive yielded Google's best results, but even they were 32 to 60 percent below what IoP's

Duplicates and triplicates were very common in many of Google's results, so its reported hits must be discounted....

native search engine produced in full-text searches of the test queries. For example, for the query "energy-loss", IoP's native search engine found 7,681 records; Google returned only 3,330.

This is not the end of the story. Sampling the results, I realized that in Google's results there is far less than meets the eye. The first hit appeared four times in Google's result list (although not all of them adjacent to each

other), and only once (as appropriate) in IoP's result list. Duplicates and triplicates were very common in many of Google's results, so its reported hits must be discounted, further weakening its results.

Don't believe it's a good thing when, in exceptional cases (like for a unique title search), you see more results from Google than from the native search. When submitting the query for a title-only search, IoP correctly yielded one record, Google lined up four "hits" for the single article. Title searching using Google's `intitle: prefix` in some archives worked very poorly or not at all (as in the Annual Reviews' archive), simply because the HTML `<TITLE>` field is not necessarily the same as the article title. Google did not bother to calibrate its indexing program, even when the publisher has the metadata tagged in Dublin Core notation, as in the case of Blackwell's archive, let alone add at least a checkbox to its search template to let the users limit their search to this special index (which is not easy to locate unless you are at one of the publishers' sites).

We know that even high-brow researchers and scientists (not to mention students) use Google to search for scholarly papers far more often than they use the publishers' digital archives and the indexing/abstracting or the full-text aggregated databases for which their library pays tens of thousands of dollars. Google's half-hearted implementation of nine publishers' archives (mostly applauded by the press) gives them false rationale to continue their habits. It gives real information professionals a Maalox moment realizing how much their patrons miss when they put all their eggs in one basket—one we now know is flimsy at best.

As the moderator of the session on "Searching for Proprietary Scholarly Content" at the 2004 Annual Meeting of the Society for Scholarly Publishing, I raised this issue. The representative from Google, who was one of the speakers, promised to look into the matter. I can only hope the situation will improve.

Péter Jacsó [jacsos@hawaii.edu] is professor of library & information science at the University of Hawaii's Department of Information and Computer Sciences.

Comments? E-mail letters to the editor to marydee@xmission.com.

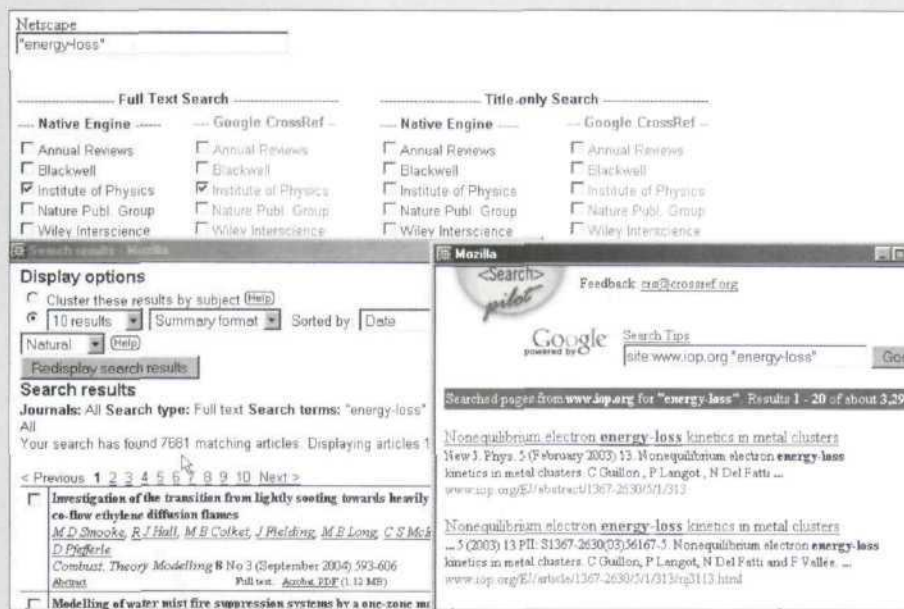


Figure 5. Full-text search submitted through my special polysearch engine using IoP's native software and Google.