Ishii and Ishii
Progress in Earth and Planetary Science (2022) 9:50
https://doi.org/10.1186/s40645-022-00508-0

Progress in Earth and Planetary Science

REVIEW

Open Access

DigitSeis: software to extract time series from analogue seismograms

Miaki Ishii* and Hiromi Ishii

Abstract

A vast amount of analogue seismograms recorded between the end of the nineteenth century and late twentieth century are often inaccessible for seismological research since they are not available as digital time series. This manuscript describes freely available software, DigitSeis, that takes a digital image of an analogue seismic record and returns waveforms either as a function of their x-y position on the image or as time–amplitude information. The overall structure and approach of the software are provided along with how they have evolved over different versions. The effectiveness of the software is demonstrated with three examples. The first example is a long-period east–west seismogram recorded at the Harvard Seismographic Observatory on photographic paper in May of 1938, which contains signals associated with a magnitude 7.7 earthquake that occurred off the coast of Northern Ibaraki, Japan. The second example is an analysis of a 35-mm microfilm copy of the short-period vertical seismogram recorded at Tucson, Arizona, on July 16, 1945, that shows blast signals from the first nuclear bomb detonation. The final example uses a 70 mm microfiche image of a long-period north–south seismogram recorded at College, Alaska in December of 1966, which shows a pair of earthquakes with nearly identical waveforms. The software is, by no means, perfect, and discussion of its limitations such as the compatibility with pen- and Develocorder-type seismograms is included, as well as comments about challenges of incorporating machine learning into the digitization process.

Keywords: Analogue seismograms, Digitization, DigitSeis software

1 Introduction

Recordings of ground motion using seismometers began in the late 1800s, and these valuable data capture any phenomenon that transfers energy into the Earth. Consequently, they can be used to study unusual earthquakes or nuclear explosions. In addition, recent developments in seismology show a trend in utilizing data beyond traditional event-based waveforms. What used to be "noise" is now becoming "signal" (e.g., Shapiro et al. 2005), and if the analogue data can be made digitally accessible, the database will expand significantly in time (by a factor of 2 to 4, depending upon how one defines the "digital" era and the start time of the analogue collection). These data will therefore open the possibility of studying phenomena that evolve over time. Geophysical processes have time scales that are much longer than the combined periods of digital and analogue recordings, but for some relatively rapid processes, having a longer time window may provide substantial insight into the mechanisms. For example, monitoring subsurface evolution before and after volcanic activity, examining the rate of inner-core super-rotation, or examining effects due to global warming would benefit greatly from having a much longer time window than just those from the digital era.

However, much of the analogue data, the predominant format up until about the mid to late 1980s, are incompatible with modern analyses that require digitized time series. The length of the analogue recording era also means that there is a staggering amount of data to be digitized. Additionally, the sheer physical volume of the analogue recordings poses a storage challenge, with increasing rent putting many analogue seismogram

*Correspondence: ishii@eps.harvard.edu

Department of Earth and Planetary Sciences, Harvard University, 20 Oxford Street, Cambridge, MA 02138, USA

collections in danger of being discarded. Finally, working with the analogue seismograms is becoming more difficult as time passes. People with knowledge of these recordings are retiring, and the medium on which they are stored (e.g., photographic paper and microfilm) is deteriorating.

Despite the immense value the analogue seismograms represent to the Earth Sciences community, there has been little progress in reclaiming these data for modern applications. The two main reasons are the lack of efficient methods to digitize the images and the substantial volume (and hence required labor) of the collections. Digitization programs have been written (e.g., Bromirski and Chuang 2003; Pintore et al. 2005), but they are difficult, complex, and time-consuming to use and, thus, have not gained general traction. Some images of significant earthquakes have been digitized by individual researchers (e.g., Kanamori and Cipar 1974; Okal and Stein 1987; Song and Richards 1994), but these are exceptions and most extant seismograms remain unused in storage. In the last decade, the Harvard Seismology group has been working to preserve and convert analogue seismograms that were recorded at the Harvard Seismographic Station. One part of this effort is the development of software that takes a digital image of a seismogram and generates time series. In this manuscript, we review the evolution and current status of this digitization software and describe some challenges and future directions.

2 Review

Harvard University operated analogue seismometers between 1908 and 1933 at the main University campus in Cambridge, Massachusetts, in the USA. These facilities were moved to a quieter location at Harvard, Massachusetts in 1933 (Leet 1934) and were operated until 1954, and many of the seismograms survived to the twenty-first century. Working with historical photograph specialists from the Weissman Preservation Center at Harvard University, the Harvard collection of analogue galvanometer-style seismograms was cleaned and scanned (Ishii et al. 2015). The seismograms were then placed into 75 boxes designed and made specifically for this purpose by staff at the Weissman Preservation Center, and transferred to the care of Harvard Archives in 2016. They are now stored in a temperature- and humidity-controlled storage space, reducing further deterioration, and are made available to the public through the Harvard Library system. Images (tif and jpg files) of all 11,339 seismograms are also available online at the Harvard Seismology Web page.

The ultimate goal of this data rescue project is to convert the analogue seismograms to digital format for use in research. We initially experimented with three digitization software that were available in 2011. The first two codes tested, SeisDig (Bromirski and Chuang 2003) and Teseo2 (Pintore et al. 2005), are well-known, freely available software for digitizing analogue seismograms. The third code, NeuraLog (NeuraLog 2013), is widely used digitization software in the oil and gas industry for well-logs. Even though all three software automatically traced some parts of the scanned seismograms, they each required a significant amount of manual interaction. For example, every software had issues identifying and tracing minute marks, which are vertically offset from the main trace, and the sampling rates caused issues with reproducing intricate features of seismograms. It became quickly evident that it would be very difficult and time-consuming to digitize the Harvard seismograms using any of the three available software.

Given such challenges with the existing tools, new software has been developed. The first version of this software, DigitSeis, was publicly released in 2016 (Bogiatzis and Ishii 2016). There have been three subsequent releases, versions 1.1 (2017), 1.3 (2018), and 1.5 (2020), and every version is available on Harvard Seismology Web page. We provide a brief description of the software and how it has evolved over versions in the following section as well as some examples of DigitSeis output. The software is by no means complete, and we outline some of the major issues that will be addressed in the near future.

2.1 Digitization software: DigitSeis

DigitSeis takes advantage of the extensive library of image processing routines that are both built into MATLAB and publicly shared by various programmers. The software is not designed to be fully automatic, i.e., it requires the presence of a human analyst. This approach allowed for the simplification of some steps compared to previous software. For example, time marks are difficult to handle in other software that tries to automatically trace everything, and the user is typically forced to manually trace individual time marks. On the other hand, in DigitSeis, by requiring user input for some key parameters, time marks can be treated semiautomatically, significantly reducing the amount of work and time required. DigitSeis also aims to obtain the most accurate time series possible, and considers often-ignored features such as distortion of the images; some of these considerations are highlighted below.

One of the main motivations for the development of DigitSeis was the challenge of working with seismograms where vertical offsets are introduced for a short time to indicate the beginning of each minute or hour (time marks). In order to effectively treat the time marks, the DigitSeis software has a classification step where different objects within the image are separated into time marks

and main (albeit disconnected) traces (Fig. 1). This is performed semiautomatically by considering the user-specified horizontal lengths of individual objects along with user identification of poorly defined or strange objects (e.g., stains and handwritten notes). Once the classification is finished, the time mark objects are amalgamated with the main trace with a vertical offset that produces as smooth a connection as possible (Fig. 2). This offset is obtained by using the derivative of the combined trace (if the vertical correction is insufficient, the derivative will be large). Because the software works under the assumption that these vertically offset segments exist, the time marks are no longer an issue in the digitization procedure.

One property of the seismogram images that came as a surprise is the level of distortion they contained. On hindsight, this should have been expected considering the significant amount of curvature and deformation that exist in the original analogue seismograms. Some of these distortions/deformations come from the poor conditions in which the seismograms were stored, but a major contribution is the difference in materials that make up the photographic paper. The trace side, with a photographic surface, has a gelatin emulsion that contracts as the surface dries over time, leading to curling of the paper inward (since there is no contraction on the back side made up of plain paper). In addition, the process of acquiring digital images of these seismograms produces distortion such as the lens effect on images taken with a digital camera (even with a wide-angle lens). Regardless of the source of distortion, it is often significant enough to jeopardize the tracing exercise by obscuring the zero-amplitude position and complicating the conversion of horizontal position to time. With these complexities, time cannot be assigned to the trace simply by use of the horizontal position, and the existence of the time marks turns into an asset. DigitSeis characterizes and corrects for image distortion, takes the positions of the time marks, determines the time associated with each mark, and interpolates between them to assign time to a corrected "horizontal" location (Bogiatzis and Ishii 2016).

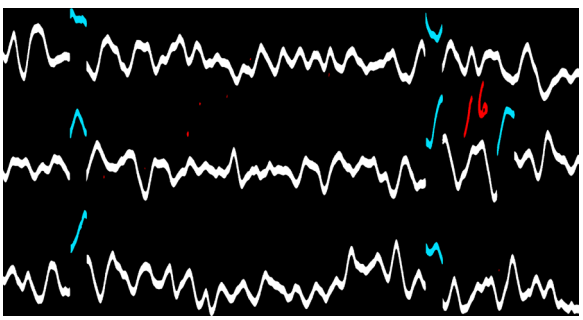

Fig. 1 An example of classification. The input image is processed to define objects that are categorized as main trace (white), time marks (cyan), or noise (red)

The above description of procedures summarizes the basic approach with which DigitSeis has been developed. There have been significant modifications in terms of how exactly each step is taken, and they are reflected in the multiple version releases. Just to illustrate the magnitude of the work that went in over the last several years, the software that was initially released (v0.53) consisted of about 4,000 lines of code. Nearly 100% of this code has been either modified, replaced, or removed, and with new features that improve accuracy, robustness, stability,

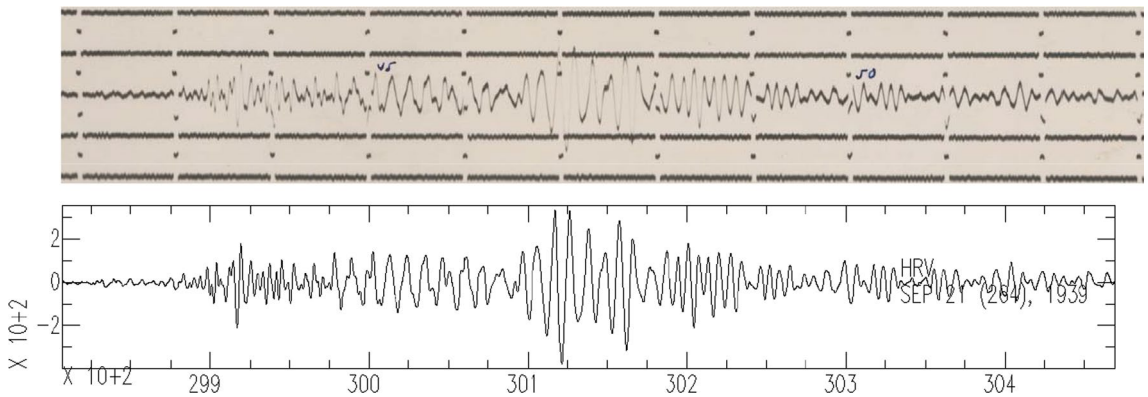

Fig. 2 An example of digitization result. Comparison of the original (top) and digitized (bottom, snapshot of the SAC window view) long-period north–south seismogram containing seismic wave arrivals. Note that the x- and y-axes do not quite match for the two plots, although the SAC display has been stretched/shrunk to mimic the original image. The numbers "45" and "50" on the original image correspond to 21:45 and 21:50 on September 21, 1939. The seismic wave arrivals are associated with a magnitude 5.8 earthquake in the Gulf of California (International Seismological Centre 2022)

and usability, the current released version (v1.5) contains nearly 50,000 lines of code. Some of the improvements for versions 1.3 and 1.5 are described below.

2.1.1 DigitSeis version 1.3

DigitSeis version 1.3 was released in 2018 and contained significant improvements over the initial version.

Stand-alone processing The initial version of software required an image that was nearly perfectly clean for processing and included a handful of tools to achieve this (e.g., remove stains, adjust contrast). All of these image-cleaning functionalities were applied to the entire image rather than to specific areas. Analogue seismograms often have regions of exposure and condition that differ in quality, that result in heterogeneities in the image, i.e., full-image processing is not ideal. Prior to version 1.3, the solution was to clean the image using another software (e.g., Photoshop) before importing it into DigitSeis. This was changed in two ways. The classification and digitization algorithms were modified so as not to require the cleanest possible images to work (in fact, version 1.3 uses the original image, including stains and handwritten notes, for the definition of objects and calculation of digital traces), obviating the need for third-party software. Another improvement was the introduction of automatic image cleanup, and tools that allow the user to process only small portions of the image when the automated results are not sufficient. These significantly reduced the work load and sped up the digitizing process. Furthermore, some images that were undigitizable with earlier versions could be processed with version 1.3.

Efficiency Many of the algorithms were rewritten to improve the efficiency of the program. Two examples are in object classification and digitization (two of the most important components of the software). The classification process used to be very cumbersome, with even a simple operation such as changing an object type (e.g., to make an object that has been misclassified as noise into a time mark) easily taking 20–30 seconds, and in the worst case, about a minute. From the user's point of view, this is frustrating; after clicking twice, the user needs to wait for an unreasonable length of time before the next object can be processed. With version 1.3, the change became nearly instantaneous. Other features needed in object classification (e.g., region removal, undo) were sped up in a similar manner. Another example of efficiency gains is in the digitization process. For simple images, it used to take about 15 minutes, and for complicated ones, it took up to an hour. With the different approach and newer algorithms, even the most complicated images could now finish in less than a minute.

Automation With the vast amount of seismograms available to be digitized, any automation is desirable. In developing version 1.3, the code was scrutinized for any tasks that could be automated, and when possible, this was done, for example, the timing calculation used to require the user to manually select the starting and ending positions, as well as the positions of the first and last time marks for each trace. All of these tasks were automated (with the option for the user to make modifications if automation resulted in imperfect location determinations) so that it could be done with a single button click.

Robustness The original version of DigitSeis had quite a number of issues, in terms of both the setup and algorithms. For example, the user could inadvertently open multiple analyses and have actions applied to wrong analysis, or if there were multiple open windows of a single analysis, the user could mistakenly click on the wrong window (e.g., the main window while working in the Classification window) leading to the next figure appearing in the wrong place. These issues were corrected in version 1.3. Another major problem with previous versions was that some algorithms resulted in problems that were either fatal, slowed the processing down, or produced numerous error messages (despite not having functional problems). All these known bugs were fixed, and practically no unintended error messages were produced, and the user is no longer forced to terminate the program. The third way in which the code was improved was by providing more flexibility. For example, the parameters that controlled the calculation of trace-zero lines (line tracing zero-amplitude position) were hard-wired in the previous version of the code, and if trace-zero lines were incorrectly determined, the user needed to make changes in the source code. These types of issues were changed in two ways. First, by providing tools so that the user can modify default settings (more below). Second, by taking advantage of the known issues and their causes, and improving the algorithms to automatically address them. The final type of modification that improved robustness of the code was the conversion into more mistake-friendly code, such that even when the user makes mistakes, in most cases they can easily correct it. In the old version, it often resulted in needing to redo major parts of the analysis.

Accuracy Accuracy of the digitization process was improved through three approaches. One was to implement algorithms that were more accurate. For example, the trace-zero lines were originally calculated using rough estimates based upon image intensity distribution. Version 1.3 takes advantage of some of the processing already done and uses the trace objects to determine the trace-zero line. This approach required reorganization of various processes, but produced more reliable results. The second way in which version 1.3 improved accuracy

was by providing the user with check tools. For example, nearly 3000 objects exist on a single analysis, and it is difficult to make sure that every single object is classified correctly (e.g., no noise object is mistakenly classified as time marks). Consequently, the ability to display locations of the time marks so that the user can easily check visually if any objects have been misclassified was introduced. Similarly, a check algorithm that allows user to see if any section of the trace has been missed was also included. Finally, additional tools that allow the user to correct results of automatic calculations were implemented. For example, in the old version, once the location of each minute position was determined by the time calculation algorithm, it was impossible to change them even if the code produced less than satisfactory positions. Version 1.3, gave the user the ability to move the minute position bars so that they can be updated easily.

Compactness The method for saving the analysis was completely rewritten to improve the data storage structure, and reduce file size (and hence time needed to save). For example, a complicated analysis with hundreds of segments requiring manual correction (e.g., the amplitudes are such that traces are touching those above and below), resulted in files that were more than 6 GB using the original algorithm and took tens of minutes to save. Using version 1.3, the same analysis could be saved in a file that is about 700 MB, with the significant reduction in size leading to quicker save and load times.

User control Many of the features and default setups were hard-wired into the original version of DigitSeis, and the user had no control over them. For example, the default color scheme for the classified objects (i.e., white for main trace, green for time marks, and red for noise objects) in v0.53 did not work for color blind people. The default color scheme was changed to take this into account, and the user was also given an option to set their favorite colors and save them as the default setup. Similarly, the user was given control over other default setups (e.g., size of the window for correcting traces) the setting for which could also be easily accessed and saved.

Clearer work flow The layout of the original version of DigitSeis (Fig. 3) was such that the user had to know exactly which buttons needed to be employed next. In version 1.3 this considerably changed in two ways. One was organization of the buttons so that they need to be pressed from top to bottom as the user made progress. The other was activation and inactivation of these buttons. Many of the functions are not made available until the user completes the previous step, and some of the functionalities are turned off after the user completes it. With this arrangement, the time spent learning the program was greatly reduced.

Simplification The DigitSeis window in version 1.3 was simplified in appearance (Fig. 3). Unnecessary plots, messages, and control buttons were removed while some of the new and old functionalities that are typically not needed were moved and hidden from the main window. Furthermore, improved programming reduced the amount of interaction with the user. For example, selecting and inputting positoin of a point used to require the user to click on the cursor button to activate cursor, click on the screen, find out the x and y positions of the cursor location, manually type in the position values into a window, click on a button to store the information, and then manually remove the data cursor symbol that remained on the figure window. In version 1.3, the need for this interaction was mostly removed (through automation), but if necessary, the user has the option to click on a button to activate a point selection algorithm, click on the screen close to the desired location, move the point around until satisfied, and then double-click on the symbol to finalize the position and to have it removed from the screen. These updates were not necessarily improvements to the digitization process itself, but nonetheless significantly reduced the amount of time the user needs to spend on a single trace.

Saved information Even though the previous version of DigitSeis generated files that were several factors larger in size, some important information was not saved (while some unnecessary data were). This made reproducing some processes difficult, if not impossible, and required redundant actions if the analysis was saved and closed. For example, association of each object with a specific trace was not saved, so whenever the user needed to digitize again, the association needed to be recalculated. Furthermore, manually corrected trace information was also not saved, so running automatic digitization again meant the manual correction got lost. Version 1.3 allowed more flexibility by saving carefully selected data that allows the user to start exactly from where they finished and to have all necessary information for future analysis.

2.1.2 DigitSeis version 1.5

The latest version of DigitSeis (v1.5) was released in September 2020 and, in addition to fixing bugs, it includes substantial improvements over version 1.3, some of which are described below (a full list of changes from one version to another is provided with DigitSeis packages).

Languages Based upon inquiries and our work with Japanese high-school students, language options have been implemented. The user can choose to run DigitSeis in English, Japanese, or Spanish. This required substantial changes throughout the DigitSeis codes, but now, a new language can be added quite simply by adding phrases in a single file.

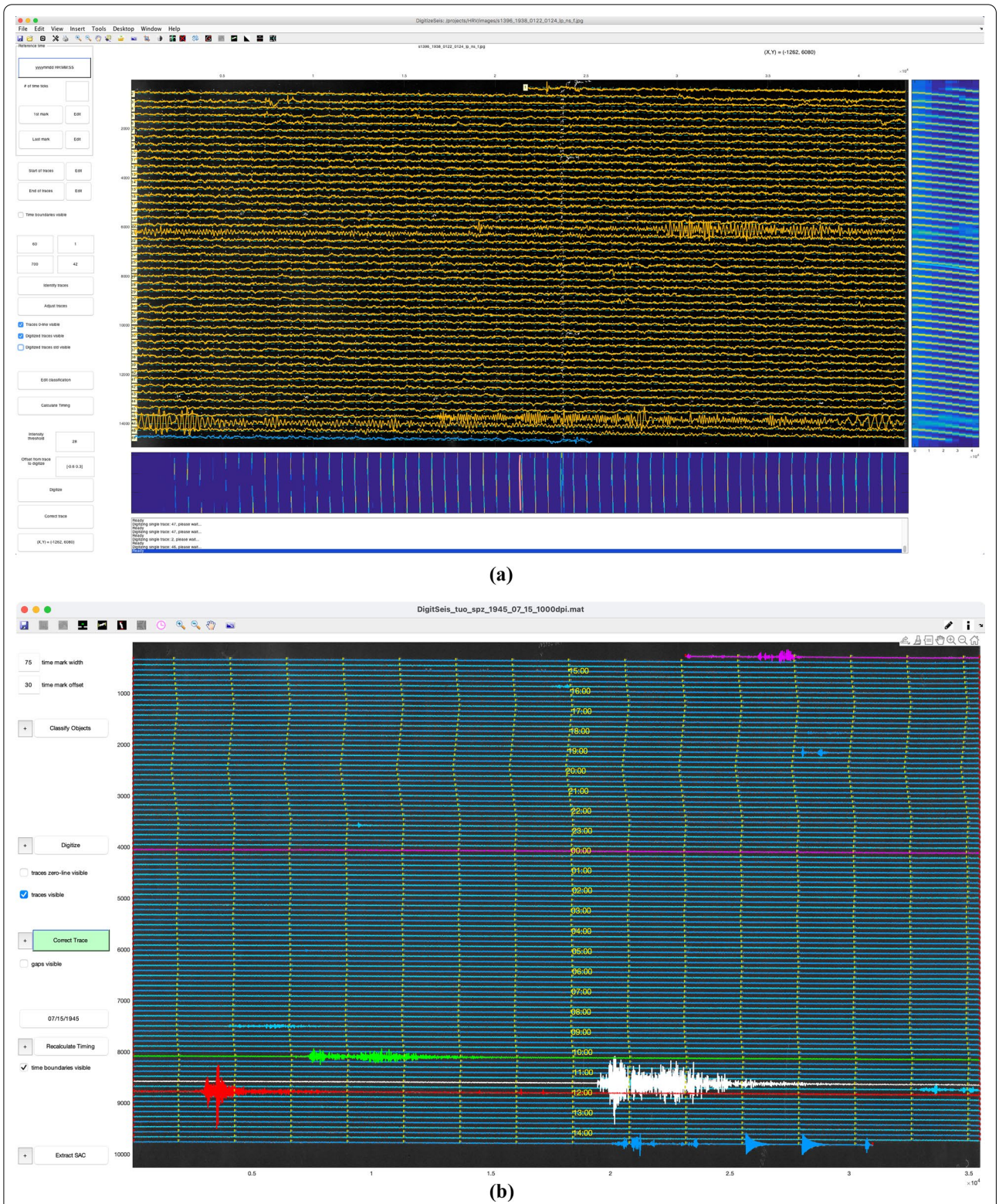

Fig. 3 Screenshots of DigitSeis. **a** Screenshot of DigitSeis v0.53. The image in the main field of the screen shows digitized (but not timed) traces of a long-period north–south recording from January 22–24, 1938. The arrivals in the middle of the window are associated with a magnitude 6.6 earthquake in Maui, Hawai'i, and the later arrivals are from a magnitude 7.0 earthquake in the Scotia Sea (ISC 2022). **b** Screenshot from DigitSeis v1.5. The image shows the original short-period vertical recording from July 15–16, 1945, that has been digitized (colored lines) and timed (yellow vertical bars with labels for every hour). This recording is from Tucson, AZ, and the large arrivals in white are those from the Trinity nuclear test

Efficiency Efficiency of object calculation was improved by implementing new algorithms for initial cleaning of the input image, and for evaluating it in portions, rather than in full. These algorithms were also sped up by parallelizing some parts of the process. Measures have been added so that classification is more automatic. For example, the user no longer needs to select an image threshold to identify objects, as this is done automatically based upon regional intensity variations. In addition, this version requires the user to measure the vertical offset of time marks, which is used in improving classification of objects. The algorithms have also been examined very carefully to reduce the number of variables stored and memory usage, leading to more stable performance (this version rarely crashes due to memory shortage with computers with at least 4 GB of RAM).

Accuracy By taking advantage of the fact that time marks should start where trace objects end (and vice versa) and the user-inputted rough vertical offset value, the accuracy of identifying time marks associated with a given trace has been significantly improved. This leads to more automatic and reliable determination of trace-zero lines as well as digitized traces. This feature is also used to distinguish between time mark objects and noise objects during classification.

Output The user now has an option to output the digitized trace information in terms of the x-y positions within the image rather than the time-amplitude information stored in SAC files. This allows more flexibility if the user wishes to set times or is concerned about the calculation of trace-zero positions or image rotation.

User interaction Since multiple reports were received of issues that turned out to be caused by users clicking on buttons or icons that should not be used at a particular time, icons and tools have been disabled until their functionalities are needed. Another problem that came up with some complicated traces is that of assigning trace or time mark objects to a specific trace. A new set of codes has been added to DigitSeis that allows the user to visualize trace assignment of objects and manually correct them if needed. Finally, the DigitSeis window has been simplified by removing buttons and options that are no longer necessary, and for functionalities that are not regularly needed, optional menus for fine-tuning the settings of the program have been created. For example, version 1.3 had a button that was clicked to calculate the trace-zero lines but version 1.5 does this calculation automatically as part of the digitization process, and hence the button has been removed.

Notes and feedback Some users suggested that a feature to add notes to the analysis would be desirable. For example, one may notice an issue with a digitized trace and want to check the classification of objects. It would be straightforward to get into the classification window and take a look at the problematic region if there is a note that propagates from the main DigitSeis window to the classification window. This has now been implemented, and the user can place a star anywhere within the image and add comment to it that can be displayed later when the star is clicked. Furthermore, a hidden functionality has been implemented where a user can place various symbols and comments as feedback to work done by a less-experienced user. For example, a teacher can point out time marks that are not correctly classified or digitized traces that need to be corrected. This feature can be activated by any user with a password. We have chosen to hide this functionality partly to reduce clutter and mostly to allow teachers to have control on the feedback items (it might become confusing if feedback items are placed by different people). However, the feedback tool can easily be made openly available if necessary.

Modular programming Significant effort has been put into separating various functions within DigitSeis so that when issues arise, corresponding codes can be more easily found as well as reused if similar calculations are needed. This was an important step in thinking ahead for restructuring and reorganization which will be needed for future improvements.

2.2 Examples of DigitSeis results

Any reliable development of software, especially those that are going to be released publicly, requires in-depth debugging. This is done by digitizing images of analogue seismograms that are available to us, and we have built a database of digitized time series, some of which are available at Harvard Seismology web pages. In this subsection, we give three examples of digitization results from three different media, photographic paper, 35-mm microfilm, and 70-mm microfiche. These examples illustrate how digitized time series provide useful insight into past events, and highlight the successful usage of DigitSeis software to extract time series.

2.2.1 1938 Off Coast of Northern Ibaraki, Japan Earthquake

A shallow-focus magnitude 7.7 earthquake struck off the coast of Northern Ibaraki prefecture in Japan on May 23, 1938, at 07:18 UTC, 16:18 local time. This event is suspected to be a precursor to an unusual swarm of magnitude 7 earthquakes, one 7.8 and two 7.7s, that occurred within a span of about a day in November of the same year (Ikeda et al. 2008). The event in May was recorded about 95 degrees away at the Harvard Seismographic Station located at Harvard, Massachusetts. The analogue seismograms on photographic paper generated by long- and short-period Benioff instruments survived years of storage, and were subsequently scanned and digitized.

We show an example of the digitization produced from the long-period east–west component recording spanning from May 22 at 13:15 to May 23 at 13:00 (Fig. 4). The analogue seismogram image has been converted to negative, i.e., dark background with whiter traces, but because the digitized lines are so successful at tracing the image, the bright pixels are difficult to see at this scale. This example also showcases the ability of DigitSeis to process crossing traces where the amplitude is so large that it goes over traces above and/or below. Finally, using the time marks, the digitized traces have been timed, i.e., the x-y positions of the lines have been converted to time series.

2.2.2 1945 Trinity nuclear test

The analogue era of seismology covers a period between 1950s and 1980s when there were frequent nuclear tests around the world. Some tests are too small or too far from a seismic station to be visible, but there are numerous tests that were captured, including the first nuclear bomb, Trinity, which was detonated on July 16, 1945, at 05:30 local time (11:30 UTC) near Los Alamos, New Mexico as part of the Manhattan Project. The short-period vertical motion of the blast was recorded in Tucson, Arizona, on photographic paper, and it was copied onto 35 mm microfilm in an effort to systematically collect, preserve, and distribute seismograms containing significant events about thirty years later (Meyers and Lee 1979). Although the original paper seismogram is still available at the National Earthquake Information Center, United States Geological Survey (USGS), we focus on digitization result from the microfilm copy (Fig. 5). Scanning the microfilm at highest possible resolution forced the seismogram image to be split in two sides, left and right. Both sides of the image are digitized using DigitSeis, and the right side is shown in this example. Even though there are only 9 time marks on this partial image of the full seismogram, the time can be correctly set with appropriate time gaps between each trace.

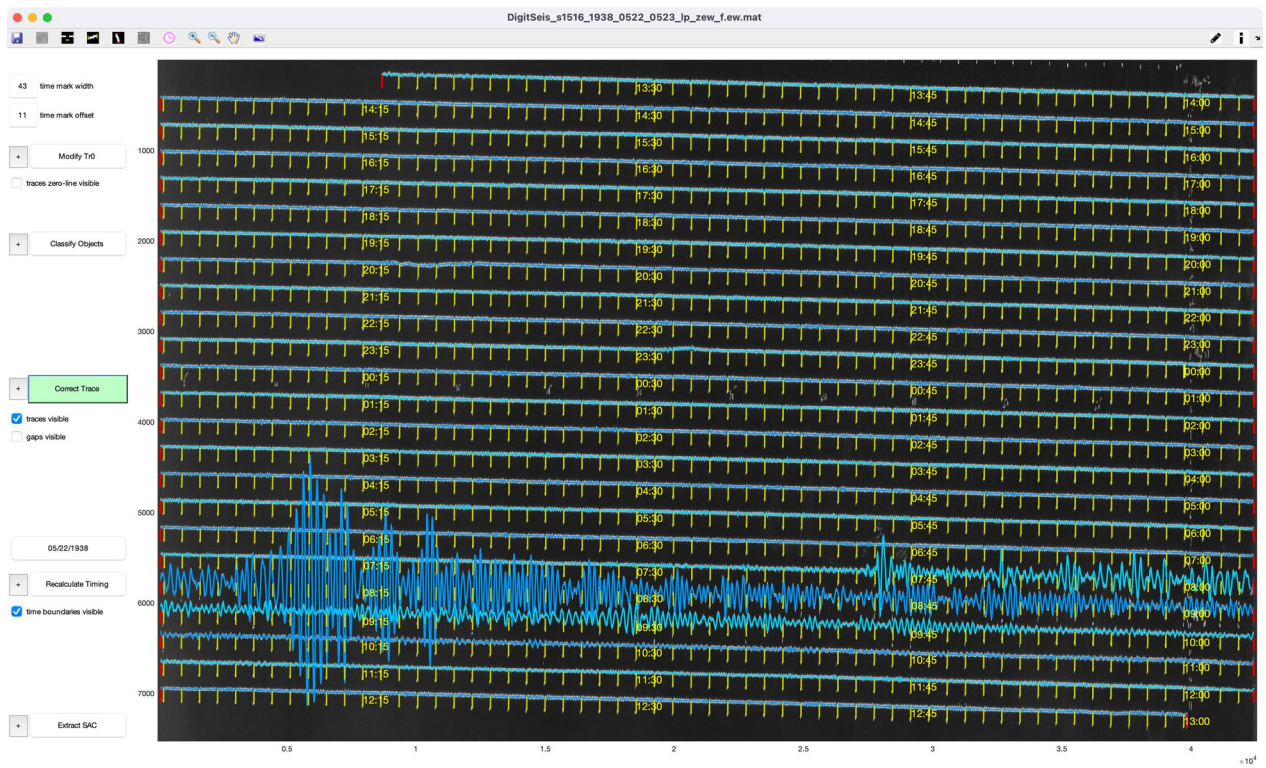

Fig. 4 Digitization Example 1: 1938 Earthquake Off Coast of Northern Ibaraki, Japan. Screenshot from DigitSeis showing the completed analysis of the May 22–23, 1938, long-period east–west seismogram recorded at the Harvard Seismographic Station. The left column contains various functionalities in the form of push-buttons and editable boxes which the user goes through from top to bottom to process the analogue seismogram digitization. Most of the screen space is taken up by the image window on the right where the digitized traces in light and dark blue are overlain on top of the analogue seismogram image. The yellow vertical bars indicate the rightmost location of each time mark, providing the position at which each minute begins, and the vertical red bars show the beginning and ending of each line. Time stamps (in UTC) are given for each 15-minute interval. Note that a beta-version of DigitSeis 1.6 is used here; hence, the layout of functionalities is not the same as the released version 1.5

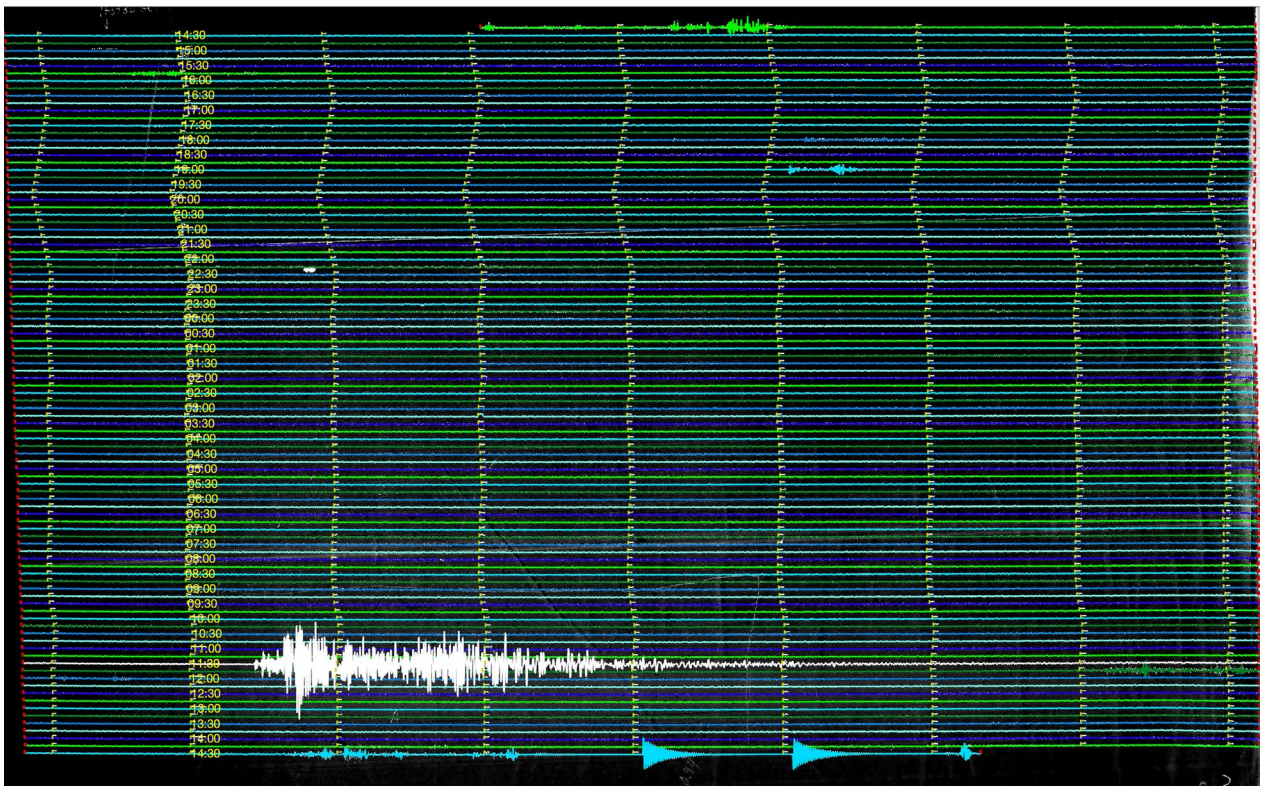

Fig. 5 Digitization Example 2: 1945 Trinity Nuclear Test. Digitization of the right side from the 35 mm microfilm copy of the short-period vertical record from Tucson, Arizona, covering the time between 14:17 on July 15 to 14:35 on July 16, 1945. The features are the same as in Fig. 4 except that only the image part of the DigitSeis screen is shown and the digitized traces are displayed with green and blue colors. There is a trace that is shown with white color to emphasize the blast signals associated with the Trinity nuclear detonation at 11:30 UTC. The time stamps are provided at 30-minute intervals. Short segments that appear regularly above the digital traces are the time marks in the background image. See Fig. 3 for the digitization result using the full photographic paper record

2.2.3 1966 Alaskan Twin earthquakes

The expansion of nuclear testing motivated the development of the World-Wide Standardized Seismograph Network to monitor and quantify nuclear tests. More than 120 stations distributed around the globe generated a vast amount of analogue seismograms, which were copied onto 70-mm microfiche for preservation and distribution (Peterson and Hutt 2014). They include nuclear blast signals, but also abundant earthquake signals. As our final example, we take the microfiche image of the long-period north–south seismogram recorded at station COL (College, Alaska) from December 19, 1966 (the time window covered by this record included expected arrivals from the nuclear test codenamed Greeley, a part of Operation Latchkey). From the image itself, we can see that there are two earthquake events that are offset by about half an hour (for this seismogram, each line of trace covers about half an hour), but the details of the arrivals are difficult to discern due to significant overlaps of the traces as well as faded lines for large-amplitude arrivals (Fig. 6a). Once the digitized traces are superimposed on the image (Fig. 6b), the striking similarities of the arrivals from the two events become evident. The waveforms can be matched almost swing by swing, and their amplitudes are nearly identical. They correspond to local twin earthquakes in Alaska that occurred about 200 km from station COL. The International Seismological Centre earthquake catalogue (International Seismological Centre 2022) lists them as magnitude 5.0 earthquakes that occurred on December 20, 1966, at 00:26:26 and 00:57:52 at 66.73°N and 148.52°W and 66.74°N and 148.34°W, respectively. The availability of digitized time series allows application of modern waveform analysis techniques, such as calculating correlation of the two waveforms, that can be used to detect earthquakes through template matching (e.g., Shelly et al. 2007).

2.3 Future directions for DigitSeis

As demonstrated in the previous subsection, the medium from which the seismogram image is taken

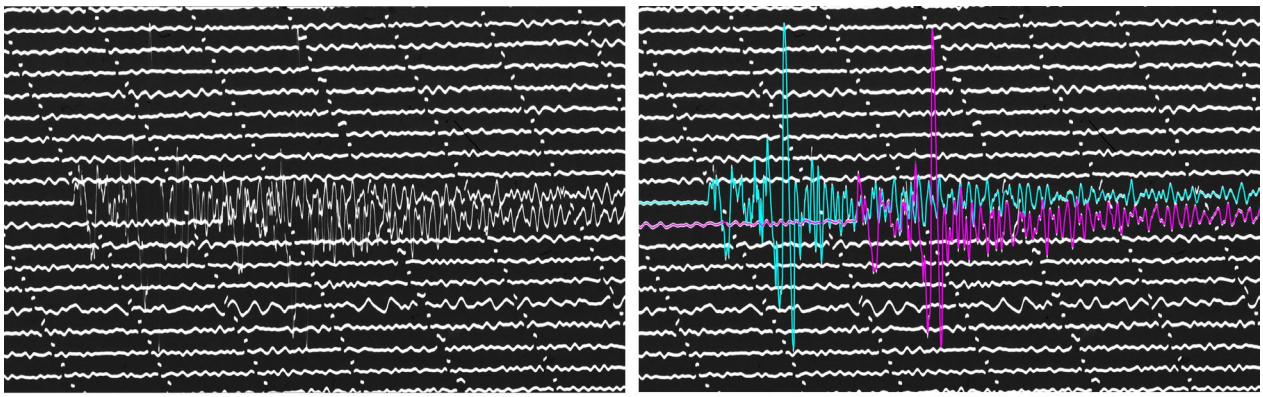

Fig. 6 Digitization Example 3: 1966 Alaskan Twin Earthquakes. A section of long-period north–south recording from COL (College, Alaska) station on December 19, 1966. **a** Scanned image of 70 mm microfiche as loaded into DigitSeis. **b** Same image as (**a**) but with digitized traces for lines 12 (cyan) and 13 (magenta) superimposed on the image

(i.e., photographic paper, microfilm, or microfiche) does not affect how the digitization process using DigitSeis is performed. However, how the ground motion was recorded on a seismogram does matter. DigitSeis has been designed with galvanometric seismogram examples (Harvard analogue seismogram collection), and it is not recommended that the software be used with pen- or Develocorder-type seismograms. Since there is a vast amount of the latter two types of recordings, future efforts to improve DigitSeis will include modifications to accommodate these data. In this section, we describe some of the major changes to DigitSeis we will be working on in the next few years.

2.3.1 Pen-type seismograms

The "pen-type" seismograms are defined as those that contain waveform distortion due to curvature introduced when a fixed-length component of the seismograph (e.g., a pen) attached to a pivot point records large-amplitude signals (Fig. 7). Some of the commonly seen examples of these seismograms are needle scratches on smoked paper, conventional ink on paper, and heated stylus and thermal paper (e.g., World Data Center 1979). The earliest of the seismogram recordings are in pen-type format, and this system continued to be used until digital recording systems became available. Consequently, a significant fraction of the analogue seismogram collections around the world are expected to be in this format. For example, the collection of nearly a million analogue seismograms from the USGS Hawaiian Volcano Observatory between 1912 and 2013 is roughly made up of about 45% on photographic paper (i.e., galvanometer system) and 55% on either smoked-drum paper (~30%) or thermal paper (~20%) (Paul Okubo, personal communications).

The feature that distinguishes the pen-type seismograms from galvanometer seismograms is the additional waveform curvature introduced by the use of pen or needles (e.g., Inoue and Matsumoto 1988). Because the pen or needle is fixed to a point and is allowed to swing to capture large-amplitude arrivals, it inevitably introduces arcing motion when it moves off from the zero-amplitude position (Fig. 7). This implies that for large-amplitude arrivals, there may be multiple y (amplitude) values for a given x position on the image. The current DigitSeis algorithm cannot handle such situations. It assumes that the amplitude information (y) is a function of horizontal position (x), i.e., there is only one y value for a given x position, and thus, it determines the y value based upon pixel positions weighted by their intensity.

Correcting for the pen curvature is more complicated than it may appear at first glance. Most analogue seismogram recordings are made using a helicorder, a drum that rotates with time to allow a long duration of observation. This drum also translates parallel to the rotation axis, resulting in multiple lines of traces being recorded. If the helicorder is only rotating, then the trace-zero line is horizontal and the time (x-axis) and amplitude (y-axis) are orthogonal on the paper (Fig. 9a). However, when translation of the helicorder is added, the time and amplitude directions are no longer orthogonal, with amplitude instantaneously measured in the vertical direction, but time axis having both horizontal and vertical components (Fig. 9b). This implies that the waveform is distorted, and assuming the amplitude to be orthogonal to the time axis gives incorrect amplitude and time information. DigitSeis takes this non-orthogonal geometry into account and measures amplitude in the y direction while determining x or time information along the trace-zero line. However,

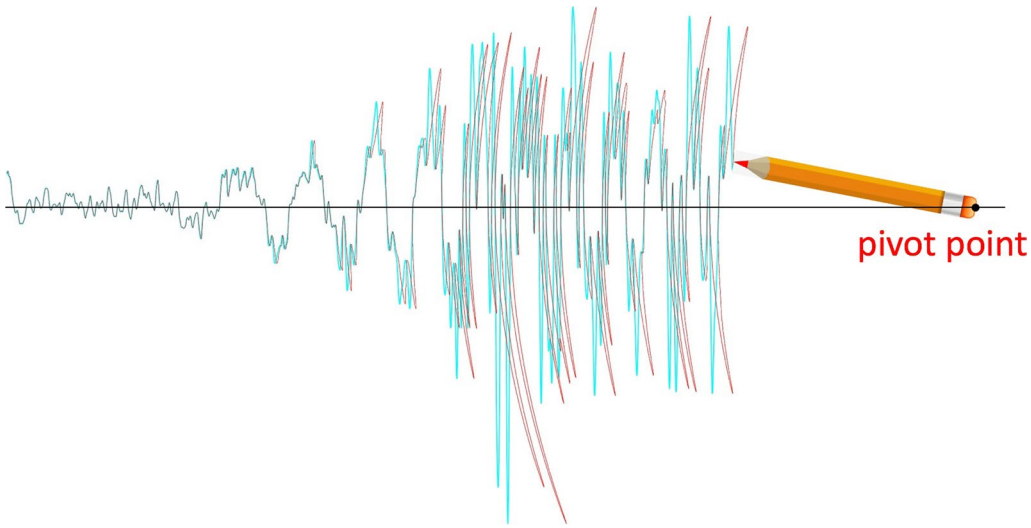

Fig. 7 Curvature introduced by pen-type recording system. Comparison of a raw seismogram showing the beginning of surface-wave arrivals at station HRV from the November 30, 2018, Anchorage earthquake (cyan line), and a version with curvature as would be generated by a pen-type system introduced (red line). The pen (pencil in the figure) swings around a pivot point, making arcs for large-amplitude arrivals, even though for small-amplitude arrivals, the cyan and red lines are practically identical

for images with pen curvature, this geometry significantly complicates necessary curvature correction (Fig. 9c). The conventional assumption that curvature can be corrected by modifying the image parallel to the trace-zero line results in overestimated amplitude. The appropriate correction must be parallel to the horizontal direction (or parallel to the rotation direction of the helicorder).

A future version of DigitSeis will address these issues as well as the need to estimate the pen length (an underdetermined problem). Contrary to the conventional approach, image modification will not be used, and this is for three reasons. First, it requires knowledge of the pen length. Not all records have this metadata information, and even though our algorithm for pen-length calculation appears to work well, it would be prudent not to rely upon this estimate. Second, the images are pixelated. Modification of the image, therefore, requires rounding to the nearest integer, leading to possible step-like output. If the image has been scanned at high resolution, this may not be a problem, but we have worked with low-resolution images which give jagged output even without image modification. Alternatively, one could use an average of pixels if the pixel translation is fraction of an integer. This provides smooth image, but does so at the expense of blurring some faint traces that often exist for traditionally interesting targets such as large-amplitude arrivals associated with significant earthquakes. For these reasons, the approach DigitSeis will take is to digitize the traces without any corrections and allow user to correct for curvature once x-y positions of the traces are extracted (a basic code to do this correction will be provided).

2.3.2 Develocorder-Type Seismograms

The "Develocorder-style" seismograms are defined as synchronized seismograms from multiple components or stations recorded on a single medium (Fig. 8). Another feature of these seismograms is that one or more of the traces are dedicated to timing, i.e., they do not record ground motion, but provide time information. Develocorder-type recordings became popular in the 1960s with development of the ability to accurately telemeter data from distributed stations back to a central location. The layout of the seismograms, i.e., multiple traces from different stations displayed at once, mean that they are quite powerful for quick visual determination of where an earthquake has occurred. Changes in the ambient "background" level are also easy to detect, for example, increases in volcanic tremors from a specific location. These features made Develocorder recordings popular, especially among institutions where monitoring was among their main duties (e.g., Okubo et al. 2014).

The most significant difference between galvanometer- and Develocorder-style recordings is that the former has time information embedded within the traces in the form of time marks while the latter has a separate time channel. This implies that the current time calculation algorithm in DigitSeis will not work with Develocorder-style records. Furthermore, in a galvanometer-type helicorder recording, the end of one

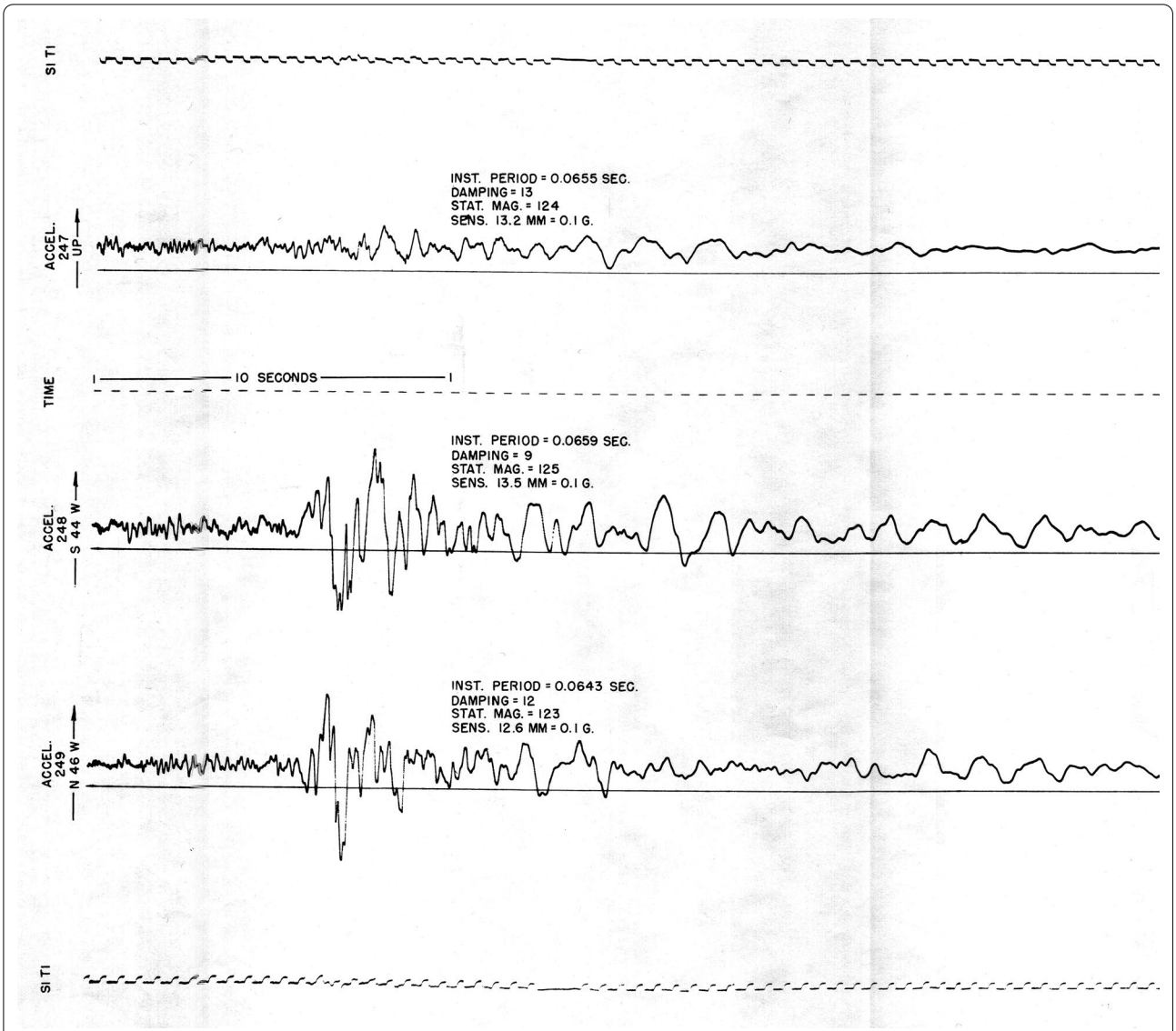

Fig. 8 An example of a Develocorder-type recording. Three-component strong-motion recordings of the Ferndale, CA earthquake, on December 21, 1954. There are three time channels, at the top, bottom, and between top and middle traces. The middle time channel indicates that a pair of line and gap covers 0.5 second

trace is connected to the beginning of the next trace. Consequently, DigitSeis assigns time to digitized traces so that it is continuous from the end of one trace to the start of the next, and when data are extracted to SAC format, filenames are assigned based upon the start and end times of each line. On the other hand, the multiple traces of a Develocorder-type record were set up so that all traces would have same time at a given horizontal position, i.e., using the current version of DigitSeis would result in incorrect time assignments, and if the user manually sets each trace to have the same start and end times, the SAC files will overwrite one another. Modifications to DigitSeis to accommodate Develocorder-style seismograms are relatively simple compared to challenges associated with pen-style seismograms, as they will mainly consist of changing the waveform-output routines and reading in a unified time (as opposed to any modification of the digitization itself).

2.3.3 Crossing Traces and Automation

The ultimate digitization software is one that is all automatic, i.e., the user loads in the seismogram image and

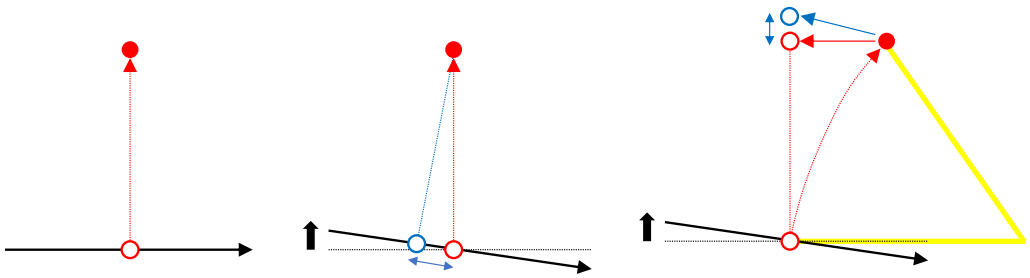

Fig. 9 Complexities associated with the helicorder geometry. **a** Illustration of a helicorder recording with only rotation. The black line is the trace-zero line that would be produced if the amplitude is always zero. **b** Same as (**a**) except that this case includes translation that results in the slope of the trace-zero line. If the amplitude is measured perpendicular to the trace-zero line, it will result in an underestimation of amplitude (dashed blue line) and a time offset (open blue circle). **c** Effect on curvature introduced by pen or needle recording. The arc is generated by instantaneous amplitude change from zero to a maximum value with the armature (yellow lines). The solid red arrow shows how this curvature should be corrected (open red circle) while the blue arrow shows correction if the correction is calculated parallel to the trace-zero line leading to overestimated amplitude (open blue circle)

corresponding time series are generated without any user input. Modifications to DigitSeis mostly aim to increase automation of processing, and we will continue to automate more components as they become feasible, although it (nor any other highly accurate digitization software) is not likely to ever achieve full automation. The greatest challenge toward automation is crossing or touching traces which cannot simply be ignored as they may be the most desirable part of the record with large swings associated with seismic wave arrivals. The current version of DigitSeis can process crossing or touching traces (e.g., Fig. 6), but this is achieved through user intervention (either tracing the lines or defining objects manually). For images with no such features, DigitSeis is near-automatic, requiring only limited user interaction to move through each step and to set the reference time. It is not clear if there will ever be a reliable, fully automatic algorithm for untangling crossing and touching traces since it is sometimes difficult, if not impossible, even for human eyes to decipher how the lines are connected (e.g., Fig. 10).

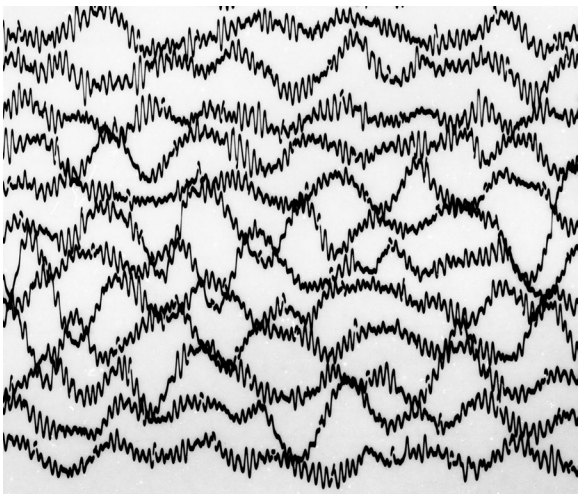

Fig. 10 An example of crossing/touching traces. A portion of the long-period north–south recording from BKS (Berkeley, CA) on February 11, 1964. One can take advantage of the time marks that are horizontally offset for each trace to unravel most parts of this image, but if time marks were occurring at the same horizontal position, one would have a hard time separating the traces

One suggestion we often receive is to utilize machine learning in the digitization process. There have indeed been studies that successfully utilized machine learning on analogue data to extract information (e.g., Wang et al. 2019). We have investigated the applicability of convolutional neural networks for recognizing and classifying trace, time mark, and noise objects from an input grayscale image. It worked well for simple images without trace crossing, but was less than successful with more complicated images. There are two main reasons for the unsuccessful result. One is that the seismograms are sparse in terms of local features. If one tries to identify a cat in an image, the object, i.e., a cat, can be defined within a small area of the image. Even if the cat is taking up the entire image, the relative information (e.g., having a head, tail, body, etc.) can be localized and distinguished. Most importantly, a zoomed-in image of a cat can be decimated down to a smaller image, i.e., we can still see the cat. Seismograms, on the other hand, are difficult to define locally. A small portion of the seismogram (basically a line) is so featureless that it is difficult to distinguish a trace or time mark from a non-seismogram segment in the image, e.g., a handwritten tilde. What allows us to identify the traces in a seismogram is their continuity and relative positions. One can train a neural

network to look for the relative information, but because we typically down-sample images when looking at a large area, faint traces tend to disappear. Also, depending upon the portion of the image being examined, the non-local aspect of seismograms requires neural networks to be fine-tuned to some specific layout of the seismogram image, losing flexibility of processing in doing so. The second issue is that we do not have a large-enough training database of digitized analyses from images with severely crossing traces. These images are difficult to digitize (i.e., need substantial manual intervention and contemplation); hence, traces such as those shown in Fig. 10 have not been digitized. This problem could partially be mitigated by creating a training database consisting of digital seismograms and corresponding synthetic helicorder or Develocorder images generated using the digital traces.

There are a number of approaches one could take to address some of the crossing/touching trace issues without machine learning. One can distinguish certain features associated with crossing traces by identifying pixels that belong to multiple objects. For example, they tend to occur around bifurcations or holes and often are thicker than a typical trace or time mark objects (i.e., cover more pixels in the y direction). Preliminary attempts show promising results (Fig. 11a). For relatively simple overlaps, this method successfully isolates overlapping areas (areas 3 through 8 in Fig. 11a). It is, however, not perfect, and for more complicated crossings, the algorithm returns vertically sliced areas as containing overlapping areas (areas 1 and 2 in Fig. 11a). This information is still useful since it ensures that resulting objects (trace and time mark segments) are isolated, i.e., belongs only to a single trace.

Accurate definition of overlap pixels is a crucial step in automating the analysis of crossing traces, but another important process is determining how these pixels should be combined with the rest of trace or time mark objects. For example, in Fig. 11b, there are 8 objects, of which 6 are individual trace segments (A through F), while 2 belong to multiple traces (a and b). Because the seismogram should always move to the right and never go back left and vice versa (this is not necessarily the case for seismograms generated using pen or needles), objects A and B must belong to separate traces and the same can be said with pairs C/D and E/F. Assuming that there are only two traces involved, one could then combine the objects with the following possibilities (top trace/bottom trace): A-a-C-b-E/B-a-D-b-F, A-a-C-b-F/B-a-D-b-E, A-a-D-b-E/B-a-C-b-F, and A-a-D-b-F/B-a-C-b-E. In this particular example, the correct combination is A-a-D-b-E/B-a-C-b-F. In order to automatically deduce the trace assignment, it becomes vital that the number of traces within the image and their trace-zero positions be known as the objects are processed. We will be exploring these approaches and implement successful algorithms into future versions of DigitSeis so that it becomes easier to process crossing/touching traces.

3 Conclusions

Seismologists have been recording ground motion since the late 1800s, and a vast amount of data covering nearly a century exists in analogue form. These data capture some unusual events that are not observed with their digital counterparts (e.g., tsunamigenic earthquakes, submarine or subaerial nuclear tests, rare volcanic eruptions), and are also essential in understanding time-dependent processes (e.g., earthquake cycle, subsurface evolution,

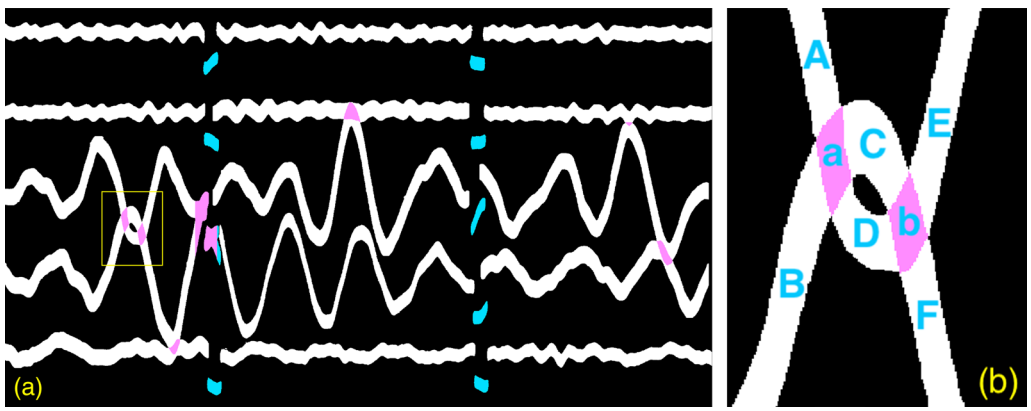

Fig. 11 Object crossing. **a** Screenshot of the classification window where trace objects are shown in white and time marks are in blue. Areas colored in magenta correspond to overlap pixels that have been automatically determined to belong to multiple objects. **b** Different objects that need to be connected within the area shown by yellow box in (**a**). The magenta regions a and b separate white regions to define independent objects A through F; in total, there are 8 objects

climate change). Conversion of these analogue data into digital time series will facilitate their analysis using modern computational techniques.

This manuscript describes freely available software, DigitSeis, that has been developed to extract digital time series from an input seismogram image, and how the software evolved over different versions. The efficacy of the software is demonstrated with three examples of images based upon three types of media, photographic paper, 35-mm microfilm, and 70-mm microfiche. These examples show that seismograms can be effectively traced even when there are crossings, and time can be correctly set using the time marks. They also indicate the potential of extracting digital time series for use in a variety of seismological research.

The development of DigitSeis is on-going, and there are multiple directions in which it will be enhanced for future releases. The current versions are not designed to process pen- and Develocorder-type seismograms, and some of the challenges associated are explained. The Develocorder issue is relatively easy to address and the next release of DigitSeis will likely include the capability to work with Develocorder records. It is more difficult to process pen-type seismograms, and it may take a while before these images can be analyzed using DigitSeis. One potential approach to digitization that is often brought up is the applicability of machine learning algorithms. While image processing using machine learning is useful for some very specific projects with well-defined features such as identifying earthquakes based upon Develocorder records, we find that there are considerable challenges facing the more general goal of tracing any type of recording. Therefore, near-term improvements to DigitSeis will be based upon traditional programming, such as unravelling of simple crossing traces based upon identification of overlapping pixels. DigitSeis will continue to evolve toward accurate, robust, and automatic processing of analogue seismograms which will be a key to building digital seismogram databases derived from the analogue instrumentation era, opening up opportunities for exciting scientific research.

Abbreviations

BKS: Berkeley, California seismic station; COL: College, ALaska seismic station; dpi: Dots-per inch; ISC: International Seismological Centre; SAC: Seismic analysis code; USGS: United States geological survey; UTC: Coordinated Universal time.

Acknowledgements

We thank Thomas Lee for feedback on the manuscript, packaging DigitSeis for public release and being its ambassador, and Katelyn Lee for preparing material for DigitSeis version 1.5 release. We also thank an anonymous reviewer and Göran Ekström for comments that improved the manuscript and Toshiro Tanimoto for handling of the manuscript.

Author contributions

MI began the work with analogue seismograms and developed DigitSeis versions 1.1 and later. HI used the software to identify bugs and to process images to obtain digital time series. Both authors have read and approved the final manuscript.

Funding

This work has been supported by the Miller Institute for Basic Science at the University of California, Berkeley, the Earthquake Hazard Program of the United States Geological Survey (grants G14AP00016, G16AP00021, G17AP00007, G18AP00018, and G20AS00042), the Department of State (contract number 19AQMM20P1475), and the National Science Foundation (award number 1822136).

Availability of data and materials

The digitization software, DigitSeis, is available at the Harvard Seismology Web page at http://seismology.harvard.edu/research/DigitSeis.html. The digitized results in Figures 4 through 6 can be obtained in both SAC and DigitSeis output file format (.mat) using links from http://seismology.harvard.edu/research/analogueData.html

Declarations

Competing interests

The authors declare that they have no competing interest.

Received: 7 July 2022 Accepted: 21 September 2022
Published online: 01 October 2022

References

Bogiatzis P, Ishii M (2016) DigitSeis: a new digitization software for analog seismograms. Seism. Res. Lett. 87(3):726–736. https://doi.org/10.1785/0220150246

Bromirski PD, Chuang S (2003) SeisDig: software to digitize scanned analog seismogram images, user's manual. UC San Diego, Scripps Institution of Oceanography Technical Report, Scripps Institution of Oceanography, p 28

Ikeda T, Kato K, Uetake T, Tsuruga T (2008) 1938 年塩屋崎沖地震群の震源モデルの特性化と地震動評価 (in Japanese). J Struct Constr Eng (in Japanese), 73(633): 1951–1958. https://doi.org/10.3130/aijs.73.1951.

Inoue R, Matsumoto T (1988) Digitization and processing of the J.M.A. strong motion records in the period of 2 to 20 sec from nine great earthquakes. In: Lee WHK, Meyers H, Shimazaki K (eds) Historical seismograms and earthquakes of the world, 390–400, Academic Press, Inc., San Diego

Ishii M, Ishii H, Bernier B, Bulat E (2015) Efforts to recover and digitize analog seismograms from Harvard-Adam Dziewoński Observatory. Seism Res Lett 86(1):255–261

International Seismological Centre (2022) On-line Bulletin. http://www.isc.ac.uk, Internatl. Seismol. Cent., Thatcham, United Kingdom

Kanamori H, Cipar JJ (1974) Focal process of the great Chilean earthquake May 22, 1960. Phys Earth Planet Inter 9:128–136

Leet LD (1934) New recording vault of the Harvard seismograph station. Bull Seism Soc Am 24:47–50

Meyers H, Lee WHK, (eds) (1979) Historical seismogram filming project: first progress report. World Data Center A for Solid Earth Geophysics, Report SE–22

NeuraLog (2013) NeuraLog: automated well log digitizing software. http://www.neuralog.com/pages/Neura.html (last visited May, 2013)

Okal EA, Stein S (1987) The 1942 Southwest Indian Ocean ridge earthquake: largest ever recorded on an oceanic transform. Geophys Res Lett 14:147–150

Okubo PG, Nakata JS, Koyanagi RY (2014) The evolution of seismic monitoring systems at the Hawaiian Volcano Observatory. In: Poland MP, Takahashi TJ, Landowski CM (eds) Characteristics of Hawaiian volcanoes. U.S. Geological Survey Professional Paper 1801, 67–94

Peterson J, Hutt CR (2014) World-wide standardized seismograph network: a data users guide. U.S. Geological Survey Open-File Report 2014-1218, 74 p., http://dx.doi.org/10.3133/ofr20141218

Pintore S, Quintiliani M, Franceschi D (2005) Teseo: a vectoriser of historical seismograms. Comput Geosci 31:1277–1285
Shapiro NM, Campillo M, Stehly L, Ritzwoller MH (2005) High-resolution surface-wave tomography from ambient seismic noise. Science 307:1615–1618
Shelly DR, Beroza GC, Ide S (2007) Non-volcanic tremor and low-frequency earthquake swarms. Nature 446:305–307
Song X, Richards PG (1994) Seismological evidence for differential rotation of the Earth's inner core. Nature 382:222–224
Wang K, Ellsworth WL, Beroza GC, Williams G, Zhang M, Schroeder D, Rubinstein J (2019) Seismology with dark data: image-based processing of analaog records using machine learning for the Rangely earthquake control experiment. Seis Res Lett 90:553–562
World Data Center A (1979) Manual of Seismological Observatory Practice. In: Willmore PL (ed) World Data Center A for Solid Earth Physics. U.S. Department of Commerce, Boulder, CO