
Citizen Science for Cuneiform Studies

Terhi Nurmikko

Web Science Doctoral Training Centre
University of Southampton
Southampton, UK SO17 1BJ
tmtn1g10@soton.ac.uk

Dr Jacob Dahl

Faculty of Oriental Studies
University of Oxford
Oxford, UK OX1 2LE
jacob.dahl@orinst.ox.ac.uk

Dr Nicholas Gibbins

Electronics and Computer Science
University of Southampton
Southampton, SO17 1BJ, UK
nmg@ecs.soton.ac.uk

Dr Graeme Earl

Archaeological Computing Research Group
Faculty of Humanities
University of Southampton
Southampton, SO17 1BF, UK
graeme.earl@soton.ac.uk

Copyright is held by the author.

WebSci 2012, June 22–24, 2012, Evanston, Illinois, USA.

ACM 978-1-4503-1228-8

Abstract

This paper examines the potential applications of Citizen Science and Open Linked Data within a critical Web Science framework. Described here is a work-in-process concerning an interdisciplinary, multi-institutional project for the digitization, annotation and online dissemination of a large corpus of written material from ancient Mesopotamia. The paper includes an outline of the problems presented by a large, heterogeneous and incomplete dataset, as well as a discussion of the potential of Citizen Science as a potential solution, combining both technical and social aspects. Drawing inspiration from other successful Citizen Science projects, the current paper suggests a process for capturing and enriching the data in ways which can address not only the challenges of the current data set, but also similar issues arising elsewhere on the wider Web.

Keywords

Digital heritage, citizen science, palaeography, cuneiform, ontology, archaeology, museums, Mesopotamia

ACM Classification Keywords

I.7.m Document and Text Processing: Miscellaneous

General Terms

Documentation, Theory.



figure 1. Example of an administrative tablet from the Old Akkadian period. Part of the collection of the Oriental Institute, University of Chicago. CDLI number P217507. [7]

Introduction

This paper is a discussion of on-going research that covers two extremes on the spectrum of written expression: cutting-edge semantic technologies for online publication and one of the oldest known examples of writing anywhere in the world. In this paper, a specialist community of practice, which is intrinsically linked to complicated information systems and to data riddled with ambiguity (Assyriology), is examined. The ways in which Citizen Science could help solve the challenges of the data set, disseminate information about the discipline to a wider audience and result in processes of learning and enjoyment for the volunteers are discussed. The question is not only which form the future online presence of Assyriology will take, but also how technological developments might facilitate changes in the discipline of Assyriology.

The project proposed in this paper builds on large scale, multi-institutional and interdisciplinary resources for the digitization and online dissemination of the archives of the ancient Near East, such as the Pennsylvania Sumerian dictionary (ePSD) [10], the Electronic Text Corpus of Sumerian Literature (ETCSL) [11] and the Cuneiform Digital Library Initiative (CDLI) [7]. Our aim is to identify possible ways that these currently largely unconnected but complementary projects could be merged with minimal or no disruption or reduplication of data, consciously stepping away from any form of restricted access. We suggest that sharing data in non-proprietary formats [18] will allow for new and innovative applications, and harnessing the potential of Citizen Science is key to disseminating and enriching cuneiform data.

Challenges of the Dataset

The dataset is a vast, heterogeneous corpus written in cuneiform, a script widely used in various forms across the Mesopotamian plateau for some three and a half millennia and for a number of different languages. Surviving tablets (some 500,000 currently known) cover a myriad of different genres ranging from creative literature to astronomy, cultic and socio-political documentation, administration, mathematics, horticulture, law codes and more.

This syllabic script owes its name to a characteristic appearance, resulting from the pressing a reed stylus into a tablet of wet clay. The correct reading of a given sign is highly context-dependent: the glyph **ŠEŠ**, for example, used to write the noun “standard”, when combined with **UNUG** and **KI** refers to the city of Ur. If preceded by the determinative **dingir** (written with the **AN** sign) and **KI**, the sign becomes the personal name of the god Nanna. Reading is further complicated by the general absence of both punctuation and spacing between separate lexical units.

Glyphs	Name of sign	Reads as	Meaning
	ŠEŠ	šeš	flagpole, brother
	ŠEŠ.UNUG. ^{KI}	uri ₂	city of Ur
	^d ŠEŠ.KI	^d Nanna	god Nanna

figure 2. The glyph **ŠEŠ** in different combinations






archaic Sumerian pictograph		c. 3500 BC – 2900 BC
Classical Sumerian glyph	or 	c. 2900 BC – 2200 BC
Middle Assyrian	or 	c. 1500 BC – 1000 BC
Neo-Assyrian	or 	c. 1000 BC – 600 BC
Unicode		c. 2000 AD

figure 3. The changing appearance of **muš₃** [17] and the appropriate Unicode character [12].

Over millennia the script went through a number of changes. Symbols of pictographic origin became increasingly abstract and polyvalent in nature - each individual sign may have a number of distinctly separate meanings and phonetic values. In addition to these typographical changes, signs could also experience semantic shifts. The example glyph **ŠEŠ** eventually became to mean “brother”.

The vast majority of cuneiform inscriptions are on fragile clay tablets. Incomplete and broken tablets are a common occurrence, and the history of Assyriology has resulted in the division of material from a single archaeological site to two or more different institutions. Since small fragments are unlikely to be transported across large distances, some broken tablets might only ever be pieced together in their digitized form on the Web.

Limitations of Existing Approaches

Although there are a number of existing online resources, a coherent, universal system allowing open, equal and simultaneous access to all of them is yet to be implemented.

Publishing Object Records Online

The willingness of many institutions to make their collections available online is a positive and increasingly popular trend in Digital Heritage. Examples include publications without copyright restrictions and the adoption of semantic technologies such as the SPARQL query point for the British Museum’s Collections Online [5]. The Open Richly Annotated Cuneiform Corpus (Oracc) [20] and those projects that link to it are an example of the adoption of open standards in the publication of new research in the field of Assyriology.

Unicode

The multitude of meanings, associations, semantic connotations and oft-used bilingual (Sumerian-Akkadian) translations linked to the polyvalence of signs means that projects such as the Unicode font are a technological reductionist approach to solve a complex and multifaceted problem. Although created with close consultation with preeminent Assyriologists and based on well-known sign lists [12, 4] the Unicode characters are limited to representing only a snippet in the timeline of palaeographical and typographical changes (as illustrated in figure 3.). Using the Unicode chart is essentially limited to a bi-directional search of number and visual representation – finding the correct sign requires extensive knowledge of the glyph-base or the names of various signs. Searching by meaning is currently not possible.

Palaeographical and Linguistic Tools

There are a number of projects aiming to aid in the translation and reading of the content of these ancient texts, such as the ePSD [10] and the ETCSL [11]. Limited interlinking means that any translation work or scholarship necessitates the use of a number of separate (even off-line) resources and the incessant swapping between tabs or windows. Additional problems include the co-constitution of known literary texts, where parts of compositions are known to originate from a number of different original pieces. The assignment of multiple unique identification numbers to any one object can also further complicate the accurate identification of a specific object. Current research with the ETCSL is seeking to represent such compositions accurately and with access to the original texts.



figure 4. High resolution photograph illustrating a line of text “over-flowing” to the edge of the tablet [7].

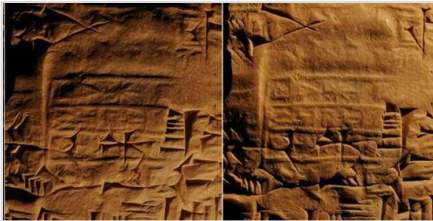


figure 5. Two RTI illustrating the same part of the same document with different light sources.



figure 6. A cuneiform tablet and its 3D model [2].

Digital Image Capture

Cuneiform inscriptions have traditionally been represented in publications as line drawings, in which glyphs on the edges or those effectively “over-flowing” to the obverse side of the tablet are awkwardly shown as extensions of the line. Problems of accurate representation of the script have been cited as the reason behind the push from Assyriologists for the digital capturing of inscriptions [7]. At the same time, this complex 3D script functions well as a case-study example for imaging projects [2]. The last decade has seen several projects in the digital capture of inscriptions, using a number of different approaches such as digital photography [8], reflectance transformation imagery (RTI) [10] and 3D imagery [19].

Ontologies

We have outlined an ontological framework (largely based on the CIDOC Conceptual Reference Model and the Text Encoding Initiative) for linking not only the metadata of objects, but also for identifying and representing explicit relationships between the places, people and events known from the content of the text. Particular attention is awarded to the complexities caused by the spatio-temporal ambiguities and the problems of formally representing them. It is possible to link this framework to general ontologies such as one describing people and interpersonal relationships (FOAF) [13] and to domain-specific ones, such as the one suggested by Jaworski [16] for representing economic tablets from the city of Umma in the Ur III period (c. 2000 BC). It may also become possible to create visualisations based on ontological timelines to represent the spatio-temporal changes taking place at a particular site [15].

URIs

Although tablets published by the CDLI are awarded unique identifiers, these are not yet utilised as URIs for online publication. Furthermore, no URIs have been assigned for the entities identified in the inscriptions. Once established, such data will allow for linking via the content (and not just the object metadata) to other non-Assyriological projects within Digital Humanities such as Pleiades [21] and other data streams. Assigning URIs to locations or events mentioned in the content of the cuneiform inscriptions is particularly complex as clear parameters and start- and end dates are difficult or impossible to assign. This has led to the conclusion that an automated identification of these specific entities is not a realistic approach at this time. Even if the knowledge base is limited to the content of the CDLI, the outlined project will require the analysis of some 1.5 million lines of text [7], thus making the task unlikely to be completed by a relatively small community (no more than 500 specialists world-wide) within a reasonable time-frame.

Citizen Science as a Potential Solution

In this paper, Citizen Science is understood as the practice of engaging the public to aid in scientific research. Where it differs from the perhaps more widely known practice of “crowd-sourcing” is that Citizen Science relies on volunteers who have been trained by experts, and who adhere to scientific research methods when completing their task. Training takes place either prior to or during the completion of the given task, and frameworks are established and put in place to monitor and help reduce error margins. Citizen Science has successfully contributed to several fields of scientific research from astronomy [14] to ornithology [8].

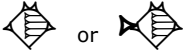



Tasks	Example
Identifying signs, akin to GalazyZoo.	 or 
Identify and record sign combinations, as with ReCaptcha.	
Transcribing cuneiform texts, similar to Old Weather.	 equates to "goddess Inanna"
Linking identified entities in cuneiform records to other data streams.	"the city of Ur" equates to "Tell Muqayyar", located 30° 57' 45.55" N 46° 6' 11.03" E
Identifying specific entities in transliterated texts.	ŠEŠ.UNUG ^{KI} equates to "the city of Ur" equates to "Tell Muqayyar"

figure 7. Examples of different tasks in a descending order of increasing complexity.

Engaging the public in such a task supports the field of Assyriology in ways beyond aiding traditional palaeographical research. It helps promote the discipline and to increasing public awareness of the various collections. A successful approach to such a project is likely to have considered:

- processes of verification and error monitoring
- intrinsic and extrinsic motivations for taking part
- incorporated reward schemes (awards, badges)
- the need for a medium of communication, promoting a sense of community [22, 23].

Citizen Science has already been shown to support transcription projects: Old Weather has successfully engaged over 25,500 volunteers to transcribe more than 870,000 pages since its launch in 2010 [23].

Prior knowledge of these ancient languages is neither necessary nor expected, but those volunteers wishing to work directly with ancient languages could identify signs or sign clusters in a project similar to the transcription of the Oxyrhynchus papyri in Ancient Lives [1]. Those able to identify words in the transliterations could in turn semantically enrich the data by linking relevant lexical items to dictionaries, translations and published papers.

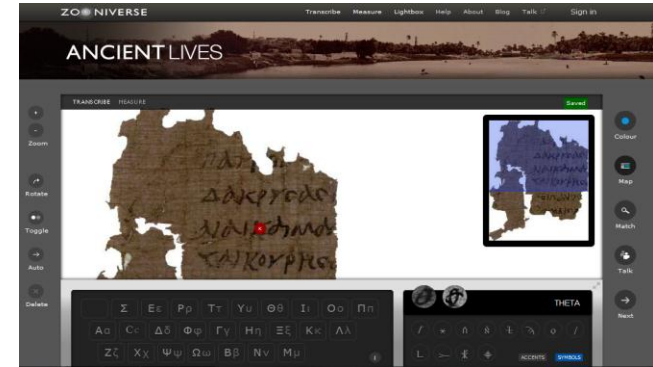


figure 8. The Oxyrhynchus papyri in Ancient Lives [1].

Training could be part of the process, or a compulsory section at the on-set, such as with GalaxyZoo [14]. The process of advancing will be merit-based: A beginner will be expected succeed at correctly identifying the correct sign as accurately as experts before moving on to more complex tasks. Progression to the next level will be dependent on a sufficiently high level of accuracy, ensuring that error margins are minimised.

The future steps of this on-going research include the design, implementation and review of a system that could be used for pilot studies. The results gained from them will in turn help in the implementation and design of the final product. Of the myriad of aims set for this project, one of the most crucial is that of successfully engaging with a vast range of new volunteers, and to spark an interest and awareness of the field beyond the existing sphere of specialists. We strive for a system where, regardless of their level of skill or even their area of interest, all volunteers will be contributing to knitting existing projects together and bringing about a more integrated, interlinked and interconnected Web.

References

- [1] Ancient Lives
<https://www.zooniverse.org/project/ancientlives>.
- [2] Anderson, S.E. and Levoy, M. Unwrapping and Visualising Cuneiform Tablets. In *IEEE Computer Graphics and Applications*, 22, 6, November/December (2002), 82-88.
- [3] Black, J., George, A. and Postgate, N., *A Concise Dictionary of Akkadian*, Harrassowitz Verlag, Weisbaden, Germany, 2000.
- [4] Borger, R. Assyrisch-babylonische Zeichenliste, Band 33. In *Alter Orient und Altes Testament (AOAT)*, Veröffentlichungen zur Kultur und Geschichte des Alten Orients und des Alten Testaments (Series), Kevelaer and Neukirchen-Vluyn (1978).
- [5] British Museum SPAQRL Endpoint
<http://collection.britishmuseum.org/Sparql>.
- [6] Cohen, J., Duncan, D., Snyder, D., Cooper, J., Kumar, S., Hanh, D., Chen, Y., Purnomo, B., and Graettinger, J. iClay: Digitizing Cuneiform. In *The 5th International Symposium on Virtual Reality, Archaeology and Cultural Heritage VAST* (2004).
- [7] Cuneiform Digital Library Initiative
<http://cdli.ucla.edu/>.
- [8] Dickinson, J.L., Zuckerberg, B. and Bonter, D. Citizen Science as an Ecological Research Tool: Challenges and Benefits. In *Annu. Rev.Ecol.Syst.*, 41, (2010), 149-172.
- [9] Earl, G., Beale, G., Martinez, K. and Pagi, H. Polynomial Texture Mapping and Related Imaging Technologies for the Recording, Analysis and Presentation of Archaeological Materials. In *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, XXXVIII, part 5, Commission V Symposium*, Newcastle upon Tyne, UK (2010), 218 – 223.
- [10] Electronic Pennsylvania Sumerian Dictionary
<http://psd.museum.upenn.edu/epsd/index.html>.
- [11] Electronic Text Corpus of Sumerian Literature
<http://etcsl.orinst.ox.ac.uk/>.
- [12] Everson, M., Feuerherm, K. and Tinney, S. Final proposal to encode the Cuneiform script in the SMP of the UC, *working document*
<http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2786.pdf>.
- [13] FOAF <http://www.foaf-project.org/>.
- [14] GalaxyZoo <http://www.galaxyzoo.org/>.
- [15] Hyvönen, E., Tuominen, J., Kauppinen, T. and Väätäinen, J. Representing and Utilizing Changing Historical Places as an Ontology Time Series. In Ashish, N. and Sheth, A.P. (eds), *Geospatial Semantics and the Semantic Web: Foundations, Algorithms and Applications*, Springer (2011).
- [16] Jaworski, W. Ontology-Based Knowledge Discovery from Documents in Natural Language, *PhD thesis*, University of Warsaw, 2009.
- [17] Labat, R. *Manuel d'épigraphie Akkadienne*. Paris, France, 1988.
- [18] Linked Data – Design Issues
<http://www.w3.org/DesignIssues/LinkedData.html>.
- [19] Mara, H., Krömker, S., Jakob, S. and Breuckmann, B. GigaMesh and Gilgamesh – 3D Multiscale Integral Invariant Cuneiform Character Extraction. In *The 11th International Symposium on Virtual Reality, Archaeology and Cultural Heritage VAST*, (2010).
- [20] Open Richly Annotated Cuneiform Corpus
<http://oracc.museum.upenn.edu/index.html>
- [21] Pleiades <http://pleiades.stoa.org/home>.
- [22] Silvertown, J. Learning from Two Citizen Science Case studies: Evolution MegaLab & iSpot. In *2nd London Citizen Cyberscience Summit*, (2012).
- [23] Tokumin, S. No Citizens No Science Lessons from 10 million Contributions. In *2nd London Citizen Cyberscience Summit*, (2012).