Development Plan for the International Seismological Centre, 1998 – 2002

This plan is subject to approval by the ISC Governing Council.
The Executive Committee recommended approval in June 1998, subject to revisions.
This version was prepared during July 1998.

Executive Summary

Rapidly growing data volume, evolving mechanisms to use and archive digital data, and uncertainty about the impact from seismological monitoring of the test ban treaty are creating new challenges. Consequently, seismologists need an evolving set of services that complement services available elsewhere. While the ISC should continue to deal mainly with parametric data, it should take advantage of advances in computer technology to become more comprehensive.

The tasks to accomplish the ISC's mission may be divided into collection, analysis and distribution of data. The growing data volume has strained the existing system. In each area, the ISC faces challenges in accomplishing the tasks well with the available resources. Data analysis is the area of greatest concern, since the algorithms are many and complex, there is no absolute standard, and the reputation of the ISC stands on reliability of the outcome.

Data are held in distinct formats during each of the collection, analysis and distribution stages. Questions arise about the consistency and relative completeness of the data in different formats. The diversity of internal formats leads to technical problems making all of the ISC's data available in a common format. The immediate development focus is to port the existing processing system from VMS to Unix. In the short term as well, access to ISC data is being improved. Need for improved automatic association of phases with earthquakes is recognised

A new system will emerge as a result of incremental changes, but an overriding consideration is that the number data will continue to grow. To handle this efficiently, data collection should be automated by aiming towards greater use of the Internet and standard formats. To provide the more timely data that are increasingly sought, some processing of data should occur soon after it is received, and be repeated regularly to include newly collected data until they are finally reviewed. As more data are received, it will become necessary to perform selective review. Many events will be small and have only the phase associations reported by the original agency while others will be sufficiently robust to have a very small chance of error. A distinction should be made between data archiving, which should include dynamic databases, and distribution of data products. Standard products should be freely available, while custom products and services are provided to subscribers.

Transitioning towards an improved system will be fitful and difficult unless data are managed in a single system for all tasks. None of the existing data storage formats is sufficient for all of the tasks. The principles of the data structures and algorithms used in the commercial systems are well known, but it is not cost-effective for a small organisation to try to implement all of these algorithms when they can be bought “off-the-shelf”. Seismological applications, however, should be developed in-house or exchanged with universities and other data centres, which would allow an abbreviated development cycle, lower cost than contracted services, better maintenance from in-house staff, and greater flexibility in future data exchanges.

Modification of the processing software should minimised during 1998, while the system is ported from VMS to Solaris. Free Internet-based data distribution should be implemented very early. A new data management system should selected and installed during 1998. DBMS-based data collection should implemented soon after installing the DBMS. Improved automatic processing should be developed before interactive editing.

The U.S. National Science Foundation has provided development funding that supplements its normal, operational support. Efforts to obtain supplemental support from the Japan Science and Technology Agency and U.K. Research Councils should continue. Governing Council members could help the ISC to identify similar opportunities with national funding agencies of other countries, and develop support for such funding within their own seismological communities. Development need not necessarily occur exclusively at the ISC, but could be joint projects involving seismologists and programmers in the country funding the development.

Significant costs may be encountered in implementing a comprehensive data management system. An important part of this cost will be purchase of software and hardware to support it. One of the most important projects is development of improved algorithms for event formation and phase association. Such changes should be implemented soon after a new data management system is put in place. The primary cost of this development would be the salary of staff devoted to the project.

A difficult aspect of funding the ISC is obtaining open-ended commitments. The ISC could seek funding for seismology projects that payoff in improved operational procedures, such as improving the historical database, computing relative locations, and studying source parameters. Ideally, special development projects would be funded regularly enough to support further staff members.

Principal extra costs of operating an improved system are salaries for a staff that is larger or includes more marketable skills. One reasonable goal is to employ a senior seismologist as a permanent staff member to maintain the continuity of editing practice and normally to have two further seismologists each serving offset, 2-year terms. A significant expansion of data services is required and the computer hardware is becoming more complex. A staff of two will be able to satisfy the ISC's data processing and management needs only if newly developed tools routinely execute without intervention or error.

Introduction

The ISC fills a critical role for seismologists – without it, no definitive source exists for data that are essential to many earthquake studies. Several developments are creating new challenges for seismologists, however, in initiating and carrying out first-rate work.

Much more data: Continued growth of dense regional and local networks can be overwhelming. In response, data centres with a narrow geographic focus have become common, while global agencies have restricted the data they accept in other ways. Use of local earth models by regional agencies has become more reliable and, for a growing proportion of studies, global-agency locations and magnitudes are not the best choice. Thus, seismologists must seek the best data from multiple agencies around the world.

Digital data: Developments in collection and analysis of digital data have created opportunities for new types of seismological research. The cost and volume of these data are so large that no universal archive of digital waveforms has been created. Ongoing development of new techniques precludes any one data centre from claiming to perform definitive inversions from digital waveforms for all source parameters.

Test Ban Treaty: If the International Data Centre (IDC) approaches its target of a global threshold of magnitude 4, it will supplant the most obvious role of the ISC. Availability of IDC results to all researchers is in question, however. Further, the IDC does not plan to distribute phases from outside of its relatively sparse network, smaller events where they can be detected, or source parameters computed by other agencies.

Because of these changes, seismologists need an evolving set of services. Computer technology has advanced sufficiently for one centre to cost-effectively distribute virtually all parametric earthquake data, simplifying individual efforts to obtain these data. Services could be attached to a seismicity bulletin to retrieve event-based waveform segments from the growing number of sometimes-obscure digital data archives and analysis centres. A conduit of parametric data and waveforms from the IDC to the academic community is required.

In order to continue serving the contemporary needs of seismologists, the ISC must reorganise its activities to complement other services that are now available. The ISC should neither impose a threshold near magnitude 4, nor reanalyse even the smallest earthquakes with a global average earth model. Instead, the ISC must be as complete as possible and base its results more on data analysis by seismologists around the world. This reliance on other seismologists is an extension of the dependence on accurate record reading at each seismic station that has existed from the outset.

Following a reorganisation, the ISC will again provide a definitive starting point for any earthquake research. With straightforward access to the required data, seismologists will undertake a wider range of studies and include more complete data in each project.

The Current Status

The primary mission of the ISC is to produce, distribute and archive a Bulletin of global seismicity. The data processing system to accomplish this is fundamentally unchanged since it was implemented in the early 1970s. Meanwhile, the number of data reported to the ISC monthly has more than tripled (Figure 1)

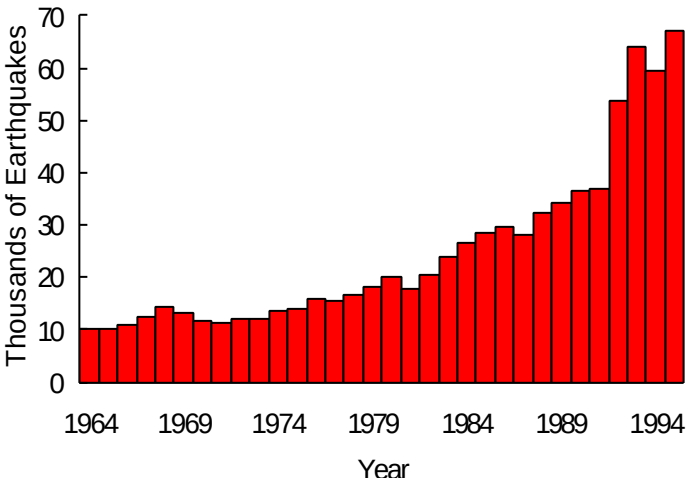


Figure 1. The growing number of earthquakes located annually by the ISC exemplifies a general increase in the number of seismic data.

despite the imposition of thresholds, and the need for prompt, flexible access to ISC data has grown tremendously. The U.S. National Science Foundation has partially funded a computer modernisation programme that will help the ISC to take advantage of advances in computer hardware and software to address these needs.

The tasks to accomplish the ISC's mission may be divided into four areas (Figure 2). In each area, the ISC faces significant difficulties in accomplishing the tasks well with the available resources.

Data Collection. New data formats and even new data types are routinely offered to the ISC. The ISC devotes programming effort in this area each year, yet still manually edits collected data, has difficulty handling duplicate and updated data, and cannot make unprocessed data available to outside users.

Automatic Processing. Increasing rates of reported seismicity are leading to many more events with arrivals that overlap in time. Partly in response, algorithmic changes are required to take advantage of all the data reported to the ISC (preliminary associations, amplitude, slowness, etc.). The earth model used by the ISC is demonstrably less accurate than models used by many seismologists.

Interactive Editing. The batch organisation of processing was unavoidable 25 years ago, and acceptable at the lower data rates that held then. Interactive processing to help editors test hypotheses should reduce the number of passes through the data and produce a better bulletin.

Data Distribution. The printed Bulletin is expensive to produce and not useful for some purposes. The 96-column format of the CD files is difficult to scan and does not hold all data types. Generating summaries of the data or making a flexible selection from it requires software development by individual users. The ISC does not provide flexible on-line access provided by other earthquake data centres.

Of these areas, the two most tightly coupled are automatic processing and interactive editing, which may be collectively labelled data analysis. The tight coupling results from iteratively editing, then reprocessing the data. These are also the areas of greatest concern about faulty execution, since the processing algorithms are many and complex, there is no absolute standard against which the results can be measured and, most importantly, the reputation of the ISC stands on reliability of the outcome. It is tempting to conclude that redevelopment of data analysis should therefore be deferred until the ISC gains experience with modern data management. Unfortunately, the growing data volume has severely strained the system, making it impossible to meet the target publication schedules, putting the accuracy and completeness of the bulletin at risk, and discouraging the staff.

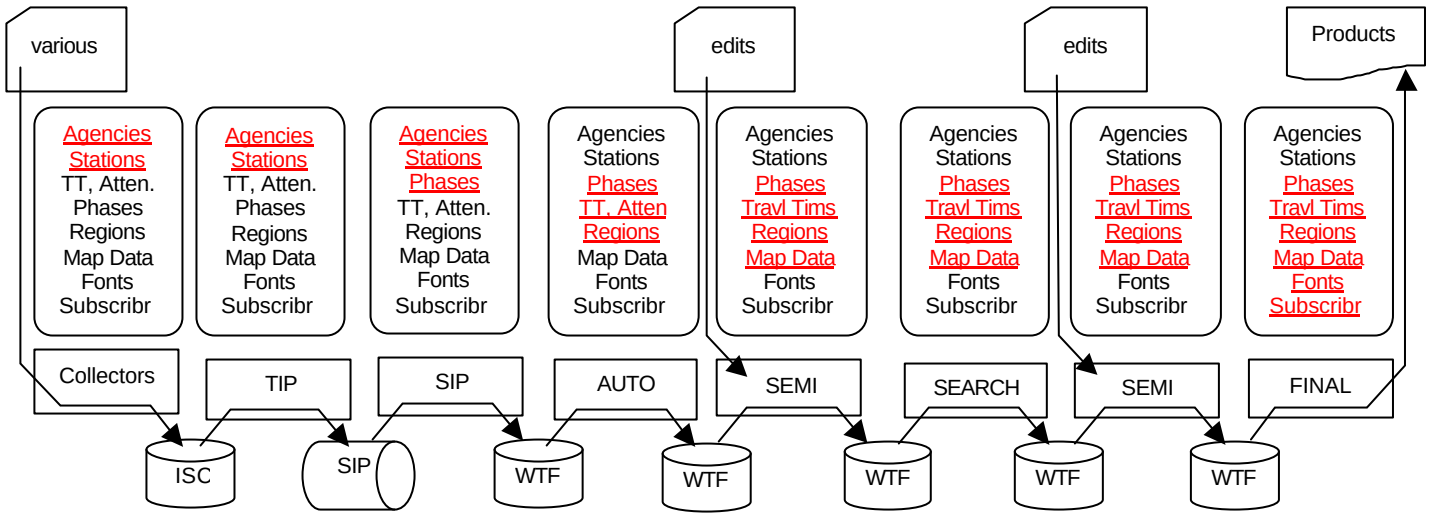


Figure 2. Processing in the current system is straightforward, but management of data is awkward. The Collect, TIP and SIP processes accomplish data collection tasks. The AUTO, SEMI, and SEARCH processes carry out data processing, which is interleaved with manual editing. FINAL is used to distribute data. Data are stored in a variety of different formats, and multiple copies are required in some formats.

Looser coupling of data collection and distribution with analysis arises from use of the collected data in analysis, and from distribution of both the collected data and the outcome of analysis. This loose coupling has allowed a system to develop in which data are held in distinct formats during each of the collection (SIP), analysis (Working Tape Format, or WTF) and distribution (96-column). Consequently, programs to copy the data from one format to another must be maintained and questions arise about the consistency and relative completeness of the data in different formats. Further, the diversity of formats leads to technical problems in making all of the ISC's data available in a single, coherent format.

The immediate focus is to port the existing processing system from VMS to Unix. The port to Unix is the first step in a change from batch processing to interactive computing that will allow ISC seismologists preparing the Bulletin to test alternative phase associations, starting locations, and depth constraints. Ultimately, the ISC hopes to use interactive graphics and decision support software to improve both the quality of the ISC Bulletin and the efficiency with which it is produced.

Also in the short term, access to ISC data is being improved. New computers purchased as part of the modernisation programme have helped us to create an ISC web site, and recent Catalogue data are available as static files.

The need for improved automatic association of phases with earthquakes is recognised. Automatic processing should take greater advantage of the associations reported in preliminary seismic bulletins and phase readings other than time, such as amplitude, slowness and polarisation, could improve association accuracy.

The ISC is an important archive of 20th century seismicity. It holds a unique collection of station bulletins and maintains a comprehensive database of seismic events since 1904.

An Envisaged System

A new system (Figure 3) will emerge as an outcome of incremental improvements. Meanwhile, computer technology will continue to advance and the ISC will gain experience in collecting, processing and distributing data in new ways. Rather than rigidly following a programme set out years in advance, it will be reasonable to alter plans in response to new opportunities. Nevertheless, it is useful to document and regularly update a coherent set of long term goals. The benefits are an understanding of how different developments will work together, and a context for considering alternative improvements of the existing system.

The overriding consideration in understanding how the ISC must evolve is that the number networks will continue to grow, and seismologists will have an increasing need for

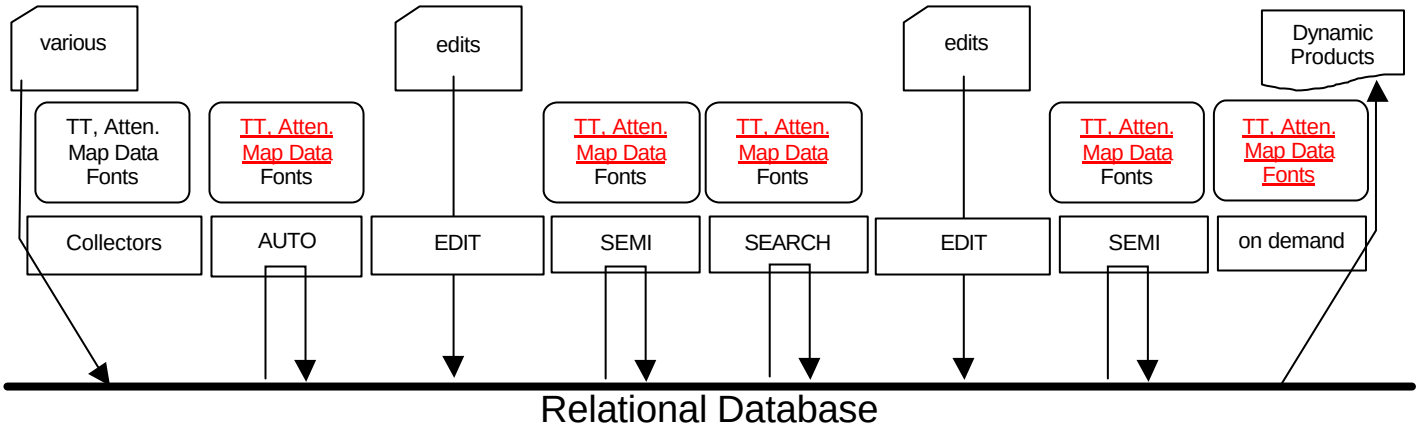


Figure 3. In an envisaged system, processing steps are similar, but happen in real time in response to receipt of data reports, scheduled event, editor actions, and user requests. A relational database allows multiple processes to connect simultaneously, while providing services required for performance and data integrity.

guidance finding data. To satisfy this need, the ISC must collect data from more agencies, and in each case collect all of the phase readings and events, with no threshold. With a world wide capability perhaps a full magnitude unit better than at present, on the order of 1 million events will be recorded by the ISC annually. Most of these will be small and have only the associations reported by a single agency, so the reported location will remain the primary location. The ISC cannot manually review so many events, but it can add value even to these small events by performing an independent (although automated) quality check and computing a secondary location and residuals with respect to a standard earth model.

Data Collection

In the envisaged system, data are generally received over the Internet. Most are complete sets of parameters, including amplitude and polarisation, measured from digital data using well-known algorithms. A few are still accepted by post from regions where stations are sparse or communications links are expensive or non-existent. Even in these cases, however, data are on disk or tape, in the same format as the Internet-delivered data. The data contain digital signatures to authenticate the source, so they can be automatically parsed into the ISC database on receipt. Parsing retains associations as well as measured values, but some data reports contain extensions with data types that are not part of the ISC schema. To preserve these other data, reports are archived in their received format. The station set includes temporary stations operated as part of well established programmes, such as PASSCAL, so some automatically parsed reports contain information required to establish a new station in the ISC database. Further, the distributed data now include mechanisms to retrieve further information, so collected data include information on availability of digital waveform data as well printed and electronic research publications. The data collection administrator, rather than spending time manually parsing routine reports,

- reviews logs and corrects errors
- mounts archive media as required
- keeps the software up to date with a few, gradually evolving standards
- checks on sources of delayed data
- helps new data providers to send data in the appropriate formats

The ISC's analysis of its own Bulletin identifies gaps in the global network of reporting stations. The ISC actively encourages better communication or other changes that lead to collection of more complete or reliable data, improving the accuracy of seismic monitoring. Where gaps are due to an insufficient number of stations, the ISC is a prominent advocate of new stations and associated facilities.

Automatic Processing

The difference between data collection and automatic processing has become less distinct, as some processing occurs soon after receipt, and is regularly repeated to include newly collected data until they are finally reviewed. Duplicate arrival reports and multiple origins (based on shared arrivals as well as proximal time and location) are detected early and, if necessary, reviewed by the data collection administrator. Phase association takes full advantage of agency-reported associations, all measured parameters, and realistic attenuation curves. Rules for recognising multiple origins and associating phases are implemented flexibly; they can be modified, tested, and updated in operations within days. Locations are computed using an earth model that includes 3-dimensional mantle and core models and local or regional crustal and lithospheric models. Searches for new events from unassociated data are exhaustive. Iterative association, location and searching occur without intervention. The reliability of each solution is automatically evaluated, possible mixed events are detected and, in the final pass before human review, alternative solutions are computed for questionable cases. The data processing administrator

- reviews logs and corrects errors
- updates earth models and recomputes tables of travel time, attenuation, etc.
- updates software in response to problems encountered in editing

Interactive Editing

Most events are not examined at all; many are small and have only the phase associations reported by the original agency while others are sufficiently robust to have a very small chance of error. Remarkable events – those causing many deaths or great damage and those with atypical epicentres, depths or magnitudes – are reviewed but rarely updated. Questionable events are examined, but usually one of several pre-computed alternatives is selected. Hypotheses can be fully tested and new locations are computed immediately, so a single pass through the data is sufficient for all but a very few earthquakes. Data are not organised into month-long batches, but into shorter blocks of only a few hundred earthquakes – similar to the size of a "slot" used in editing now. Within each slot, events are ordered from most important to be reviewed to less important. Events above some threshold of uncertainty must be reviewed, but further down the list events are reviewed only if time permits.

Data Services

Data Archiving. BULLETIN is a database of edited events. As each slot is completed, processes are launched that retrieve digital waveform data from remote archives, supplement the existing parametric data with additional information (Appendix A), and insert the parametric data in BULLETIN. BULLETIN includes a primary estimate (possibly null) for each parameter of an event, which is never updated. ISC BULLETIN values (even those originally from another agency, quality-checked by the ISC) can be cited without further qualification. Other ISC databases include

CURRENT, a database of events that have not yet been edited and so are subject to revision by reprocessing and manual review.

HISTORICAL, a database of events that are copied regularly from CURRENT, replaced from BULLETIN when it is completed, and revised thereafter when the ISC incorporates new information or reprocesses data.

METADATA, information on stations, archived waveforms, and related publications (Appendix B). Data are inserted in METADATA routinely but information in METADATA is never changed, so processing to create the events databases can be reproduced.

Data Distribution. Data are distributed as "products". Each data product is derived from one of the databases and may include

- all origins for each event, primary origins only, or no origins at all
- all phases, phases from selected stations, or no phases at all
- any types of metadata

Each product is available as plain text in one of a few formats (96-col, IMS x.x, Seisan) or can include layout instructions and hypertext links for interpretation by general-purpose software (Postscript, PDF, HTML). Regardless of the format, each product is available printed, over the Internet (via http, ftp, and e-mail), and on computer media (disk and tape). For example, one user might request "primary Historical origins with bibliographic information as PDF on CD" while another requests "all Bulletin origins and phases with station and waveform information as IMS x.x on ftp". If the product includes metadata then, in formats that support them, hypertext links are embedded to retrieve waveform data segments, station-book pages, and electronic abstracts or publications related to particular earthquakes.

One or two standard products that include all data are available on the ISC web site free of charge and as a mass-produced item (printed Bulletin and CD) at the cost of reproduction. Copies of the standard products are retained for later distribution and as backups of the database archives. Other products are available to users from institutions that pay annual subscription fees. Most subscriber requests are sent over the Internet. Subscribers have been assigned authentication codes, in which case the product is automatically generated and returned to the user or queued for writing to the appropriate medium. The data distribution administrator

- reviews logs and corrects errors
- executes queued product requests
- mounts media to write queued products
- keeps the software up to date with a few, gradually evolving formats

- registers new users and assigns authentication codes

Special Services. Earthquake searches, seismicity statistics, maps, and several other services are provided to subscribers from the ISC web site by software that generally runs without ISC staff intervention using input supplied by the subscriber. A few users, however, require more elaborate services, which are supplied for fees set to maximise revenue to the ISC. These include authoritative statements of locations and magnitudes for insurance claim purposes, and computing new absolute or relative earthquake locations using special-purpose travel time models.

Development Considerations

Seismological Outcome

The objective of development during this period is to improve efficiency in handling an ever-growing data volume. As nearly as possible, the seismological outcome should be unchanged, pending further consideration. For example, the initial development will include neither replacement of the Jeffreys-Bullen travel-time tables, nor computation of relative earthquake locations in place of absolute locations.

Data Management

One feature of the envisaged system is that several task areas have become less distinct than at present. Data are processed soon after they are collected. Data may be distributed at any time, as custom products prepared on request. Processing includes editor-like generation of alternative hypotheses, while editing includes real-time reprocessing. The implication is that transitioning towards something like the envisaged system will be fitful and difficult unless data are managed in a single system for all tasks.

None of the existing data storage formats is sufficient for all of the tasks. SIP lacks some of the data types in WTF that are required to support analysis, and 96-column is even more limited. Data are not collected directly into WTF because of the difficulty checking and updating data. Checking could be addressed by developing tools that derive distribution products from WTF and the products would then be checked. But the requirements to insert new data as they arrive and to update data as a result of reprocessing or editing would be a fundamental change of WTF. Currently, WTF files are never altered; instead, a new file is written to replace an existing one. Each WTF file conceptually covers only a short time interval (up to 1 month) and there is no defined organisation of WTF files into larger units for searching across longer time periods. WTF files lack indexes and other data access structures that allow retrieval of the data that will be required both for interactive analysis and on-line data distribution.

The alternatives are to modify one of the existing formats (almost certainly WTF) or to adopt a new system. The primary reason to consider adopting a new system, rather than elaborating an existing format, is the availability of commercial database management systems (DBMS). While the principals of the data structures and algorithms used in the commercial systems are well known, it simply is not cost-effective for a small organisation to try to implement all of these algorithms when they can be bought “off-the-shelf”.

Purchase of a data management system would not mean that there would be no development costs. Tools to perform the tasks need to be developed in any case. But if a modified WTF were adopted as a management system, for example, new tools to collect data directly into WTF would be needed.

One of the important tasks in selecting and configuring a data management system is defining the typical interactions with the database, including inserts, updates and queries. A synopsis of current database interactions is included as Appendix C.

Different Task Areas

If a well-defined data management system that serves the needs of all phases were implemented, the current loose coupling would allow data collection, analysis, and distribution to be redeveloped independently. Temporarily, data could be collected directly to this data management system, and transferred once to the legacy system when processing begins. Similarly, results could be transferred to the data management system after analysis is completed. Known shortcomings in the collection system can be tolerated temporarily since the existing system for collection to the legacy format is effectively a backup. Users would probably tolerate known shortcomings in the distribution system, provided the problems were clearly stated, and corrected within a reasonable period. On-line data distribution, either after processing and editing or promptly on receipt by the ISC, would have an immediate payoff visible to outside users and funding agencies.

Development of the programs to transfer data between the existing data management systems and the new system would be part of the cost of independent redevelopment. This is not an open-ended task, however, since the only formats involved are SIP, WTF, 96-column and the new system. Transfer between formats can be rigorously evaluated, since the number of data and their values should be unchanged in each step.

In-house vs. *Contract Development*

The ISC has purchased general-purpose software (operating systems, networking software, and business applications), but has never funded commercial development of seismological applications. Instead, seismological applications are developed in-house or, very occasionally, obtained (with a source code and enough information for subsequent in-house maintenance) as part of an exchange with universities and other data centres. The perceived advantages include

Abbreviated Development Cycle: Many seismologists are smart enough and have enough programming experience to contribute to development, obviating the need to take time to rigorously formulate requirements. The ISC is continuously prototyping.

Lower cost: Apart from a brief period in the late 1970s (following increases in oil prices) there have always been more seismologists trained than seismology jobs offered. Thus, the prototypers generally work for less than people formally trained in computer science. Sometimes they even work for universities and offer their software without charge.

Better maintenance: The seismologist/ programmers are inexpensive enough to keep around after they develop the software, providing immediate and informed responses to needs for software modification.

Greater flexibility: The ISC is free to exchange its software with other seismic data centres and research seismologists.

Given the likelihood that applications will be developed in-house, the skills of computer staff and the time that they have to devote to development are important considerations in planning development.

A Development Outline

The considerations above do not lead inevitably to particular milestones, much less the some specific sequence. Instead, this outline is intended as another tool for considering what development should be undertaken, what trade-offs are necessary, and generally how soon different results can be expected.

Development Priorities

Modification of the processing software is minimised during 1998, while the legacy system is ported from VMS to Solaris. The first priority in development is to create a stable

basis for implementing new features. The current system is strained because it depends on a computer that is incapable of supporting the goals of the ISC. New hardware has been purchased, but the existing system is yet to be ported to the new operating system. Postponing updates of the existing system frees time for porting, and reduces the need to implement updates in both VMS and Solaris versions.

Internet-based data distribution is implemented very early, but very simply. The newly purchased Sun workstations make it easy to post static data files to the Internet. Although this method of providing data differs from the envisaged system, it provides visible improvement of services and the ISC begins to gain experience with on-line data distribution. Since this on-line data distribution is based on the current products, it may not include unprocessed data.

A new data management system is selected and installed during 1998. If data management is based on a commercial system, the newly hired system analyst should be able to take a large role in its implementation. Guidance from the existing staff on data attributes in the current system will still be required, but this should not have a large impact on their time, allowing other development to continue in parallel.

DBMS-based data collection is implemented soon after installing the DBMS. As discussed above, this requires temporary use of a program to copy data from the new DBMS. The benefits of early implementation include realising simplifications expected from DBMS-based collection sooner and making recently collected data are available to outside users as soon as DBMS-based distribution is implemented.

Improved automatic processing is developed before interactive editing. Greater efficiency gains are expected from better processing than from interactive editing (Appendix D).

The data set is loaded into the database over many months. A large database risks unacceptable response time, which might alienate some users. Gradual addition of data will provide an opportunity to judge performance and, if necessary, modify the system or postpone growth until solutions are found.

Development Stages

In this section, stages of development are described in somewhat more detail. An important aspect of this development is step-by-step reorganisation of processing around a new data management system (Figure 4)

The sequence of stages reflects both technical constraints and development priorities. An example of a technical constraint is that interactive editing by multiple seismologists should follow use of the relational data management system to take advantage of its data integrity features. An example of a development priority is that dynamic products could be implemented any time after stage 2, but are postponed to stage 10 while more urgent development proceeds.

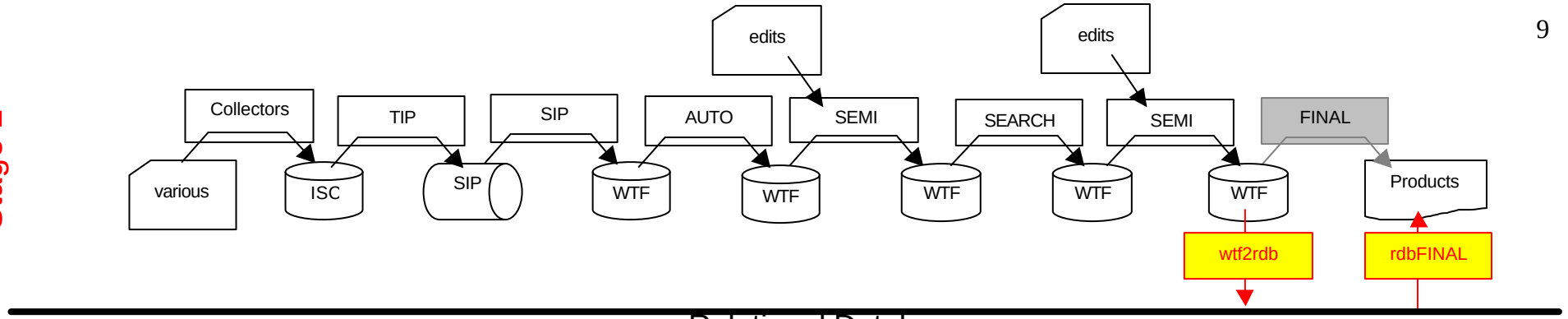
The development goals are realistic but ambitious. They can be accomplished over the next few years only with a concerted effort, including employment of an additional software analyst. The development described in these stages does not conclude with the envisaged system, which could be achieved only over a longer period or with a higher level of effort.

Stage 0 Newly processed data on-line

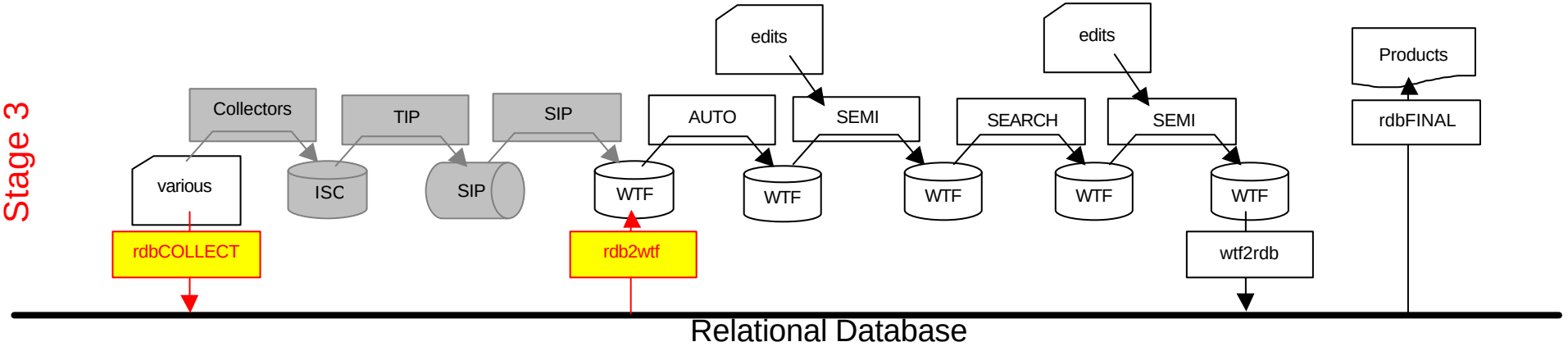
This has already been successfully completed by creating a single, static file for each day's Catalogue from the "96-column" format data used on the CDs. Users interested in events from throughout the time period posted must first download the complete set of data. Although unworkable in the long run, complete downloads are acceptable for the relatively small amount of data involved in the initial stage.

Figure 4 (next two pages). Programs used in operational processing are successively replaced with new versions that use the relational database management system. In each stage, new or modified processes and data flow are are highlighted in red and yellow. Processes and data formats that are dimmed to grey in the first stage where they are no longer used, and excluded thereafter. Development stages that do not involve re-organisation of processing are not shown in the figure

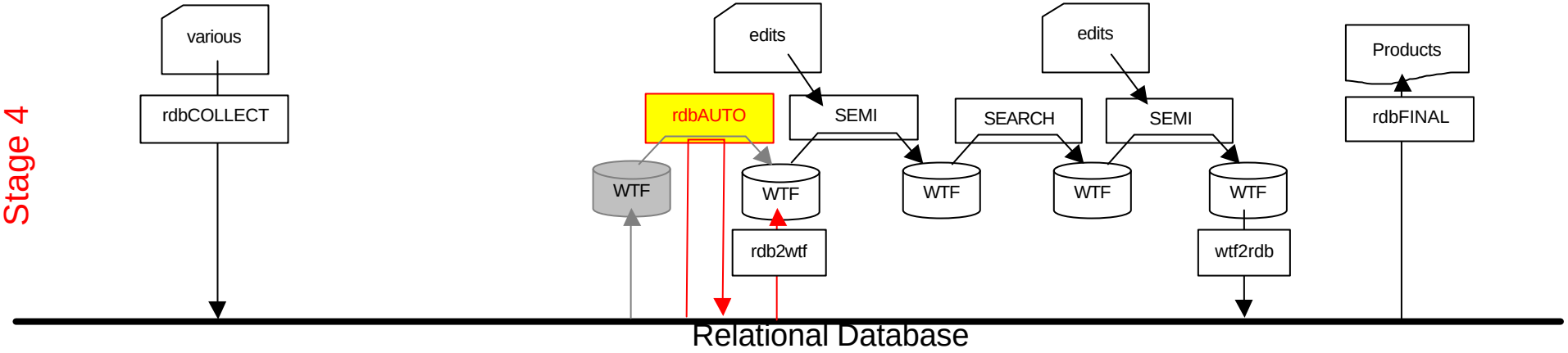
Stage 2

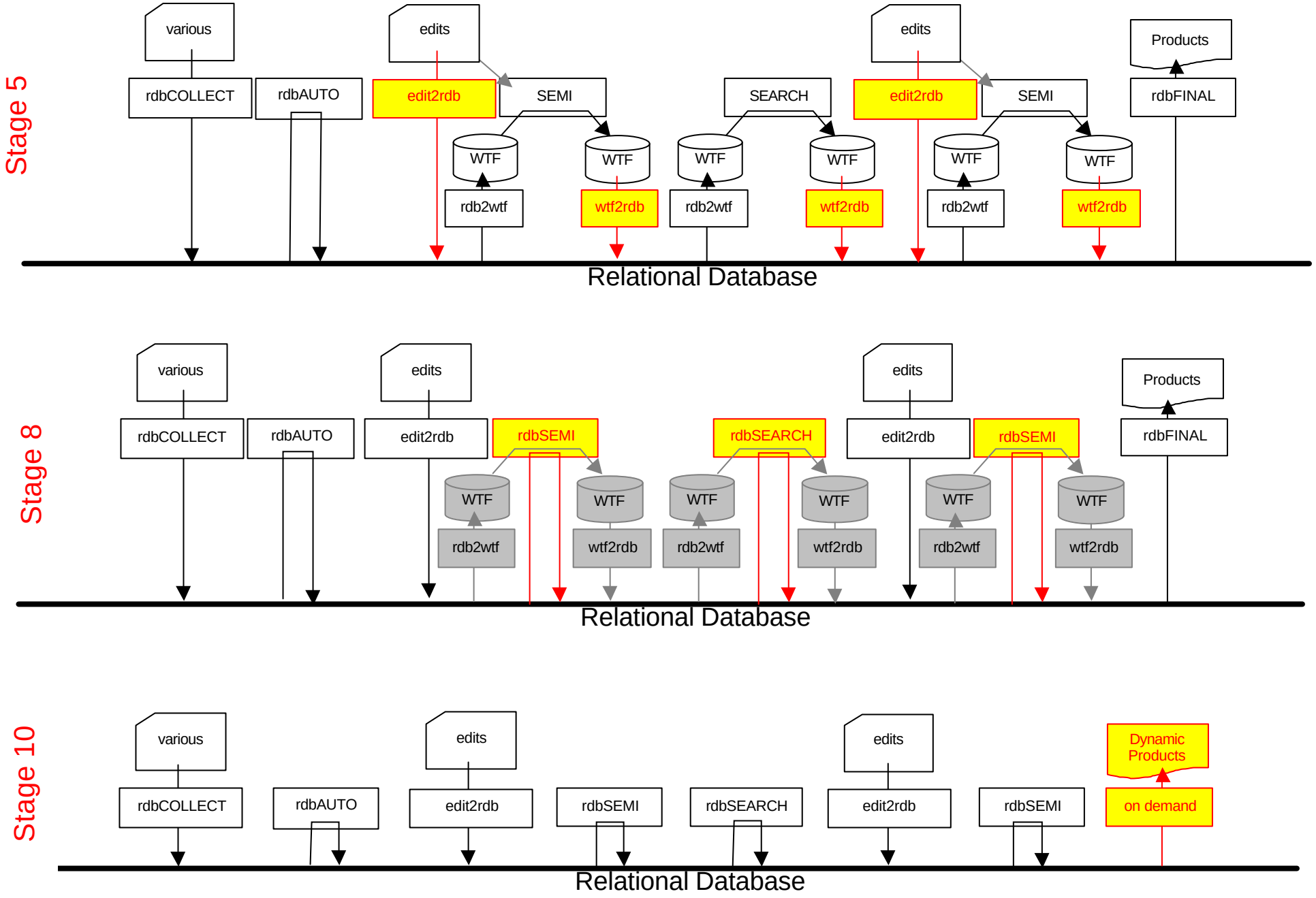


Stage 3



Stage 4




Stage 5
various
edits
edits
Products
rdbCOLLECT
rdbAUTO
edit2rdb
SEMI
SEARCH
edit2rdb
SEMI
rdbFINAL
WTF
WTF
WTF
WTF
WTF
WTF
rdb2wtf
wtf2rdb
rdb2wtf
wtf2rdb
rdb2wtf
wtf2rdb
Relational Database
Stage 8
various
edits
edits
Products
rdbCOLLECT
rdbAUTO
edit2rdb
rdbSEMI
rdbSEARCH
edit2rdb
rdbSEMI
rdbFINAL
WTF
WTF
WTF
WTF
WTF
WTF
rdb2wtf
wtf2rdb
rdb2wtf
wtf2rdb
rdb2wtf
wtf2rdb
Relational Database
Stage 10
various
edits
edits
Dynamic Products
rdbCOLLECT
rdbAUTO
edit2rdb
rdbSEMI
rdbSEARCH
edit2rdb
rdbSEMI
on demand
Relational Database

Stage 1 Process on Unix workstations

Run programs for data collection (to SIP), processing (SIP, AUTO, SEARCH and SEMI), editing (FIXER, etc.) and distribution (FINAL, CAT) without use of the VAX computer. This stage is underway; processing and distribution under Solaris is planned to begin in August.

Stage 2 Generate products from RDBMS

Specify an initial relational schema, implement a relational database system, and insert the initial set of data. The objective is to test the functionality of the schema, probably based on CSS 3.0. Loading of the complete data set may be postponed until the schema is more stable. Also during this stage, FINAL is modified to insert data in the relational system. Ideally, FINAL will be replaced during this stage by **wtf2rdb**, which writes from a WTF to the relational database, and **rdbFINAL**, which derives fixed format files and a PostScript Bulletin from the relational database. Writing each of these programs during this stage would support changes later.

Stage 3 Collect data to the RDBMS

Write a program, **rdb2wtf**, to extract the data from the database to WTF for AUTO and SEMI runs. Ideally, also re-code the **collectors** to write directly to the relational database. Otherwise, sip2rdb must be written to stand in place of it for some agencies on a temporary basis. The re-coded collection programs will be designed to retain some data that are lost under the current system, most importantly the associations supplied by agencies that report the phases. Collecting to the relational database will have the immediate pay-off of making it easier to determine which agency reported a given phase observation and, where duplicates phase reports are collected, which report was retained.

Stage 4 Use agency-supplied associations

Use of agency-supplied associations is generally expected to provide the single largest improvement in the outcome from automatic processing. Since data collection to the relational system includes retention of agency associations, this stage involves writing a program to update the relational database. The new program, **rdbAUTO**, will begin by identifying origins from different agencies for a single event and duplicate reports of phase arrivals at a station. If time permits, development of more complete identification of duplicates and more appropriate selection of a preferred report would significantly reduce editing time. The next step in rdbAUTO will be to associate phases, but to update reported associations only where there is good evidence of a misassociation. In operations, rdbAUTO will be followed by rdb2wtf, which was developed in the previous stage. Ideally, event relocation code is incorporated into rdbAUTO during this stage. If improvement of duplicate processing were given priority, however, a run of SEMI with a null edit file would precede initial editing.

Stage 5 Revision updates RDBMS

In this stage, **edit2rdb** is written to update the relational database based on editor commands. This would be a 2-stage process. First edits are inserted into a table, which could be used to review the editing sequence. Then the edits committed in this table are used to update the table of associations and to mark “deleted” events and phases in other tables. The event locations are not updated by edit2rdb. Instead, each pass would involve updating the relational database based on edits, extracting data from the relational system to WTF, running SEMI with a null edits file to update locations and write a new WTF file, and updating the RDBMS from the new WTF file. This stage will probably require modification of **wtf2rdb** to support repeated updates of the database from successive WTF files

Stage 6 Improve editing utilities

Replace existing utilities such as **daz** and **stn** with programs that return information in more useful formats with less manual keying. One objective is to allow Bulletin editors to refer to events by their number rather than re-keying their coordinates. This would require the utilities to access the current versions of event

locations as well as static information such as station locations. This access will be more straightforward with data managed in a single system rather than the variety of data files that exist currently. With the appropriate development tools, it may be straightforward to incorporate the existing algorithms into a forms-based interface, which might allow Bulletin editors to choose events and phases from a menu rather than keying in event numbers.

Stage 7 Further improve processing

Identifying a larger fraction of duplicate arrivals (allowing for small time differences, or phase misidentifications) and improving selection of a preferred observation could reduce editing time. Improving automatic association (beyond use of agency associations) would also be helpful. A context for further improvements could be developed, for example metrics that measure how well the outcome from a set of association rules matches the edited Bulletin or a flexible system for configuring new rules. A heuristic to identify origins sufficiently robust not to need manual review might be developed, but is unlikely to be used operationally.

The algorithm to exclude association with small magnitude events at large distances could be extended. Magnitude could be estimated (for association purposes) from distances at which phases are observed, or typical station noise and attenuation curves could be used. S-P times can indicate an observation of a local event, in which case teleseismic associations could be excluded. Even if insufficiently reliable for location, vector slowness may indicate need to consider alternative associations.

Stage 8 Eliminate WTF files

Integrate wtf2rdb, SEMI and rdb2wtf to create **rdbSEMI**, eliminating the need for intermediate WTF files. From stage 5, WTF files are used only as temporary buffers between the relational database and the SEMI process. Eliminating buffering to disk files would simplify processing and allow straightforward use of data integrity features in the database management system, which is important to support interactive editing.

Stage 9 Interactive editing

Develop an **interface** to run edit2rdb and rdbSEMI on short time segments or a few events, and show the outcome. As with the new editing utilities developed in stage 6, creating a forms-based interface may be straightforward with the appropriate development tools. Modification of the programs connected to the interface may be necessary to achieve the necessary response time. For example, it may be necessary to maintain an open connection to the relational database. If each Bulletin editor requires several continuously open database connections, the relational database license may have to be upgraded.

Stage 10 Dynamic data products

Modify **rdbFINAL** to allow more flexible selection of time periods and geographic regions, allowing the relational database to generate data products on-demand. For example, users might generate a PostScript Bulletin file for their own use that includes events from an arbitrary geographic region or time period. In addition, shorter time periods could be released; the Bulletin data available to users might be appended weekly rather than monthly. Standard data products would still be needed for the printed Bulletin and to produce the CD, but extraction of these products could be deferred until needed and would be accomplished using the modified rdbFINAL with appropriate parameters.

Stage 11 Further improve processing

Modify **parameters** used for duplicate identification or for automatic association by rdbAUTO. This will involve evaluating new weights or rules using the context established in Stage 6. There may be an iterative process in which a new set of parameters appears to be better based on the evaluation, but in practice must be further refined to actually reduce editing time. There could also be an evaluation of the performance of the origin robustness heuristic, and a decision on whether or not to proceed with excluding manual review of some events.

Provisional Schedule

Stage	Target	Object	Method	Metric
0	1998 May	Make newly processed data available on-line	Create daily files in the web server tree	Static, web-accessible files for completed months since Jan 95
1	1998 Sept	Process data on Sun work stations	Port existing collection, SIP, REVISE, FINAL, utilities from VMS to Solaris	One month processed and edited without the VAX
2	1998 Dec	Generate products from a relational database	Purchase and install RDBMS. Write programs to insert and select data	Postscript and Fixed-Format files for one month generated from RDBMS
3	1999 March	Collect data to the relational database	Rewrite collection programs. Write a program to select from RDBMS to WTF	Create a WTF file for a month without using a SIP file
4	1999 June	Use agency-supplied phase associations in automatic association	Replace AUTO with a program that updates the RDBMS	Create a WTF file consistent with agency associations for one month using rdbAUTO
5	1999 Sept	Revision of event locations updates the relational database	Update RDBMS direct from edits. Modify wtf2rdb to repeatedly update RDBMS	Analyse one month without using edit files in REVISE.
6	1999 Dec	Editing utilities with less keying, more useful output	New programs that select stations and current origins from the RDBMS	No need to key in event or station locations.
7	2000 March	Further improved automatic processing	Develop new algorithms and performance tests	TBD
8	2000 Sept	Eliminate WTF files	Integrate RDBMS select and update tools with SEMI	TBD
9	2001 March	Interactive editing	Write an interface to run existing programs	TBD
10	2001 Sept	Dynamic data products	Modify database extraction tools for greater flexibility	TBD
11	2002 March	Further improved automatic processing	Modify parameters of existing algorithms	TBD

Costs and Funding

Software Development

Software development has a cost, in terms of more staff time devoted to computing than the ISC has had in recent years and purchase of computers and commercial software. The U.S. National Science Foundation has demonstrated a willingness to provide development funding that supplements its normal, operational support. Efforts are being made to obtain supplemental support from the Japan Science and Technology Agency and the U.K. Natural Environment Research Council. Governing Council members could help the ISC to identify similar opportunities with national funding agencies of other countries, and develop support for such funding within their own seismological communities. Development need not necessarily occur exclusively at the ISC, but could be joint projects involving seismologists and programmers in the country funding the development.

One of the most important projects is development of improved algorithms for event formation and phase association. Such changes are introduced in stages 4, 7 and 11 of the development outline. A separately funded project, perhaps carried out off-site, would pay off as bigger improvements when the changes are implemented. The ISC is unlikely to find software elsewhere for these tasks, since associating arrivals from a global network is distinct from the local or regional association done by many other agencies. The algorithms planned for use at the CTBT IDC are known, but implemented only in proprietary programs that are likely to be prohibitively expensive. What's more, they are directed towards forming new events rather than associating arrivals with reported events, and have been tested only against sets of phase arrivals that are many times smaller than those typically used at the ISC. The primary cost of this development would be the salary of a seismologist/ programmer devoted to the project. The project is risky in the sense that it is difficult to estimate in advance how much improvement will result from new algorithms.

Significant costs may be encountered in implementing a data management system. An important cost of this project would be purchase of software and hardware to support it. It is a low-risk project in the sense that the properties of the purchases can be evaluated in advance. The requirements depend on both the size of the database and the level and patterns of use. The size of the databases can be estimated in advance, but the volume and manner of use can have an enormous impact, and these will be incompletely known until after implementation. For example, it is conceivable that so many casual users will make use of the ISC web site that service to subscribers is significantly degraded. A solution to this problem is to purchase multiprocessor computers and multi-disk storage systems. For optimum use, this hardware is complemented by "enterprise" versions of web server and database server software. These versions can be configured to dedicate processors to internal use and subscribers, so that they are well served no matter how many casual users connect. Such purchases are pointless, however, unless experience shows that they are required.

Seismology Development

A difficult aspect of funding the ISC is obtaining open-ended commitments. Stable, long-term funding is essential due to the ongoing nature of ISC operations, but many tasks of finite duration have languished for years because they could not be accomplished by staff supported by open-ended funding. In the future, the ISC could seek funding for seismology projects that payoff in improved operational procedures, such as,

Improving the historical database. Apart from adding recent seismicity, the historical file has not been updated in years, while there have been many improvements in our knowledge of 20^{th} century seismicity. Phase data from ISS and other bulletins could be added to the file. Printed station bulletins at the ISC could be catalogued and preserved.

Relative locations. The ISC could compute relative locations for particular historical

earthquake sequences to improve estimates of rupture area, and then incorporate these procedures into routine operations.

Source parameter studies. The ISC could compute stress drops in particular geographic regions to improve estimates of peak ground acceleration, and then incorporate these procedures into routine operations.

Operating the Envisaged System

Statements of the operating cost of the ISC have normally excluded the cost of replacing computer equipment. This is not necessarily inevitable, as the ISC could depreciate equipment over just a few years, and so demonstrate the need for open-ended funding to cover these costs in its financial statements. Apart from the possibility of funding computer replacement from the operating budget, however, the principal extra costs of operating the envisaged system are salaries for a staff that is larger or includes more marketable skills.

Seismology Staff. Eventually, there will be a many-fold increase in the number of phase reports if thresholds are eliminated. A proportionate increase in the number of ISC seismologists is out of the question. Thus, the size of the seismology staff is linked to the capability of the software to accurately process a sufficiently large fraction of the data, and then to set aside well processed events as not in need of review. A reasonable goal is to employ a one seismologist permanently to maintain the continuity of editing practice, and normally to have two further seismologists each serving for 2–3 years. If the year terms were offset, there would be less disruption while a replacement is being trained. With this staff, the ISC could adopt a transition strategy of incorporating additional data from dense networks into processing only as quickly as software development allows. If a sufficient number of special seismology projects are funded, a further seismologist might be employed.

Computer Staff. The purpose of purchasing a data management system is to reduce the total effort compared with in-house development of the same capabilities. On the other hand, a significant expansion of data services is required for repeated reprocessing of unedited data, real-time reprocessing during editing, and generation of custom data products in response to user queries. Simultaneously, the computer hardware is becoming more complex with introduction of multiple Unix and NT workstations, each with at least some maintenance requirements. Currently, a staff of two fills the roles of system, data collection, data processing and data distribution administrators. This size staff will be able to satisfy the ISC's data processing and management needs only if newly developed tools routinely execute without intervention or error. Even so, it is likely that regular retraining will be necessary, that higher salaries will be required to retain staff with marketable data management skills, and that occasional use of outside consultants may be necessary. Ideally, special development projects will be funded regularly enough to support one further computer staff member.

Clerical and Administrative Staff. The Finance and Administration officer is an essential position that must continue indefinitely. Ideally, data entry would be eliminated at some time, but it is difficult to see when this time will be reached. Keying in edits that seismologists mark on paper is the principal clerical task at present. If the need for this keying is not reduced by introduction of on-line editing, an increased need for data keying lies one or two years ahead due to NEIC's recent cancellation of its data entry contract. Further clerical support would be required if certain projects, such as cataloguing and scanning ISS or station bulletins, were funded or to support resumption of the Bibliography. If enough such projects were funded, one further clerical staff member might be employed.

Including the Director, the projected total staff is 8 to 11, depending on the number special projects that are funded. This is 1 to 4 more people than at present, due to the addition of a seismologist for operations and one further person in each area where special project funding is sufficient.

Appendix A: Readings from Digital Data

Before the advent of digital waveform data, seismologists most often used a few "readings" from seismograms to study earthquakes, partly because the waveform data themselves were difficult to share. Unlike onset times and amplitudes, most information in the seismogram simply could not be used to learn about either the source or the propagation medium due to limits of then-current knowledge.

Many important studies are still conducted this way, but others involve measuring some previously uninterpreted feature of the observed waveforms, and adjusting a model to fit these feature new measurements. On an even more advanced basis, in some studies observed digital waveforms are directly compared with waveforms computed from theoretical models.

There are two potential objectives in retrieving digital waveform data to the ISC and measuring features from them. The first is to expand the set of stations and phases for which the ISC has the traditional set of readings. The second is to provide measurements not commonly offered in the past, but now potentially useful to a large number of seismologists.

Additional Stations and Phases

Although not well documented, many seismologists believe that there is a growing number of stations from which data are digitally recorded and archived, but not routinely read. These data are used for improving studies of particular earthquakes or earth structure in particular regions, but much of their potential value is lost as a result of the absence of systematic reading. They do not contribute to either improvement of the detection threshold or studies based on standard phase readings.

The problem is that shrinking resources for phase reading cannot cover salaries for personnel required to read all of these records. One way to partially address this is to make automatic readings from digital waveforms. This might not be an STA/LTA phase picker, but could be based on more sophisticated signal processing algorithms run at the predicted arrival times from known earthquakes. This would not improve the detection threshold, but could make readings of later phases more common in the Bulletin. It could be used to read amplitudes of surface waves at stations where body wave phases are not detectable.

One potential problem is that automatic readings might differ systematically from manual readings. Because of this potential bias, it would be prudent to clearly distinguish between automatic readings and manual readings in the ISC databases. Furthermore, might prefer that the ISC continue to compute estimates of origin parameters from manual readings alone, and to supplement them with origin parameters computed from the joint set of readings.

New Features

There is a wide variety of potential new readings, and each one would be useful for at least a few types of studies. Spectral ratios would be useful for earthquake/explosion discrimination. Shear wave splitting measurements would be useful for some types of structure studies. Corner frequencies could be used to estimate stress drop, and thus aid some types of hazard analysis.

One important question is whether or not any new measurements would be widely enough used to justify measurement and archiving by the ISC. In addition to spectral ratio and corner frequency, measurements likely to be widely used might include slowness measured at arrays and polarisation at 3-component stations. To some extent, the question can only be answered after the fact; some features might have uses that will be recognised only after they are readily available. The implication of this is that some measurements might be included on a trial basis, and dropped several years later if they are not sufficiently popular. Signal-to-noise ratio might be widely used, not as data in a study, but to decide which waveform data are most likely to be useful when they are retrieved.

Another question is if ISC measurements would be considered sufficiently trustworthy by individual researchers to dispense with measurement themselves. The ISC would not necessarily have to satisfy all researchers in this regard. For some types of studies, for example, ISC arrival times are the best data because they are so numerous even though individual researchers re-read arrival times for other studies in which consistency and care in reading a relatively small number of times is more important.

Development Schedule

Automated reading of digital data at the ISC is not included in the provisional schedule. This type of project would require considerable investment in development by ISC staff qualified in both seismology and programming. However, it is ideally suited to treatment as a separate project. The software could be developed over a finite duration, which is a requirement of many funding agencies. The software could even be developed at sites in other countries, which might broaden the range of funding agencies that would consider the project.

Appendix B: The Bibliography

The Bibliography of Seismology has been a valuable tool for some seismologists in seeking published information on various topics in seismology and in references related to particular earthquakes. But general-purpose indexing and abstracting services are now widely available, and are providing an increasing range of services. For example, hypertext links to sources are becoming more important as the number of on-line abstracts and full text on-line publications continues to grow. Examples of on-line abstract already available include *Geophysics* from SEG, *Geophysical Journal International* from Blackwell, *Earthquake Engineering* from Wiley and all AGU publications. Computer searches have made poring through the printed Science Citation Index a thing of the past for most scientists. Progressively more sophisticated tools allow searches to be both more selective, reducing the number of inappropriate references returned, and more comprehensive, for example by searching the full text of papers rather than simply titles and keywords. More flexible downloading of search outcomes saves users more time by delivering results directly into personal bibliographic databases.

The Bibliography was suspended for financial reasons, and resuming it would have a cost to the ISC. Time would be required for data entry, and the computer staff time would have to maintain the programs used to compile it and develop tools necessary to encourage use by many, or by anybody for very long. Some have argued that the data could be compiled electronically. One approach to this would be to compile a fixed list of keywords, to retrieve citations and abstracts from publishers, and to index based on searches for the fixed list of keywords. The task is conceptually straightforward, but this saves only data entry time and requires development and maintenance of further software. What's more, the outcome is of

uncertain value compared with general-purpose indexing such as WAIS and fails to provide the indexing with respect to particular earthquakes discussed below. What's more, while AGU, SEG and RAS have offered positive preliminary responses to enquiries, the availability of low-cost electronic text to the ISC from for-profit publishers is questionable.

Reasons cited for resuming the Bibliography are that searches of other, general-purpose bibliographic databases

- are expensive
- return non-seismic references
- omit some sources that seismologists need to search
- cannot be used find references related to particular earthquakes

Cost. If the ISC is to offer the bibliography less expensively than general-purpose bibliographic databases, then at least one of these must be true:

- The ISC's costs can be better hidden from users of its database.
- The ISC's production cost per search is less than other services.
- Prices of other bibliographic services are disproportionate to production costs.

The first statement is false; the ISC is not in a financial position to subsidise development of a bibliographic service comparable to others already available. The second statement is also false. Because they spread costs over many more users, well-known general-purpose bibliographic services will have much lower costs per search than the ISC if it develops comparable services. An attempt to keep production costs very low would likely lead to inadequate service compared with general-purpose bibliographic databases, which would cause most potential users to ignore the Bibliography.

There is little evidence of overpricing among bibliographic services. Certainly, university libraries have made a persuasive case that some journal publishers are charging so exorbitantly that the wide dissemination of scholarly work is at risk. However, bibliographic services compete directly since each can form a database from all widely subscribed journals. This limits overpricing, especially when a non-profit organisation such as AGI plays a major role in the field by producing GeoREF.

Irrelevant References. Computer-based bibliographic services provide a wide variety of ways for users to restrict searches. Many allow searches to be restricted to a selected set of journals, which would be similar to, although more flexible than, the filter that the ISC provides. Some allow restricting a search to references related to particular fields, for example GeoREF allows users to restrict searches to one or more "category codes", including "general geophysics", "solid earth geophysics", "applied geophysics" and "seismology". Bibliographic database users would serve themselves best by learning to use all of the features of existing services or working with a librarian to formulate their searches.

Limited Sources. The Bibliography includes "grey literature" publications, mostly related to weapons test monitoring, but it is incomplete in this area. For example, many references to IAEA publications can be found in GeoREF that are absent from the ISC Bibliography. What's more, it seems unlikely that cataloguing this type of publication is the principal reason that most ISC users might want to see the Bibliography resumed. Some have argued that other bibliographies typically have a bias in favour of US and western European publications compared with the ISC Bibliography. However, Elsevier's GeoBASE, AGI's GeoREF, and ISI's Science Citation Index each probably include all of the journals indexed in the 1995 Bibliography. Certainly, no specific examples have been found by or reported to the ISC.

Earthquake Indexing. In preparing the Bibliography, keywords for significant earthquakes are introduced and publications mentioning these earthquakes are indexed with these keywords. This is the aspect of preparing the Bibliography most closely related to the ISC's primary mission of producing a global seismicity Bulletin. This indexing is a unique feature of the Bibliography, and it is plausible that searching the general-purpose Bibliographic databases for papers related to particular earthquakes would be difficult.

Earthquake indexing could be viewed as an updated version of the citations found in the Shide Circulars and in the earliest ISS Bulletins. Probably, better established methods for routinely gathering the data required to locate earthquakes is part of the reason that citations do not appear in ISS Bulletins after the early 1920's. Perhaps improved timeliness of the Bulletin, so that important papers were more often still in preparation, contributed as well. If so, this situation no longer prevails as an on-line Bulletin can easily be updated to include new citations whenever they are published.

Two important and related questions are how widely citations within the Bulletin might be used and what is the minimum cost to compile these citation indexes. The implication of the preceding discussion is that compiling a database is prohibitively expensive if the goal is to produce a bibliography that can be used to for a wide variety of searches. On the face of it, it would appear that preparing a database for a narrower purpose would be even more difficult to justify. Perhaps, however, seismologists elsewhere can be enlisted to index publications for this focused end, leaving it for the ISC only to gather and archive their indexing. This would be analogous to the way that record reading, which is the majority of work required to produce the Bulletin, is done by the globally distributed seismology community.

Appendix C: Typical Data Access

Data Collection

Phase readings, preliminary earthquake locations, and associations of phase readings with earthquakes are provided by operators of seismic stations and regional networks, national earthquake agencies, and other agencies preparing preliminary global bulletins. Data in an individual report may comprise as few as dozens of earthquakes and hundreds of phase readings, or as many as thousands of earthquakes and tens of thousands of phase readings. Some data arrive within a few days of real time, while others arrive as much as 18 months behind real time and include data for intervals as long as 6 months.

Some data are duplicates from multiple sources (*e.g.* an individual station operator and the national agency to which it reports). In the case of duplicates, the ISC has an internally maintained set of preferences based on number of data attributes or presumed reliability. Some data are updates of previous reports, *e.g.*, refined earthquake locations based on readings from additional stations. Tracking of data sources and stated accuracies is critical. Many data will be missing some attributes, which might be recorded as null values in the ISC database. Updates of previously reported data are not necessarily flagged, so the ISC needs to recognise updates based on comparison with previously accepted data.

Data collection software undergoes continuous change to adapt to new data formats and data types. Less frequently, data are accepted to update information about seismic stations or provide information about newly registered stations.

Automatic Processing

Data are traditionally processed in 1-month batches. Recently processed months include as many as 8000 earthquakes and 200,000 phase readings. Since the bulletin is released monthly, there has been no advantage to completing processing of the first day of a month before the last day. Shorter batches probably could be processed; other earthquake data centres automatically analyse intervals as short as 1 hour. Earthquakes could not be processed one at a time since association of a reading with several earthquakes must be considered.

Automatic processing comprises

- Identify duplicate earthquakes and phase readings, and identify the "best".
- Associate phases with reported earthquakes based on predicted versus observed times, amplitudes and other properties.
- Compute refined earthquake locations and magnitudes from observed times and amplitudes, using models of travel time and signal attenuation, including provision for detecting insufficient or conflicting data.
- Identify previously unreported earthquakes by associating otherwise unassociated phase readings.
- Prepare listings and other materials used for "editing", *i.e.*, manual examination and correction of the results from automatic processing
- Prepare materials for release, including postscript files for publications and fixed-field ASCII files. Publications are the Bulletin (one month of earthquakes and associated phase data) and the Catalogue (six months of earthquake locations *without* phases).

Processing requires use of all attributes of all data within the time interval being processed, including uncertainties, and some attributes of data about selected stations. Recording the results of processing includes inserting or updating earthquake locations and associations.

Interactive Editing

Data are traditionally edited in lots of approximately 500 earthquakes. Smaller lots probably could be edited; other earthquake data centres interactively edit intervals as short as 4 hours. Editing comprises the same tasks as automatic processing, except preparing materials for editing and release. Editing requires use of all attributes of earthquakes, phases and associations within the time interval being analysed, including uncertainties, and some attributes of stations from which arrivals are reported. Recording the results of processing includes inserting or updating earthquake locations and associations. Editing is done now from tables of printed data, but graphical displays (maps using projections specific to earthquake or station locations, phase reading attributes versus association attributes) might improve the speed and accuracy of editing. Editing now involves reprocessing of data only in batch mode, but interactive reprocessing (based on hypotheses that might be rejected almost immediately) might improve the speed and accuracy of editing.

Data Distribution

ISC data users may be divided into two classes – users of the Catalogue, which contains earthquake *without* phase data, and users of the Bulletin, which contains both the earthquake locations and the phase data. Until a few years ago, the printed Bulletin and Catalogue were the only standard products with new data from the ISC. Many users requested the data in computer readable format, however, which were distributed as ASCII files in a Fortran-influenced format on 9-track tapes written in response to individual orders. The complete data collection of the

ISC is available now on five CDs, and a new disk is planned annually. The most recently distributed CD includes, in addition to the 1994–1995 Bulletin, the complete 30 year Catalogue.

The Catalogue contains summary statistics of associated phases, *e.g.* number of associated phases and average misfit of the observed phase times to the predicted times. These statistics might be seen as an *ad hoc* implementation of data warehousing concepts, since they allow selection of the most reliable earthquake locations based on the associated phases without accessing the phase records themselves. Catalogue users are interested in earthquake locations and sizes, often for estimating seismic hazard or for academic research in tectonics, usually in a restricted geographic range. The Catalogue includes "region numbers" under a geographic indexing system developed in the 1960's from seismicity patterns. Most Catalogue queries are restricted to just one of the 50 "geographic regions", and many may be restricted to just a few of the hundreds of smaller "seismic regions". However, a front-end would be required to convert user-supplied latitude and longitude bounds to region numbers. Earthquakes are often treated as a stationary process, so many queries are not restricted by time. Some queries will have very tight time restrictions to select a particular earthquake, moderate time restrictions to select an aftershock sequence, or broad time restrictions to select earthquakes occurring after certain types of instruments were installed.

Bulletin users are almost exclusively academic researchers. A bulletin user might

- Relocate earthquakes using phase times and different wave propagation models.
- Refine wave propagation models using differences between observed and predicted phase times.
- Investigate earthquake properties using phase data attributes other than time.

Most Bulletin users probably begin with queries similar to Catalogue users, employing statistics of phase data to select the most reliable earthquake locations. Depending on the application, the selected earthquakes might number from one to thousands. A Bulletin user would then retrieve associated phases restricted not by time misfits, since they will likely relocate the earthquake and so change the misfits, but by phase type or by distances between earthquakes and stations.

Appendix D: Interactive Editing

There are two goals in introducing interactive editing. One is to improve the quality of the Bulletin by providing editors with an opportunity to consider more alternative associations and inverting from more starting locations and with a wider variety of constraints (*e.g.*, fixed depth or location). The other goal is to reduce the time spent editing each event, so that the current number of staff can handle the load. Deadlines cannot be missed indefinitely, so the two goals are coupled, since the inevitable result of insufficient time is more errors.

One ISC seismologist's description of editing an event is:

1. Work out how the ISC estimate was calculated, whether it was calculated at all and, if so, that it is reasonable for the data. In some instances, work out a better starting point for the location algorithm using additional information not taken into account by the location algorithm.
2. Decide whether the location and depth are plausible, meaning extra care with data revision if they are not.
3. Look through the external locations submitted to the ISC. Make sure all of them were properly associated. If not, put them into the proper event or remove them.

4. Look in the nearest time and space vicinity to identify the possibility of a split event, cause by different agency mislocation. Merge two events if it is the case.
5. Make sure that current event doesn't consist of two events, accidentally merged together.
6. Make sure all comments provided are properly associated and do not carry logical and geographical errors or redundancies. Correct errors encountered.
7. Scan through the list of stations reported, phase identifications provided and residuals calculated. Decide which readings were not properly associated and whether they should be put into different events or just removed.
8. In some instances, look into the unassociated data stream to use the data not associated with the event due to poor original determination of the hypocentre.
9. Remove duplicate station readings.
10. Modify station comments that are too long to fit in the Bulletin format.
11. Remove outrageously wrong amplitude readings to avoid artificially high or low mean magnitude value.
12. Confirm doubtful cases with the local, regional or global bulletins provided for the ISC in both printed and electronic forms.
13. Review automatically rejected readings in order to use them properly in the current or different event.
14. Move promptly to the next event.

When trying to envisage appropriate interactive tools to aid this work, it is useful to consider the amount of time spent on each event. After subtracting holiday leave (6 weeks), bank holidays (2 weeks), typical medical leave (1 week), and conference participation (1 week) a full-time seismologist spends 42 weeks or, presuming 37½ hours per week, 1575 hours editing the Bulletin. With two full-time seismologists, the average time available to edit each of the 67,000 Bulletin events in the most recently completed year is less than 3 minutes. Since the first pass for each month typically takes about half of the total time to process it, no more than 1.5 minutes must be spent on average per event in the first pass.

ISC seismologists routinely handle most small events (with no more than a dozen or so associated phases) in the time available. Indeed, many small events are accurately processed and these, seismologists claim, receive less than 10 seconds of attention in the first and only pass in which that they are reviewed. The claim is credible since as many as 2000 earthquakes from a late pass are occasionally reviewed by a seismologist in a single day, which works out to an average of 15 seconds per earthquake sustained for 8 hours.

Larger events, even if accurately processed, take longer simply because there are more phase readings to review. Minor delays arise if automatic processing "splits" an event, *i.e.*, produces multiple locations for single events based on disjoint sets of arrivals. Further delays can arise if the existence of an event is plausible, but it appears to be badly mislocated by the reporting agency due to insufficient data. In these cases, perhaps using multiple location hypotheses, unassociated arrivals are manually searched for additional data. Troublesome delays also arise when events are "mixed" – *i.e.*, when phase readings from two or more events are associated with each other, leaving the seismologist to work out where the events really are and which phases to associate with each.

How can software help to speed editing? With respect to routine events,

- Most duplicate phase readings are already flagged. The only significant improvement here is likely to come from greater confidence in automatic detection of duplicates, so that a seismologist need not check them at all.
- Associations at large distances from small events are easy to spot. Automatic rejection of some of these has already been implemented, and extending this automatic processing is the most likely route to further improvement.
- Phase readings with large time residuals from events with many dozens of readings might be more quickly spotted with a graphical display of travel time residual, but such large events are a small minority the total.
- Other inconsistent data, such as vector slowness, might be useful for finding false associations. As for large time residuals, however, a graphical display seems unlikely to speed things much except when many readings are associated with a single event.

Finally, selecting constraints and starting locations is an important issue only for a small fraction of locations that are poorly constrained. An opportunity to test location runs at this stage might help to reduce the number of passes through the data. But the time spent in each pass would increase, especially if seismologists came to routinely invert for locations for every event. In short, it is hard to see how time spent editing routine events can be reduced, except by improving automatic processing so that many more routine events need no editing at all.

But speeding review of the routine events is not really the point. In most recent months, approximately one-third of all events to require no modification after the initial processing. Supposing that 2000 correct events each require 20 seconds to review, approximately 10 hours are spent confirming accurate events in each month's data. Since each month is currently requiring 5 weeks work by two full time editors, or 400 hours of editing, confirmation of accurate events is clearly not the most important cause of delay in the Bulletin.

In recent months, a little more than one-half of all events are modified in the first pass, then reviewed and accepted without further change in the second pass. From these events a few phases must be taken or duplicates deleted. Perhaps some reported origins need to be moved from one event to another, or changes are required in the magnitude, felt reports, moment tensors or other comments. An editor spends, say, 2 minutes with each of these events in the first pass, and a further 30 seconds confirming the changes in the second. Collectively, then, these events consume approximately 125 hours of editing.

More than 250 further hours are spent editing events from each month. The majority of this time, apparently, is devoted to fewer than 1000 events that have mixed phases, are split, or for other reasons require special effort. Thus, each of these troublesome events requires an average of approximately 20 minutes of editing time, and it may be here that an interactive editing tool can achieve significant reductions in time. In these cases, the seismologist is judging how well a phase is likely to fit a location hypothesis. The existing tools for this require keying of station codes, event locations and arrival times. Significant editing time might be saved with some simple improvements to these tools. One could imagine, however, a sophisticated system in which seismologists move icons representing phase arrivals between fields representing association with different events. This action might prompt relocation of the effected events, after which new residuals would be displayed and large misfits flagged.

For a given level of effort, an alternative to developing such sophisticated tools is to invest it in improving the automatic processing, so that there are fewer mixed and split events to be edited. Using the associations from the final pass as “ground truth”, it should be possible to test new association rules and count the number of mixed events and misassociated arrivals that result from each set of rules. The more successful we are in discovering rules that correct misassociations, the smaller the benefit will be from implementing interactive editing tools. Given that little time has been available to devote to such improvements and that editors recognise the same, routine association problems recurring each month, it is plausible that straightforward rule changes might halve the rate of split and mixed events. In contrast, the effort required to develop a graphical interface that halves the time spent editing each incorrectly processed event might be substantial.