DIGITAL ANTIQUITY

Reports in Digital Archaeology Number 1 **June 2011**

Building tDAR: Review, Redaction, and Ingest of Two Reports Series

Joshua Watts, Digital Data Curator, Center for Digital Antiquity

In the summer of 2010 I was given the task of adding two sets of legacy archaeological publications to the Digital Archaeological Record (tDAR) database. This short report summarizes the work that went into preparing the documents, entering relevant metadata, and uploading the documents to the tDAR server. Shelby Manney and Sophia Kelly also carried out some of the report reviews and file uploading for the project. One set of documents was a series of reports published and distributed by the Western Archeological and Conservation Center (WACC) of the National Park Service (NPS) in Tucson, Arizona. The other set was a collection of reports generated by archaeological salvage work in central Phoenix, completed under contract with the Arizona Department of Transportation.

Many of the reports discussed here are from the WACC publication series *Papers in Anthropology* (PIA), which at the time of ingest to tDAR numbered 106 documents. The documents were available to Digital Antiquity (DA) as a set of image PDF's on four compact discs. For the most part, the PDF's were high-resolution scans of the original documents, but more recent publications were converted to PDF format from the original word processing files. Average length of the reports was about 220 pages, though lengths ranged from 3 pages (a small set of over-sized maps) to 852 pages. Most were between 200 and 400 pages. Because the WACC documents represent over 30 years of archaeological work, from 1975 to 2008, they present a variety of difficulties for us that rarely occur with recently published materials. These issues are discussed thoroughly below. Note that a few of the PIA series reports were unavailable in PDF format, presenting us with different challenges.

Watts, J. Building tDAR: Review, Redaction, and Ingest of Two Reports Series. *Reports in Digital Archaeology* Number 1, Center for Digital Antiquity, Arizona State University, Tempe AZ, 2011.

The second set of reports discussed in this report are from a compilation of archaeological monographs and technical reports called *Intersections: Pathways Through Time*, which was a digital archive of 37 reports from 11 archaeological projects conducted at Hohokam sites associated with the ancient Canal System Two in Phoenix, Arizona.

Fieldwork and most associated analyses were conducted between 1981 and 1994. The CD-ROM archive of the reports was published in 2003. The *Intersections* compilation and CDs were made possible by an ISTEA (Intersurface Transportation Enhancement Act) grant from the Department of Transportation to Arizona Department of Transportation. The CD compilers and editors (Shears et al. 2003) also included a few papers synthesizing some of the information from the monographs and technical reports. These synthetic papers also were included among the documents uploaded to tDAR. In total, 39 Intersections documents were uploaded to tDAR (two of the original 37 volumes were split into two parts before publication). The length of these reports varied widely, from 20 pages to 610 pages, but the average length of the Intersections documents was 314 pages. Because these reports had been previously digitized and contained no sensitive or confidential content, in general they presented no difficulties when adding them to tDAR.

As of fall 2010, the tDAR web software was in a relatively early stage of development – meaning that the core functionality was in place, but the data entry forms for defining projects and entering metadata were rather rough. Screen shots of the metadata entry form are pasted below (Fig 1). Not long after this project was completed, the forms were significantly upgraded, so while most of the process described below was unaffected by the changes to the forms, some details about the metadata entry would need to be slightly revised if a user were to adopt this process for their own work.

In tDAR, individual digital objects (such as reports) usually are organized into “projects”, which are analogous to folders containing related files. Sometimes the tDAR project is associated with a well-bounded archaeological research project, such as documentation related to the excavation of a specific site. Alternately, the tDAR project may also be used more flexibly, to organize resources in other ways that may be intuitive for researchers or archivists. For the WACC documents, we determined that projects would be defined according to the national park in which the research was conducted. The 106 documents were split unevenly among 25 new projects. Each new project was created to cover the area of the national park unit within which the archaeological investigations reported had been conducted. For example, projects were created for “The Archaeology of Joshua Tree National Park,” “The Archaeology of Saguaro National Park,” etc. Alternately, the 39 Intersections documents were grouped into a single project: “Phoenix Basin Archaeology: Intersections, Pathways Through Time” because all the reports were from the same geographical area.

Legacy archaeological reports, as opposed to recent publications that conform to modern standards and laws, present a variety of challenges. For example, many reports from the 1970's and 1980's sometimes contain information and images that may be either confidential (e.g. detailed maps with site locations) or culturally sensitive (e.g. images of Native American human remains). Also, the content of some of the earlier-scanned documents was not searchable without performing optical character recognition (OCR) scans on the PDF files. The following paragraphs summarize the process used to overcome those challenges and upload the older reports to tDAR.

The first step in the process, when necessary, was to perform OCR on a PDF image document. The Adobe Acrobat engine for OCR was used, in this case Acrobat 9 Pro Extended. The time required to complete this step varied depending largely on the length of the document, but averaged between seven and eight minutes per report. While imperfect, the accuracy of the OCR was surprisingly good – depending somewhat on the report fonts and image quality of the original PDF.

Acrobat flashes each page on the screen for a couple seconds as it performs the OCR, providing a shortcut to the next step of the process: identifying potentially confidential or sensitive content. As each page appeared on the screen, it was quickly reviewed for problematic content such as detailed maps, tables with site locations, or images of human remains.

For the purpose of managing tDAR documents, "confidential" was defined as information that was restricted because posting the information posed a reasonable risk that it would lead to the vandalism or destruction of the archaeological resource. The Archaeological Resource Protection Act (ARPA) restricts the release of information about specific locations or characteristics of sites covered by the act if the public release of such information might result in damage or loss of the site.

"Sensitive" was defined as information that may be offensive or inappropriate to some readers, e.g. images of human remains or other culturally sensitive materials. Pages noted as potentially have either confidential or sensitive content during the OCR pass were further examined to determine whether the potentially confidential content was detailed enough to require its removal from the document before upload to tDAR. More recent reports did not require OCR scanning were paged through manually to check for the confidential or sensitive content.

Confidential information was removed using the Redaction Tool in Acrobat (under the Advanced menu in the Acrobat 9 Pro Extended version). From the two sets of documents discussed here, 44 of them required that some content be redacted (about 30 percent of the documents included in this study) – mostly older WACC reports that included detailed maps or other location information. In the case of redacted figures (usually maps), the whole map was selected – but not the caption – and replaced with a

white box with the text "Confidential content removed" along the upper boundary. In cases where the confidential information was in tables, such as UTM coordinates, only the problematic numbers or text were redacted – with a gray box to indicate where that content was removed. Potentially sensitive information was not removed from the documents, but a warning of this potential content was added to the document (see below.) The amount of time required to complete the redaction varied widely depending on how much information had to be removed from the document, but the average was just over seven minutes per document.

Cover pages were prepared for each document that was found to have either confidential or sensitive content. These pages provide an explanation for why content was redacted or flagged as sensitive. The new pages (Figure 2) were inserted after the front title page of the report. Where information was removed from a document, a list of affected pages numbers was added to each cover page. Reports with sensitive content were given a blanket disclaimer on a cover page that was inserted after the title page of the document. Typically, adding a cover page to a document required three or four minutes.

While only a subset of the reports required all of these extra preparation procedures (i.e., the OCR scan, review followed by redaction activities, and preparation of cover pages), most of the documents required one or more of these steps before ingesting them into tDAR. Once the initial review and related actions were completed, however, the steps of entering metadata and uploading the document to tDAR were relatively simple and straightforward.

To upload resources to tDAR, you must be logged into the site using an account set up with permission to contribute (an option selected when the account was initially registered). Entering metadata consists of two primary efforts: first, to enter a complete citation so that the resource can be identified and authors credited, and second, to use the form to populate a keyword list so that researchers can find the document using the tDAR search function.

Most of the citation information from the WACC documents was usually available on the title page of each report, e.g. title, author (editor, contributor, etc.), and publisher. The WACC reports, particularly the older examples, were inconsistent as to whether or not they included an abstract. Where there was an abstract, it was copied directly from the report and pasted into the tDAR form. For the rest, an abstract, based on the report's introduction, table of contents, and other relevant information, was prepared at the same time as the other metadata information was entered. In a rare few cases, the documents had an ISBN (International Standard Book Number), which was entered into the tDAR form. This section of the form also has a field for selecting a PDF to upload with the metadata. At the time of this writing, only PDF files are supported and only one file can be associated with the metadata.

The bulk of the upload form is designed to encourage a thorough keyword list. It provides a series of check-box menus, text boxes, and a map to indicate where the research was conducted. Entry of geographic information may be entered as keywords, but the tDAR upload form provides a tool that allows the user to manually highlight the project area on a map. To avoid the release of confidential site locations to later tDAR viewers, cases where the highlighted project area is smaller than one square mile the location is obfuscated. If a project area is larger than one square mile, it will be accurately shown to users. Most of the information relevant to completing this part of the form can be found in the document's table of contents, abstract, and large-scale maps usually located in the opening pages of the report. Readers should note that the form has been developed further and some of the metadata categories and the controlled values within categories have been refined since the reports discussed here were ingested.

At the time this project was underway (September 2010) the rules for inheritance of keywords from the "project" level were very much in flux. At the time of the WACC and Intersections ingest, a list of project-level keywords were shown on the document metadata form – but those keywords had to be re-entered in order to tie them to the report. Some categories (e.g. radiocarbon dates) were only relevant to small subset of the reports, while other categories (e.g. investigation type) were used for almost all the reports. Once the form was filled out as completely as possible, clicking the submit button ended the process. A unique tDAR resource identification number was assigned to the document and its associated metadata. Total time required to enter the metadata into this version of the form averaged between ten and eleven minutes. A summary of the time needed to complete the various tasks described in this report is provided in the table below.

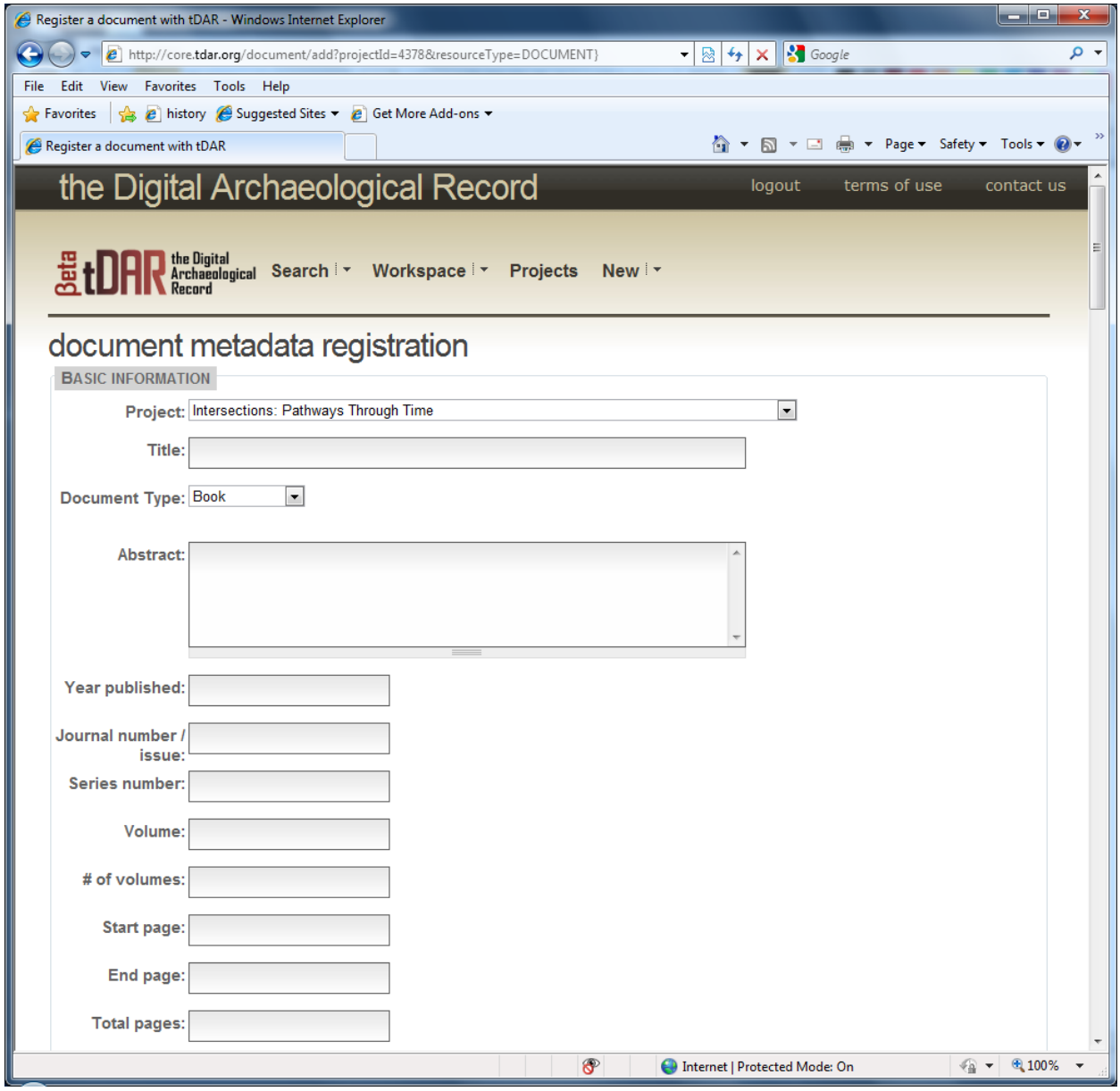
Activity	Number of documents	Average time	Median time
Redaction	33	7:10	5:15
OCR scanning	27	7:40	5:45
Insert cover page	42	3:45	3:15
Metadata entry	127	10:26	9:45

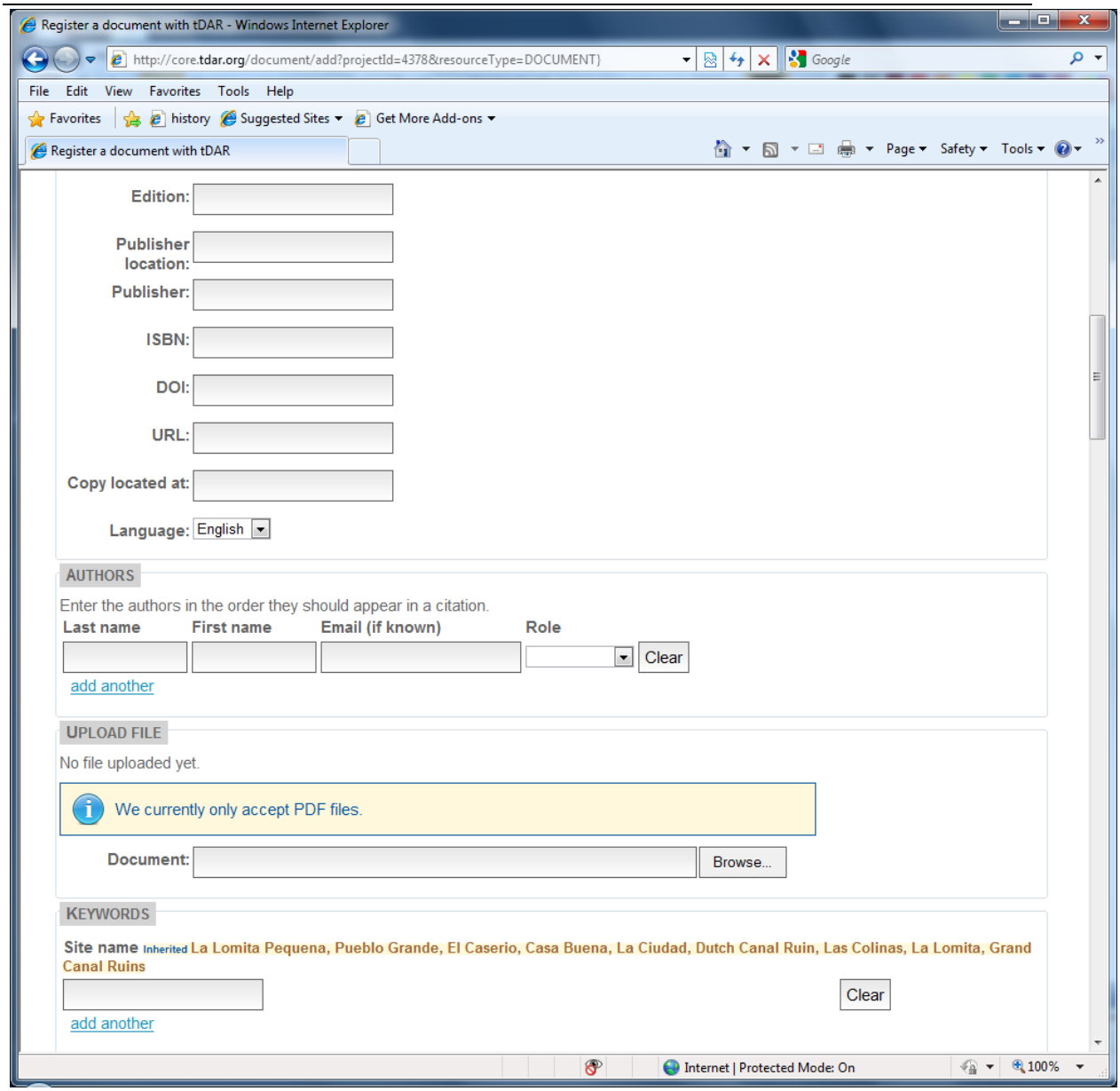
While most of the reports in the WACC series and Intersections documents conformed to the process above, there were a handful of exceptions. The last step of the ingest process was to tie up loose ends related to those irregular documents, including: a handful of over-size map packets; volumes that consisted entirely of site records had the metadata entered into tDAR, but the documents were not uploaded (due to large amounts of confidential content); "missing" reports that are not available in a PDF

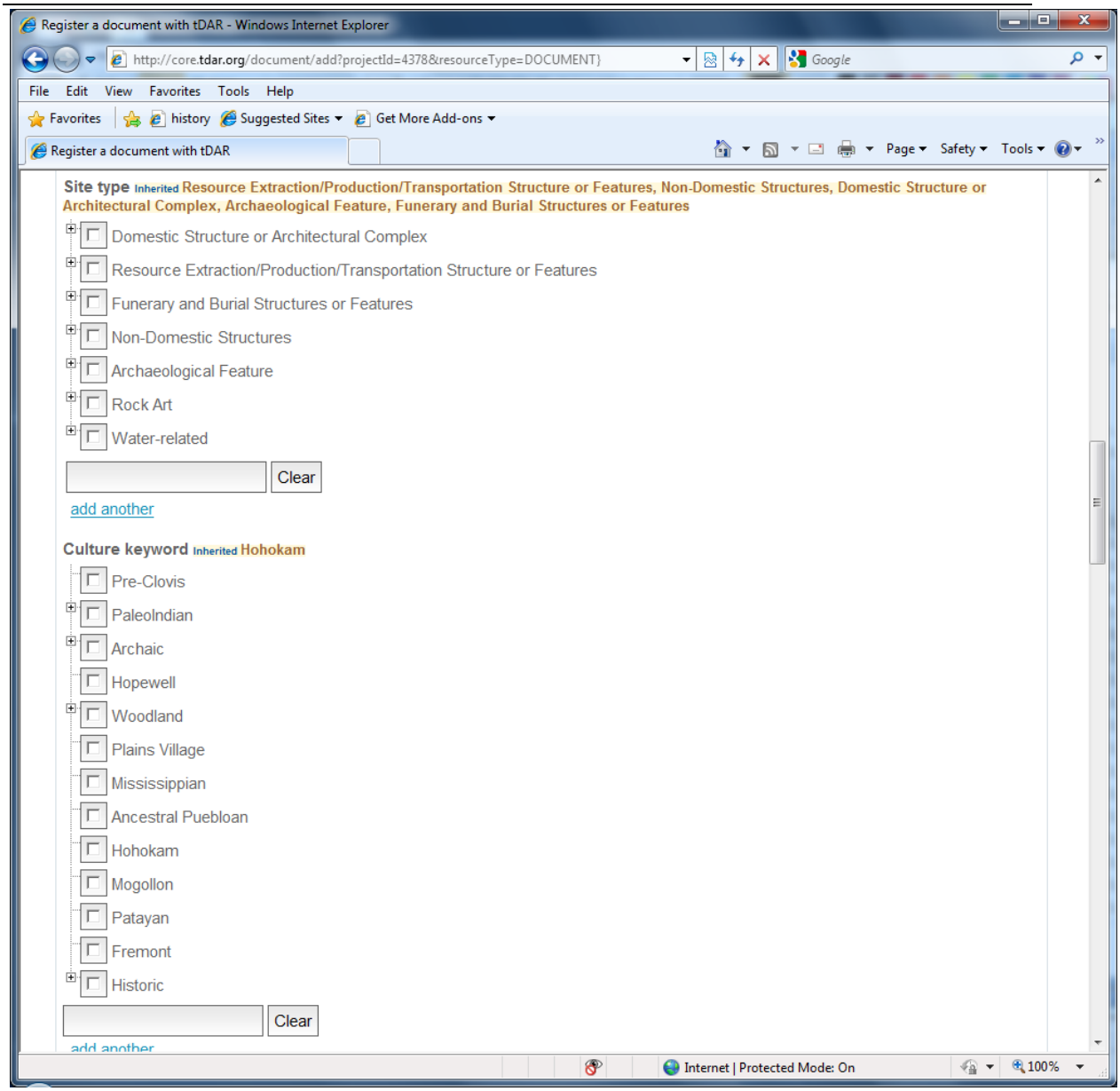
format and/or are out-of-print – two cases were not found, while another was located on the NPS website in an html format. That last report was converted by a third party to a somewhat clunky PDF (from the html pages) and added to tDAR.

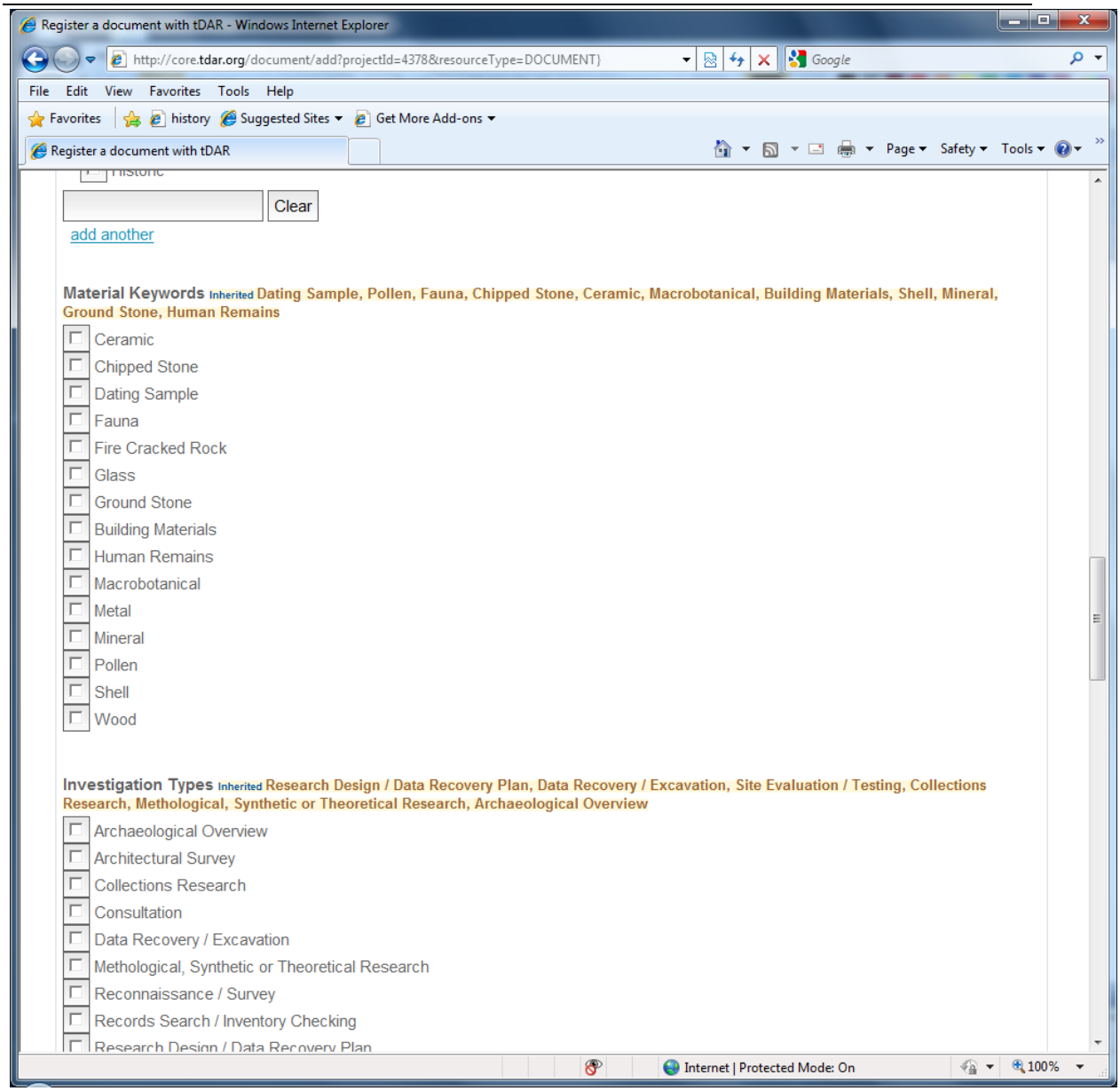
Figure 1 –The tDAR document upload form in September 2010 (split over 7 pages)

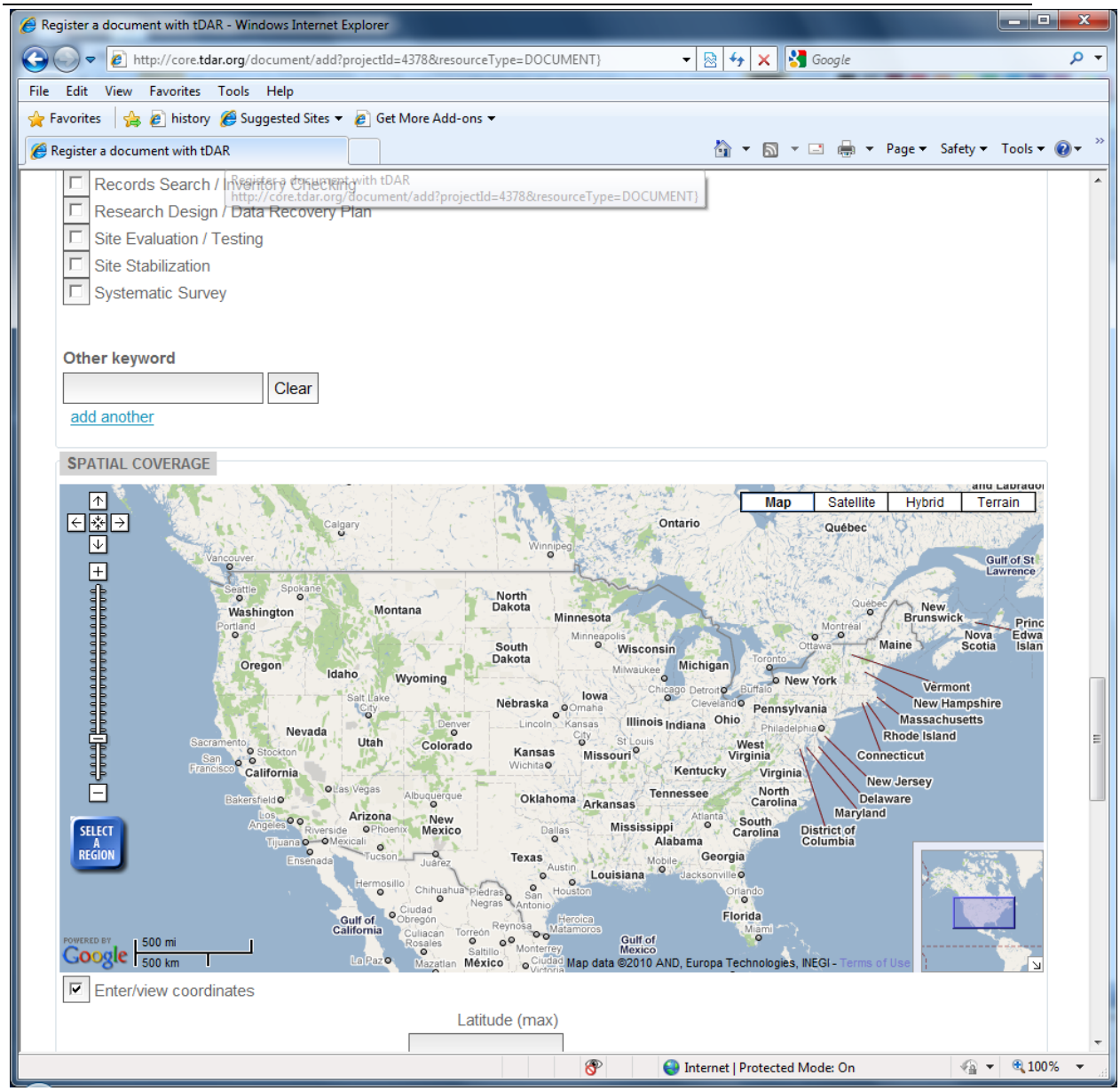
.

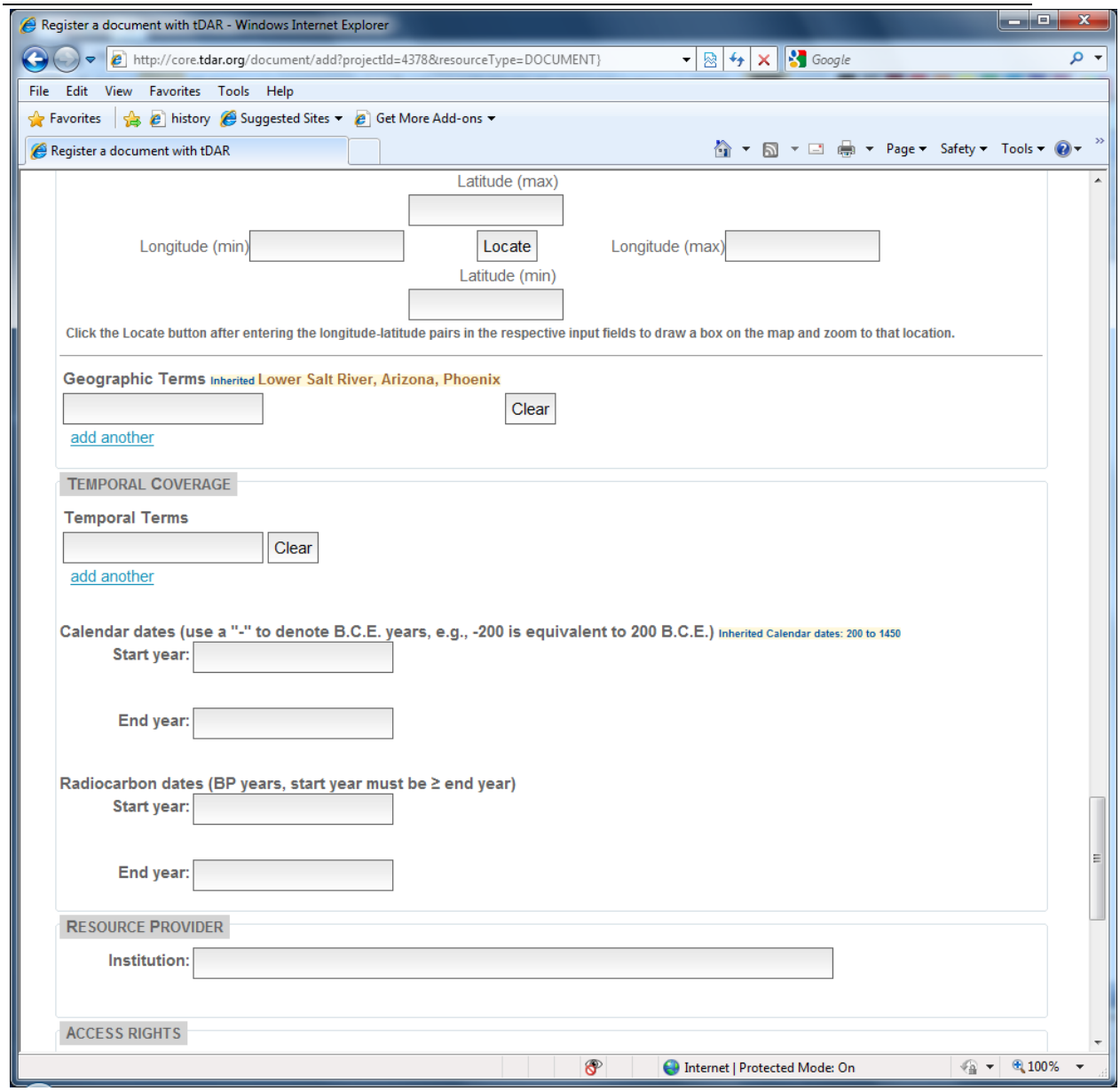



Register a document with tDAR - Windows Internet Explorer
http://core.tdar.org/document/add?projectId=4378&resourceType=DOCUMENT}
File Edit View Favorites Tools Help
Favorites history Suggested Sites Get More Add-ons
Register a document with tDAR
Page Safety Tools
Edition:
Publisher location:
Publisher:
ISBN:
DOI:
URL:
Copy located at:
Language: English
AUTHORS
Enter the authors in the order they should appear in a citation.
Last name
First name
Email (if known)
Role
Clear
add another
UPLOAD FILE
No file uploaded yet.
We currently only accept PDF files.
Document:
Browse...
KEYWORDS
Site name Inherited La Lomita Pequena, Pueblo Grande, El Caserio, Casa Buena, La Ciudad, Dutch Canal Ruin, Las Colinas, La Lomita, Grand Canal Ruins
Clear
add another
Internet | Protected Mode: On
100%


Register a document with tDAR - Windows Internet Explorer
http://core.tdar.org/document/add?projectId=4378&resourceType=DOCUMENT}
Google
File Edit View Favorites Tools Help
Favorites history Suggested Sites Get More Add-ons
Register a document with tDAR
Page Safety Tools
Site type Inherited Resource Extraction/Production/Transportation Structure or Features, Non-Domestic Structures, Domestic Structure or Architectural Complex, Archaeological Feature, Funerary and Burial Structures or Features
Domestic Structure or Architectural Complex
Resource Extraction/Production/Transportation Structure or Features
Funerary and Burial Structures or Features
Non-Domestic Structures
Archaeological Feature
Rock Art
Water-related
Clear
add another
Culture keyword Inherited Hohokam
Pre-Clovis
PaleoIndian
Archaic
Hopewell
Woodland
Plains Village
Mississippian
Ancestral Puebloan
Hohokam
Mogollon
Patayan
Fremont
Historic
Clear
Internet | Protected Mode: On
100%


Register a document with tDAR - Windows Internet Explorer
http://core.tdar.org/document/add?projectId=4378&resourceType=DOCUMENT}
Google
File Edit View Favorites Tools Help
Favorites history Suggested Sites Get More Add-ons
Register a document with tDAR
Page Safety Tools
Clear
add another
Material Keywords Inherited Dating Sample, Pollen, Fauna, Chipped Stone, Ceramic, Macrobotanical, Building Materials, Shell, Mineral, Ground Stone, Human Remains
Ceramic
Chipped Stone
Dating Sample
Fauna
Fire Cracked Rock
Glass
Ground Stone
Building Materials
Human Remains
Macrobotanical
Metal
Mineral
Pollen
Shell
Wood
Investigation Types Inherited Research Design / Data Recovery Plan, Data Recovery / Excavation, Site Evaluation / Testing, Collections Research, Methological, Synthetic or Theoretical Research, Archaeological Overview
Archaeological Overview
Architectural Survey
Collections Research
Consultation
Data Recovery / Excavation
Methological, Synthetic or Theoretical Research
Reconnaissance / Survey
Records Search / Inventory Checking
Research Design / Data Recovery Plan
Internet | Protected Mode: On
100%


Register a document with tDAR - Windows Internet Explorer
http://core.tdar.org/document/add?projectId=4378&resourceType=DOCUMENT}
File Edit View Favorites Tools Help
Favorites history Suggested Sites Get More Add-ons
Register a document with tDAR
Page Safety Tools
Records Search / Inventory Checking
Research Design / Data Recovery Plan
Site Evaluation / Testing
Site Stabilization
Systematic Survey
Other keyword
Clear
add another
SPATIAL COVERAGE
Map Satellite Hybrid Terrain
SELECT A REGION
500 mi
500 km
Map data ©2010 AND, Europa Technologies, INEGI - Terms of Use
Enter/view coordinates
Latitude (max)
Internet | Protected Mode: On
100%


Register a document with tDAR - Windows Internet Explorer
http://core.tdar.org/document/add?projectId=4378&resourceType=DOCUMENT}
File Edit View Favorites Tools Help
Favorites history Suggested Sites Get More Add-ons
Register a document with tDAR
Page Safety Tools
Latitude (max)
Longitude (min)
Locate
Longitude (max)
Latitude (min)
Click the Locate button after entering the longitude-latitude pairs in the respective input fields to draw a box on the map and zoom to that location.
Geographic Terms Inherited Lower Salt River, Arizona, Phoenix
Clear
add another
TEMPORAL COVERAGE
Temporal Terms
Clear
add another
Calendar dates (use a "-" to denote B.C.E. years, e.g., -200 is equivalent to 200 B.C.E.) Inherited Calendar dates: 200 to 1450
Start year:
End year:
Radiocarbon dates (BP years, start year must be ≥ end year)
Start year:
End year:
RESOURCE PROVIDER
Institution:
ACCESS RIGHTS
Internet | Protected Mode: On
100%

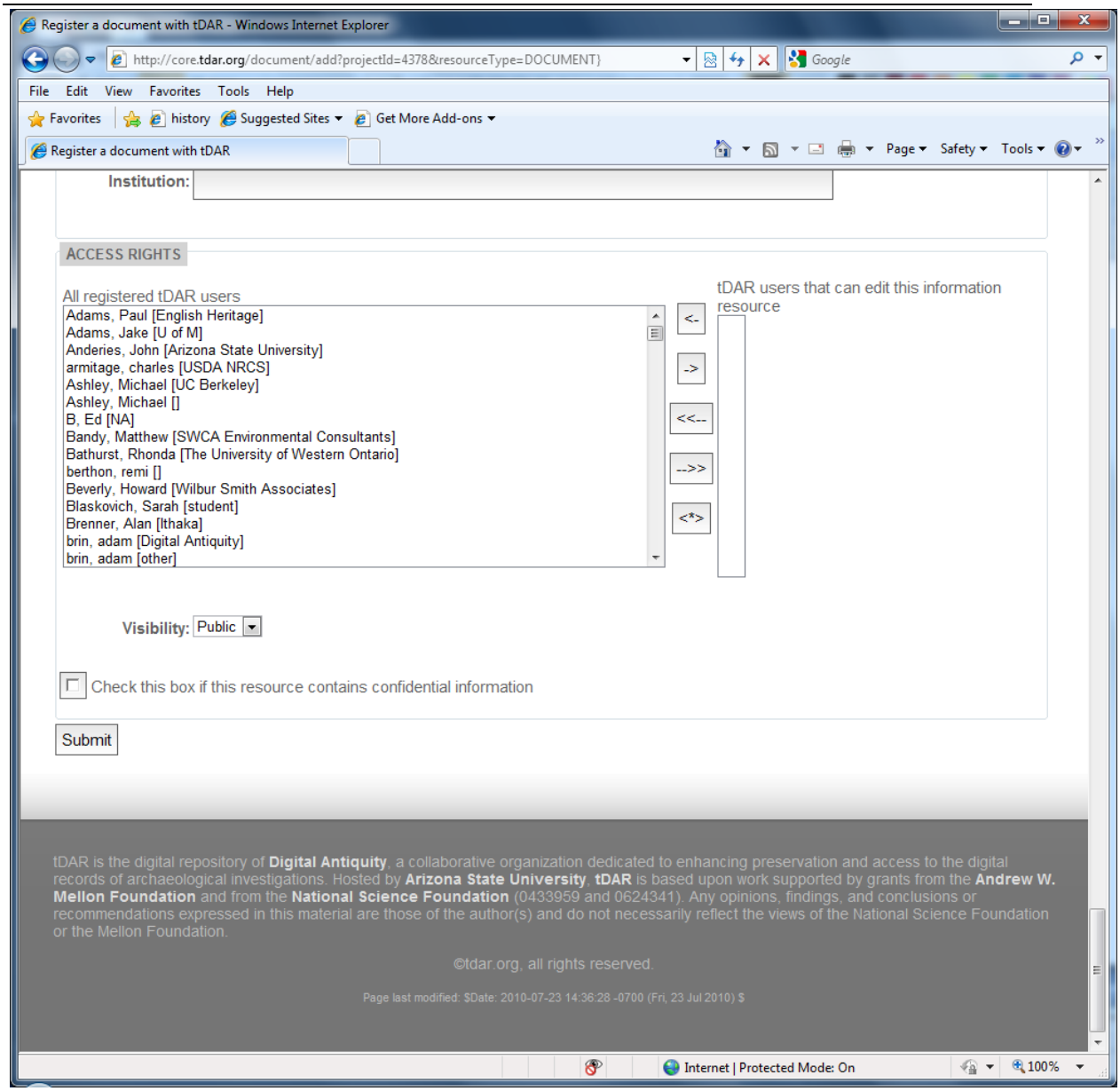

Register a document with tDAR - Windows Internet Explorer
http://core.tdar.org/document/add?projectId=4378&resourceType=DOCUMENT)
File Edit View Favorites Tools Help
Favorites history Suggested Sites Get More Add-ons
Register a document with tDAR
Page Safety Tools
Institution:
ACCESS RIGHTS
All registered tDAR users
Adams, Paul [English Heritage]
Adams, Jake [U of M]
Anderies, John [Arizona State University]
armitage, charles [USDA NRCS]
Ashley, Michael [UC Berkeley]
Ashley, Michael []
B, Ed [NA]
Bandy, Matthew [SWCA Environmental Consultants]
Bathurst, Rhonda [The University of Western Ontario]
berthon, remi []
Beverly, Howard [Wilbur Smith Associates]
Blaskovich, Sarah [student]
Brenner, Alan [Ithaka]
brin, adam [Digital Antiquity]
brin, adam [other]
<-
->
<<--
-->>
<*>
tDAR users that can edit this information resource
Visibility: Public
Check this box if this resource contains confidential information
Submit
tDAR is the digital repository of Digital Antiquity, a collaborative organization dedicated to enhancing preservation and access to the digital records of archaeological investigations. Hosted by Arizona State University, tDAR is based upon work supported by grants from the Andrew W. Mellon Foundation and from the National Science Foundation (0433959 and 0624341). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Mellon Foundation.
©tdar.org, all rights reserved.
Page last modified: $Date: 2010-07-23 14:36:28 -0700 (Fri, 23 Jul 2010) $
Internet | Protected Mode: On
100%

Figure 2 – Cover pages for reports with "confidential" and/or "sensitive" content.

tDAR the Digital Archaeological Record
Digital Antiquity

Notification of Confidential Content Deleted

Information has been deleted from this digital copy of this document because it was judged to be confidential information, the general release of which might result in harm to the archaeological resources described.

> [Section 9 of the Archaeological Resources Protection Act [16 USC 470aa-mm] prohibits the general public release of "…information concerning the nature and location of any archaeological resource for which the excavation or removal requires a permit or other permission under this Act or under any other provision of Federal law…" unless the release of this information will:
>
> > (1) further the purposes of this Act or the Act of June 27, 1960 [the Reservoir Salvage Act, as amended, 16 U.S.C. 469-469c-1] and
> >
> > (2) not create a risk of harm to such resources or to the site at which such resources are located."]

Text or figures on the following pages of the original document have been deleted: 3

tDAR the Digital Archaeological Record
Digital Antiquity

Notification of Potentially Culturally Sensitive Content

This document contains material that may be considered culturally sensitive to some readers. Please be aware that potentially sensitive information related to human remains and associated burials, including images and/or detailed descriptions, may be present in this digital copy.