# ArchaeoSRP

## An R Package for Extracting and Synthesizing Federal Cultural Resources Data for Research and Management

*Sean Bergin and Grant Snitker*


## ABSTRACT

For much of its history, archaeological research has relied on site-specific projects, regional comparisons, and theory building from case studies. However, recent research themes concerning the emergence of complex social-ecological systems and long-term land-use legacies require a new approach to archaeological data. Large-scale syntheses of archaeological data provide an effective way forward to address these new research themes. In more concise terms, "big questions" require "big data" to help answer them. The archaeological information collected by the USDA Forest Service is one such "big dataset" and represents an incalculable investment in time, resources, and expertise. This article explores this concept and presents an R package (ArchaeoSRP) designed to extract archaeological information from USDA Forest Service site record files. We demonstrate the functionality of this R package through a case study examining the archaeological data for the Cle Elum Ranger District, within Central Washington's Okanogan-Wenatchee National Forest. Our results reveal the efficiency of using automated methods to extract, organize, and synthesize district-level archaeological data, which, in turn, reveal patterns of precontact and historic land use that were otherwise not distinguishable.


**Keywords:** archaeological synthesis, R, cultural resource management, USDA Forest Service


Durante un gran parte de su historia, investigaciónes arqueológicas se basaban en proyectos del sitios particulares, comparaciones regionales y el desarrollar de teorías desde estudios de casos. Sin embargo, los temas de investigaciónes recientes relacionados con la emergencia de sistemas socio-ecológicos complejos y de la utilización de la tierra a largo plazo necesitan una nueva metodología para los datos arqueológicos. Las síntesis de datos arqueológicos de escalas amplias proporcionan una manera eficaz para encargarse de tales nuevos temas de investigación. Es decir más concisos, los "grandes cuestiones" demanden "grandes datos" para contestarlas. La información arqueológica recolectada por el Servicio Forestal del USDA es un ejemplo de tales "grandes conjuntos de datos" y representa una inversión incalculable del tiempo, recursos y experiencia. Este manuscrito explora este mismo concepto y presenta un paquete en el programa de R (ArchaeoSRP) diseñado para extraer información arqueológica desde los archivos de sitios del Servicio Forestal del USDA. Demostramos la funcionalidad de este programaa través de un estudio de caso que examina los datos arqueológicos del Cle Elum Ranger District, dentro del Okanogan-Wenatchee National Forest del centro del Estado de Washington, EEUU. Nuestros resultados demonstrar la efectividad del uso de métodos automatizados para extraer, organizar y sintetizar datos arqueológicos al nivel de distrito, lo que igualmente dar a conocer patrones de la utilización de la tierra prehistóricas e históricas que no eran identificable de otra manera.


**Palabras clave:** síntesis arqueológica, R, gestión de recursos culturales, Servicio Forestal del USDA

Recent research themes concerning the emergence of complex social-ecological systems (Barton et al. 2004; Silva et al. 2022), long-term land-use legacies (Stephens et al. 2019), and the development of the Anthropocene (Ellis et al. 2021) require new approaches to collecting and interpreting archaeological datasets. Large-scale syntheses of archaeological and paleoenvironmental information provide one possible way forward in addressing new research themes, given that they provide new means of assessing emergent patterns of social and ecological change over large spatial and temporal scales that may not be visible from individual sites alone (Ellis et al. 2021; Stephens et al. 2019). In more concise terms, the "big questions" in archaeological research often require "big data" to answer them (McCoy 2017). Operationalizing big data has the potential to inform theory and interpretations to address macroscale archaeological questions, landscape-level management of archaeological and other cultural resources, and strategies for preservation (Altschul 2016; Doelle et al. 2016; Huggett 2020; Perreault 2019; Wilshusen et al. 2016). One of the most prevailing challenges in archaeological synthesis is access to large datasets that are curated, comparable, and spatially/temporally explicit. The

*Advances in Archaeological Practice* 11(4), 2023, pp. 402–412
Copyright © The Author(s), 2023. Published by Cambridge University Press on behalf of Society for American Archaeology. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.
DOI:10.1017/aap.2023.22



https://doi.org/10.1017/aap.2023.22 Published online by Cambridge University Press

nature of cultural resource management (CRM) and other archaeological compliance data collection, which account for most archaeological data generated in the United States today, further complicates this challenge because the scope, methods, and reporting for CRM are highly variable and dependent on the managing agency's protocols and requirements. This can be particularly true of federal agencies, such as the USDA Forest Service, National Park Service, Bureau of Land Management, and others.

The USDA Forest Service (USFS) currently manages databases of archaeological material for the National Forest System, which encompasses over 93 million ha (230 million acres) of forests, grasslands, wilderness areas, and research units throughout the United States. These data include spatial information and detailed descriptions of sites, isolates, and cultural landscapes from the late Pleistocene to the mid-twentieth century. The research potential for these data is astounding, particularly if they can be efficiently synthesized through space and time. Such an endeavor would not only strengthen this dataset's role in management but also reposition cultural resources as a valuable tool in creating policy and restoration efforts with heritage resources and archaeology at their center (Foster et al. 2003; Helmer et al. 2020). The potential of these archaeological datasets is recognized by the research and management communities (Schlanger et al. 2015), and efforts to integrate or synthesize CRM databases and other large archaeological datasets are currently underway (e.g., Halford and Ables 2023; Ortman and Altschul 2023).

More than a century of archaeological research continues to shape how we interpret the past through ever-increasing catalogs of the data we collect and refinement of new methods and techniques we develop. This "deluge of archaeological data" has the potential to both answer new questions and overwhelm the analysts asking them (Bevan 2015). The term "big data" refers to large-scale datasets that can be used to answer questions that are often unanswerable with smaller-scale data. In archaeology, big data can be described in a multitude of dimensions, ranging from global-scale spatial datasets derived from remote sensing (VanValkenburgh and Dufton 2020), to regional-scale reconstructions of social networks (Mills et al. 2013), and everything in between. In our view, regardless of the dataset's scope or reach, its curation and potential for synthesis remains a vital component of the "big data" movement within archaeology (Altschul 2016).

Archaeological research from federal repositories, state site files, and State Historic Preservation Offices (SHPOs) constitute incredibly valuable "big" datasets. However, many of these databases still exist as paper forms or digital copies of paper forms, which limits their potential for data extraction, analysis, and comparison. Making these datasets more accessible for management and research would require a substantial investment in digitizing and making text from paper forms machine readable to enable researchers to extract and synthesize information from these sources (Fletcher 2023). These efforts are currently underway in many federal repositories (e.g., USDA Forest Service Heritage Natural Resource Manager [NRM] Database) and through nonprofit and university-affiliated repositories (e.g., tDAR [McManamon et al. 2017], Open Context [Kansa et al. 2020], Archaeology Data Service [Wright and Richards 2018]); however, they are primarily focused on data management or preservation rather than dataset synthesis or analysis. Keyword extraction, site record indexing, and other means to make archaeological data standardized, searchable, and accessible in these large bodies of gray literature remain elusive goals.

Federal archaeological repositories, such as the USDA Forest Service (USFS) NRM database, are also subject to unique constraints that limit sharing the location and content of many of the archaeological sites they manage. Unlike other "big data" initiatives in the natural sciences to make data more accessible and more transparent (e.g., Neotoma Paleoecology Database [Williams et al. 2018] and the Global Paleofire Database [Aleman et al. 2018]), federal laws and programmatic agreements with Tribal Nations mean diligently protecting sensitive archaeological information yet also ensuring that federally funded projects generate knowledge that is available and useable for the benefit of the public. Additionally, archaeological sites managed within the National Forest System often do not have a single site form associated with them; instead, they have a series of site updates and monitoring reports that build on an initial inventory and investigation. In some instances, an archaeological site originally recorded in the 1980s has been revised and its record updated on numerous occasions over the last 40 years. The subsequent site record is more akin to a living document rather than a definitive report, given that new site boundaries, artifact concentrations, and interpretations may have been added to the record during the duration of its management life. These challenges, coupled with the prohibitive amount of time and resources needed to digitize and index archaeological site records manually, highlight the need for new tools that are designed to extract archaeological data from paper site-record and update forms quickly and efficiently.

In response to these challenges, we present and outline the functionality of the Archaeological Site Record Processor (ArchaeoSRP), an R package of functions for reading, parsing, and extracting pertinent site information from USFS archaeological site records to create rich datasets for large-scale research syntheses or management. The package was developed using (1) R (ver. 4.2.0), a free and open-source language and environment for statistical computing (R Core Team 2023); and (2) RStudio (ver. 9), a freely available interface for coding and visualization in R (Posit Team 2023). R can be used to conduct a wide range of statistical analyses useful for archaeologists (see Marwick 2023) and excels in data visualization while lending itself well to generating reproducible research (Marwick 2019). The R language is also highly extensible through downloadable "packages," which are created and managed by members of the R community. Archaeologists have contributed numerous packages on topics such as dating (e.g., Dosseto and Marwick 2022), artifact analysis (e.g., Carlson and Roth 2021), spatial analysis (e.g., Plutniak 2022), and visualization (e.g., Steinmann and Weissova 2021). The ArchaeoSRP package is freely available and updated via GitHub (Bergin and Snitker 2023a), and it can be installed via R using the *remotes* package (example: remotes::install_github("seanbergin/archaeosrp")).

# METHODS

## ArchaeoSRP Package Description

Simply scanning paper documents to create digital copies does little to make the information accessible and usable or to increase



https://doi.org/10.1017/aap.2023.22 Published online by Cambridge University Press

its longevity, particularly when they are rendered as non-machine-readable formats (Clarke 2015). Although digitizing archaeological documents and other information is a commendable first step in data preservation, we, and others, envision digital data as a part of larger goals for creating opportunities for greater access, replicability, and stewardship in the archaeological community (Marwick et al. 2017). New tools are necessary to accomplish these objectives, especially with datasets from public lands.

Given the issues and goals at hand, ArchaeoSRP systematically parses archaeological site records and saves specific information to a tabular form for future management or analyses. This type of package may be designated an extract, transform, and load tool. Once the dataset is recorded in a standardized format, it becomes possible to work within it to either search for specific keywords or bring the data to bear on specific research questions.

The ArchaeoSRP package is designed to extract information from documents that use standardized recording forms. Such forms have long been used by state site files, museums, historic preservation authorities, and federal archaeological repositories to ensure that the proper information is recorded and to enable the comparison of information from different sites or surveys. Obviously, a single or universal form is not used by all research organizations or managing agencies. Therefore, the first task for the ArchaeoSRP package is to identify a given form type so that the type and location of information recorded on a form can be accurately identified.

## How to Download and Use the ArchaeoSRP Package

The ArchaeoSRP package was created with the intention of providing it to the wider archaeological community, including land managers, practitioners, and researchers. The package is available for use or modification from GitHub, and information on how to use or modify this package is accessible from documentation and a vignette included there (Bergin and Snitker 2023a). The package is fully customizable, which allows other archaeologists to alter this package to record different information from scanned documents or identify new types of site forms for their project-specific needs. The package can be installed to R directly from the GitHub repository as described in the package's "Read Me" and illustrated in a CRAN-style vignette (Bergin and Snitker 2023a).

In order to use the ArchaeoSRP package, a user must supply a directory that contains all of the site forms as individual PDFs. The ArchaeoSRP package converts from PDF to text by way of an optical character recognition (OCR) engine called Tesseract, which was initially developed by HP and subsequently maintained as an open-source project by Google until 2018 (Smith et al. 2022). In short, Tesseract converts an image of a word to the actual characters. The Tesseract engine can be trained to identify new character types, which is useful in converting handwritten notes (although see Fletcher 2023). However, most records in our case study contain typed script, so the default character sets were used. The ArchaeoSRP package also makes use of the Magick (Ooms 2023a), Pdftools (Ooms 2023b), and Stringr (Wickham 2022) packages to aid in image processing and text identification. Once each PDF in the directory has been scanned, a dataframe is returned with information identified and recorded by the package. A directory with example site forms containing "dummy" information is included so that users can test the package.

## ArchaeoSRP Form Identification

Although standard forms are often used by agency archaeologists and contractors, forms vary significantly between regions, by the types of archaeological materials being recorded, or throughout the life of an organization or program (See Figure 1). Therefore, the challenge is identifying similar types of information across disparate types of forms or recording styles. For example, a form might refer to a site's chronological information or date as a "Period of Use," another form type might refer to a "Chronological Period," and a third form type might use the term "Estimated Age." All three forms contain the same information, but identifying them all as the same data type is difficult. For this reason, ArchaeoSRP's first step is identifying the type of form it is analyzing.

Phrases unique to a site record form are initially documented, then subsequently used as identifier key so that the form type can be recalled. In many cases, the form types contain unique headers that can be used to identify them. For instance, the phrase "Department of the Interior" is used to identify a specific document type, whereas "Cultural Resource Isolated Find" indicates another. Not all document types can be easily identified by a single phrase, so multiple phrases were used to identify them. Examples of form identifiers are highlighted in Figure 2: Step 1. Currently, 21 form types have been identified and included in ArchaeoSRP. New document types can be added as they are identified, and users are encouraged to identify and submit new form types via GitHub (Bergin and Snitker 2023a; see Bergin and Snitker 2023b for instructions on how to add new document types).

## ArchaeoSRP Information Processing

Once the form type has been determined, the package identifies the type and location of information that should be recorded. We use a "bookend" approach to search a document for information by defining the words that come before and after information that should be recorded. Using this approach allows for the targeting of specific information when the exact length or amount of information that was originally recorded is unknown. It is difficult to record the exact word or words that follows a key phrase such as "site name" because we cannot predict the particular length of a given site name. For instance, the information after "occupation period:" could simply include a two-word phrase such as "early archaic." On the other hand, a record may include much more detailed information, such as "early archaic with a historical homestead nearby that records indicate was abandoned in 1920." In this case, we are able to capture all of the original information because the "bookend" approach records everything written in answer to a given prompt. An example site record is shown in Figure 2: Step 2. ArchaeoSRP records the information that follows "any):" and is before "location" to identify the site name listed on the form. Using this approach, ArchaeoSRP is able to record location information, site use information, and chronological information. Specific "bookends" are determined after the form type has been identified and the information is recorded. This process is repeated for each site record and recorded as a



https://doi.org/10.1017/aap.2023.22 Published online by Cambridge University Press

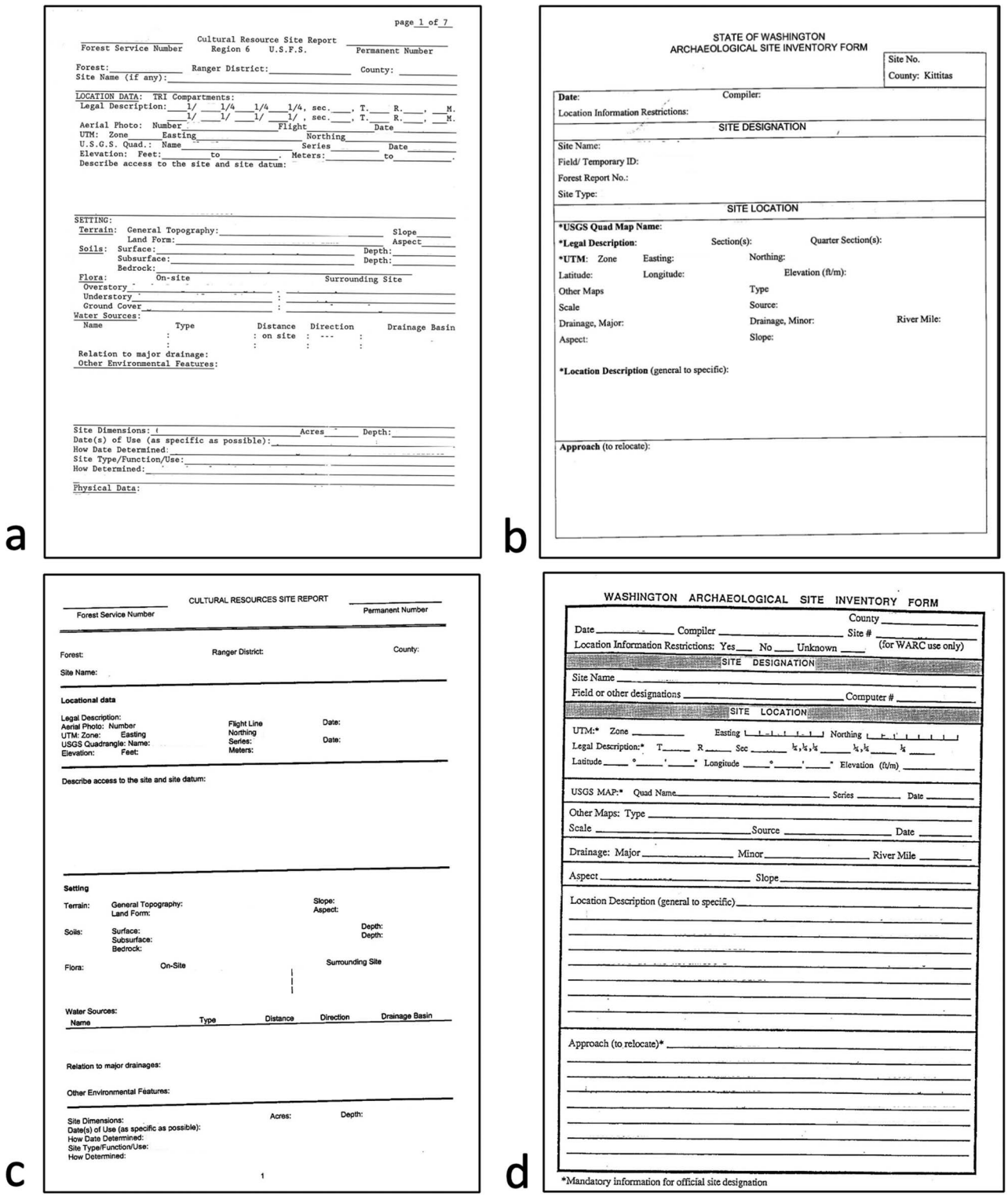


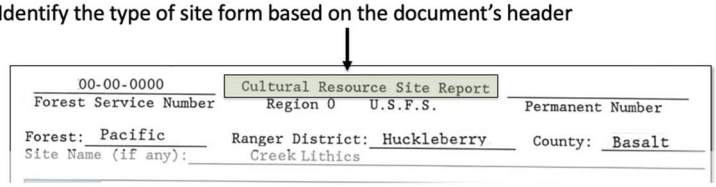

**FIGURE 1.** A sample of the first page from distinctive site form types: (a) Site Document Type 3, (b) Site Document Type 4, (c) Site Document Type 8, and (d) Site Document Type 10.

dataframe in R that can be further modified or exported to other formats, as illustrated in Figure 2: Steps 3 and 4. Currently, ArchaeoSRP extracts information on location, site attributes (e.g., site number, size, chronology), and site description; however, the "bookend" approach allows us to extract any information that is recorded on a given form.

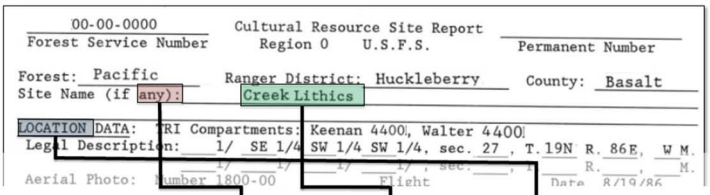


https://doi.org/10.1017/aap.2023.22 Published online by Cambridge University Press

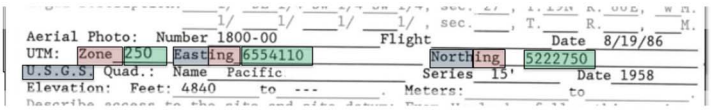
1. Identify the type of site form based on the document's header



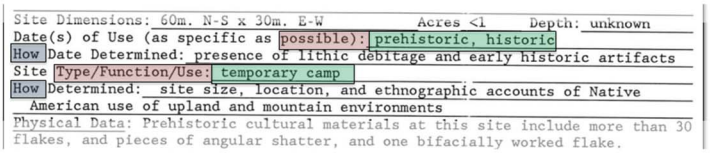

2. Save information based on the words that come before and after the information we wish to record based on a "bookend" approach:




| | Word 1 | Word or Phrase | Word 2 |
|---|---|---|---|
| Site Name | any): | Creek Lithics | Location |




| | Word 1 | Word or Phrase | Word 2 |
|---|---|---|---|
| Zone | Zone | 250 | Easting |
| Easting | Easting | 6554110 | Northing |
| Northing | Northing | 5222750 | U.S.G.S. |




| | Word 1 | Word or Phrase | Word 2 |
|---|---|---|---|
| Period | possible): | prehistoric, historic | How |
| Site Use | Type/Function/Use: | temporary camp | How |

3. Record Information from the document to a database of all the scanned sites

| Site Number | Site Name | Zone | Easting | Northing | Period | Use |
|---|---|---|---|---|---|---|
| 00-00-00-0000 | Creek Lithics | 250 | 6554110 | 5222750 | Prehistoric, historic | Temporary camp |

4. This process continues with each site record in the specified directory.

**FIGURE 2.** Overview of the ArchaeoSRP workflow and extraction procedure using fictitious data.



https://doi.org/10.1017/aap.2023.22 Published online by Cambridge University Press

## Considerations and Challenges

The ArchaeoSRP package was designed with the Cle Elum case study in mind, and it has only been applied to this test case thus far. The aim of designing this R package around the concept of document types is to make it possible for future users to define their own document types. However, in customizing this package for our case study, we noted the challenge of identifying words or phrases that make a given form unique. For example, form types 3 and 8 (see Figures 1a and 1d) both use the title "cultural resources site report." In this case, the presence of a second title, "Location Data," is used to distinguish document type 3 from document type 8. Recognizing potential "bookends" and unique identifiers can make the initial identification of a new document type a tedious process, but it is one that ultimately saves time when processing multiple documents. A similar consideration is the repetition of words used as "bookends" within a document. If a word used as a "bookend" occurs multiple times within a document, then further efforts are necessary to designate which occurrence to use.

An important issue that ArchaeoSRP does not address is the accuracy of the text scanned by the OCR engine Tesseract. Here, we only use OCR on text that has been typed, which minimizes the number of errors, but some errors do occur, and in any application such as this, there is a possibility that the OCR engine misidentifies a character. In many cases, the errors made by the OCR are consistent for a document type and obvious when further attempts are made to use the data. For instance, when performing the OCR on document type 7, the number "10" is often imported as "jo" for the UTM Zone. Given that a number is expected, the error is clear, and in this case, the identification of "jo" as the UTM zone is automatically changed to "10." Nevertheless, errors can be included when conducting the OCR, and for this reason, we recommend that users implement project-specific quality assessment and quality control protocols to ensure that the data in their site forms match the information recorded using ArchaeoSRP.

# APPLYING ARCHAEOSRP TO RECORDS FROM THE CLE ELUM RANGER DISTRICT, OKANOGAN-WENATCHEE NATIONAL FOREST

In collaboration with the Cle Elum Ranger District of the Okanogan-Wenatchee National Forest in central Washington State, we developed a case study to evaluate the utility of the ArchaeoSRP package in extracting and synthesizing archaeological information from paper site and isolate recording forms. The Cle Elum Ranger District is composed of 419,554 acres within Washington State's Central Cascades and Eastern Slope, encompassing multiple ecological zones that include the crest of the Cascade Range, ponderosa pine ecosystems with the eastern foothills, and the western portion of the Yakima River Basin (Figure 3). Archaeological compliance work has occurred on this forest continually since the late 1970s, resulting in almost 800 sites and isolates described over the last four decades. For the last several years, archaeological site and isolate records have been generated using electronic forms; however, most of the archaeological records exist as paper copies that are housed at the Cle Elm Ranger District and at the Okanogan-Wenatchee National Forest Supervisor's Office. Prior to developing this case study, a complete digital archive of all known site forms for the Cle Elm Ranger District did not exist.

The cultural history of the Cle Elum Ranger District follows chronologies similar to the adjacent Puget Sound, Columbia River Basin, and Columbia Plateau regions. Although no specific cultural chronology has been established for the Central Cascades, the generalized chronology is developed from regional datasets that mark general changes in land-use, settlement, and subsistence strategies as reflected in the material record (Kirk and Daugherty 2007; McManamon et al. 2009). Table 1 presents the regional archaeological periods used for the Cle Elum Ranger District in this case study.

## Extracting Data from Paper Records Using ArchaeoSRP

The ArchaeoSRP package was used to extract information from archaeological record forms from the Cle Elum Ranger district of the Okanogan-Wenatchee National Forest. A total of 770 site records were scanned and saved as PDFs by the authors. Archaeological record forms exhibited a range of recording dates that matched the history of the cultural resources program for the district—from 1977 to 2020—which is when this case study was completed. The PDFs were used to identify all possible site record form types used by the district over the previous 43 years, resulting in a total of 21 unique form types. Each form type was used to create a template from which pertinent information on each site or isolate could be extracted. For this case study, the specific information extracted by ArchaeoSRP includes USFS site number, location, any chronological information, site type, and the interpreted use of the site.

Site location is a key piece of information extracted by the ArchaeoSRP package. However, the specificity and type of spatial information associated with each site record varies by the date of recording and the frequency of subsequent site condition updates conducted at the site. For example, some sites recorded in the 1970s and 1980s include site location information derived from the Public Land Survey System (PLSS), including township, range, and nested quarter sections. Later site records include generalized UTM coordinates derived from approximate locations on paper maps, or latitude and longitude coordinates rounded to the nearest second. Finally, modern site records contain UTM coordinates recorded in the field using handheld, recreational, or professional-grade GPS units. Regardless of the method used to collect spatial information for a site, all of the relevant metadata pertaining to a site's location (e.g., datum, coordinate system) is also extracted and collected in the ArchaeoSRP package. A user is able to convert or compare locational data with either a standard geographic information system (GIS) or any of the spatial packages available within R.

The ArchaeoSRP package was able to successfully scan and extract information from 94% of the forms ($n = 721$, total = 770). Forms that were not successfully extracted by ArchaeoSRP were those that contained handwritten information in multiple fields (training Tesseract to read handwriting is possible but difficult [see Fletcher 2023]) or that were recorded on unique forms with nonstandard formatting. The computation time for this procedure



https://doi.org/10.1017/aap.2023.22 Published online by Cambridge University Press

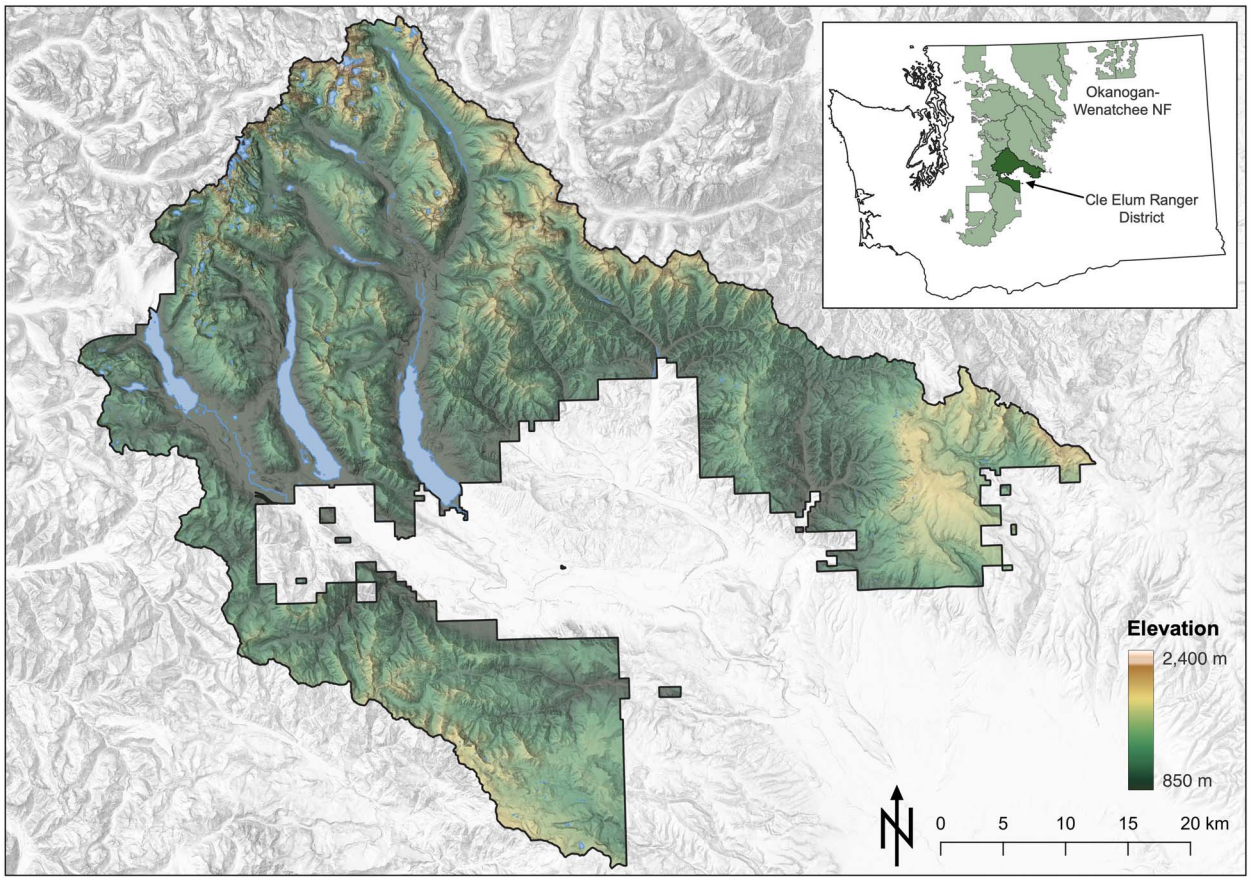



**FIGURE 3.** Location of the Cle Elum Ranger District within the Okanogan-Wenatchee National Forest in central Washington.

took approximately five hours on a desktop computer, which represents a substantial improvement in time expenditure over digitizing and data entry by hand. All extracted site information was output as a tabular dataset. The site records analyzed and transcribed by the ArchaeoSRP package formed the basis for the following data exploration of archaeological site density and temporal patterns in the Cle Elum Ranger District.

## Mapping Archaeological Site Distribution and Density

The spatial information extracted from each site record by ArchaeoSRP was used to map archaeological site density across the entirety of the Cle Elum Ranger District. Site locations were used to aggregate sites into 1 × 1 km grid cells to evaluate broad trends in density and to accommodate for differences in spatial accuracy in locations collected over the last 40 years. Figure 4a demonstrates the density of archaeological sites and isolates recorded by the ranger district and extracted using ArchaeoSRP.

The density of archaeological sites is variable throughout the district, but the information from site forms alone does not indicate if this is due to an absence of archaeological material or an absence of archaeological reconnaissance in these areas. We evaluated this question by linking extracted archaeological information to pedestrian archaeological survey coverage to map potential archaeological site density based on these two inputs. Survey coverage data were compiled from digitized reports or through the Washington Department of Archaeology and Historic Preservation (DAHP) WISAARD online repository. Due to limited data availability, only archaeological survey polygons after 1995 were available. Survey polygons data were used to calculate percent coverage in the same 1 × 1 km grid cells that were used to map site density (Figure 4b). Potential site density was calculated by dividing actual site density by the percent survey coverage for each cell. The resulting metric indicates a potential density of archaeological sites within the grid cell if the entirety of the cell were to be surveyed. For example, a 1 × 1 km grid cell with 100% survey coverage and 10 archaeological sites would have a potential site density of 10 sites per km$^2$. A grid cell with 20% survey coverage and 10 archaeological sites would have a potential site density of 50 sites per km$^2$. When taken together, site density and potential site density illustrate which portions of landscape are the highest priority for archaeological survey and site recording within the Cle Elum Ranger District, and they reflect areas requiring archaeological survey as mandated by the federal NEPA regulatory process (Figure 4c). However, this exercise also identifies areas that lack adequate survey coverage but may nonetheless have high site density, which provides resource managers and researchers with expectation for where to locate new survey projects or to prioritize project objectives during future field campaigns. It is important to note that this exercise assumes all portions of the landscape as having equal likelihood of containing either precontact or historic archaeological sites—a



https://doi.org/10.1017/aap.2023.22 Published online by Cambridge University Press

**TABLE 1.** Regional Archaeological Periods Adapted from the Archaeological Chronologies Currently Used by Cle Elum Ranger District for Managing Cultural Resources within the District.

| Archaeological Period | Estimated Date Range |
|---|---|
| *Postcontact* | |
| Post-War Recreation | AD 1945–1970 |
| Civilian Conservation Corps (CCC) / Great Depression / WWII | AD 1930–1945 |
| Commercial Period (Mining/Logging/ Trapping) | AD 1880–1930 |
| Contact / Euro-American Exploration / Fur Trapping | AD 1720–1880 |
| | |
| *Precontact* | |
| Late Archaic | 2000 BP–AD 1720 |
| Middle Archaic | 5000–2000 BP |
| Early Archaic | 8000–5000 BP |
| Paleoarchaic | 11,000–8000 BP |
| Pre-Clovis/Clovis | +15,000–11,000 BP |

condition that does not reflect the reality of the cultural history of this region. However, this is not an attempt at predicting site locations, evaluating survey effectiveness, or modeling land use. Rather, we utilize the data extracted with ArchaeoSRP to simply visualize the relationship between survey coverage and sites encountered (from both precontact and historic periods) for the district as a whole in order to provide baseline information to guide future research and inventories, which can help us understand how factors such as landform, access to resources, and cultural values influence archaeological site location.

## Mapping Temporal and Spatial Changes in Encountered Archaeological Sites

Use of the ArchaeoSRP package to extract chronological and spatial data allows us to conduct a preliminary evaluation of changes in archaeological sites encountered within the area covered by the documents we processed. Here, we conduct a simple data exploration of site chronology and distribution throughout the Cle Elum Ranger District using information from the cultural chronology established for the region (Table 1). Archaeological chronology, site type, and information pertaining to the interpreted use of the site are extracted from each site record and used to develop chronological maps of archaeological site distribution across the Cle Elum Ranger District. For each 1 × 1 km grid cell, the number of archaeological sites with specific chronological markers (e.g., diagnostic artifacts, features, primary documents, dating methods recorded in the site form) are aggregated by archaeological period. For example, a site record that documents a single Cascade Point, which dates from 9000 to 5700 cal BP based on established typologies for the region (e.g., Carter 2017), is added to the tally of Middle Archaic sites for its corresponding grid cell. The same logic is applied to sites with multiple components, such as a site with both a Columbia Corner-Notched arrow point (2000 BP–Contact) and twentieth-century commercial

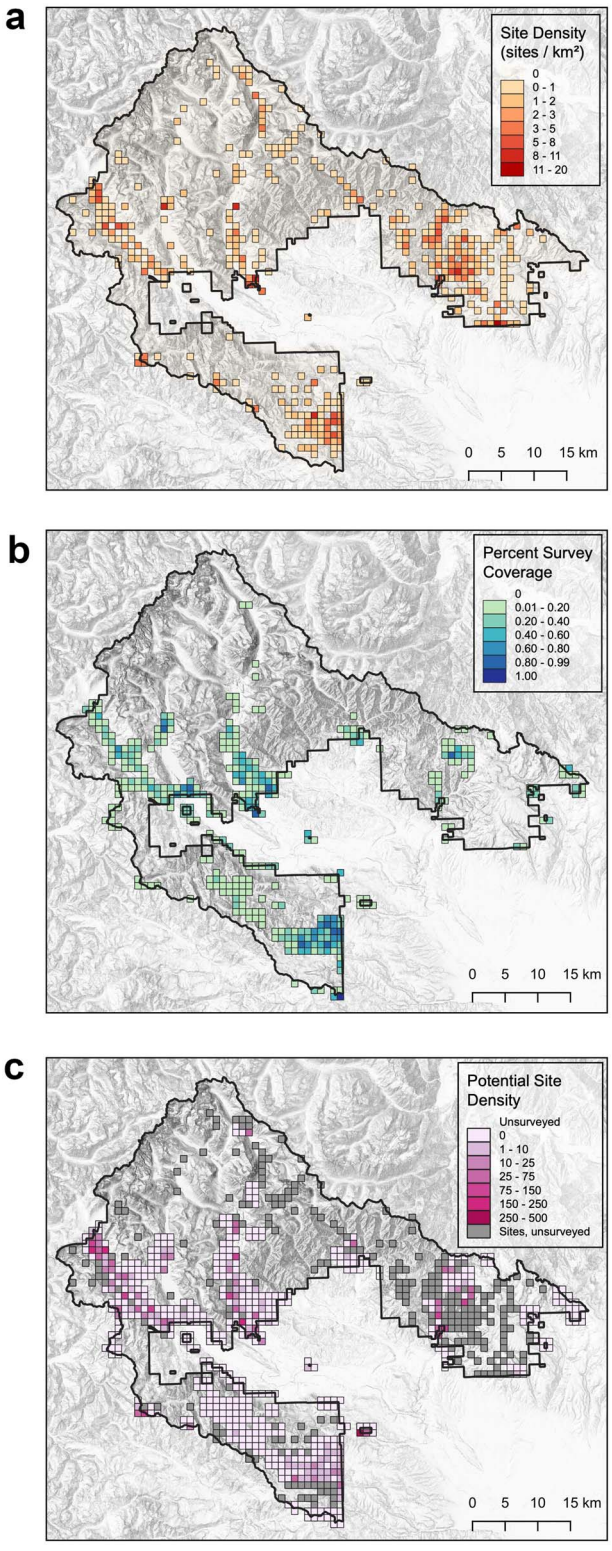



**FIGURE 4.** Archaeological site densities, including (a) archaeological site density per km$^2$, as compiled from the ArchaeoSRP dataset; (b) pedestrian survey coverage (%) per km$^2$ synthesized from the Washington DAHP WISAARD database; (c) calculated potential hypothetical site density for the Cle Elem Ranger District.



https://doi.org/10.1017/aap.2023.22 Published online by Cambridge University Press

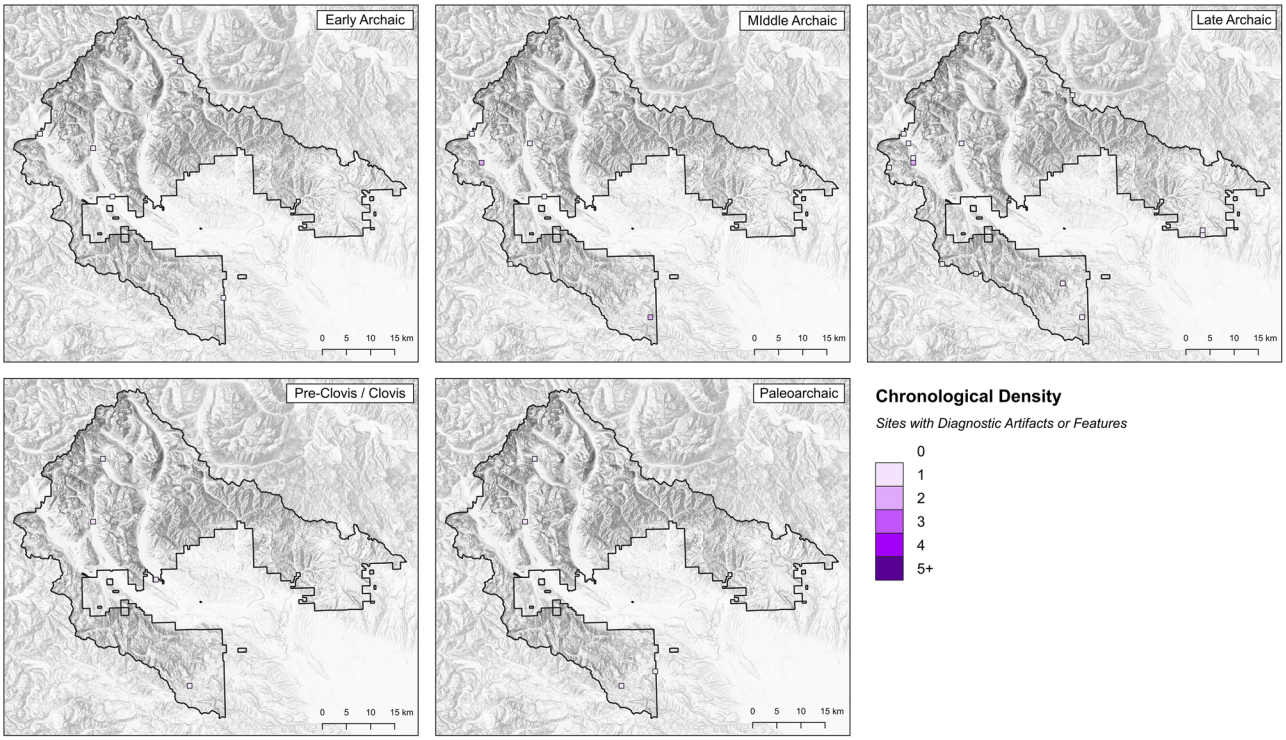



**FIGURE 5.** Chronological density of archaeological sites with diagnostic artifacts, diagnostic features, or chronological information for precontact periods within the Cle Elum Ranger District.

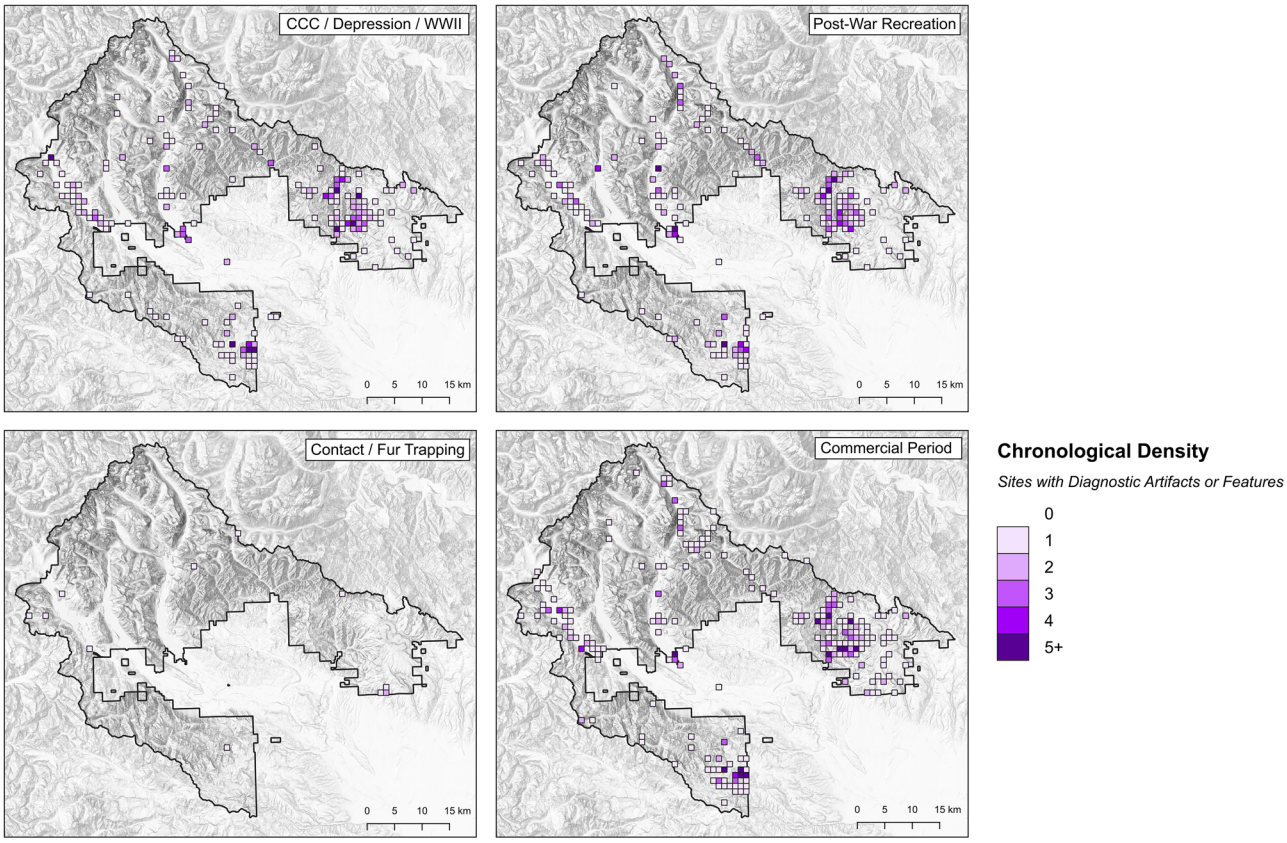



**FIGURE 6.** Chronological density of archaeological sites with diagnostic artifacts, diagnostic features, or chronological information for postcontact periods within the Cle Elum Ranger District.



https://doi.org/10.1017/aap.2023.22 Published online by Cambridge University Press

logging artifacts. This site would be added to the tally of both Late Archaic and Commercial periods. Sites without diagnostic chronological markers are not tallied in this exercise. Figures 5 and 6 illustrate site densities by chronological period across the Cle Elum Ranger District.

Site densities during all precontact periods (Pre-Clovis/Clovis–Late Archaic) exhibit low values across the Cle Elum Ranger District, except for isolated diagnostic artifacts found as isolates or within sites (Figure 5). This trend can be explained by the presence of nondiagnostic lithic materials (e.g., lithic flakes, nondiagnostic bifaces, cores) in many of the surface archaeological sites in the region. Conversely, the postcontact periods (Contact—Post-War Recreation) exhibit much higher densities across the ranger district. This is most evident in the periods that span the late nineteenth and twentieth centuries, given that the artifacts associated with each period are highly diagnostic (Figure 6). Changes in land use across the ranger district are evident throughout the precontact and historical periods, highlighting the diversity of cultural and economic activities that have occurred on these landscapes for millennia. Constructing and visualizing these chronological data from site records creates the link between the goals of the original investigators and the research and management objectives employed on the ranger district today.

## CONCLUDING REMARKS

Preserving records of past archaeological data collection—and focusing that evidence on new questions—is a vital step in developing the next generation of archaeological research. The ArchaeoSRP package serves not only as a tool to digitize, import, and extract information from nondigital site reports but also as a foundation for the development of future tools to automate data collection and enhance the accessibility of archaeological information. In the Cle Elum Ranger District case study, ArchaeoSRP allows us to expand how we interpret land use through time by unlocking the decades of data collection and archaeological expertise contained in paper and other nondigital formats. The ability to convert nondigital information expeditiously and accurately into digital datasets paves the way for new strategies in data synthesis, comparison, and management for federal, state, and tribal agencies.

We encourage readers to download ArchaeoSRP via GitHub and explore its functionality using the mock site forms that we have included with the distribution. Additional updates to the source code will be tracked via the GitHub repository (Bergin and Snitker 2023a). Included in the GitHub repository is a detailed user guide, instructions for creating new site form templates, and a CRAN-style vignette. We also encourage feedback from users so that we can continue to develop this tool for a wide variety of archaeological applications in North America and beyond.


### Acknowledgments

We thank Pete Cadena, Heather Davis, and the rest of Okanogan-Wenatchee National Forest Heritage Program for their assistance in digitizing these site records and insights into the operation of their cultural resource management program. We also thank the three anonymous reviewers for their comments and suggestions to improve this manuscript. All opinions expressed in this article are those of the authors and do not necessarily reflect the policies and views of USDA, DOE, or Oak Ridge Associated Universities (ORAU) / Oak Ridge Institute for Science and Education (ORISE).

### Funding Statement

Snitker's participation in this research was supported in part by an appointment to the USDA United States Forest Service Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the US Department of Energy (DOE) and the US Department of Agriculture. ORISE is managed by ORAU under DOE contract number DE-SC0014664.


### Data Availability Statement

The R code and package described in this article is available from Zenodo (https://doi.org/10.5281/zenodo.7065505) and on GitHub (https://github.com/seanbergin/archaeosrp). Data are curated by USDA Forest Service at the Okanogan-Wenatchee National Forest Supervisor's Office.

### Competing Interests

The authors declare none.

## REFERENCES CITED


Aleman, Julie, Andy Hennebelle, Boris Vannière, Olivier Blarquez, and the Global Paleofire Working Group. 2018. Sparking New Opportunities for Charcoal-Based Fire History Reconstructions. *Fire* 1(1):7. https://doi.org/10.3390/fire1010007.

Altschul, Jeffrey H. 2016. The Society for American Archaeology's Task Forces on Landscape Policy Issues. *Advances in Archaeological Practice* 4(2):102–105. https://doi.org/10.7183/2326-3768.4.2.102.

Barton, C. Michael, Joan Bernabeu, J. Emili Aura, Oreto Garcia, Steven Schmich, and Lluis Molina. 2004. Long-Term Socioecology and Contingent Landscapes. *Journal of Archaeological Method and Theory* 11(3):253–295.

Bergin, Sean, and Grant Snitker. 2023a. ArchaeoSRP v.1.0.1. Electronic document, https://github.com/seanbergin/archaeosrp/, accessed July 5, 2023.

Bergin, Sean, and Grant Snitker. 2023b. How to Add New Types of Archaeological Site Records. Electronic document, https://github.com/seanbergin/archaeosrp/blob/main/How_to_Add_New_Types.md, accessed July 5, 2023.

Bevan, Andrew. 2015. The Data Deluge. *Antiquity* 89(348):1473–1484. https://doi.org/10.15184/aqy.2015.102.

Carlson, David L., and Georg Roth. 2021. archdata: Example Datasets from Archaeological Research. Electronic document, https://CRAN.R-project.org/package=archdata, accessed July 5, 2023.

Carter, James A. 2017. A Typological Key for Projectile Points from the Central Columbia Basin. *Archaeology in Washington* 18:25–46.

Clarke, Mary. 2015. The Digital Dilemma: Preservation and the Digital Archaeological Record. *Advances in Archaeological Practice* 3(4):313–330. https://doi.org/10.7183/2326-3768.3.4.313.

Doelle, William H., Pat Barker, David Cushman, Michael Heilen, Cynthia Herhahn, and Christina Rieth. 2016. Incorporating Archaeological Resources in Landscape-Level Planning and Management. *Advances in Archaeological Practice* 4(2):118–131. https://doi.org/10.7183/2326-3768.4.2.118.

Dosseto, Anthony, and Ben Marwick. 2022. UThwigl—An R Package for Closed- and Open-System Uranium–Thorium Dating. *Quaternary Geochronology* 67:101235. https://doi.org/10.1016/j.quageo.2021.101235.

Ellis, Erle C., Nicolas Gauthier, Kees Klein Goldewijk, Rebecca Bliege Bird, Nicole Boivin, Sandra Díaz, Dorian Q. Fuller, et al. 2021. People Have Shaped Most of Terrestrial Nature for at Least 12,000 Years. *PNAS* 118(17):e2023483118. https://doi.org/10.1073/pnas.2023483118.




https://doi.org/10.1017/aap.2023.22 Published online by Cambridge University Press

Fletcher, Emily C. 2023. Creating a Software Methodology to Analyze and Preserve Archaeological Legacy Data. *Advances in Archaeological Practice* 11(2):139–151. https://doi.org/10.1017/aap.2022.44.

Foster, David, Frederick Swanson, John Aber, Ingrid Burke, Nicholas Brokaw, David Tilman, and Alan Knapp. 2003. The Importance of Land-Use Legacies to Ecology and Conservation. *BioScience* 53(77):77–88. https://doi.org/10.1641/0006-3568(2003)053[0077:TIOLUL]2.0.CO;2.

Halford, F. Kirk, and Dayna M. Ables. 2023. The National Cultural Resources Information Management System (NCRIMS): New Horizons for Cultural Resources Data Management and Analyses. *Advances in Archaeological Practice* 11(1):52–62. https://doi.org/10.1017/aap.2022.39.

Helmer, Matthew, Jennifer Lipton, Grant Snitker, Steven Hackenberger, Mallory Triplett, and Lee Cerveny. 2020. Mapping Heritage Ecosystem Services in Ecological Restoration Areas: A Case Study from the East Cascades, Washington. *Journal of Outdoor Recreation and Tourism* 31:100314. https://doi.org/10.1016/j.jort.2020.100314.

Huggett, Jeremy. 2020. Is Big Digital Data Different? Towards a New Archaeological Paradigm. *Journal of Field Archaeology* 45(sup1):S8–S17. https://doi.org/10.1080/00934690.2020.1713281.

Kansa, Sarah W., Levent Atici, Eric C. Kansa, and Richard H. Meadow. 2020. Archaeological Analysis in the Information Age: Guidelines for Maximizing the Reach, Comprehensiveness, and Longevity of Data. *Advances in Archaeological Practice* 8(1):40–52. https://doi.org/10.1017/aap.2019.36.

Kirk, Ruth, and Richard D. Daugherty. 2007. *Archaeology in Washington*. University of Washington Press, Seattle.

Marwick, Ben. 2019. Chapter 3 Writing Reproducible Research. In *Archaeological Science with R*. Electronic document, https://benmarwick.github.io/aswr/writing-reproducible-research.html, accessed January 5, 2023.

Marwick, Ben. 2023. CRAN Task View: Archaeological Science. Electronic document, https://github.com/benmarwick/ctv-archaeology, accessed January 5, 2023.

Marwick, Ben, Jade d'Alpoim Guedes, C. Michael Barton, Lynsey A. Bates, Michael Baxter, Andrew Bevan, Elizabeth A. Bollwerk, et al. 2017. Open Science in Archaeology. *SAA Archaeological Record* 17(4):8–14.

McCoy, Mark D. 2017. Geospatial Big Data and Archaeology: Prospects and Problems Too Great to Ignore. *Journal of Archaeological Science* 84:74–94. https://doi.org/10.1016/J.JAS.2017.06.003.

McManamon, Francis P., Linda S. Cordell, Kent G. Lightfoot, and George R. Milner. 2009. *Archaeology in America: An Encyclopedia*. Greenwood Press, Westport, Connecticut.

McManamon, Francis P., Keith W. Kintigh, Leigh Anne Ellison, and Adam Brin. 2017. tDAR: A Cultural Heritage Archive for Twenty-First-Century Public Outreach, Research, and Resource Management. *Advances in Archaeological Practice* 5(3):238–249. https://doi.org/10.1017/aap.2017.18.

Mills, Barbara J., Jeffery J. Clark, Matthew A. Peeples, W. R. Haas, John M. Roberts, J. Brett Hill, Deborah L. Huntley, et al. 2013. Transformation of Social Networks in the Late Pre-Hispanic US Southwest. *PNAS* 110(15):5785–5790. https://doi.org/10.1073/pnas.1219966110.

Ooms, Jeroen. 2023a. magick: Advanced Graphics and Image-Processing in R. Electronic document, https://CRAN.R-project.org/package=magick, accessed July 5, 2023.

Ooms, Jeroen. 2023b. pdftools: Text Extraction, Rendering and Converting of PDF Documents. Electronic document, https://CRAN.R-project.org/package=pdftools, accessed July 5, 2023.

Ortman, Scott G., and Jeffrey H. Altschul. 2023. What North American Archaeology Needs to Take Advantage of the Digital Data Revolution. *Advances in Archaeological Practice* 11(1):90–103. https://doi.org/10.1017/aap.2022.42.

Perreault, Charles. 2019. *The Quality of the Archaeological Record*. University of Chicago Press, Chicago.

Plutniak, Sébastien. 2022. Archeofrag: An R Package for Refitting and Spatial Analysis in Archaeology. *Journal of Open Source Software* 7(75):4335. https://doi.org/10.21105/joss.04335.

Posit Team. 2023. *RStudio: Integrated Development for R*. RStudio, PBC, Boston, Massachusetts. Electronic document, hhttp://www.posit.co/, accessed August 5, 2023.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Electronic document, https://www.R-project.org/, accessed August 5, 2023.

Schlanger, Sarah, Richard Wilshusen, and Heidi Roberts. 2015. From Mining Sites to Mining Data: Archaeology's Future. *KIVA* 81(1–2):80–99. https://doi.org/10.1080/00231940.2015.1118739.

Silva, Fabio, Fiona Coward, Kimberley Davies, Sarah Elliott, Emma Jenkins, Adrian C. Newton, Philip Riris, et al. 2022. Developing Transdisciplinary Approaches to Sustainability Challenges: The Need to Model Socio-Environmental Systems in the Longue Durée. *Sustainability* 14(16):10234. https://doi.org/10.3390/su141610234.

Smith, Ray, Ahmad Abdulkader, Rika Antonova, Nicholas Beato, Jeff Breidenbach, Samuel Charron, Phil Cheatle, et al. 2022. Tesseract. Electronic document, https://github.com/tesseract-ocr/tesseract, accessed July 5, 2023.

Steinmann, Lisa, and Barbara Weissova. 2021. datplot. A New R Package for the Visualization of Date Ranges in Archaeology. *Advances in Archaeological Practice* 9(4):288–298. https://doi.org/10.1017/aap.2021.8.

Stephens, Lucas, Dorian Fuller, Nicole Boivin, Torben Rick, Nicolas Gauthier, Andrea Kay, Ben Marwick, et al. 2019. Archaeological Assessment Reveals Earth's Early Transformation through Land Use. *Science* 365(6456):897–902. https://doi.org/10.1126/science.aax1192.

VanValkenburgh, Parker, and J. Andrew Dufton. 2020. Big Archaeology: Horizons and Blindspots. *Journal of Field Archaeology* 45(sup1):S1–S7. https://doi.org/10.1080/00934690.2020.1714307.

Wickham, Hadley. 2022. stringr: Simple, Consistent Wrappers for Common String Operations. Electronic document, https://CRAN.R-project.org/package=stringr, accessed July 5, 2023.

Williams, John W., Eric C. Grimm, Jessica L. Blois, Donald F. Charles, Edward B. Davis, Simon J. Goring, Russell W. Graham, et al. 2018. The Neotoma Paleoecology Database, A multiproxy, International, Community-Curated Data Resource. *Quaternary Research* 89(1):156–177. https://doi.org/10.1017/qua.2017.105.

Wilshusen, Richard H., Michael Heilen, Wade Catts, Karyn de Dufour, and Bradford Jones. 2016. Archaeological Survey Data Quality, Durability, and Use in the United States: Findings and Recommendations. *Advances in Archaeological Practice* 4(2):106–117. https://doi.org/10.7183/2326-3768.4.2.106.

Wright, Holly, and Julian D. Richards. 2018. Reflections on Collaborative Archaeology and Large-Scale Online Research Infrastructures. *Journal of Field Archaeology* 43(sup1):S60–S67. https://doi.org/10.1080/00934690.2018.1511960.

## AUTHOR INFORMATION

**Sean Bergin** ■ Arizona State University, School of Complex Adaptive Systems, Tempe, AZ, USA (sbergin@asu.edu, corresponding author)

**Grant Snitker** ■ New Mexico Consortium, Cultural Resource Sciences, Los Alamos, NM, USA (gsnitker@newmexicoconsortium.org, corresponding author)



https://doi.org/10.1017/aap.2023.22 Published online by Cambridge University Press