LA-UR- 97-2710

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36

TITLE: **PHYSICAL REVIEW ONLINE ARCHIVES (PROLA)**

AUTHOR(S): T. Thomas, J. Davies, D. Kilman, F. Laroche, C. McEvilly, M. Nichols, R. Rivenburgh, M. Yan, D. Carstensen, R. Kelly

SUBMITTED TO: External Distribution - Hard Copy

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

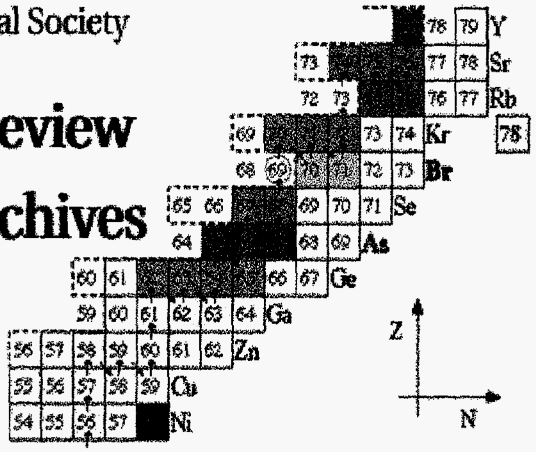
Los Alamos

Los Alamos National Laboratory
Los Alamos New Mexico 87545

7/3/97

American Physical Society

Physical Review OnLine Archives (PROLA)



PROLA: An OnLine Image Archive for the Journal Physical Review

Timothy Thomas, James Davies, David Kilman, Francois Laroche, Carlos McEvilly, Mojo Nichols, Reid Rivenburgh, Meilin Yan.

Los Alamos National Laboratory

Dale Carstensen

Los Alamos Microcomputer Professionals, Inc.

Robert Kelly

American Physical Society

May, 1997

1. INTRODUCTION

In cooperation with the American Physical Society, the Computer Research and Applications Group (CIC-3 -- see Section 13 for an acronym glossary) at Los Alamos National Laboratory has developed and deployed a journal archive system called, The *Physical Review OnLine Archive* (PROLA) It is intended to be a complete, full service on-line *archive* of the existing issues of the journal*Physical Review* from its inception to the advent of a full-service electronic version. The fundamental goals of PROLA are to provide screen-viewable and printable images of every article, with full-text and fielded search capability, good browsing features, direct article retrieval tools, and hyperlinking to all references, errata, and comments. The research focus is on transitioning large volumes of paper journals to a modern electronic environment.

2. BACKGROUND

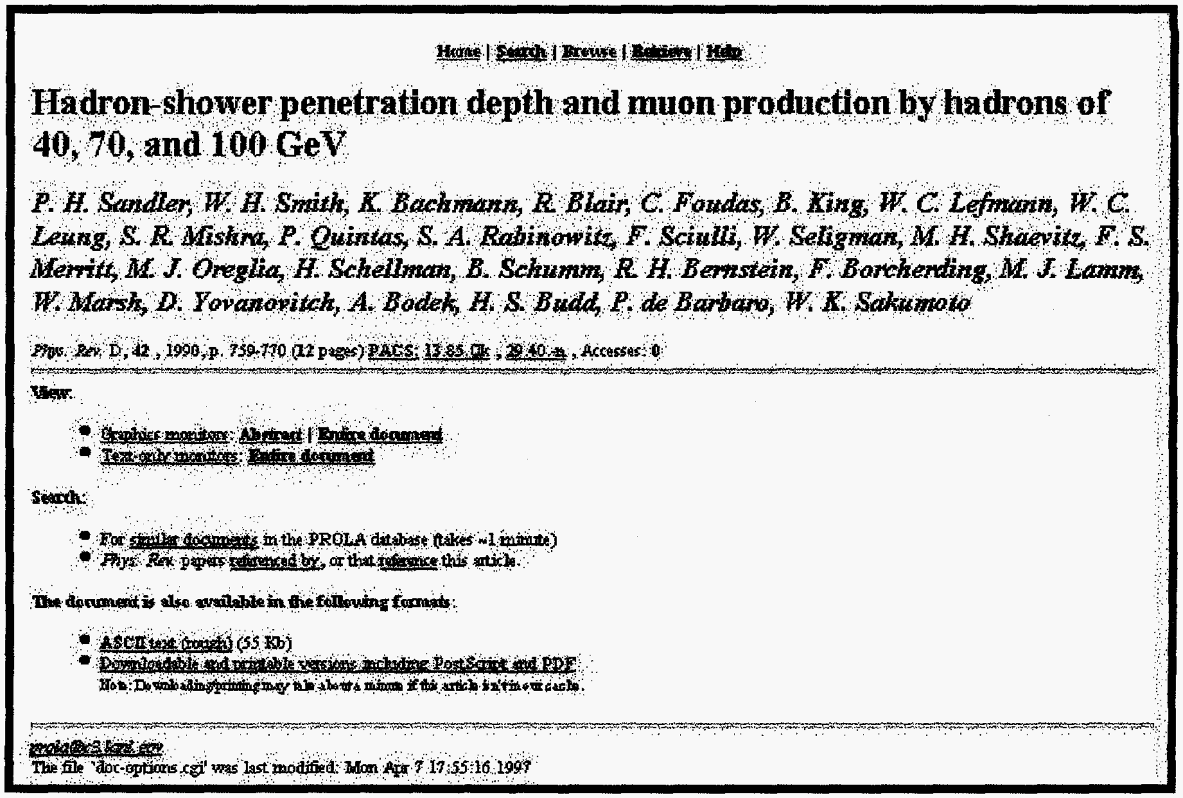
Work on PROLA began in the fall of 1994. Funding from the American Physical Society (APS) was in place by the Spring of 1995. The very early versions of PROLA preceded the World Wide Web (WWW) and were based on Wide Area Information Servers (WAIS). However, it was established very early on that the WWW was the distribution method of choice for images, and WAIS was retained only as the basic search engine. This decision was made primarily to insulate PROLA from the problem of fitting a system onto the widely divergent user systems, and to retain the collection at a single point thus providing easier control of quality and security. Since there was virtually no commercial software to support the early effort, it was decided to focus on adapting free-ware in the UNIX environment to fit PROLA's specific needs. Where no appropriate software was available, task specific code was written in C or Perl. This decision permitted the precise tailoring of PROLA to meet the unique requirements of the user community and to fit the peculiarities of the particular collection. It was decided that the problems of maintaining the huge number of files and managing the very large storage requirements would best be served by using the Los Alamos Common File System (CFS). That system is well maintained, operates 24 hours a day, and is very secure. In the interest of advancing Physics, Los Alamos volunteered the use of CFS free of charge during the developmental process. If PROLA became successful, and generated income for APS, then normal usage fees would be paid.

The problem of producing the scanned images was given to the Naval Research Laboratory (NRL). They were interested because they had a major in-house library image project called TORPEDO well underway and felt they could advantageously use their existing scanning facility. They also expected to learn something from collaborating with Los Alamos, and they would get access to certain *Physical Review* journals for inclusion in TORPEDO. The PROLA project took on the technical problem of producing screen viewable images for use within the WWW, and for creating user defined print image packages for delivery to the user's computer.

There are two versions of PROLA currently on-line --- an*alpha* version used to develop and debug the system, and a*beta* version (**http://www.c3.lanl.gov/prola**) that is used to test user reaction to the system. The beta system was made available to Los Alamos researchers during the Fall of 1996 covering sections A, B, C, D and E through the years 1989 to 1994. Since that time the years have been extended to include 1985 to 1995, and sections Letters has been added. This coverage totals approximately 125,000 articles, and about 700,000 pages.

3. Basic Design Features of PROLA

PROLA is a page-based system, with each article centered around a central document information (doc-info) page (see Fig. 1). Off the doc-info page hangs all the associated data attached to any particular article. This data is generally the page images, the references both forward and backward, the errata, the print versions, the ASCII versions, and the pdf versions. To say you are going to connect to an article, means you connect to that doc-info page, and then decide what type of specific information or format you require. Naturally, that page has complete bibliographic information. Like almost all pages in PROLA the doc-info page is constructed on the fly upon demand. This is primarily done to permit error correction without having to regenerate all the pages repeatedly, since errors which occur in one article, usually also occur in an unknown number of additional articles.



Home | Search | Browse | Reviews | Help

Hadron-shower penetration depth and muon production by hadrons of 40, 70, and 100 GeV

P. H. Sandler, W. H. Smith, K. Bachmann, R. Blair, C. Foudas, B. King, W. C. Lefmann, W. C. Leung, S. R. Mishra, P. Quintas, S. A. Rabinowitz, F. Sciulli, W. Seligman, M. H. Shaevitz, F. S. Merritt, M. J. Oreglia, H. Schellman, B. Schumm, R. H. Bernstein, F. Borcherding, M. J. Lamm, W. Marsh, D. Yovanovitch, A. Bodek, H. S. Budd, P. de Barbaro, W. K. Sakumoto

Phys. Rev. D, 42 , 1990 , p. 759-770 (12 pages) PACS: 13.85.Qk , 29.40.-n , Accesses: 0

View:

- Graphics monitors: Abstract | Entire document
- Text-only monitors: Entire document

Search:

- For similar documents in the PROLA database (takes ~1 minute)
- *Phys. Rev.* papers referenced by, or that reference this article.

The document is also available in the following formats:

- ASCII text (rough) (55 Kb)
- Downloadable and printable versions including PostScript and PDF
  Note: Downloading/printing may take about a minute if this article isn't in our cache.

prola@c3.lanl.gov
The file 'doc-options.cgi' was last modified: Mon Apr 7 17:55:16 1997

Figure 1. Doc-info page showing links to references, abstract, printables, etc.

All PROLA articles also have a special associated file which contains a brief, abbreviated, unparsed version of the bibliographic and reference information we have managed to obtain. This file has a variety of functions, but is primarily used to rapidly create reference pages, hits pages, and the doc-info page.

In principle PROLA offers any version of any article or page that the user community may want. For example, pdf was not a common format when PROLA was designed. But because of PROLA's open design, it was a very simple matter to add it to the doc-info page and produce it on the fly when it became popular. This design strategy frees PROLA from the vagaries of the market, just as its choice of WWW frees it from the demands of any particular type of user hardware.

Navigation features play a critical role in PROLA. Every page has a navigation bar at the top, and often the bottom, which permits the user to move through high level functions as well as functions specific to the area in which the user is currently located. Image navigation is done by both thumbnail images, and by entering image sequence numbers in a dialog box. Basic browsing functions are offered via a complete list of all journals by year, volume, section, issue, and page range. PROLA offers a more or less normal table of contents (TOC) page for each issue, containing the title, authors, and pages of each article. That TOC is navigable by type of article, such as brief report, errata, normal article, etc. A separate function called *retrieval* is used to go directly to a doc-info page when the full particulars of a reference are known.

Searching in PROLA is currently done by the WAIS search engine, but if warranted by user demand, additional search engines can be easily installed. PROLA actually has ready to go the GLIMPSE search engine should users demand it. It was temporarily deleted from PROLA because very few users choose it over WAIS. PROLA has defined fields for title, author, section, year, and full text.

4. Functional Features

4.1. Searching Features (See Fig 2)

•Boolean searches based on words in the text body, title, author, year or section (A,B,C,D,E or L), fields. The operators AND, OR, NOT and ADJ are supported.

•Display of the Boolean translation of the query submitted.

•Wild card searches based on truncation of terms.

•Adjacency operators that permit searches on words that must appear in a specific sequence, such as "plutonium oxide".

•Stemming techniques in which suffixes such as "ing", and "ed" are removed prior to searching. Thus a search on the term "operate" will match "operating", "operated", etc.

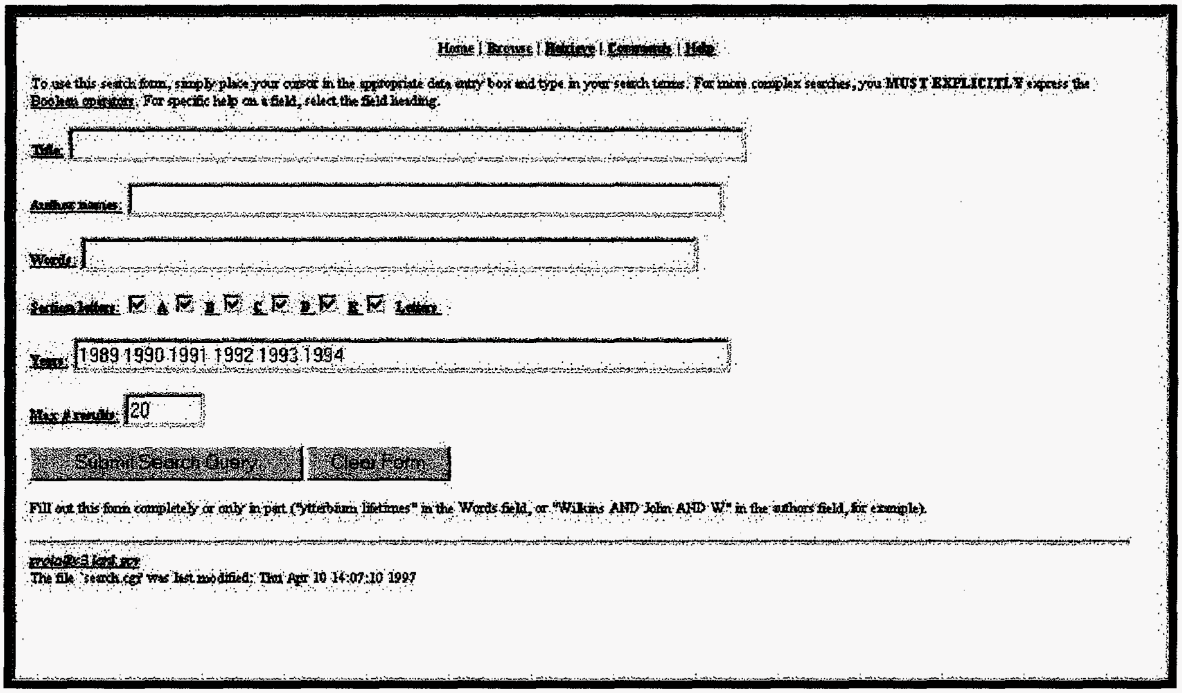
•Frequency weighting techniques so that each search term is weighted by the frequency of its occurrence, with rare terms having a greater influence on the document's ranking.

•Both Natural Language Queries and Keyword searches are acceptable.

•Relevance Feedback provides a list of documents that are similar to a given query document, based on a co-occurrence of significant words in the full text.

•The maximum number of documents that are returned after a search can be varied from 1 to 250.

•All documents found during a search are ranked according to the degree of match to the query terms.



To use this search form, simply place your cursor in the appropriate data entry box and type in your search terms. For more complex searches, you **MUST EXPLICITLY** express the Boolean operators. For specific help on a field, select the field heading.

Title:

Author names:

Words:

Section letter: ☑ A ☑ B ☑ C ☑ D ☑ E ☑ Letters

Years: 1989 1990 1991 1992 1993 1994

Max # results: 20

Fill out this form completely or only in part ("ytterbium lifetimes" in the Words field, or "Wilkins AND John AND W" in the authors field, for example).

prola@e3.lanl.gov
The file `search.cgi' was last modified: Thu Apr 10 14:07:10 1997

Figure 2. Search page interface to WAIS search engine.

4.2. Retrieval Features

•Any article in PROLA can be directly retrieved if its year, volume, section and page number are known. This helps remote retrieval, linking to other collections, and eliminates unnecessary searches.

•Articles within PROLA that are referenced by a PROLA article, can be retrieved from a "reference" list.

•Articles within PROLA that reference a particular article can be retrieved via a "referenced by" list.

•Errata can be retrieved directly from the article to which they apply. Similarly, the article corrected is retrievable directly from the erratum article.

4.3. Navigation Features

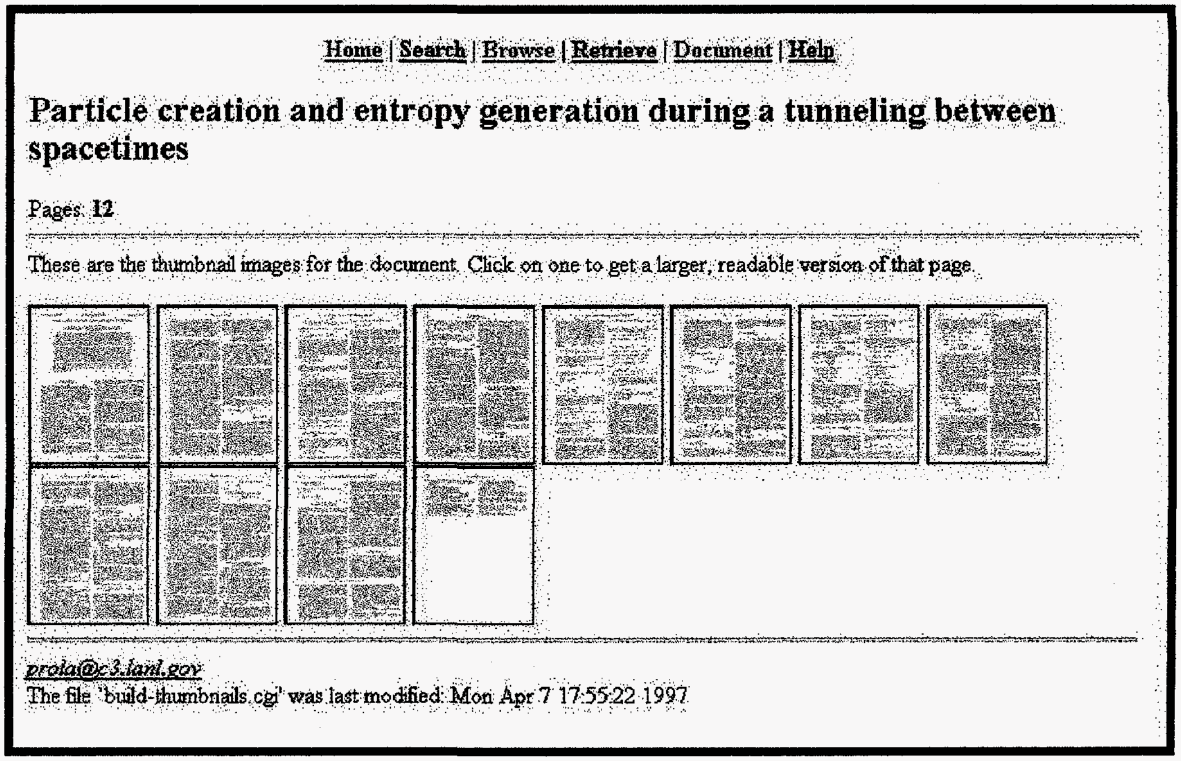
•All pages in PROLA have a navigation bar at the top that lets the user move effectively between functions.

•A set of Table of Contents pages lets the user browse authors and titles by the traditional means.

•Previous and Next links are added to many pages, where those concepts are clearly defined, such as browsing issues, or screen viewable images.

•Dialog boxes permitting the user to jump to specific pages in a sequence are available.

•Thumbnail versions of the pages are used to navigate between page images of an article. (See Figure 3)



Particle creation and entropy generation during a tunneling between spacetimes

Pages: **12**

These are the thumbnail images for the document. Click on one to get a larger, readable version of that page.

prola@c3.lanl.gov

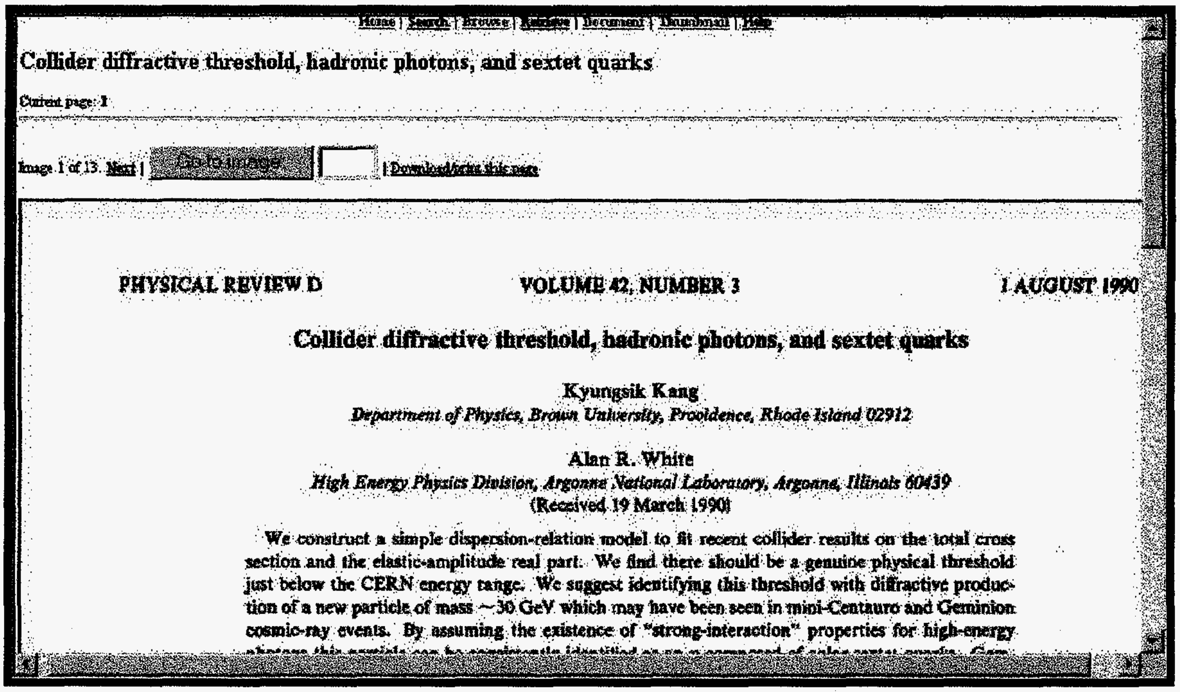
The file "build-thumbnails.cgi" was last modified: Mon Apr 7 17:55:22 1997

Figure 3. 8-bit thumbnail intradocument navigation aids.

4.4. Display Features

•Screen viewable images are provided at a resolution of 100 dpi for easy full page display on common monitors. (See Fig. 4)

•Portable Document Format (pdf) versions of the article page images are provided for those who prefer this method of viewing images on screen. These images are 300 dpi, but pdf viewers provide full zooming capabilities.


Home | Search | Browse | Retrieve | Document | Thumbnail | Help
Go to image
| Download/print this page
Next

Collider diffractive threshold, hadronic photons, and sextet quarks

Current page: 1

Image 1 of 13.

PHYSICAL REVIEW D VOLUME 42, NUMBER 3 1 AUGUST 1990

Collider diffractive threshold, hadronic photons, and sextet quarks

Kyungsik Kang
Department of Physics, Brown University, Providence, Rhode Island 02912

Alan R. White
High Energy Physics Division, Argonne National Laboratory, Argonne, Illinois 60439
(Received 19 March 1990)

We construct a simple dispersion-relation model to fit recent collider results on the total cross section and the elastic-amplitude real part. We find there should be a genuine physical threshold just below the CERN energy tange. We suggest identifying this threshold with diffractive production of a new particle of mass ~30 GeV which may have been seen in mini-Centauro and Geminion cosmic-ray events. By assuming the existence of "strong-interaction" properties for high-energy

Figure 4. 100-dpi, screen viewable greyscale image of the scanned article.

4.5. Format features

•Printable Postscript versions (300 dpi) are provided. The user can download a package containing compressed images of any combination of article pages.

•A variety of compression methods are available to the user.

•ASCII versions of every article are available primarily as a convenience for "cutting and pasting".

4.6. Other Features

•A browser linked from the article lets the user interpret the subject index code (PACS) used to classify every article.

•An access count is available where appropriate letting the user know how many times a particular article has been touched.

•A cache system is used for printable and screen-viewable versions. This permits rarely touched articles to be stored remotely on tape, while offering quick access to commonly touched articles.

5. Technical Description

5.1. Hardware

PROLA hardware consists of development machines, a web server, a local disk storage system, a centralized disk/tape storage system, tape readers, a suite of UNIX workstations used occasionally for image processing, and a bank of old workstations used exclusively for PROLA image processing.

5.1.1. Development Machines

PROLA uses 5 development workstations. All are SPARC UNIX Machines ranging from 12.5 MIPS to around 60 MIPS. Four are Tatung 385 machines, one a Sun machine. One is assigned to the problem of creating the browse tables and the ASCII versions; one to maintaining the web server and developing cgi scripts, one to managing the image processing problems, one to managing the project, and one (without a monitor) to running cgi scripts.

5.1.2. Web server

PROLA shares its web server with the Computer Research Group at Los Alamos. It is a SPARC 10 with 47 Meg of memory. It is connected to the internet via the Los Alamos LAN, a 10 base T minimum, to a fiberoptic FDDI connection to the Internet.

5.1.3. Local Disk

Local disk storage consists of four 4 gigabyte disks on the an Auspex LAN server. PROLA also has disks on each of its development machines. One Auspex disk is partitioned into a single 4-gig partition and is used only for temporary storage since it is reserved for reindexing the dataset when required. PROLA also reserves 2 gigabytes for image cache storage, and keeps all thumbnails, static pages, ASCII versions etc. on local disk. Altogether PROLA has 22.5 gigabytes of local storage.

5.1.4. Central File System

All large page images are stored on CFS. These include the 300 dpi scanned tiff images for printing, the 100 dpi byte images for screen viewing, backups, and copies of tapes received from APS and NRL. When images are requested, they are retrieved from tape by a mechanical robot, moved to CFS disk and then to our local cache. There they remain until removed when they become the oldest untouched image in the over-full cache.

5.1.5. Image Processing Machines

PROLA has developed an in-house method of doing the image processing via distribution of the processing tasks to all unused machines in the local LAN. This is a maximum of 64 workstations, but at any given time a much smaller number are actually available (i.e. not being used by someone else). PROLA has taken sole possession of 12 old SPARC 1 machines and 3 IPX's that were retired from normal usage, and uses them without monitors to do image processing.

5.2. Data Elements

PROLA has a wide variety of data elements, including information and navigation pages, article components, indexes, browsing elements, help pages, and pages for searching, retrieving, printing, and downloading. It also has pages for converting formats, interpreting subject code keyword (PACS) numbers, etc. PROLA is page based and each page is identified by its year, volume, section, and page number. Where there are two articles on a page, they are identified by a subscript to the page number.

5.2.1. Article components

Each article offered in PROLA has a number of data elements associated with it. Together they make up the "document". There is no single element that defines a document, and not every document has every element. However, they all have a set of images associated with them, and some basic bibliographic information.

5.2.1.1.Gif Images

Each page has a 100 dpi byte image produced by down sampling from the scanned 300 dpi tiff bit image with an anti-aliasing technique. This is a computationally intensive effort, and is responsible for the need to use a distributed processing technique to process images. Typical size is 81 Kbytes/page.

5.2.1.2.Tiff Images

Each page has a 300 dpi tiff image that is compressed and bundled in articles. This bundle is retrieved when someone wants to print the article or view it in a pdf viewer. Typical size is 89 Kbytes/page.

5.2.1.3.Thumbnail Images

Each page has a 10 dpi thumbnail byte image produced by anti-aliasing down sampling. These images are used for page navigation. Typical size is 1 Kbytes/page.

5.2.1.4.Formatted ASCII version

Most articles from 1985 to the present have a formatted ASCII file that was used at some stage of the publication process to produce the page image. The formats used are versions of TROFF, TeX, SGML, and xfs. Prior to 1985 this file will be replaced by the raw OCR output file. Typical size is about 5 Kbytes/page.

5.2.1.5.Abbreviate Bibliographic Info file

Each article has a file containing the title, author and list of references from the article. These are in the original format supplied by APS. By keeping this file in its original version (of course with the full text removed), PROLA is always referring directly to the most accurate data we have, and any parsing errors found and corrected are immediately implemented for all instances of the error. These files are read by the cgi scripts that need that information, such as the "hits" page returned from a search and the Doc-info page. This technique serves to help in correcting errors, and in quickly generating any page that needs bibliographic information. Typical size is 21 Kbytes/article.

5.2.1.6.Display ASCII version

Each article has an ASCII file that contains an approximation of the published article. It does occasionally have decent tables in it, and some slightly useable formulas, but in general all complicated formatted elements are not useable in this form. It is offered primarily as a convenience for cutting and pasting text. Of course it is also the basis for the full text search. It is produced in a number of different ways, depending on the type of formatted ASCII we received from APS. Typical size is about 4 Kbytes/page.

5.2.1.7.Reference to list

Each article has a list all other articles in PROLA that reference it. These are all hyperlinked back to the referencing article. These are created on the fly from a file created for each section/volume. Typical size is 40 Kbytes/section/volume.

5.2.1.8.Reference by list

Each article has a list of all Physical Review articles that it references. Those that exist in PROLA are hyperlinked. These are created on the fly.

5.2.1.9.Errata backward reference links

Each article that has an errata posted for it, has an entry in a separate file for each volume/section containing the linkage between the articles and their errata. Typical size is 3 Kbytes, and there are 133 such files so far.

5.2.1.10.Bibliographic Search Fields

Each article has a list of the fields indexed by the WAIS search engine, they are placed at the top of the ASCII version, along with a warning that the ASCII version shown is NOT accurate.

5.2.1.11.Access Count File

Each article has a file that contains the count of the number of times the Doc-info page has been created for users. This lets a user see how popular any article is. This count appears on the hits list, where it will do the most good, but also appears on the Doc-info page. It does not appear on the TOC since that page is batch generated at infrequent intervals. This file is only a few bytes in size.

5.2.2. Index Elements

PROLA is indexed with a commercial WAIS system. It is very efficient at producing a small, compressed index and is very fast at completing the search. Producing the hits page (which is a PROLA, not WAIS function) however is not as fast as we would like, but PROLA feels it is important to give the user a nice chunk of info on that page. The completed index is now 1.08 Gig in size. As we start indexing OCR output we expect it to grow, since OCR errors will produce lots of "words" that we have not seen before.

WAIS should scale very nicely to the entire collection. Producing the WAIS index however is a major problem. It currently takes about 3 days, and requires a single partition of 4 gig. WAIS has poor recovery techniques, so it often takes us several tries to get a finished index. We now keep the old index in place while we produce the new one, thus requiring more available disk space. We also keep one copy of the index in back-up on CFS.

5.2.3. Browsing Functions

The table of contents pages are the most important parts of the browsing technique. In contrast to most important PROLA pages, these pages are not generated on the fly, but are generated by a batch mode, then stored as static files. This is because they took too long to generate on the fly. They serve not only to let the users find their way around, but also lets PROLA correctly unbundle the unlabeled tiff image files produced by NRL. Any error in these pages causes the image processing to fail since it creates an equivalent error in identifying which image goes with which article. We have scripts designed to help us locate and fix any such errors.

5.2.4. PACS Explanation

We have constructed a PACS tree that links from the Doc-info page. It lets the user interpret PACS numbers in terms of topics. We plan to add this field to the WAIS index, but have not yet done so.

5.2.5. Help Files

Most PROLA pages have an associated help file which explain the referring page and addresses what we felt would be common problems.

5.2.6. Printing Tools

Almost all printing techniques are done on the fly. They all begin with a tar file that contains all the tiff images of the article's pages. We consider pdf to be a printing aid, since we feel our provided Gif images are much better than pdf for reading page images on the screen. The main function needed for printing is the selection of specific pages, and the ability to download only those. Without this functionality, most PC users would be severely handicapped by slow modems, and limited storage space. Also, formats specific to the various common printers must be available.. This means that a relatively complex set of pages are needed to produce this functionality. Printing the Gif image with the Browser print function is unsatisfactory because of the low resolution.

6. Current Coverage

6.1. ASCII Versions

This section refers to the alpha server (see Section 2), since that is most advanced. Basically we have the ASCII versions of 11 years in place for A,B,C,D,E and L, 1985 to 1995. An as yet unknown number or articles have only the abstract and bibliographic info in ASCII. We believe there are about 8000 of these, spread mostly in 85 to 89 and Letters.

Where the complete full text of the article is missing, but the abbreviated ASCII is present, everything looks normal, including the images. The only problem is that the full text cannot be searched, and the cut and paste function cannot be used.

That said, we still have some serious problems in the ASCII area. Namely we do not always do the best possible job in creating the ASCII versions. With effort the presentation of both tables and figures could likely be improved. Presenting the TeX or TROFF versions received from APS is not a viable option, since in very many cases those files do not contain a complete version of the published article (which was actually produced by a photo-offset process by cutting and pasting from a variety of printed files), and in many cases the macros and style files necessary to accurately interpret the markups have been lost.

6.2. Images

At the present time we have images installed for sections A, B, C, D, and E for the years 1985-1994. We are in the process of obtaining additional images from NRL covering Letters and 1995. Plans to obtain pre-1985 images are being formulated.

7. Quality Assurance

PROLA works very hard to produce a high quality product. One aspect of this effort is providing as much functionality as possible, another is to provide continuous prompt service, and the last is to provide error free content. At the present time we are quite happy with the functionality, reasonably satisfied with the service level, but not yet comfortable with the content errors.

Considerable effort has been put into the system to provide a mechanism for correcting errors. In 1995 PROLA was completely redesigned specifically for error correction purposes. At that time we began to fully appreciate the magnitude of the error problem. Basically the redesign focused on creating all pages dynamically so that we could continuously add error detection and correction code to the Perl scripts that formatted each page. In general this has been a success, and the overall level of quality is greatly improved.

It is anticipated that quality control will be a continuous problem. It will need more or less constant attention for the foreseeable future. However, as time goes on it will wrap more and more into the problem of keeping the references updated and correct, and adding new features as user preferences and technology change. The very rapid rate of technology change is likely to be a constant, and adjusting to it will require a small team of knowledgeable people continuously working on the new problems.

7.1. Content Errors

7.1.1. Individual Errors

Some errors are very specific to a particular article, such as a misspelling of a name, an error in the page number, or a mislabeling of a field. These, when we discover them, are corrected quickly and easily with a standard procedure that lets us simply edit in the correction, while saving the original. We do this whenever we think the error is not general. For example the last one we fixed was a misspelling in the title which occurred because the TROFF file had multiple copies of the title. The version our script found happened to have a misspelling, the original typesetter used the correctly spelled version. In this case we just corrected the spelling, and the next time the title was created for a page it was correct. If we had felt this was a general problem, we would have addressed the issue of multiple titles.

7.1.2. General Errors

General errors are those that we have reason to believe occur in a substantial number of cases. They must be fixed by correcting the Perl scripts to recognize the case and adjust for it. These are things like mishandling of long author fields when different name separators are used, particularly when the names are separated by institutional identifiers in novel ways. Other common remaining errors concern how to handle odd words slipped into references, or odd macros used to tag non-Latin1 characters. What is really needed now is some thought directed at errors that have not yet been detected by users. Fixing errors we don't yet know about is a major problem.

7.2. Service Problems

We have occasionally been plagued with a rash of web server problems of unknown origin. We have repeatedly addressed this problem by upgrading the server or the network, only to have new problems emerge later. Therefore we may have come to the reluctant conclusion that maintaining a heavily used working site, will require a continuous effort at improvements and error correction.

For our current loads, the connection speeds and compute times are reasonable --- with one exception. We strongly believe it takes too long to retrieve and prepare a print package from tape archive. We imagined our cache system would solve this problem, and it may, but meanwhile it is a real issue for which we do not have an obvious solution.

One problem in this area we did solve, was the time it took to create and display the TOC. Normally this page would be created on the fly in order to get the latest error correction codes, but now we to the creation in batch mode whenever we get the urge, and display it from a static copy. This has greatly increased the speed of browsing.

8. Security

PROLA is protected by the security systems in place at Los Alamos for commercial, proprietary information. While this offers reasonably good protection, it is no real guarantee that intrusions and destructive attacks cannot occur. We are not aware of any intrusion problem, and our system is continuously monitored for the purpose of detecting any such attack.

A no-robot policy is currently in place, but that policy is currently under review, since it restricts harvest indexes from indexing PROLA, which may not be desirable.

We have designed a system for authenticating the images in PROLA to prevent alteration. and to prevent massive downloads. This will be installed as soon as it is finished.

9. Storage Requirements

9.1. CFS storage.

Computations for storage requirements depend to a large degree on the number of pages to be archived. Unfortunately we do not have a fully accurate count. It has not been critical up to this point to resolve the different numbers that the different counting methods produce. We could do so, but it would be a serious effort, and for now I choose to just use approximate numbers that sort of split the difference between methods. For example for the number of pages archived from '85 to '95, APS reports 717,007 based on retrieval from the APS database. By hand counting the actual pages in the library, we come up with 686,706. Obviously there are probably errors in both methods, but for current purposes I just estimate that there are 700,000 pages during this period. Counting the number of articles in the TOC with a script (which surely has specific errors here and there) we come up with 105,000 articles in the 11 years. The estimated total number of pages from 1893 to 1995, based on the library counting technique, is 1,511,483.

In addition to page elements CFS also stores all backups, which are created automatically whenever changes are made to any part of the PROLA system. CFS also stores backups of the original files received from NRL, and a copy of the index.

Using this information we can produce the following tables that estimate storage requirements for various data elements.

9.2. Table 1: CFS storage requirements (1985-1995)

Item	*K bytes/page*	*pages*	*storage (Gigabytes)*
GIF	81	700,000	56.7
TIFF	89	700,000	62.3
TROFF/TeX	5	700,000	3.5
Backups	-	-	89.8
Total	-	-	212.3

9.3. Table 2 CFS Storage Requirements (1893-1995).

Item	*K bytes/page*	*pages*	*storage (Gigabytes)*
GIF	81	1,511,000	122.4
TIFF	89	1,511,000	134.5
TROFF/TeX	5	1,511,000	7.6
Backups	-	-	188.1
Total	-	-	452.6

However, when evaluating Tables 1 and 2, it should be remembered that so far the storage has been cost free, we have not compressed nearly as much as we could. However, compression would increase the computational requirements and degrade performance a bit. There is not much advantage to further compressing the tiffs or gifs, but the ASCII could be very effectively compressed.

We currently have 105 gigabytes of data stored on CFS. These include some of our back-ups (which we don't know the size of, since they are handled by another system), tiff's, gif's, and old versions and residuals which could be cleaned up. PROLA also currently has 319,896 files on CFS. These are primarily directories for each article, and tar files with the previews and the tiffs. There is a minimum 3 files/article required.

9.4. Local Storage

We currently have 22 gigs, with another 4 gig disk on order. This will bring us to about 24 gigs for our local use, which is mostly indexing space, cache space, development space, thumbnails, and ASCII versions. This should be adequate for the 11 years. We would expect that another 10 or 12 gig would cover the full collection.

10. Costs

Current, un-negotiated prices for CFS usage are: $6.00/month/gig; $0.03/month/file; and $0.10/file-access. On top of these charges are overhead and burden charges that are assessed for any services purchased in the Lab. These can be substantial, currently amounting to 38% to support the group, and 6% for the Computing Division, and 43% of the total for the Laboratory, producing approximately 106%.

Working backwards from the number of pages and articles we can estimate a current minimum for the 11 years under consideration, as detailed in Table 3.

10.1.1.1.Table 3. CFS Estimated Costs (1985-1995)

item	*amount*	*price*	*cost/month*	*full overhead*
files	315,000 files	$0.03/month	$9,540	$10,112
storage	212.3 gig	$6.00/month	$1,273	$1,349
access	10,000 downloads	$0.10/access	$1,000	$1,060
Total			$10,813	$11,461

We see that CFS charges of about $20K/month would be incurred. However, those charges are almost entirely from the per-file charge. The storage charges are minimal. By redesigning the system, to place identifying information in the file name, we could eliminate the directory "file" charge reducing the cost by 1/3. We could also bundle articles together, since the delivery time is not affected by file size much, mostly by the transfer setup time. With this in mind, it seems likely that the high per file charge is due to a misconception on the part of CFS, and we could negotiate it to a much smaller figure. I suspect that we could get the CFS charges to about $80K/year without much trouble, but then we have a tax of 106% to deal with. So my best current estimate is about $170K/year for CFS charges for the 11 years. We could surely do better if we used our local disk more effectively and put our minds to reducing cost.

The full realization of this cost, combined with the slow retrieval from tape, and the constantly decreasing cost of self maintained storage systems, has recently encouraged us to reconsider our long held opinion that CFS was the way to go. At the present time we are uncertain as to the best way to handle the huge storage problems confronting PROLA.

Using the same method of estimation, Table 4 shows estimates for the entire collection.

10.1.1.2.Table 4. CFS Estimated Costs (1985-1995

item	*amount*	*price*	*cost/month*	*full overhead*
files	667,000 files	$0.03/month	$20,010	$21,211
storage	452.6 gig	$6.00/month	$2,716	$2,878
access	15,000 downloads	$0.10/access	$1,500	$1,590
Total			$24,226	$25,679

10.1.2. Personnel Costs

To maintain PROLA on a continuous, professional footing will likely require 1 half time project manager at about $100K, and two technicians at $40K. One technician will specialize in the Physical Review Collection, and the other on the WWW, the various media formats, and the cgi scripts. If the system is not at Los Alamos, access to a system operations, networking person will be required. At Los Alamos that cost is included in the personnel and equipment costs.

10.1.3. Equipment Costs

We currently have adequate equipment. We may have to add another 10 gigs or so to local storage, but that cost is relatively minor. The only major equipment costs will be related to substituting our own storage for CFS, or upgrading our server.

10.1.4. Network Access Costs

We have not explored this area since it is provided as part of the Los Alamos environment.

10.2. Charging for PROLA access.

At the present time we at PROLA imagine that there will be two methods of recovering the costs of PROLA: 1) by subscription along with library subscriptions or on-line journal subscriptions, and 2) by a pay-per-view method for access to the full articles for those physicists without access to a subscription. The subscription method could be authorized by password, or as now, by IP address. Pay-per-view technology is not as mature, but we have begun negotiations with Carnegie Mellon University to act as a test site for their NetBill software.

11. Future Issues

11.1. Optical Character Recognition

The big issue for the immediate future is how we are going to handle the OCR. In particular how we are going to get the bibliographic information extracted. We have discussed this issue extensively with NRL, and would like to begin a practice run on the "abstract only" problem. We would send NRL our images of the pages for which we do not have ASCII text, and NRL would OCR them and return the ASCII version.

11.2. Search Engines

Search engine technology is rapidly improving and users are becoming more sophisticated. Therefore we anticipate the need to purchase a more modern search engine within the next 3 years. In the short term we want to modify the way "stop words" are handled, by eliminating most of them. We would also like to be able to highlight the location in either the images or the ASCII of the search terms used.

11.3. Linking to other Physics Articles

Of particular interest is the collection of the forward references from all APS on-line journals, so that the "reference this article" list on the doc-info page is as up-to-date as possible. Of course, with reciprocal arrangements, it would be nice to include the preprint archives, the SLAC databases, and other physics journals in this system as well.

11.4. New Features

We would like to build several new functions, including a PAC's browser, a reference checker, a better monitoring tool, an image authentication system, an enhanced cache system, a better security system, and a billing system.

11.5. WWW Design

PROLA is beginning to age in its design. We would like to upgrade the design to modern web standards, using frames, Java, and more images.

12. Special Considerations

PROLA benefits enormously from the support of the Los Alamos National Laboratory. Much of this support is given freely in the interest of advancing Physics. Much of this support is in consulting and advice, much is in dedicated hardware, and some is in un-billed services. But beyond that, it is in the general willingness to see the project succeed. Without that support PROLA could not happen at the costs currently incurred.

13. Acronym Glossary

1. **10baseT** - Twisted-pair 10-megabit/sec. Ethernet
2. **ADJ** - Adjacent. Boolean operator indicating two strings adjacent to each other.
3. **ASCII** - American Standard Code for Information Interchange.
4. **Auspex**: Computer Manufacturer, specifically a UNIX file server computer.
5. **CFS** - Common File System. A LANL designed, disk/tape storage system that uses robots to retrieve tapes, and transfers files to the users workstation.
6. - **CIC-3** - Computer and Information and Communications Group 3. The Computer Research and Applications group at LANL.
7. **FDDI** - Fiber Distributed Data Interface. A 100-megabyte/seconds token ring network medium.
8. **GLIMPSE** - A search engine based on the UNIX grep command, with modifications to allow fuzzy searches.
9. **Gig** - Gigabyte, ~1,000,000,000 Bytes. Also GB.
10. **IP** - Internet Protocol
11. **IPX** - Sum computer model, 40 MHz clock and 28.5 MIPS
12. **K** - Thousand
13. **Kbytes** - Kilobytes, ~1,000 bytes
14. **LAN** - Local Area Network
15. **LANL** - Los Alamos National Laboratory
16. **Latin1** - Standard character set for Western European Languages
17. **LX** - Sum computer modes, about 30 MIPS
18. **LZW** - Lempel-Ziv-Welch. Welch's modification of Lempel-Ziv compression algorithm.
19. **Meg** - Megabyte. ~1,000,000 bytes. Also MB.
20. **MIPS** - Million Instructions per Second
21. **NRL** - Naval Research Labs

22. **OCR** - Optical Character Recognition
23. **PACS** - Physics and Astronomy Classification Scheme. A subject code originated by the American Institute of Physics.
24. **PC** - Personal Computer
25. **PROLA** - Physical Review OnLine Archives
26. **SGML** - Stands Generalized Markup Language
27. **SLAC** - Stanford Linear Accelerator Center
28. **TOC** - Table of Contents
29. **TROFF** - Typesetter Runoff. Early UNIX text formatting language
30. **TeX** - Greek: Tau, Epsilon, Chi. A text formatting language with advanced equation formatting capabilities.
31. **UNIX** - An efficient multi-tasking operating system developed at Bell Labs.
32. **WAIS** - Wide Area Information Server
33. **dpi** - Dots per inch
34. **Gif** - Graphic Interchange Format, an image coding scheme that originated with Compuserve and uses LZW compression.
35. **pdf** - portable document format. Associated with Adobe Acrobat.
36. **tar** - tape archive
37. **tiff** - tag image file format. Originated by Aldus
38. **xfs** - Proprietary text formatting language used by the American Institute of Physics to format text.

14. References

Schatz, B. R. Information Retrieval in Digital Libraries: Bringing Search to the Net. *Science*, **275**, 1997, 327-334. -- This is a decent review of the state-of-the-art in digital libraries, which will put this work in context.

There are many, many efforts that place document image archives on-line. Interested readers could profitably start their exploration at any of the following sites:

http://www.jstor.org -- A similar effort to PROLA applied to Humanities and Social Science journals.

http://cs-tr.cs.cornell.edu -- This is one place to enter the Dienst system which displays indexed images of technical reports at a large selection of university libraries.