White Paper Report

Report ID: 104273

Application Number: HD5142511

Project Director: Eric Kansa (ekansa@ischool.berkeley.edu)

Institution: Alexandria Archive Institute

Reporting Period: 9/1/2011-2/28/2013

Report Due: 5/31/2013

Date Submitted: 7/31/2013

The Gazetteer of the Ancient Near East

White Paper

Prepared for the NEH Office of Digital Humanities
Digital Humanities Start-Up Grant, Level II
Grant #: HD-51425-11
Project Director: Eric C. Kansa
Grant Period: Sept. 2011 – Feb. 2013
Report Date: July 31, 2013

Eric C. Kansa
GANE Project Director
Alexandria Archive Institute & UC Berkeley
kansaeric@gmail.com

Sarah Whitcher Kansa
Executive Director, Alexandria Archive Institute
skansa@alexandriaarchive.org

Francis Deblauwe
GANE Content Editor
Alexandria Archive Institute
fdeblauwe@gmail.com

1. INTRODUCTION

The Gazetteer of the Ancient Near East (GANE) project launched in September of 2011, with a Digital Humanities Level II Start-Up grant from the National Endowment for the Humanities (NEH) in the amount of $49,707. The GANE project proposed, over an 18-month period, to develop an authoritative, free and open access compendium of more than 8,000 gazetteer records from the Ancient Near East. The project's ultimate goal is the establishment of a geospatial index of archaeological sites and historical places spanning some twelve thousand years, from the Natufian period through the Iron Age (c. 12,500 – 600 BCE). From this foundation built by the GANE project, researchers can link historical events, historical persons, and archaeological evidence through notions of place and time. By developing a free and open corpus of ancient places, the GANE project makes it easier for scholars to bring together disparate lines of historical and archaeological evidence on the Web. This critical piece of infrastructure will greatly benefit research in the Ancient Near East by adding greater breadth, depth, and precision to the research process. The GANE project employs software developed by the Pleiades project (http://pleiades.stoa.org/), an extant and successful model for open access publication of easy to reference authoritative place and time information. Pleiades offers unique URLs to reference specific places, locations, and toponyms. This project, thus, serves as an example in the humanities and beyond of how adoption of extant, simple tools and collaboration with similar efforts can expand the reach and impact of open data.

This report provides details of activities and accomplishments over the duration of the project, from September 2011 to February 2013.

2. PROJECT ACTIVITIES

2.1 Goals

A primary goal of the GANE project is to help researchers cope with the growing information explosion by providing a critically needed open reference corpus that can be used to link diverse scholarly content through concepts of place and time. A critical concept in online communication of archaeological (and related) data is context. Disassociated from its context, an artifact loses key information essential to its interpretation and use. As they become accustomed to the Digital Age, scholars are starting to appreciate the importance of linking place and time with content they publish[1]. Where print and digital publications may contain such information, they often express geographic and chronological context in haphazard ways. Such inconsistencies make it very difficult for researchers to find and relate geographically relevant content from different sources. As scholarly publishing, museum collections, and excavation datasets all migrate to the Web, the need for simple but authoritative ways to represent geographic context becomes more urgent.

2.2 Implementation

The Pleiades gazetteer, developed by the Institute for the Study of the Ancient World (New York University), serves as this project's model for how authoritative place and time information can be made freely available for researchers and institutions to easily reference and use. Pleiades (Figure 1) is largely based on the Barrington Atlas, arguably the key reference work on Greco-Roman geography. Pleiades has amassed a database of some 35,000 places, which people sharing content online can attach to their data in order to ensure that others clearly understand the context of their content. Pleiades serves the classical world and, thus, is useful to archaeologists working within that disciplinary scope.

The GANE project proposed to extend Pleiades coverage to sites and places relevant to scholars of the Ancient Near East. Use of Pleiades, an established and working system, minimizes technical costs and delays. Pleiades has a mature architecture ideally suited for developing a reference work like the one proposed. Pleaides has three fundamental concepts:

[1] Michael K. Buckland, Fredric C. Gey, and Ray R. Larson. Access to Heritage Resources Using What, Where, When, and Who. (Presentation at Museum and the Web conference, San Francisco, April 11-14, 2007.)

1. Place: Each place has a unique web address (URL) and may link to one or more names and zero or more locations (see below). By having a concept of place distinct from location and names, scholars can unambiguously reference a common place entity even if its location is disputed and if it carries multiple appellations. An example is "Byzantium," which has its own unique URL identifier and one location, but multiple names (Constantinopolis, Istanbul, etc.)
2. Location: A physical location for a place may not be known, may be in dispute, or may even change over time (for example, the course of the Euphrates has changed considerably since the 3rd millennium BCE). Thus, locations are distinct from concepts of places or place names in Pleiades.
3. Name: Places may have multiple names. Similarly, the same toponym (place name) may be associated with different places. For example, the place name "Alexandria" refers to different places, including the city on Egypt's Mediterranean coast and a town in Virginia, US.

Individual places, names, and locations in Pleiades have unique and stable URLs, making it easy to reference specific items. Stable links are core to best practice in developing a web-based reference work. In addition to stable linking, Pleiades makes all data available under Creative Commons licenses and in machine-readable formats (Atom + GeoRSS, KML, JSON), so that researchers can programmatically obtain critical geospatial and other data for use in other software especially in "Linked Data" or "Semantic Web" applications. Thus, a key aspect of this project included close collaboration with the Institute for the Study of the Ancient World to build the Gazetteer of the Ancient Near East upon the existing Pleiades software and infrastructure.

The GANE project proposed to incorporate content spanning 12,000 years, from the start of the Natufian period (c. 12,500 BCE) through the end of the Iron Age (c. 600 BCE), at which point it would be contiguous with Pleiades' coverage (Pleiades' chronological coverage begins in the Archaic Period, roughly corresponding with the Iron Age in the Near East). The geographic span was the Ancient Near East, from Cyprus to Iraq and from Turkey to Jordan. This geographic region was selected because UCLA's Encyclopedia of Egyptology is developing a similar resource for Egyptology, but, to the best of our knowledge no such comparative digital resource exists for the Levant or Mesopotamia.

The GANE project draws on multiple extant web resources, and thus involves a high degree of coordination across project teams. Over the course of the project, Project Director Eric Kansa and Content Editor Francis Deblauwe had frequent conversations with Pleiades team members Tom Elliott and Sean Gillies about how to reconcile the different models used by Pleiades and the GANE resources (see below) for describing places. Feedback has given Pleiades a better sense of future requirements related to representing multiple languages in Pleiades.

3. ACCOMPLISHMENTS

NEH funds supported the 18-month development period for the Gazetteer of the Ancient Near East. A goal of the GANE project was to serve as an example of the potential of Linked Open Data to improve scholarship on the Web. We have achieved this goal by successfully preparing 1,372 place records and an additional 3,777 place-names (plus 2,988 romanization variants) for the Pleiades gazetteer. In addition, we have prepared a dataset of approximately 65,000 additional place names to be added in the future.

The work reported here draws on two rich resources: the Tübingen Atlas of the Near and Middle East (TAVO), whose digitization yielded an unexpectedly massive database that, though not yet fully ready for incorporation into Pleiades, can be invaluable to future researchers; and the West Bank and East Jerusalem Searchable Map (WBEJSM), a smaller but richly-documented dataset that highlights how Linked Open Data can connect places with relevant scholarship about those places. In this section, we describe these two resources, the process by which we incorporated them into the GANE project, and the current disposition of their content with regards to the ultimate goal of linking them to Pleiades.

The data prepared from the TAVO index covered paleolithic through modern periods, with a geographic extent ranging from Egypt to the Iranian plateau. The WBEJSM dataset had similar chronological coverage, but as its name indicates, focused exclusively on the West Bank of the Jordan River, in territories now under control of the Palestinian Authority and modern Israeli state.

3.1 Data from the Tübingen Atlas of the Near and Middle East (TAVO)

3.1.1 Relevance to the GANE Project

We identified the TAVO as the richest source of place names, site descriptions, and map locations for this project. The 3-volume TAVO index consists of over 1,900 pages with approximately 40 places or place-names on each page. The high level of consistency in the printed TAVO index records made the atlas even more appealing as a resource for the GANE project. The consistency allowed us to efficiently transform raw optical character recognition (OCR) results of the volumes into structured data. Such efficient data processing promised to vastly increase the speed at which we could enter place names (rather than entering all place data manually).

3.1.2 Process

1. Scanning and OCR: The Internet Archive scanned all three volumes[2], resulting in 5.4GB of images representing all 1,900+ pages of the TAVO atlas in double-column format (Figure 2). The double-column format of the atlas pages required us to establish a method to preprocess the images to standardize them before we could perform OCR. To do this, we set up an automated process to break apart each page into two columns. We performed OCR using OmniPage, a proprietary OCR program that can handle many of the linguistic diacritica in TAVO atlas (Figure 3). Although OmniPage captured some diacritica, the OCR results still required substantial manual editing.
2. Scraping: Project Director Kansa wrote a custom scraper to translate the OCR text into structured data to load into a MySQL database. One problem we encountered in this process is that OmniPage is too "smart" and attempts to format OCR output. This formatting cannot be overridden and it results in jumbled line positioning of the OCR data, making it more difficult to parse and load into our database. Since this unwanted formatting cannot be overridden, we had to perform the clean-up manually. This delayed later data clean up and preparation for Pleiades.
3. Editing: Once the scanning, preprocessing, and OCR tasks were completed, Content Editor Deblauwe compared the OCR data against the original scans and performed a rough correction of recognition errors, including but not limited to merged columns and misinterpreted diacritics. Editing also included place names in the major languages of the general region as well as sub-regions (Hebrew, Arabic, Turkish, French, English, German, and others as applicable, such as Persian and Kurdish and transliterated ancient languages, such as Akkadian)[3]. Kansa developed a simple web-based interface for Deblauwe to conduct this editing, author new romanization variants, and perform other quality control work (Figure 4). The interface allowed Deblauwe to link together place-name references in the TAVO into the more abstract notion of "place" that organizes Pleiades data. The interface also includes a simple JSON web-service that Kansa designed in collaboration with Sean Gillies, the software developer for Pleiades. The JSON service simplifies transfer of cleaned TAVO data to Pleiades. We will maintain our TAVO web interface and web-service for at least one year beyond the project period to enable continued clean up of the TAVO data.
4. Prioritizing entries: Due to the sheer amount of data and difficult data quality challenges, we chose to prioritize place records that have the most metadata (that is, the places with a long

[2] The TAVO 3-volume set cost $66.25. Scanning all three volumes cost the project $197.00. The project donated the hard copy to the Internet Archive, which allows them to loan out the digital copy (legally), to other members of the community. This digitization contributes substantially to the digital infrastructure of the humanities.

[3] Pleiades is fully UTF-8 compliant, meaning it can manage international fonts and character sets.

temporal duration, with many map references, and multiple languages). While this means the work was slower at first, it populated the gazetteer with the most richly-documented and, thus, most significant, places (Figure 5). Therefore, though we did not add many new "places" to Pleiades, we focused efforts on toponymns and romanizations and successfully processed the most complex and metadata-rich records that required the most specialized expertise (Deblauwe's domain knowledge) to edit.

5. Sharing the TAVO dataset: The TAVO index yielded a massive dataset, far beyond our capacity to fully clean and prepare for incorporation into Pleiades. Nevertheless, we invested substantial effort in digitizing, OCRing, and parsing / scraping the TAVO index to create structured data. Though the dataset requires additional cleanup, it can be an invaluable resource to other research efforts. To this end, we released the TAVO derived data via GitHub openly under a CC-Zero public domain dedication. To avoid copyright infringement, we carefully redacted information needed to fully reconstruct the original text. The full TAVO dataset is available on GitHub at: https://github.com/ekansa/gane.

3.1.3 Results

The TAVO dataset available in GitHub includes:

- Approximate latitude / longitude coordinates
- Feature type (e.g., archaeological site, region, geographic features)
- Associated time periods, including chronological periods (e.g., Chalcolithic, Early Bronze I) and general date ranges, as defined in the previously published reference gazetteers
- Original page number from the TAVO, as well as identifiers for different lines of text and different index entries (each entry has associated toponymns, map references, feature types, language codes, and geospatial data)
- TAVO language codes reconciled with the ISO standard
- Toponymns

At the end of the project period, the place records edited and uploaded to Pleiades represent the most significant, complex, and challenging TAVO place records to edit. We measured significance and complexity by the number of TAVO map references associated with a given place-name record. Records with more map references had more metadata associated, likely reflecting greater significance. Making such determinations was only possible because we had generated structured data from the raw OCR of the TAVO.

Much of the effort in this project focused on correcting OCR errors. In addition, Deblauwe identified and needed to resolve occasional inconsistencies in the TAVO index. Furthermore, some of the TAVO records (about 7% of toponyms, 60% of "place" concepts) already had Pleiades identifiers (that is, they already exist in Pleiades), so the TAVO dataset had to be reconciled with the Pleiades dataset. The user interface created for this work allows Deblauwe to associate TAVO records with Pleiades records (Figure 4). In those cases, the TAVO-derived data supplemented and extended existing Pleiades places records with additional toponymns, romanizations, and metadata.

We should also note that the reason a high percentage of the TAVO dataset could be associated with existing Pleiades place URIs stems from our prioritization of "significant" places from the TAVO. As discussed above, we focused data clean-up efforts on the TAVO records with the most associated metadata. It is not surprising therefore that such significant places where already in Pleiades. In this sense, the main contribution of the GANE project with respect Pleiades and the TAVO is the addition of much more metadata and toponymn records (we estimate only about 7% of TAVO toponymns where already in Pleiades) to existing Pleiades places, rather than the creation of new Pleiades place records.

Furthermore, the lack of precision in the geospatial information in the TAVO index further justifies our decision to focus efforts on topoymns and romanizations rather than on creating new "places." The TAVO index only offered coordinate information to a maximum of 2 significant digits. Originally, Pleiades had similar levels of precision in its geospatial information, since most of Pleiades derived from the paper maps of the Barrington Atlas. It has taken years of collaborative curation

work to improve the geospatial data in Pleiades, and such efforts will need to be continued to improve the geospatial data deriving from the TAVO index. Thus, we decided that emphasizing the linguistic and toponymn data from the TAVO, rather than the geospatial data, would be much more immediately useful to the wider research community.

In addition to extending Pleiades with additional place and place-name records from the TAVO, Deblauwe also collected rich metadata further documenting the data derived from the TAVO. Deblauwe developed a dataset about TAVO's chronological periodization and reconciled TAVO's language codes with the ISO standard set of language codes. He also provided additional links to Wikipedia entries and other Linked Data resources to further enhance the utility of the TAVO derived data for Linked Data applications.

Because the TAVO index has a different organizational structure than Pleiades, it is difficult to estimate the number of "places" remaining in the TAVO that have not yet been edited, cleaned, and prepared for Pleiades. There are some 20,276 unique sets of geographic coordinates that we scraped from the TAVO, and some 90,391 unique toponyms. Thus far, we have cleaned and at least partially edited about 15,880 toponyms. These numbers demonstrate that the scale of the TAVO rivals the scale of Pleiades itself (with 27,000 toponymns and 35,000 places). Pleiades has been in continuous development and expansion since 2006. Thus, full incorporation of the TAVO data into Pleiades will require significant continued effort (see section 6. Grant Products & Continuation of the Project).

The less significant place records digitized from the TAVO (but not edited to date) are available as open data (under a Creative Commons Zero – Public Domain dedication) for future use beyond the period of this startup grant. All unedited places have been marked as such and the database archived with the California Digital Library so others can continue this work. The data correction form has been shared (via GitHub) and linked to the original TAVO index scans hosted by the Internet Archive. Unfortunately, we cannot make the raw scans available open access, since the TAVO is still under copyright. The Internet Archive, however, does make these scans available for limited distribution via inter-library loan.

3.2 Data from the West Bank and East Jerusalem Searchable Map (WBEJSM)

3.2.1 Relevance to the GANE Project

The West Bank and East Jerusalem Searchable Map (WBEJSM)[4] was created to document archaeological sites surveyed or excavated since Israel occupied the West Bank and East Jerusalem in 1967 (Figure 6). The system provides information for over 6,000 survey records. This information may include site name(s), location, period of occupation, major components (such as village, tomb, church), the names of the excavators or surveyors who gathered data about the site, and bibliographic information relevant to the site. However, many of the survey records only have arbitrary identifiers and some chronology metadata. Thus, they lack sufficient information to qualify as Pleiades "places," so we did not prepare these for import to Pleiades. Instead, we focused on the WBEJSM records with toponymn and bibliographic information. Working closely with GANE project consultant Adi Keinan, a principal developer of the WBEJSM who specializes in GIS, we prepared some 980 place records and 1,345 toponymns for Pleiades. These data were supplemented with rich chronological and bibliographic information. Most effort focused on finding and adding stable web URI information for the bibliographic records associated with WBEJSM places. Stable URIs are a fundamental aspect of any Linked Data resource, and stable URIs for the bibliographic resources enable link information about places with associated scholarship about places.

3.2.2 Process

Adding place data from the WBEJSM involved using ArcGIS to change the WBEJSM coordinates to the WGS 84 coordinate system (the standard used by Pleiades and most web-based mapping systems). Part of the task of preparing the 980 (qualifying) WBEJSM places to Pleiades was determining which sites are already represented in Pleiades and linking them to the sites in the WBEJSM. We employed

[4] http://digitallibrary.usc.edu/wbarc/map.html

Google Refine to check on the consistency of the WBEJSM data and to associate WBEJSM places with existing Pleiades places using the Pleiades reconciliation service. About 8% of the WBEJSM records can be associated with existing Pleiades URIs. The remainder of the WBEJSM dataset appears to consist of entirely new (to Pleiades) place records.

3.2.3 Results

Until its incorporation into the GANE project, the WBEJSM was a stand-alone resource. By the project's completion, we finalized 980 place records and 1,345 toponymns from the WBEJSM for incorporation into Pleiades. Linking these resources means that the WBEJSM data will be vastly more discoverable and can be further documented and enriched through linking with other open web resources in future digital humanities initiatives.

4. CHANGES TO THE PROJECT

The following beneficial change to the project plan occurred early on in the project and was reported in the first interim report (March 31, 2012):

- TAVO digitization: Scanning the TAVO atlas and performing OCR on the c. 1,900 pages of the TAVO atlas allowed for vastly more sites to be documented during this project (up to ten times the original number proposed). Though correcting the OCR takes some time, this is time that would have been spent entering place names manually, so this involved a simple revision of tasks rather than a significant change requiring additional funding or personnel.
- TAVO data cleanup: The most significant change to the project involved the development of a web-based interface to clean and edit data obtained by digitizing the TAVO index. We tried to limit software development costs as much as possible. Indeed, this was one of our main motivations for using Pleiades, a well established and existing platform. However, because the TAVO dataset turned out to be massive, highly complex, and requiring a great deal of cleanup, we had to develop a specialized interface for cleaning the data. The time it took for us to develop this interface and for Content Editor Deblauwe to use it to clean the data took significant amount of the budgeted time for the project. Thus, the GANE project mainly contributed toponymns, romanizations and metadata to enrich existing Pleiades "places," rather than creating wholly new places. In order to continue the work and add new place-entities to Pleiades, we have made arrangements with colleagues at academic institutions to enlist student help in continued cleanup of the TAVO data.
- Changes in approach to dissemination: Because of the continued need to clean TAVO derived data, we decided to share the remaining TAVO dataset via the GitHub version control system (see above).

5. AUDIENCE & EVALUATION

The results of this project have a wide audience—essentially, anyone with access to the Web seeking information about places. The body of people who actually implement the Pleiades place names (that is, the active users of the code) is growing. Pleiades has already done much to spark the development of Linked Open Data resources and systems for Classical archaeology and history. The Pelagios project (http://pelagios-project.blogspot.com/), funded by the JISC (United Kingdom) and the Mellon Foundation, is probably the key example of how the Pleiades Gazetteer unites more than a dozen major museum, digital archive, and digital library collections in Classics. The addition of Ancient Near East toponymns and chronological metadata will be an invaluable contribution in encouraging the growth of Linked Open Data resources for the archaeology and history of the Ancient Near East.

We have presented the GANE project in a variety of conference forums and online, via the project webpage, the Heritage Bytes weblog and via discussion forums in Pleiades.

- Project Webpage: Updates to the GANE project are posted on the project webpage at: http://alexandriaarchive.org/projects/gane/.
- Presentation: American Schools of Oriental Research Annual Meeting (November 2011, San Francisco): Eric Kansa (AAI) and Charles E.

Jones (NYU / ISAW) gave presentations about the GANE project and Pleiades in the forum *Topics in Cyberinfrastructure, Digital Humanities, and Near Eastern Archaeology (I)*.

- Presentation: American Schools of Oriental Research Annual Meeting (November 2012, Chicago): Eric Kansa and Sarah Whitcher Kansa presented on the GANE project and how it relates to publishing Linked Open Data with Open Context. This presentation "From Data to Knowledge: Organization, Publication, and Research Outcomes" occurred in the forum *Topics in Cyberinfrastructure, Digital Humanities, and Near Eastern Archaeology (II)*.
- Presentation: American Schools of Oriental Research Annual Meeting (November 2013, Baltimore): Eric will discuss the TAVO dataset, Pleiades, and the potential application of the GANE project outcomes to facilitate text mining in Biblical studies and other areas of Near Eastern studies in the forum *Topics in Cyberinfrastructure, Digital Humanities, and Near Eastern Archaeology (III)*.
- Blog post (Heritage Bytes): "The Red Sea Is Arabian, Erythraean, …" (http://ux.opencontext.org/blog/2012/09/12/the-red-sea-is-arabian-erythraean/)
- Blog post (Pleiades): "Sorting out the Red Sea" (http://pleiades.stoa.org/docs/content-development-projects/sorting-out-the-red-sea).

6. Grant Products & Continuation of the Project

Digitization of the TAVO atlas resulted in a dataset of over 74,000 place names. This project focused on the most complex places in the TAVO, resulting in approximately 8,000 well-described places for incorporation into the Pleiades gazetteer. Because the remaining 65,000+ toponymns in the TAVO dataset still require review and clean-up, we have focused our efforts on finding ways to continue the project now that grant period is over. All content related to this project benefits from digital archiving services provided by the California Digital Library (CDL). The DOI for the TAVO dataset is: doi:10.6078/M7QJ7F7K.

In addition, we continue to host and maintain the TAVO editing interface for use by groups interested in contributing to the documentation effort. We are pleased to have commitments from colleagues who will enlist the help of undergraduates to continue clean up of the TAVO-derived data in the context of digital humanities courses. The TAVO dataset will provide the students with an excellent opportunity to explore the complexities of working with different data models all working version control systems like GitHub, all while contributing invaluable data to Pleiades.

7. Long-Term Impact

Future applications of an open Gazetteer of the Ancient Near East include innovative research programs using text-mining and Semantic Web technologies. For example, Pleiades is an integral part of a number of collaborative digital humanities initiatives, including the Pelagios project, to establish Linked Data models and vocabularies for use in historical and archaeological studies. By building on Pleiades, the GANE project immediately benefits from Pleiades' established international collaborations. In addition, Project Director Eric Kansa is a co-investigator with the Google-funded Google Ancient Places[5] (GAP) project and is currently working to use the Pleiades gazetteer of Classical period places to index some of the millions of books digitized by Google. GAP uses standard techniques in text-mining to find references to ancient places compiled by the Pleiades gazetteer in scanned books. In doing so, GAP develops map-based interfaces that allow users to zoom in specific places on a digital map and then see all books related to that place. The project is also exploring ways of visualizing possible non-spatial conceptual relationships that may be expressed in the literature.

GAP helps illustrate the significance of the GANE project's contributions of toponymns and romanizations. The additional toponymns and romanizations of significant places will greatly improve use of Pleiades for applications in the archaeology and history of the Ancient Near East in the following ways:

[5] http://googleancientplaces.wordpress.com/gapvis/ and http://googleancientplaces.wordpress.com/

- Pleiades search interface: Scholars of the Near East will be more likely to reference Pleiades because the new toponymns (in Near Eastern languages) will be available to query, making it easier to find records of places. The different romanizations Deblauwe provided help mitigate some of the difficulties in querying for Near Eastern toponymns using the Latin alphabet.
- Entity reconciliation: Similarly, the additional toponymns and romanizations will make it easier to apply automated and semi-automated techniques to look up Pleiades identifiers to gazetteer records. The additional toponymns and romanizations should greatly improve Pleaides' performance in responding to reconciliation requests relating to Near Eastern places.
- Entity identification: The additional toponymns and romanizations are invaluable to text-mining projects such as GAP. The additional toponymns and romanizations will facilitate string matching, and enable a wider corpus of texts to be related to the Pleiades gazetteer.

By expanding the scope of Pleiades to the Ancient Near East, this project enables similar text-mining initiatives on corpora of literature relevant to Assyriology, Hittitology, Biblical studies, and Ancient Near Eastern archaeology. Furthermore, the frequent occurrence of multi-period occupations at many of the sites included in this project resulted in the documentation of many places from the Islamic period. While this falls outside of the scope of this project, these results will be useful to Middle Eastern Studies, in general, and to Islamic studies, in particular. Beyond text-mining, this project provides needed common points of reference that enable researchers to link museum collections, cuneiform archives, and archaeological datasets together. Finally, by openly releasing all data, free-of-charge, and free of copyright or other intellectual property restrictions, this project serves as an urgently needed benchmark for making Ancient Near Eastern scholarship more open, transparent, and collaborative.

Figure 1: *A Pleiades entry, showing the various attestations for the place "Adana"*

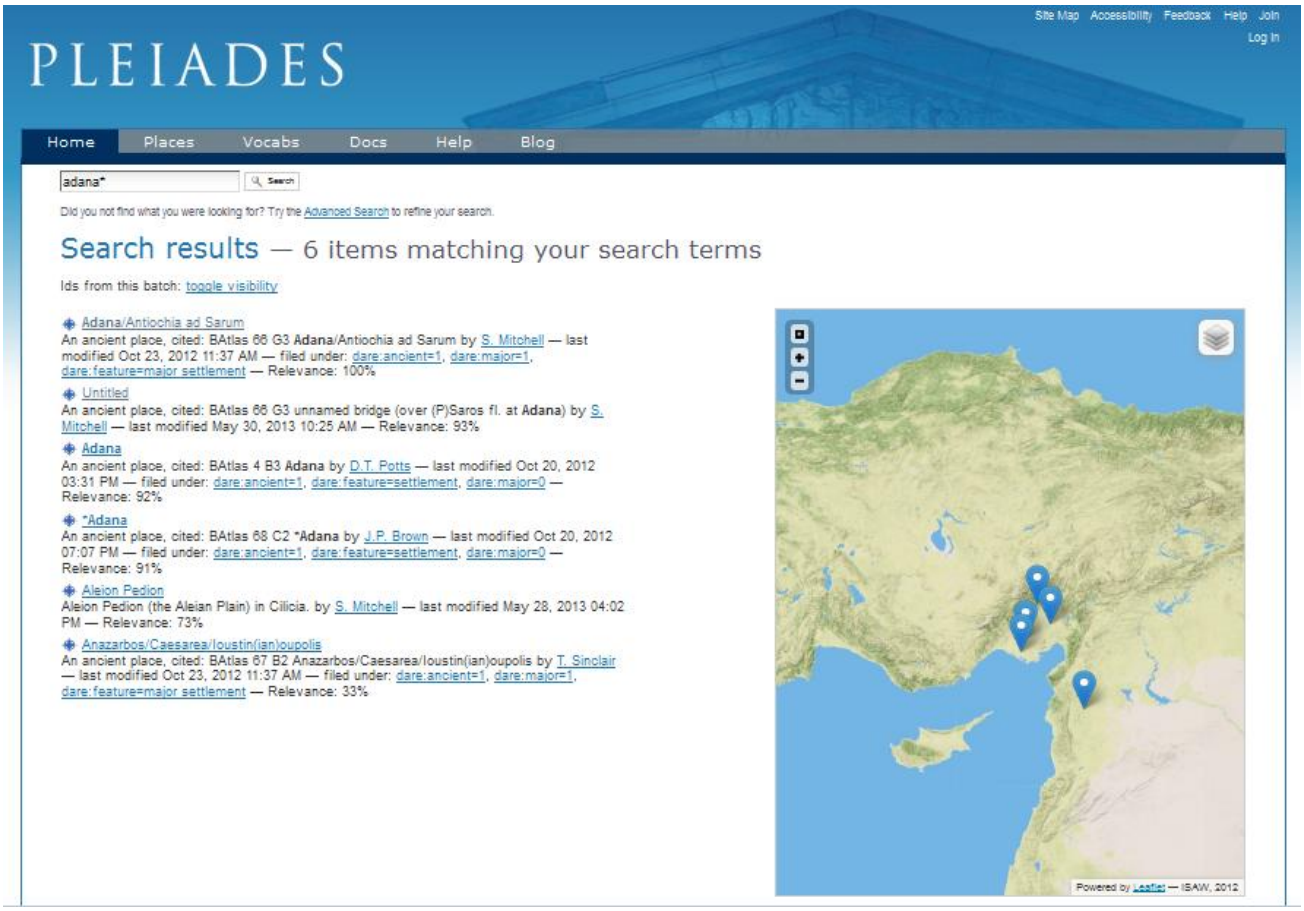

Figure 2: *A page from the scanned TAVO index. The text highlighted in gray indicates all names and map references pertaining to the place "Adana"*

Figure 3: *Example of how the portion of the TAVO index page with references to the place "Adana" looked after undergoing OCR*

```
Adan Siedl. 12.45/45.00
AVI1
- c Adan
c Adan Siedl. 12.45/45.00 ar
A 1 1, A II 2, A II 3, A II 6, A III 5, A IV 1, A IV 2, A IV 3,
A IV 4, A V 1, A VII 1, A VIII 1, A VIII 13, A IX 1, A IX
5, A IX 6, A X 1, A X 11, A X 18.1, A X 18.2, A X 19, B V
22, B VI 1, B VI 7, B VII 1, B VII 3.2, B VII 7, B VIII 1, B
IX 1, B IX 2, B IX 7, B IX 8, B IX 15, B IX 22, B IX 23, B
X 4, B X 6, B X 11.1, B X 11.2,
-»Aden, Arabia Eudaimön, Eudaimönes Polis, Madlnat
as-Sai), Nesoi Eudaimönes
-Adan (A VI 1), Aden (B X 1)
c Adan Admin. ar
A VIII 1 (12.30/45.00)
--Madmat as-Sat
Adana Siedl. 37.00/35.20 os/ tu/
ar/ la
A 1 1, A I 2, A II 2, A II 4, A III 2, A IV 1, A IV 2, A IV 3,
A IV 4, A V 1, A VI 1, A VI 4, A VII 2, A VIII 1, A VIII 2,
A VIII 3, A VIII 8, A VIII 10, A VIII 13, A VIII 14, A IX
I, A IX 4, A IX 5, A IX 6, A X 1, A X 2, A X 6, A X 18.1,
A X 18.2, A X 19, A X 20.2, B II 14, B III 6, B IV 13, B IV
14, B IV 23, B V 3, B V 4, B V 5, B V 6, B V 7, B V 9, B V
II, B V 12, B V 13, B V 15.1, B V 16.1, B VI 1, B VI 2, B
VI 4, B VI 8, B VI 11, B VII 10, B VII 19, B VIII 8, B VIII
10, B VIII 12, B VIII 15, B VIII 19.1, B IX 2, B IX 3, B IX
7, B IX 8, B IX 9, B IX 10, B IX 11, B IX 12, B IX 13, B IX
15, B IX 24, B IX 25, B X 1, B X 2, B X 3
-»Adana, Antiocheia, Antiocheia am Saros, Antiocheia
```

Figure 4: *View of a place record for Adana, showing that it is fully documented and edited, is linked to an appropriate place record in Pleiades, and is ready for export to Pleiades.*

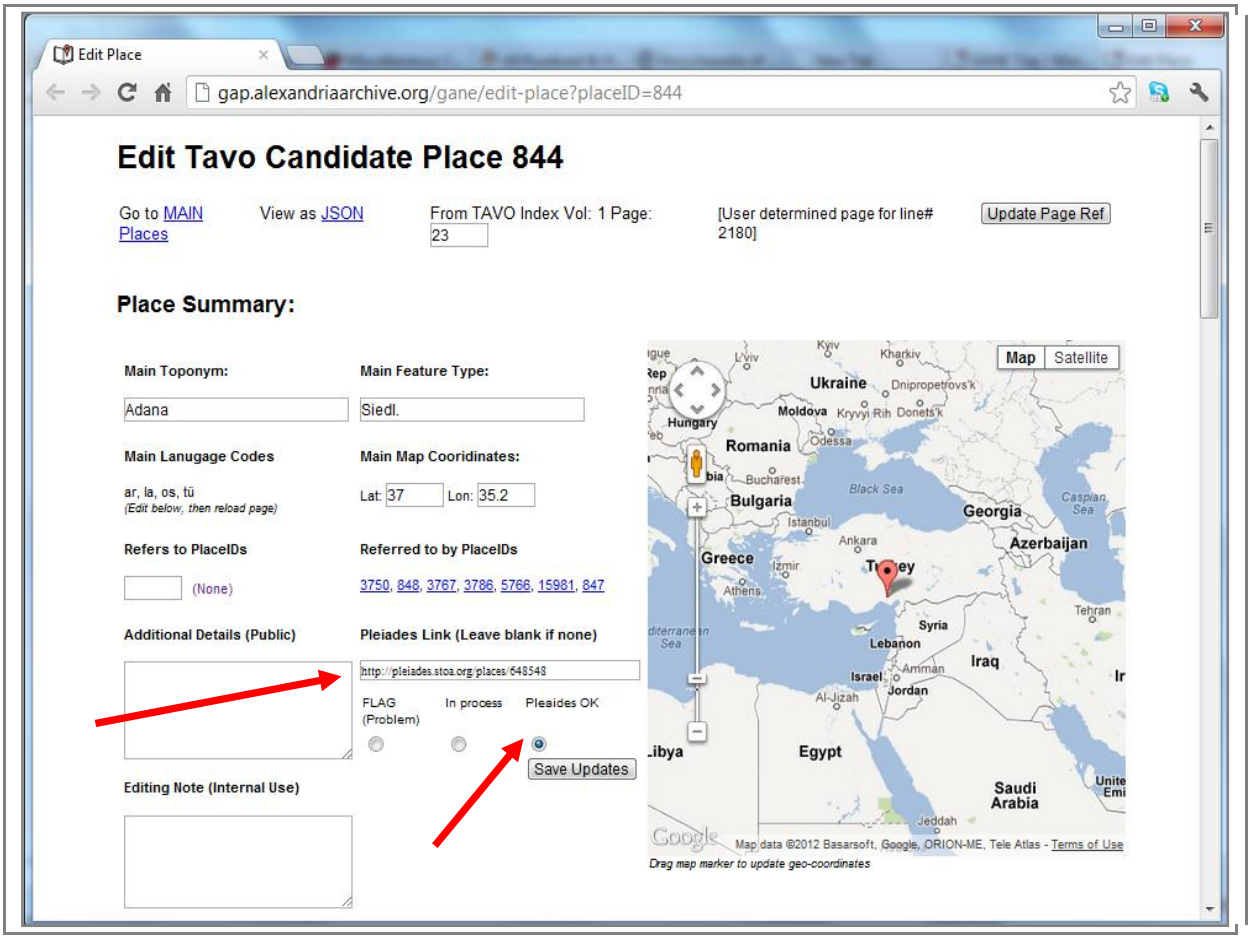

Figure 5: *View of the database for editing site metadata for the GANE project. The Content Editor prioritized sites based on their metadata richness, which were used as proxy evidence for scholarly significance. This web-based interface illustrates place records ranked in metadata richness (from greatest to least). Color-coding helps determine which tasks remain.*

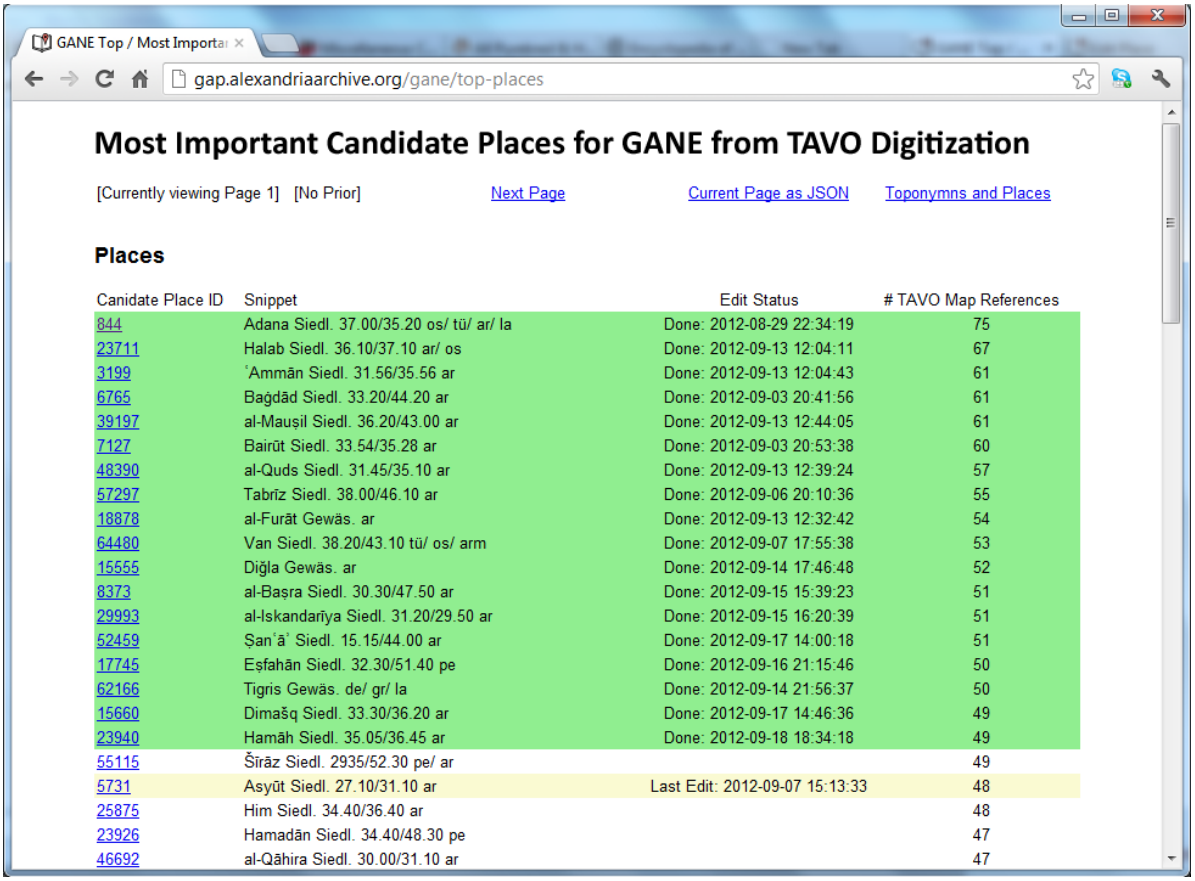

Figure 6: *The West Bank and East Jerusalem Searchable Map*