The Global Village Come True: High-Tech Information Network in High Energy Physics

Ann-Sofi Israelsson
CERN Scientific Information Service

Mogens Sandfær
CERN Scientific Information Service

Stephan Schwarz
CERN Scientific Information Service

THE GLOBAL VILLAGE COME TRUE: HIGH-TECH INFORMATION NETWORK IN HIGH ENERGY PHYSICS

Ann-Sofi Israelsson, Mogens Sandfær, and Stephan Schwarz
CERN Scientific Information Service
Geneva, Switzerland

PREAMBLE: THE GLOBAL VILLAGE

In times of fast development, actors in the center of events have a distinctive advantage. Effective participation requires current knowledge of what is going on ("the right information in the right form at the right time"), reliable channels of feedback, and visibility. The idea of a "global village", when introduced in the sociology of communication some decades back, expresses the confidence in modern communications and computing technology to overcome the obstacle of distance and time delays. This, as we all know, is an oversimplification: the complexities of human communication cannot be reduced to signals and processes in cables and computers. Yet, there are areas, for example in well-defined scientific areas with coherent research communities and a "lingua franca", where access to, and dissemination of current information makes global participation possible. It is in this sense that recent developments in information services in High Energy Physics (HEP) make the label of a "global village" justified by covering the entire research community, including groups in the Third World. The solutions found may prove to be relevant to many other areas as a pilot study.

THE HEP COMMUNITY

The area is High Energy Physics (or Particle Physics) which is concerned with the study of the fundamental properties of the physical world: how and why things are the way they are, right back to the origin of the universe. To achieve experimental conditions revealing such properties, one has to build particle accelerators of ever greater energies. This requires cooperation between physicists and technicians in a wide range of high-tech areas, including materials, vacuum, superconductivity, electronics, computers. In fact, recent advances in accelerator and detector technology contribute a lot to such areas. A brief but very informative and readable historical account is given in "The Hunting of the Quark" by Michael Riordan. [1]

A recent directory [2] lists about 1000 HEP institutes and laboratories from 70 countries, but only few of these have their own accelerators. In most cases, experimental physicists participate in huge international collaborations, related to experiments at the major accelerator laboratories. One of these is CERN, the European Laboratory for Particle Physics near Geneva on the border between Switzerland and France. CERN has an annual budget of about $700 M, 3500 employees, 5000 associates not on CERN payroll, and over 200 universities/ institutions participating in its projects. Another is SLAC (Stanford Linear Accelerator Center), our principal US counterpart in information services.

TIMELINESS, GREY LITERATURE, AND THE "PREPRINTS CULTURE"

Scientific communication has always had an important element of informal communication, the real progress happening within "invisible colleges" in the sense of Derek Price [3] by direct contacts, letters, and transmission of manuscripts (nowadays even by Fax or E-mail) long before publishing . However, this process can work only on a limited scale, providing access by privilege. In any dynamic science the time factor is of prime importance for all participants and may call for new ways to disseminate information, when the

delays of the traditional publishing system are unacceptable. The advent of electronic journals is a case in point for sophisticated solutions, but it requires technical coordination and compatibility of input and advanced equipment for access. A more conventional but highly effective method (although with obvious financial implications) is the production of local print runs for directed distribution to colleagues and institutions.

Curiously, it is through the dependence of the HEP community on the rapid and reliable distribution of this kind of "clandestine publishing" (a typical case of "grey literature", normally not manageable by libraries), that a service could be designed and established, meeting the rather extreme information requirement of the community. To explain this peculiarity and its implications, a few remarks on the concept of grey literature are needed.

The term "grey literature" covers a wide range of quasi-publications, with the feature in common that they are not really public, i.e. easily accessible on the publishing market and in libraries other than in highly specialized collections. There are different reasons for avoiding the regular publishing market: reasons of cost, commercial interest, confidentiality, a limited audience, the speed of dissemination, etc.. New techniques for storage, access, and reproduction have precipitated the erosion of any demarcation lines: the shades of grey multiply, and what appears grey to some, may be ivory or icy white to others .

A particular feature of communication in HEP is the dominance of the "preprints culture". The preprint, i.e. the manuscript ready to be submitted to conferences or journals, is by far the most important source for information transfer on the research front in HEP, to the extent that there has developed a veritable Preprints Culture with its own social structure. Transgressing the concept of an "invisible college", the shortcutting of the lengthy publishing process involves the entire community. The preprint has become the main medium for communicating recent results. Therefore, the flow of preprints has become highly organized, to the point of "whitewashing the grey literature" so that it effectively becomes public/published the day it leaves the author's desk.

The flow of preprints in the HEP community in the majority of cases compares quantitatively with the distribution of specialized journals. A print run of 1000 copies is not unusual. About 12,000 preprints are issued annually, and most of them end up as published articles in journals or proceedings of conferences. However, direct mailing to institutions and individuals, and the additional self-service photocopying in a few places where there are comprehensive collections (approaching 1 million pages a year in the CERN library), obviously does not grant access to all concerned. Scientists in smaller HEP-environments, the home institutions of most HEP-scientists - whether they participate in experiments in the main labs or not - normally would not have access to all this material. They would be on a number of mailing lists, and for the rest they would have to write to the authors to receive specified preprints.

A necessary condition for awareness of recent material and for orientation in the full literature of the area is the existence of an informative database, fully up-to-date, and easily accessible to the community, effectively backed up by full-text document delivery service. It is in this process that the elimination of the greyness really takes place.

The main problem is timing. Users in the main centers want the database to be updated within a day or two of arrival of the preprint, and the document available on display at the same time. This requirement of actuality has made it difficult to avoid a considerable overlap of work: the same preprints are processed in several HEP centers in parallel, to produce essentially the same records, and the manual creation of catalogue records (often of considerable

size), should be done automatically from the full-text file of the preprint. These are a non-trivial problem of coordination and compatibility, and require a tightly knit network of the information services involved, combined with a true ambition to cooperate rather than to compete.

INFORMATION SERVICES TO THE INTERNATIONAL HEP COMMUNITY

CERN information system diagram

Diagram 1 shows the structure of the information services in HEP, as set up at CERN, (where some of the elements are in the implementation stage or in various stages development).

THE DATABASE PRODUCTION SYSTEM

The PREP (preprints) database is the backbone of information services in HEP. The international effort dates back to the 1960's, and the most complete database now counts about 200 000 records. The problems of maintaining this operation are many:

- the comprehensive information required: Authors with affiliations, title, report number, locus of expected publication, keywords added from controlled vocabulary, added free subject terms, and citations.

- the big experimental collaborations in this archetypal "Big Science", resulting in author lists of up to 500 authors from as much as 50 institutions.

- once the preprint has been published, the PREP record has to be updated with the reference to the relevant journal or proceedings.

- the users require immediate access to the information, so that the records have to be inserted within days of arrival of the material.

- several centers are doing database input independently or subject to "weak interaction" using different systems, conceived before the possibilities of data transfer and automated merging were regarded as feasible.

The last point contains the real challenge: a lot of multiple effort could be avoided - and the time delays of entry even shortened - by routine online procedures handling the exchange and merge of catalogue records. This, however, requires a high degree of agreement on input conventions and standards, compatibility of data, and control of data flow.

Diagram 2 shows an outline of a system under development at CERN for automated data exchange. In principle, a document identification key is created from bibliographic data elements in the record (somewhat along the lines of the "Universal Standard Bibliographic Code" investigated in the context of the EEC DOCDEL project for published articles.[4] For grey literature, the unique identification is more difficult since the unique publication data are missing, and since the input conventions differ. This is taken care of by a sequence of nested comparisons, sorting the genuine matches from spurious ones, and finding genuinely new records. In the testing process, the coordination problems are disclosed, to be ultimately solved by agreements.[5]

Diagram 3 outlines a system for automated creation of bibliographic records from full-text documents. A preferred text processing system for preprints

production in HEP is LaTeX. The text processing file is translated into SGML, using the Document Type Definition from the American Association of Publishers.[6] At this stage, the data elements are readily identifiable by the system. The appropriate data are thus extracted, reformatted according to the cataloguing rules, and combined to a new record. The abstract can, of course, be added. Since the same full-text database is used by publishers like AIP or Elsevier, the publications data (so-called anti-preprint data) could be downloaded as soon as it has been decided in the journal editing process, i.e. long before the corresponding issue appears in print. The conversions are made by means of the software package XTRAN (from Exoterica Inc.).

This leads us to a research project on "computer-aided indexing", which is being pursued in cooperation with the University of Darmstadt. It is mainly based on extension of the AIR/X system, applied to the Physics Briefs database in FIZ/Karlsruhe, and uses abstracts in machine-readable form for linguistic analysis.[8]

THE LIBRARY AND DATABASE MANAGEMENT SYSTEM

In 1990 the previous CDS/ISIS system, used for management of the library's database, was replaced by ALICE (ALEPH Library Information for CERN). This is why we use the Tenniel rabbit as a logo.[10] The ALEPH Library System is an international and commercially available integrated library system, originally developed at the Hebrew University in Jerusalem. The new, radically improved version 3 is developed in cooperation with the National Technological Library (DTB) in Copenhagen.[11] In the course of the implementation at CERN, a number of further enhancements are introduced.

The main strength of the system, which runs on DEC equipment, but will soon be available in a UNIX version, is its flexible tool-box approach to functional extension, its adherence to standards, and its orientation towards the needs of remote users. Examples of standards implemented are ISO 8777 (Common Command Language), ISO 8859 (character set), ISO 2709 (data export), and Postscript output for printing. Among special features added are the handling of very big bibliographic record, special links between main records and sub-records (e.g. Proceedings and contributed papers), an E-mail retrieval facility, a link to full-text storage devices, and special retrieval modes for document archive catalogues. Some of these developments have been made by direct systems development cooperation between ALEPH and CERN.

THE USER ACCESS CHANNELS

In order to make the systems and resources of the Scientific Information Service available to users on campus and world-wide, special attention has been given to access channels compatible with the normal computer environment at CERN (IBM/VM and VAX/VMS systems).

Diagram 4 shows the elements of a sophisticated device created at CERN by Lian Yachun, a Computer Scientist from ISTIC, Beijing, the QALICE system with SDI function. It is inspired by design ideas of QSPIRES (developed at SLAC for a VM environment), providing access to database retrieval without normal password- controlled logon to the platform. The design for a VAX environment turned out to meet with unexpected difficulties, to which elegant solutions were found. Essentially, access is gained by using a MESSAGE or TELL command, or by an E-mail message sent to the computer. This message contains a single or a chain of search strings, which are picked up and sent to the retrieval system. The retrieval is then carried out, and the resulting file is sent to the requestor's local E-mail account.[12]

This system therefore makes it possible for anyone on any network reaching out to CERN, to access the library files and carry out searches. The system is enhanced with a module for SDI services (Selective Dissemination of Information) to the international community. Presently, about 300 profiles are run on the PREP preprints database.

Another effort to facilitate the exploitation of CERN preprint records is the MicroPREP database management system, an application of the flexible MicroISIS system from Unesco, running on IBM compatible PCs.[14] It is geared to operate on a subfile of the CERN PREP file, which is currently updated either via data transfer or on diskettes mailed. Since the software is in the open domain, this is an ideal product for installation in LDCs (Less Developed Countries) and in other remote institutes without easy online access to CERN. For document delivery, until now the requestor had to contact the originator of the preprint, but the creation of an international clearinghouse will change this situation.

The Directory of HEP institutes is a file on ALICE, providing comprehensive up-to-date information on institutes throughout the world: Name and address, all telecommunications data, short descriptions of research programs and accelerators. It is currently updated from information received. A FILEMAKER-II version for MACs has been designed and is available on diskette, for local installation and operation of the file for retrieval, address label printing, etc.[13] There is also a printed version.[2]

THE LIBRARY, WITH AUTOMATED ACCESS TO FULL-TEXT DOCUMENTS

The CERN library was created as a modest institutional library in the mid 1950's. It is specialized in Particle Physics, with areas related to accelerator and
detector construction. The collection consists of about 40000 volumes (weeded regularly to provide space), about 1,000 current periodicals, and over 100,000 preprints. One main problem is the inconsistent requirement of open access 24 hours per day every day, and staffing a maximum of 40 hours per week. The theft rate is astounding, particularly of recent material, including recent preprints (an average of 15% of the current week's display disappears, mainly items in high demand). The use of the collections is illustrated by the production of self-service photocopies approaching one million per year.

The ideal solution to this problem is furnished by the optical disk (WORM) storage technique. Recent disks carry 6.4 GB of data, corresponding to 6 months of preprints. Two disk stations match the obsolescence of preprints, related to the lead time for journal publishing, which is rarely more than a year. New preprints will be scanned within a day of arrival, and logically linked to the corresponding bibliographic record in ALICE. A system from the French firm DORODOC (related to Thomson-France), using French, US and Japanese products, is presently being installed at CERN.

A typical retrieval and display sequence is shown in a few images. A bibliographic record is found, the full-text image is requested and displayed. The present access time on campus for the first page of a preprint is about 10 seconds (for retrieval, decompression, and transmission), but as the whole document is stored locally on the display device, the following page shift is "immediate". A central automatic printing facility will generate hard copies equivalent to the original, to be mailed to the requestor. Central printing can be ordered online by any remote user of the ALICE system. There is an option for automated Fax transfer. With Group-3 technique, some loss of resolution is inevitable, but with broadband networks in preparation, Group-4 technology will provide high speed document delivery of quality equivalent to the original. (This document delivery system is in line with current projects in France.[17] It is also possible to modify the QALICE system to allow for

printing commands to be sent. Thereby an international clearinghouse for preprints is established.

THE GLOBAL VILLAGE REVISITED

This brings the presentation to a conclusion. It is seen that the various components if the "grand scheme" jointly form a pattern of information resources management (IRM) and service production that reaches out world-wide, with minimal delays even for remote users in poor institutions: The distributed network of regional PREP data input centers (foreseen in the design), the controlled data flow for a common comprehensive database in several main laboratories, the various modes of database access, the international clearinghouse for preprints, etc.

By having access, timely and comprehensively, to new information, the disadvantage of distance experienced by HEP physicists all over the world will be reduced, sometimes dramatically. This is particularly, but not exclusively, the case for many LDCs and small groups. Effectively, massive distribution of preprints can be reduced, and replaced by services from the Clearinghouse. The Global Village has an electronic library. Hopefully, this scheme will have the resources needed and the international interest and cooperation indispensable to bring it to fruition.

REFERENCES

1. RIORDAN, M. The hunting of the quark. New York, Simon & Schuster, 1987.

2. Directory of Research Institutes in High Energy Physics (CERN AS-SI). Geneva, 1990.

3. PRICE, D. J. DE SOLLA. Little science, big science ... and beyond. New York, Columbia University Press, 1986.

4. AYRES, F. H. Electronic document delivery - the linkage between bibliographic and full-text databases. (CEC Report EUR 10677 EN), Luxembourg, December 1987.

5. FALCOZ, F. Automated comparison and merging of bibliographic records. (Draft project report, CERN AS-SI), 1991.

6. Standard for Electronic Manuscript Preparation and Markup, Washington, D.C., Association of American Publishers, 1987.

7. OXNARD, B. Automatic CERN Preprint Upload. (Draft project report, CERN AS-SI), June 1991.

8. BIEBRICHER, P. ET AL. The automatic indexing system AIR/PHYS - from research to application. Y. Chiaramella (ed.). Proceedings of 1988 ACM Conference on R&D in Information Retrieval, Presses Universitaires de Grenoble, 1988.

9. RENFREW, K. Research on Automatic Indexing System AIR/X with Respect to a High Energy Physics Database. (Draft report, CERN AS-SI), May 1991.

10. CARROLL, L. Alice in Wonderland.

11. SANDFÆR, M. Remote access - the true test of the OPAC as a front-end of library services (to be published by the British Library R&D Division, 1991).

12. YACHUN, LIAN. QALICE Design Description. (CERN AS-SI) 90-01, 1990:
-QALICE User's Guide (CERN AS-SI 90-02, 1990)
-QALICE Manager Guide (CERN AS-SI 90-03, 1990)
-QALICE Programs (CERN AS-SI 90-04, 1990).

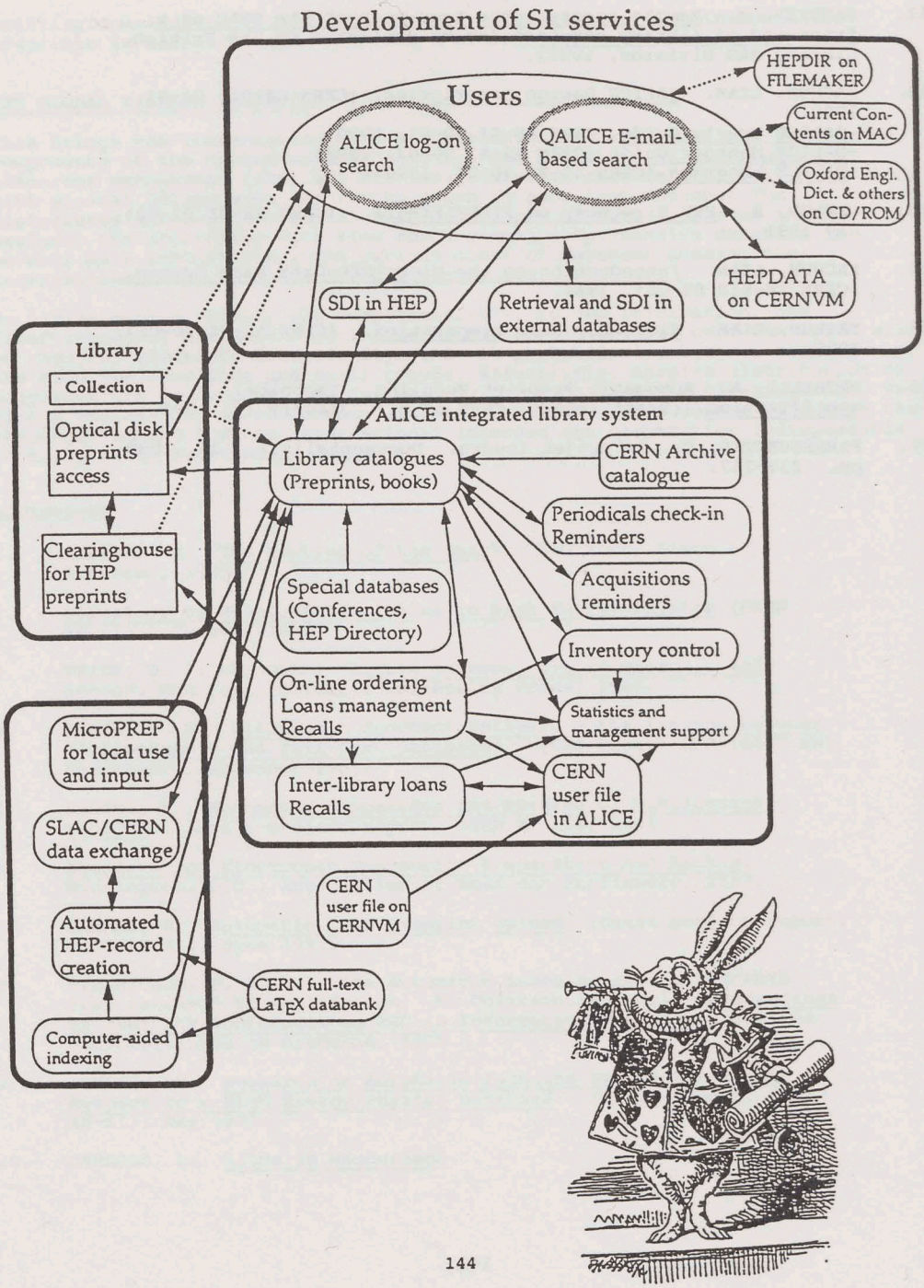
13. OXNARD, B. HEP Directory on FILEMAKER-II. (CERN AS-SI 91-02), May 1991.

14. YACHUN, LIAN. Introduction to the MicroPREP Data Base System. (CERN TH-SIS 89-02), 1989.

15. YACHUN, LIAN. MicroPREP File Preparation. (CERN AS-SI 90-05), 1990.

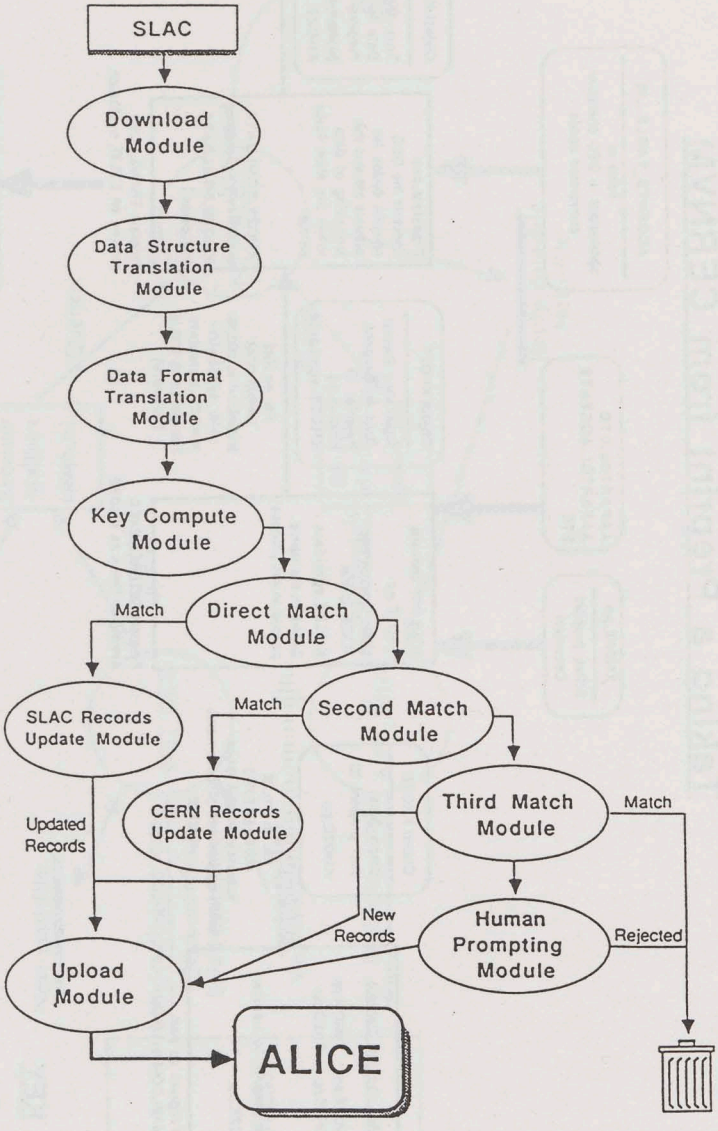
16. PETRILLI, A. Automatic Preprint Handling - Technical specification. (Draft report, CERN AS-MI), January 22, 1991.

17. FABREGUETTES, C. Le projet foudre. Documentaliste, 26, 1989: pp. 239-247.

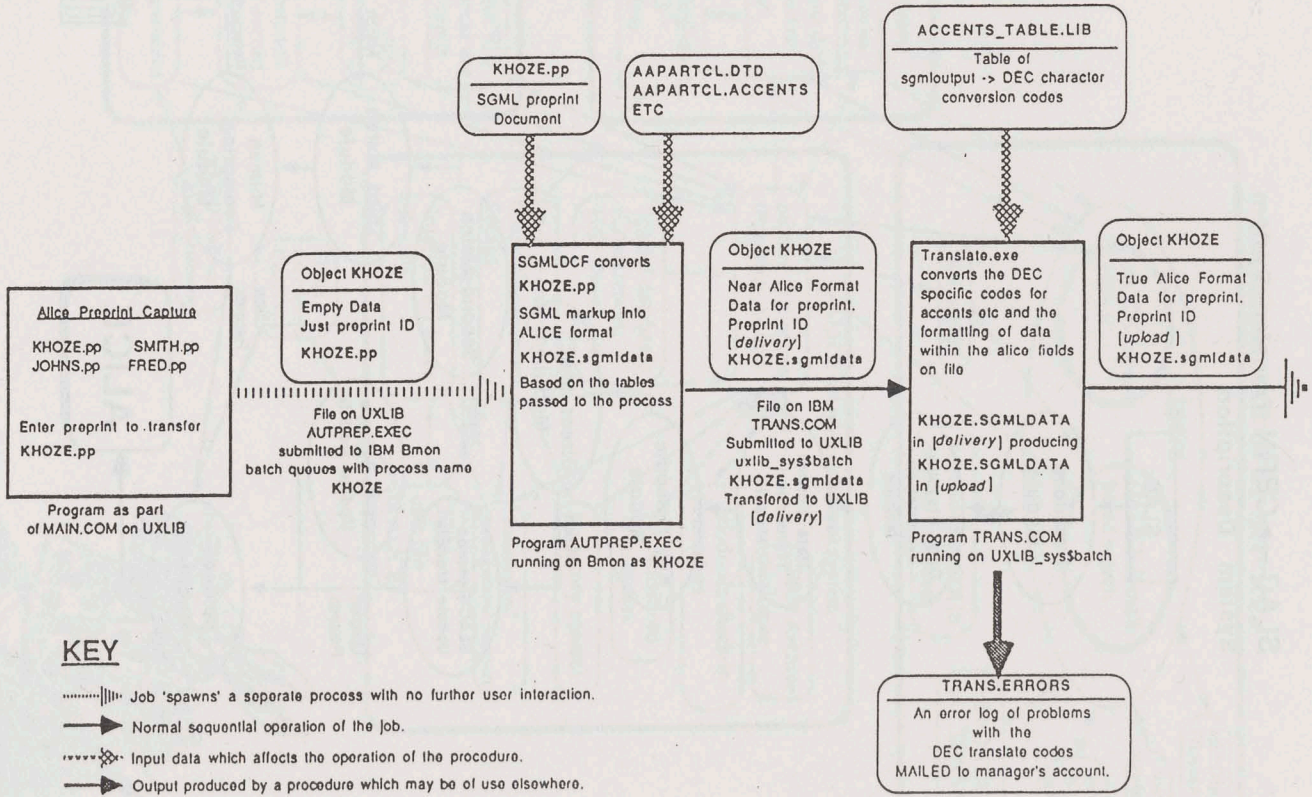

Development of SI services
Users
ALICE log-on search
QALICE E-mail-based search
HEPDIR on FILEMAKER
Current Contents on MAC
Oxford Engl. Dict. & others on CD/ROM
HEPDATA on CERNVM
SDI in HEP
Retrieval and SDI in external databases
Library
Collection
Optical disk preprints access
Clearinghouse for HEP preprints
ALICE integrated library system
Library catalogues (Preprints, books)
CERN Archive catalogue
Periodicals check-in Reminders
Acquisitions reminders
Special databases (Conferences, HEP Directory)
Inventory control
On-line ordering Loans management Recalls
Statistics and management support
Inter-library loans Recalls
CERN user file in ALICE
CERN user file on CERNVM
MicroPREP for local use and input
SLAC/CERN data exchange
Automated HEP-record creation
CERN full-text LaTeX databank
Computer-aided indexing

SLAC → CERN Transfer System.

System Description



Taking a Preprint from CERNVM



KEY

- Job 'spawns' a seperate process with no further user interaction.
- Normal sequential operation of the job.
- Input data which affects the operation of the procedure.
- Output produced by a procedure which may be of use elsewhere.
- Process terminates.

NOTE Processing of the preprint SMITH.pp at the same time is perfectly acceptable.

NOTE Processing of the preprint SMITH.pp at the same time is perfectly acceptable.