THE DIGITAL CLASSICIST 2013

EDITED BY STUART DUNN
& SIMON MAHONY

INSTITUTE OF CLASSICAL STUDIES
SCHOOL OF ADVANCED STUDY UNIVERSITY OF LONDON

THE DIGITAL CLASSICIST 2013

BULLETIN OF THE INSTITUTE OF CLASSICAL STUDIES SUPPLEMENT 122

DIRECTOR & GENERAL EDITOR: JOHN NORTH

DIRECTOR OF PUBLICATIONS: RICHARD SIMPSON

THE
DIGITAL CLASSICIST
2013

EDITED BY
STUART DUNN
AND SIMON MAHONY

INSTITUTE OF CLASSICAL STUDIES
SCHOOL OF ADVANCED STUDY
UNIVERSITY OF LONDON

2013

The cover image is of a torso of Pothos (Roman 1st century BC – 1st century AD) in the Museu Calouste Gulbenkian, Lisbon, Portugal.
Photo © Simon Mahony 2013. All rights reserved.

This PDF edition published in 2019
First published in print in 2013

Available to download free at https://www.humanities-digital-library.org

ISBN: 978-1-905670-80-2 (2019 PDF edition)

DOI: 10.14296/917.9781905670802

ISBN: 978-1-905670-49-9 (2013 paperback edition)

Designed and typeset at the Institute of Classical Studies

This volume is dedicated to the memory of two people whose untimely death marks a great loss, both personally and to our communities.

Elaine Matthews (died 26 June 2011): one of our esteemed contributors, ambassador and advocate of the Digital Humanities and the place there for Classics, thank you for all your many contributions to scholarship, to this volume, and your generous words on the cover of the earlier *Digital Classicist* (Ashgate 2010) volume.

Gerhard Brey (1954-2012): a valued friend, colleague, and collaborator with whom we shared intellectual ideas as well as coffee and biscuits. Gerhard was always willing to seek out new areas of 'interest' and so could be willingly called upon to review chapters in this and the earlier Ashgate volume.

TABLE OF CONTENTS

Modelling

Texts

Infrastructure

ACKNOWLEDGEMENTS

The editors would like to thank the Institute for Classical Studies, and specially the Deputy Director and Administrator Olga Krzyszkowska, for their continued support and generosity in hosting and supporting the Digital Classicist seminars. Thanks are also due to all the members of our community who have presented papers at our seminars and conference panels as well as all those who have come along to listen and support these events.

We are grateful to the following scholars for comments, criticism, and advice on individual chapters in this volume; it is through their input that we are able to ensure the high quality of the final work: Chris Blackwell; Gabriel Bodard; John Bodel; Kalina Bontcheva; Gerhard Brey; Hugh Cayless; Graeme Earl; Michael Fulford; Sebastian Heath; Tim Hill; Kathryn Piquette; Dot Porter; Julian Richards; Matteo Romanello; Charlotte Roueché; Melissa Terras; Notis Toufexis; Charlotte Tupman; Michelle Wienhold.

ABSTRACTS

Andrew Bevan *Travel and interaction in the Greek and Roman World. A review of some computational modelling approaches* pp. 3-24

Inferring dynamic past behaviours from the static archaeological record is always a challenge, but computational and quantitative techniques can be helpful. In particular, they can provide useful insight on patterns of movement and interaction, by better characterising existing archaeological evidence, suggesting simple models of mobile decision-making or proposing expected patterns against which the observed record can be compared. This paper reviews the range of modelling options now available for understanding the movement and interaction behind the archaeological and historical record. There are increasing opportunities not only to pick and choose between different modelling approaches, but also to integrate them in a more theoretically and practically satisfactory way.

Vince Gaffney, Phil Murgatroyd, Bart Craenen, and Georgios Theodoropoulos

'Only individuals': moving the Byzantine army to Manzikert pp. 25-43

Traditionally, history has frequently emphasized the role of the 'Great Man or Woman', who may achieve greatness, or notoriety, through the consequences of their decisions. More problematic is the historical treatment of the mass of the population. Agent-based modelling is a computer simulation technique that can not only help identify key interactions that contribute to large scale patterns but also add detail to our understanding of the effects of all contributors to a system, not just those at the top. The Medieval Warfare on the Grid project has been using agent-based models to examine the march of the Byzantine army across Anatolia to Manzikert in AD 1071. This article describes the movement model used to simulate the army and the historical sources on which it was based. It also explains why novel route planning algorithms were required in order to surmount problems with standard solutions.

Elton Barker, Leif Isaksen, Nick Rabinowitz, Stefan Bouzarovski, and Chris Pelling

On using digital resources for the study of an ancient text: the case of Herodotus's 'Histories' pp. 45-62

Involving the collaboration of researchers from Classics, Geography, and Archaeological Computing, and supported by funding from the AHRC, *Hestia* aims to enrich contemporary discussions of space by developing an innovative methodology for the study of an ancient narrative, Herodotus's *Histories*. Using the latest digital technology in combination with close textual study, we investigate the geographical concepts through which Herodotus describes the conflict between Greeks and Persians. Our findings nuance the customary topographical vision of an east versus west polarity by drawing attention to the topological network culture that criss-crosses the two, and develop the means of bringing that world to a mass audience via the internet.

-In this chapter we discuss three main digital aspects to the project: the data capture of place-names in Herodotus; their visualization and dissemination using the web-mapping technologies of GIS, Google Earth, and Timemap; and the interrogation of the relationships that Herodotus draws between different geographical concepts using the digital resources at our disposal. Our concern will be to set out in some detail the digital basis to our methodology and the technologies that we have been exploiting, as well as the problems that we have encountered, in the hope of contributing not only to a more complex picture of space in Herodotus but also to a basis for future digital projects across the Humanities that spatially visualize large text-based corpora. With this in mind we end with a brief discussion of some of the ways in which this study is being developed, with assistance from research grants from the Google Digital Humanities Awards Program and JISC.

Marco Büchler, Annette Geßner, Monica Berti, and Thomas Eckart
Measuring the Influence of a Work by Text Re-Use pp. 63-79

Over the centuries an incredible amount of ancient Greek texts have been written. Some of these texts still exist today whereas other works are lost or are available only as fragments. Without considering intentional destruction, one major question remains: why did some texts remain and others get lost? The aim of this chapter is to investigate this topic by trying to determine the influence of certain ancient Greek works through detecting text re-use of these works. Text re-use measures if and how an author quotes other authors and in this chapter we differentiate between *re-use coverage* and *re-use temperature*.

Tobias Blanke, Mark Hedges, and Shrija Rajbhandari
Towards a virtual data centre for Classics pp. 81-90

A wide variety of digital resources have been created by researchers in the Classics. These tend to focus on specific topics that reflect the interests of their creators; nevertheless they are of utility for a much broader range of research, and would be more so if they could be linked up in a way that allowed them to be explored as a single data landscape. However, while the resources may be reusable, the variety of data representations and formats used militates against such an integrated view. We describe two case studies that address this issue of interoperability by creating virtual resources that are independent of the underlying data structures and storage systems, thus allowing heterogeneous resources to be treated in a common fashion while respecting the integrity of the existing data representations.

Ryan Baumann *The 'Son of Suda On-line'* pp. 91-106

The Son of Suda On-Line (SoSOL) represents the first steps towards a collaborative, editorially-controlled, online editor for the Duke Databank of Documentary Papyri (DDbDP). Funded by the Andrew W. Mellon Foundation's Integrating Digital Papyrology Phase 2

(IDP2), SoSOL provides a strongly version-controlled front-end for editing and reviewing papyrological texts marked up in EpiDoc XML.

Elaine Matthews and Sebastian Rahtz

The Lexicon of Greek Personal Names and classical web services pp. 107-24

This chapter documents the data resources of the long-term classical research project, *The Lexicon of Greek Personal Names* (LGPN), published in six volumes since 1987. It explains and demonstrates the web interfaces and services which now make available online the bulk of the LGPN, providing both powerful searching tools for scholars and an interface to allow other systems to link to LGPN data. Making the data available online provides direct, unmediated access to the material and supports exploitation of the data for further research both individual and collaborative.

We describe the work that went into creating the Lexicon, detail the granularity of the data structures, and explain the history of the project's record management. We then move onto the work undertaken in recent years to provide an archival XML-based format for the Lexicon's long-term preservation, and show how this has allowed us to build new web services, including exposure of Resource Description Framework (RDF) metadata, using the ontology of the CIDOC Conceptual Reference Model (CRM) ontology for semantic web applications.[1]

Simon Mahony: HumSlides *on Flickr: using an online community platform to host and enhance an image collection* 125-46

Moving a teaching and research image collection from an analogue to a digital medium for delivery brings with it many advantages but at the same time it also presents many new problems and ones probably not previously considered. This chapter discusses the move of a departmental slide collection, firstly to a proprietary in-house format, and then subsequently to the online community platform Flickr. It draws on the experience and model of the Library of Congress in partnership with Flickr and *The Commons*, as well as initiatives at Oxford and at New York University, and in doing so critically analyses and evaluates the possibilities for the future development of this collection. It asks why this collection is not currently being used to its potential and examines how the development of a user community would help to enrich the collection and ensure long term sustainability and future growth.

[1] CIDOC CRM is an ISO standard (21127:2006) that ‘provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation’ <http://www.cidoc-crm.org/>.

Valentina Asciutti and Stuart Dunn

Connecting the Classics: a case study of Collective Intelligence in Classical Studies pp. 147-60

One of the great potentials of the internet is its capacity to aggregate and unify information from diverse sources. Information in the Classics, and data generated by classicists, is inherently fragmented, and organized according to different standards. This paper describes a project at King's College London which sought to provide a set of aggregating services to humanities scholars. www.arts-humanities.net provides a platform, a library, and a taxonomy to organize and present data: we describe its facilities for supporting a multi-source dataset tracing the paths of Romano-British inscriptions, both in space and conceptually. Itinerant geographies of metrical versus text inscriptions are discussed, including how these can be published in a variety of non-conventional platforms, such as Twitter. We argue that, in the future, these platforms will come to play a critical role in the wider scholarly discourse of the Classics.

ABBREVIATIONS

ABM	Agent-based modelling
ADS	Archaeology Data Service
AHRC	Arts and Humanities Research Council
API	Application Programming Interface
APIS	Advanced Papyrological Information System
AWIB	Ancient World Image Bank
CC	Creative Commons
CI	Collective Intelligence
CIDOC	International Council of Museums
CRM-CIDOC	CIDOC Conceptual Reference Model
CSV	Comma-separated data fields
DANS	Data Archiving and Networked Services
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DDbDP	Duke Databank of Documentary Papyri
DPI	Dots per inch
DVCS	Distributed Version Control Systems
EDM	Europeana Data Model
GAP	Google Ancient Places
GIS	Geographical Information Systems
HEA	Higher Education Academy
HEFCE	Higher Education Funding Council for England
HESTIA	Herodotus Encoded Space-Text-Image Archive
HGV	*Heidelberg Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens*
IAph	*Inscriptions of Aphrodisias*
IDP	Integrating Digital Papyrology
ISAW	Institute for the Study of the Ancient World
JDI	Image Digitization Initiative
JISC	Joint Information Systems Committee
JSON	JavaScript Object Notation
KML	Keyhole Markup Language
LaQuAT	Linking and Querying Ancient Texts
LCCW	Longest Common Consecutive Words
LGPN	*The Lexicon of Greek Personal Names*
LoC	The Library of Congress
MDID	Madison Digital Image Database
OAI-ORE	Open Archives Initiative Object Reuse and Exchange
OER	Open Education Resources
OGSA-DAI	Open Grid Service Architecture–Data Access and Integration
OGSA-DQP	Distributed Query Processing
PDF	Portable Document Format

PN	Papyrological Navigator
PRM	Probabilistic Road Map
RDF	Resource Description Framework
RIB	*Roman Inscriptions of Britain*
SGML	Standard Generalized Markup Language
SOL	Suda On-Line
SoSOL	Son of Suda On-Line
SQL	Structured Query Language
SVG	Scalable Vector Graphic
TEI	Text Encoding Initiative
TLG	*Thesaurus Linguae Graecae*
URI	Uniform Resource Identifiers
V&A	Victoria and Albert Museum
VRE	Virtual Research Environment
WFS	Web Feature Service
WMS	Web Map Service
WYSIWYG	What-You-See-Is-What-You-Get
XML	Extensible Markup Language

INTRODUCTION

The Digital Classicist has run a summer seminar series at Senate House London, generously supported by the Institute of Classical Studies, every year since 2006. The listings for these seminars, along with abstracts, audio files (since 2008), and slides from the presentations, are available on the Digital Classicist website.[1] The topics are wide and varied but in all cases the content must be of interest to classicists, ancient historians, or archaeologists, as well as information specialists or digital humanities practitioners, and must demonstrate a research agenda that is relevant to at least one of those disciplines. Ideally they should present work that drives forward the research interests of both humanists and the technologists. Where possible we seek to publish a selection of these: the first was online in a special edition of the *Digital Medievalist*,[2] and the second as part of the Ashgate series, 'Digital Research in the Arts and Humanities'.[3] The chapters in this volume are all developed from presentations at those seminars, with the exception of two (Mahony and Dunn) which are taken from Digital Classicist panels run at the Classical Association Annual Conference in Durham 2011.

The Digital Classicist, established in 2004, is a network, a community of users, and has become defined by what we, as a community, do. The seminar series has become central to our activities, giving focus as well as a voice to our members, a platform for the dissemination and discussion of their work, a place for inspiration and introductions to be made. Our seminars promote the research activity of our members and allow the promotion of the Digital Classicist itself, as well as raising the profile of our speakers. Much of this is achieved by effective use of our social media networks and the Stoa Consortium blog. The online audio files and slides, but more especially the immediacy afforded by social media channels such as Twitter and live-tweeting (#digiclass), allow the spatial dimension to become less important, so that distance is no obstacle to developing that sense of belonging which is a foundational aspect of community. It allows us to be inclusive rather than exclusive. Indeed, the most striking and successful aspect of the Digital Classicist is its sense of community and collaboration. Practitioners do not work in isolation; they develop projects in tandem with colleagues in other humanities disciplines or with experts in technical fields – engineers, computer scientists, and civil engineers.

[1] Digital Classicist seminars: <http://www.digitalclassicist.org/wip/>.

[2] G. Bodard and S. Mahony, ed., '"Though much is taken, much abides": recovering antiquity through innovative digital methodologies', Digital Classicist special issue, *Digital Medievalist* 4 (2008), available at: <http://www.digitalmedievalist.org/journal/4/>.

[3] G. Bodard and S. Mahony, ed., *Digital research in the study of classical antiquity* (Farnham 2010).

A volume such as this cannot aim to give a broad overview of the state of play of scholarship in Digital Classics at any one time, but what it can do is present, in a coordinated way, a snapshot of the varied research conducted by its diverse membership. As such this volume collects together papers on a wide range of classical subjects, exemplifying multiple technical approaches, and taking a variety of forms. We show that this diversity of scholarship all contributes in a coherent way to the academic agenda that makes Classical Studies a leader in the use of innovative methods. Collectively, this volume illustrates and explores the highly collaborative nature of research in this field, the interdisciplinarity that has always been core to Classical Studies, the importance of innovation and creativity in the study of the ancient world, and above all the fact that digital research in Classics relies just as heavily as the mainstream upon rigorous traditional scholarship.

All of the chapters in this volume are research papers in their own right which engage with and contribute to the development of scholarship in the study of both classical antiquity and of the Digital Humanities more broadly.

TRAVEL AND INTERACTION IN THE GREEK AND ROMAN WORLD. A REVIEW OF SOME COMPUTATIONAL MODELLING APPROACHES

ANDREW BEVAN

1. Introduction

The inferential leap from static archaeological patterns to dynamic past behaviours is always challenging, especially in the case of the Greek and Roman world, where there is a real temptation to let the historical record do the heavy interpretative lifting. However, computational and quantitative techniques can also be of great assistance and this chapter offers a brief overview of their potential, as well as of some continuing problems associated with their effective application.

A useful preliminary question to pose, however, is simply this: why, in the first place, should we seek to build method and theory about travel and transport behaviours using classical archaeological datasets? In one sense, an easy but fairly narrow answer is: to understand better the dynamics of this particular period of the human past and this specific geographic area. However, it might also be argued that the Greek and Roman world offers a host of modelling advantages that have wider relevance, as exemplars in other, less evidentially advantaged archaeological and historical contexts worldwide. First, we can think of unusual levels of inter-regional connectivity (at several spatial scales) as being a defining feature of Mediterranean landscapes in particular.[1] Second, Mediterranean, European, and Middle Eastern paleo-landscapes are comparatively well understood and formal movement networks such as roads have been intensively studied, as well as less formally articulated ones such as patterns of pathways, transhumance, and migration. Third, we possess excellent evidence for the economic logistics of human and animal movement in certain parts of the classical world (*e.g.* Hellenistic and Roman Egypt) that allows us to speculate about the spatial and temporal scale of particular activities as well as their social and political structure. Fourth, we have a comparatively developed understanding of assisted movement technologies in this period and region, such as chariots, carts, boats, *etc*. Fifth, we have a good feeling for the contexts in which information about routing and territory are exchanged, such as via itineraries and geographic exegesis. Sixth and finally, there are clearly also interesting differences to be explored in terms of the directionality and scale of travel in this world (*e.g.* sailing *versus* rowing or paddling; grain fleets *versus* tramping trade;[2] individuals *versus* armies), as well as how it might change over time.

[1] For examples see: P. Horden and N. Purcell, *The corrupting sea* (Oxford 2000) 123-72.

[2] Paddling is a form of assisted human movement involving people in small- to medium-sized boats who face forward and propel their vessel via a one- or two-bladed flat paddle. It can be contrasted with rowing which often occurs in similar sized craft, but where people face backwards to the

These advantages argue for an approach to model-building in this region that does not circumscribe its relevance only to questions of interest to Greek and Roman specialists, but seeks to explore which features are contingent and which ones are of wider convergent interest. The discussion offered below cannot hope to address much of that larger agenda, but contributes by considering some cross-culturally relevant issues to do with modelling travel and transport, such as: (a) the range of cost-benefit trade-offs typically involved, and (b) the impact of different modes of travel. Thereafter, I will take a brief look at a range of different computational modelling approaches that all have complementary, spatially-explicit contributions to make for our understanding of past movement and interaction behaviours.

2. Some initial provisos: cost trade-offs and transport modes

One of several theoretical criticisms that might be levelled at the computational models of movement used by archaeologists so far, might be that most applications have been agnostic about, or willfully ignorant of, the kinds of costs and benefits they are seeking to measure. A good example, discussed in more detail below, is that of 'cost surface' analysis and 'least cost path' delineation. These two related methods typically either do not define the units with which they measure (*i.e.* costs have values without direct real-world correlates), or they blur several kinds of costs together, as if they were the same. In fact, it is useful to think explicitly about the interaction of at least three broad domains of movement cost: time, effort, and uncertainty. We can for example consider (a) the time taken to get from point A to point B (*e.g.* in hours), or (b) the effort (in terms of metabolic energy or money spent), or (c) the risk associated with such travel (in terms of uncertain travel conditions, personal danger, loss of cargo, unwanted observation by another party, *etc.*). As the above examples suggest, there are sub-divisions within these three broad domains of cost as well, and we cannot assume that any of them are related in a simple linear way to one another. Rather, we must seek to establish likely relationships and trade-offs between different kinds of cost cross-culturally and/or empirically from the archaeological and historical record. In some cases, we may still wish to offer a blurred, aggregate model of movement (*e.g.* via the circuit theory approaches introduced below), but we should at least do so as a part of a carefully thought-out and strategic choice, encouraged for example by an especially large spatial scale and/or *longue durée*.

In any case, where more specific routes and costs are to be modelled, it is important to understand the relevant trade-offs involved. Two good examples are the zig-zagging behaviour exhibited by both humans and animals on mountain paths[3] and the

direction of travel and use oars that are usually fixed into oarlocks. Sailing refers to the use of boats in which the dominant form of propulsion is a sail that harnesses the strength of the wind (though some sailing ships can still be rowed on occasion). Grain ships in the Roman and Byzantine period in particular could be very large vessels (*e.g.* the 50+m long *Isis* described in Lucian's *The ship, or the wishes,* even if there is some literary exaggeration) and followed very direct routes, typically from Carthage or Alexandria to Rome or Constantinople. In contrast, tramping involved much smaller sailing ships whose routes and destinations were far more irregular, multi-stop, and flexible to changing market conditions.

[3] See: A. Bevan, C. Frederick, and N. Krahtopoulou, 'A digital Mediterranean countryside: GIS approaches to the spatial structure of the post-Medieval landscape on Kythera (Greece)',

coast-hugging behaviour of many pre-modern sailing ships.[4] Zig-zagging is the propensity for mountain paths to follow indirect 'hairpin' routes. It is mainly engendered by the curve of a human's metabolic energy expenditure on slopes of different steepness, but important features of the problem also relate to the chosen velocity of the traveller. Likewise, sailing ships tack because of a relationship between optimum speed forward in the direction of the destination and the optimum direction for effectively harnessing the force of the wind.[5] In addition, it is worth noting the propensity for much Mediterranean shipping to hug the coastline for several reasons: (a) to trade off speed of more direct travel against potential risks incurred by travel out of sight of land, and (b) to trade off speed of more direct travel against the greater certainties often provided by diurnal shifts in near-shore breezes.

As the above should emphasize, there may not always be easy answers to these potentially complex cost-benefit calculations, and indeed at times the simpler rule-of-thumb decisions that any given individual might make could conceivably render much of the complexity irrelevant for the modeller (or at least dramatically change the modelling agenda). Even so, we should at the very least be far more explicit about these issues than we typically have been so far, either by declaring our starting assumptions, or by testing whether the implications of different costs are significant or not (*e.g.* by sensitivity analysis).

A further issue that has hitherto been addressed in only a very limited way by archaeological modelling is the need to differentiate modes of travel involving different numbers of individuals (*e.g.* single person, large group), different kinds of human traveller (*e.g.* adult, child, female, male), different kinds of assisted travel and transport technology (*e.g.* on foot, on horseback, in a boat), different seasonal constraints, and different travelling agendas.[6] Related to this is also the issue of intermodal shifts – for example, the process of switching from boat to cart to foot transport – for which we have little, if any, coherent archaeological and ancient historical theorizing, in contrast, for example, to the attention this subject has received in modern transport planning.[7] In fact, it is to this critical threshold issue of 'transferral', or breakage-in-bulk for cargo, that future computational modelling will need to pay substantially greater attention, particularly with regard to the movement of goods. In addition, different group sizes (*e.g.* individual

Archeologia e Calcolatori 14 (2003) 227-29, and M. Llobera and T. J. Sluckin, 'Zigzagging: theoretical insights on climbing strategies', *Journal of Theoretical Biology* 247 (2007) 206–17.

[4] For example: for its relevance later in the Medieval period, see: F. Braudel, *The Mediterranean and the Mediterranean world in the age of Philip II* (London 1972).

[5] For a modelling initiative, see: A. Philpott and A. Mason, 'Optimising yacht routes under uncertainty', *Proceedings of the 15th Chesapeake Sailing Yacht Symposium* (2001).

[6] For examples of the latter two in Roman times see: R. Duncan-Jones, *Structure and scale in the Roman economy* (Cambridge 1990) 7-29, and C. van Tilburg, *Traffic and congestion in the Roman Empire* (Oxford 2007) 41-89.

[7] But also see some archaeological work on portage phenomena: A. Sherratt, 'Portages: a simple but powerful idea in understanding human history', in *The significance of portages. Proceedings of the first international conference on the significance of portages,* ed. C. Westerdahl (Oxford 2006) 1-13.

donkey or ship transport *versus* large caravans or fleets) obviously have different consequences for each of these different modes.

Table 1 provides some basic parameters for different kinds of travel speed and load, as relevant to Greek and Roman contexts.[8] It is surprising how rarely such variables as these are actually part of a modelling agenda (and indeed the computational examples provided later in the chapter do not do so either), but there are plenty of opportunities to so incorporate them in the future. Perhaps one of the best current examples is the work by Alberto Minetti on trade-offs between time, speed, and physiological stress on horses in equine postal services, where surprising cross-cultural regularities are present in the spacing of staging posts and the average adopted speed along the route.[9]

Type of Transport	**Speed (kmph)**	**Load (kg)**
Human pedestrian	4-5	n/a
Human porter	2.5-3	30-60
Pack donkey	2.5	60-100
Pack camel	3-4	150-250
Pack horse	4-8	80-180
Riding camel	15-30	n/a
Riding horse	10-30	n/a
Wagon	2-4	200-1000
Small rowboat/canoe	3-8	<1,000
Oared galley	7.5-15	<100,000?
Large riverboat	2-40	10,000-100,000
Sailing ship	2-7.5	10,000-1,000,000

Table 1. Some rough performance estimates for travel speed and load-carrying in the Greek and Roman world. Loads and speeds assume those sustainable over a full day's work, and could otherwise be higher for short trips. Speeds assume relatively flat terrain.[10]

3. Computational methods

Computational models can be more or less complex in their design and the number of parameters they consider. For archaeologists or historians, it is tempting to model as close a fit to observed reality as possible, building in all of the rich conceptual detail that we

[8] See more generally: B. Cotterell and J. Kaminga, *Mechanics of pre-industrial technology* (Cambridge 1992) 193-225 and C. Adams, *Land transport in Roman Egypt* (Oxford 2007) 49-69.

[9] A. E. Minetti, 'Efficiency of equine express postal systems', *Nature* 426 (2003) 785-86.

[10] Sources for the data: Adams, *Land Transport* (n. 8 above); J. Rennell, 'On the rate of travelling, as performed by camels; and its application, as a scale, to the purposes of geography', *Philosophical Transactions of the Royal Society of London* 81 (1791) 129-45; L. Casson, 'Speed under sail of ancient ships', *Transactions and Proceedings of the American Philological Association* 82 (1951) 136-48; P. J. Sijpesteijn, *Customs duties in Graeco-Roman Egypt* (Zutphen 1987); W. Habermann, 'Statistiche Datenanalyse an den Zolldokumenten des Arsinoites aus römischer Zeit II', *Münstersche Beiträge zur Antiken Handelsgeschichte* 9 (1990) 50-94; Cotterell and Kaminga, *Mechanics* (n. 8 above) table 8.1; A. E. Minetti, 'Efficiency' (n. 9 above) 1698-1703; van Tilburg, *Traffic and congestion* (n. 6 above).

believe is present in the real world case under study. With respect to human travel, this would involve, for example, assessing the full balance of cost and benefit calculations discussed above (travel time, effort, risk in myriad combinations and types), as well as any perceived cultural attitudes, special places in the landscape that might repel or attract, *etc.*[11] For many commentators therefore, the use of only very basic features of the landscape such as topography and access to roads, rivers, or the sea vastly under-appreciates the encultured nature of travel and transport decisions and ushers in a kind of casual environmental determinism. However, a reverse argument can also be made to the effect that, done well, there is an analytical elegance to highly simple models that is conceptually useful. They are, in theory, easy to understand, because only a limited number of parameters are involved, and easy to test, because we can offer robust predictions of how they should behave under controlled conditions. It is therefore often better, I would argue, to provide simple models of movement and interaction prior to exploring more complex ones, as the former often offer useful null hypotheses from which the real archaeological record can be observed to depart (or not) in interesting ways.

The rest of this chapter presents a review of the range of possible models that might be used for understanding movement and interaction, offering examples for some of them based on the landscapes in Greece, and spanning the Bronze Age, early Iron Age, and Graeco-Roman periods. While remaining aware of the need to build theory and explore method in the future with respect to types of travellers, cargoes, or agendas, the examples below are stripped free of such subtle differentiations, but do thereby remain simple to grasp.

3.1 Cost surfaces

Cost surfaces and so-called ‘least cost paths’ are by now very well established and closely related methods for exploring movement in a Geographical Information System (GIS), typically based on little more than a skeleton model of topography (with the latter usually being in a ‘raster’ or pixel-based format). The methodological stages involved are: (i) to define a set of costs for each cell in a raster map, (ii) to create a ‘cost surface’ by accumulating these costs out from a fixed point of departure (A), and (iii) if required, to trace a route from another point (B) back to the departure point (A) and thereby define a ‘least cost path’ between them. There are different kinds of cost surface. Some assume that (i) all of the costs incurred along the way are not altered by the direction of travel (*i.e.* they are isotropic, such as the cost of moving through different types of land cover) while others assume that (ii) costs are operating from a particular direction (partially anisotropic, such as the effect of a strong wind on a cyclist), (iii) costs are entirely direction dependent (*i.e.* fully anisotropic, such as the cost of moving across a slope which varies depending on the direction in which you walk across it), and/or (iv) that a combination of isotropic and anisotropic costs are in operation.

[11] For a good discussion on this, see: M. Llobera, ‘Understanding movement: a pilot model towards the sociology of movement’, in *Beyond the map: archaeology and spatial technologies*, ed. G. Lock (2000) 71-75.

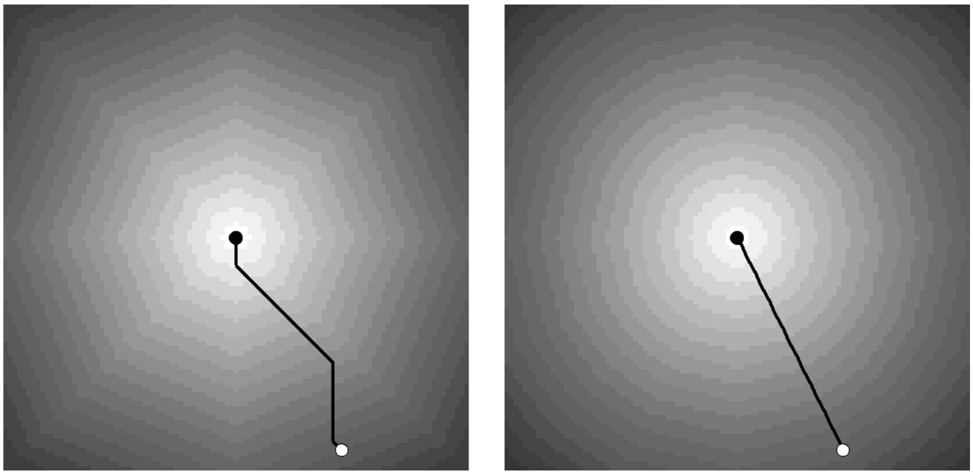


Figure 1 Simple models and cost surface problems. Cost surfaces and a least cost path calculated on a flat 100 x 100 surface, for two kinds of spreading algorithm: (a) D8 in ArcGIS, and (b) D16 in GRASS GIS.

Unfortunately, despite a great deal of early enthusiasm with archaeological applications of cost surfaces in the 1990s, there continue to be a host of difficulties associated with cost surface analysis.[12] We can list them as being: (i) a failure on the part of the vast majority of implementations to address anisotropic costs, (ii) the widely varying computational 'curves' used to measure the cost of passage across slopes of differing steepness in a landscape,[13] (iii) the specific kind of 'spreading' algorithm used to accumulate cost, and (iv) the degree to which the results are ever formally calibrated or validated.

A further, well-known but often-ignored problem is associated with the search neighbourhood used by a cost surface algorithm. The default in many software packages is a queen's case search neighbourhood (or 'D8', where the routine moves through each cell in the raster map, and for each one, checks the costs of the 8 immediately neighbouring cells, effectively allowing it to search in the manner that a queen on a chessboard would move).[14] If this method is applied to the simplest case of an entirely flat topography of uniform costs, however, the resulting surface from any point (*e.g.* A in Figure 1a) has a faceted appearance,

[12] See: D. H. Douglas, 'Least cost path in GIS using an accumulated cost surface and slope lines', *Cartographica* 31.3 (1994) 37-51; T. Bell and G. Lock, 'Topographic and cultural influences on walking the ridgeway in later prehistoric times', in *Beyond the map*, ed. Lock (n. 11 above) 85-100; W. Collischonn and V. Pilar, 'A direction dependent least cost path algorithm for roads and canals', *International Journal of Geographical Information Science* 14 (2000) 397-406; J. Conolly and M. Lake, *Geographical Information Systems in archaeology* (Cambridge 2006) 215-25; I. Herzog, 'Theory and practice of cost functions', in *Computer applications and quantitative methods in archeology. Computing applications in archaeology*, ed. F. Javier Melero and P. Cano (Granada 2010).

[13] For example: Herzog, 'Theory and practice' (n. 12 above) fig.1; and also A. E. Minetti, 'Optimum gradient of mountain paths', *Journal of Applied Physiology* 79.5 (1995) 1698–1703.

[14] See Conolly and Lake, *Geographical Information Systems* (n. 12 above) fig.10.11.

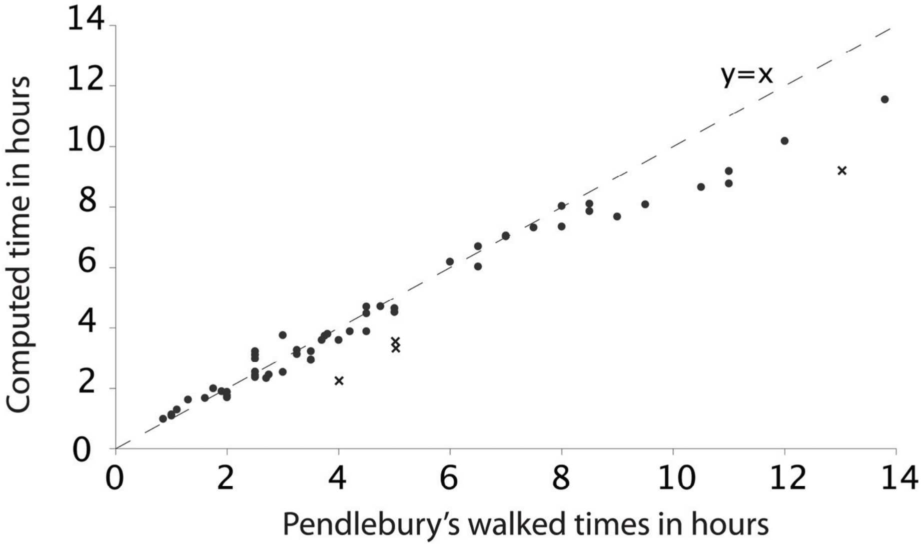


Figure 2 Comparison of the times recorded by John Pendlebury for his walks between Cretan sites in the 1930s (n=60 trips) and those computed by anisotropic cost surface analysis with GRASS *r.walk* using default parameters and D16 search (suggested outliers are marked as crosses).

when it should appear as perfectly concentric rings, and any path delineated to a second point B risks being a dog-leg, when it should be a straight line. One way to reduce this problem is to expand the search to a knight's case neighbourhood ('D16', where the routine searches 16 neighbouring cells, including both the queen's case cells and those to which a knight on a chessboard could theoretically move), or larger (Figure 1b). In any case, it is also surprising how rarely cost surface results have been tested against a known set of journey times. One such opportunity is provided for the Greek island of Crete by John Pendlebury, due to journey times he recorded for travel by foot during the 1930s, prior to major mechanized transport or modern road-building.[15] Figure 2 compares the times that Pendlebury recorded for sixty of these journeys with the results from anisotropic modelling with a D16 search (using GRASS GIS' *r.walk* module).[16] The only costs involved are the direction-specific ones imposed by terrain of varying steepness.[17]

The correlation between the two sets of times (observed and computed) is very good, particularly for journeys of less than about eight hours. In fact, such a correlation is

[15] J. D. S. Pendlebury, *The archaeology of Crete* (London 1939).

[16] GRASS: Geographical Resources Analysis Support System: <http://grass.fbk.eu>, free Geographical Information System (GIS) software.

[17] As measured on a 15m cell digital elevation model: N. Chrysoulakis, M. Abrams, H. Feidas, and D. Velianitis, 'Analysis of ASTER multispectral stereo imagery to produce DEM and land cover databases for the Greek islands: the REALDEMS project', in *e-Environment: Progress and Challenge*, ed. P. Prastacos, U. Cortés, J.-L. Díaz de León and M. Murillo (Mexico City 2004) 404-17.

potentially deceptive as even straight-line, ‘as the crow flies’ distances are already highly correlated with Pendlebury’s estimates (r^2=0.88, *i.e.* sites further away as the crow flies will typically take longer to walk to as a matter of course). However, it is reassuring to note that the anisotropic calculations of less than eight hours offer significantly improved explanatory power (r^2=0.96, likely to be different from the above at $p < 0.005$). Thereafter, the predicted times for journeys over eight hours are still useful, but often a little too rapid, probably reflecting the fact that overnight travel requires extra time for rest-stops, the burden of extra baggage, *etc*.

Compared with many other methods of this kind, the one implemented in GRASS GIS offers two further advantages of: (i) working directly from rate of change in elevation values rather from a derived slope map, and (ii) being based on a well-established rule-of-thumb used by hikers (Naismith’s rule)[18] and estimating costs explicitly as travel time. Some continuing theoretical problems with it, however, are: (i) that its function for estimating travel time is discontinuous,[19] (ii) that in some instances, its knight’s-case moves potentially leapfrog intervening costs or barriers, and (iii) the fact that it still significantly under-estimates the kinds of times, and mis-delineates the kinds of hairpin route, that we might expect in extremely steep terrain (see below). Despite these provisos, however, the above results should give us confidence (albeit certainly not blind faith) that the times and paths modelled here offer interpretative added value when considering interaction across Crete’s often tyrannically rugged landscape. I therefore make use of them in several examples below.

Figure 3a offers a useful, albeit earlier and prehistoric, example of how an anisotropic cost surface might be compared to documentary evidence to get an idea of possible travel and/or administrative thresholds. It takes as a departure point the Bronze Age centre of Knossos on Crete (although the shorter distance thresholds perhaps also provide relevant points of comparison for the territory of the later and smaller city-state located here in the Classical to Roman periods), and suggests the travel times out from this centre to all other parts of the island. Overlaid on these are the likely locations of toponyms found in the fourteeth- to thirteenth-century BC Linear B tablets from Knossos,[20] which suggest that this centre was organizing substantial activity across much of the island. The toponym groups shown here reflect those locations that regularly co-occur in the tablets and suggest qualitatively different kinds of interaction with Knossos in different regions and at different removes from the centre. Figure 3b then considers the possible effect of sea voyages (modelled here very grossly indeed as encouraging twice the speed of typical pedestrians). The key point to note here is both geographic and political: simply that political, economic, and social connections between Knossos and the far west or east of the island would be far better sustained via maritime travel that by terrestrial means.

[18] See: E. Langmuir, *Mountaincraft and leadership* (London and Edinburgh 1995) 39-43.

[19] See: Herzog, ‘Theory and practice’ (n. 12 above) fig.1.

[20] J. Bennet, ‘The structure of the Linear B administration at Knossos,’ *American Journal of Archaeology* 89.2 (1985) 231-49.

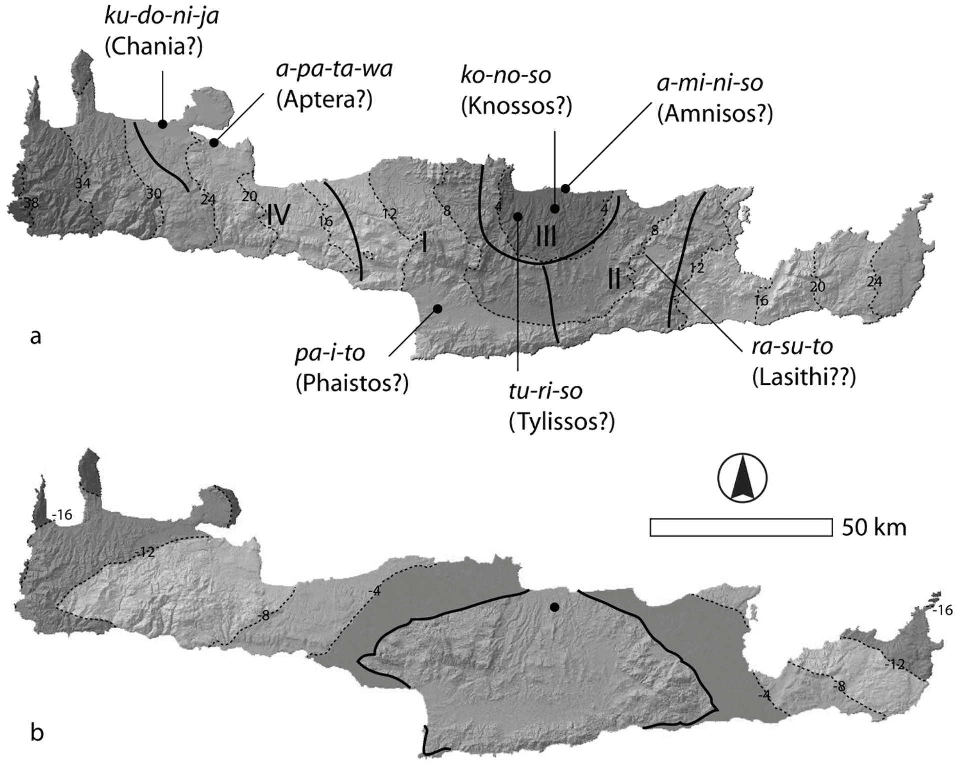


Figure 3 Anisotropic cost surfaces from Knossos: (a) terrestrial travel times (dotted lines are 4 hour contours outwards from Knossos), along with the toponym groups suggested by the Linear B archives (solid lines are very rough group divisions after Bennet 1985: fig.iii.4), (b) a rough impression of the time saved by including a maritime leg in the trip from Knossos. The negative values refer to the number of fewer hours needed to travel to that location by comparison to figure 3a above (assuming no embarkation delays). Dotted lines are 4 hour contours and the solid line marks the area with no change.

3.2 Network analysis

Thinking of the world in terms of network connectivity is a very popular approach nowadays, prompted by its conceptual simplicity in one sense, and potentially great analytical and interpretative complexity in another. Networks, can be used to think about social relationships (*e.g.* 'small worlds'),[21] physical relationships (*e.g.* terrain morphology),[22] or conceptual ones (*e.g.* ancient itineraries,[23] and religious ideas),[24] to name

[21] For example: 'small worlds', D. J. Watts and S. H. Strogatz, 'Collective dynamics of "small-world" networks', *Nature* 393 (1998) 440-42.

[22] See: J. Wood, 'Constructing weighted surface networks for the representation and analysis of surface topology', in *5th international conference on GeoComputation*, (2000): www.soi.city.ac.uk/~jwo/Geocomputation00.

[23] S. Graham, and J. Steiner, 'Travellersim: growing settlement structures and territories with agent-based modelling', in *Digital discovery: exploring new frontiers in human heritage. CAA 2006.*

but a few applications. In a sense, modern network science is a triangulation of many established approaches such as space syntax, graph theory, and social network theory.[25] At its heart is the idea of a set of nodes as an abstraction of real-world sources of interaction, with a set of edges as connectors among them. One key issue with network analysis is the degree to which it is or is not sensitive to uncertainty in: (a) the location or number of defined nodes, and (b) the connection matrix (in terms of which nodes are connected, in what directions, and with what weights).[26] All of these issues can in theory be explored by considering a range of different scales of connection, by assessing a range of input parameters (*e.g.* parameter sweeps), and by randomized Monte Carlo perturbation of the network (*e.g.* slightly moving existing nodes, altering connectivity, and/or addition or subtraction of nodes). A further possible route is to consider not observed nodes (*i.e.* not observed archaeological settlements if we are talking about a settlement network), but hypothetical candidates (*e.g.* plausible settlement locations).

Many commonly used networks are single-mode (*i.e.* only one kind of connection is being considered at any one time), unweighted (*i.e.* where edges all have the same cost of passage along them) and undirected (*i.e.* the passage from A to B and from B to A is equivalent), but we can make them more realistic (without necessarily overcomplicating) by adding edge weights based on likely travel times. Just to take an example, settlement on the island of Crete has, over the last nine thousand years or so, clearly been affected by environmental affordances (*e.g.* access to better land for certain activities, to preferred trans-insular routes, or to off-island maritime connections) as well as by a range of historically contingent and culturally specific influences. A useful first step is to consider some very simple baseline models of likely demographic connectivity across the island. For example, we can start by dropping 100 random settlements down onto the Cretan landscape (taking inspiration from the classical tradition of 'hundred-citied Crete'),[27] with a preference for more agriculturally-favourable parts of the island (Figure 4a). For our purposes here, a map of such favourable areas was formally defined as follows: (a) identify all cells in the map that are ≤ 10° slope and hence might initially be preferred for agriculture (*i.e.* without terracing), (b) exclude all such flatland cells that are more than 1000m above sea-level (*i.e.* the approximate tree line), and (c) for each cell in the map (*i.e.* a neighbourhood operator), calculate the mean number of such flatland cells that are

Computer applications and quantitative methods in archaeology. Proceedings of the 34th conference, Fargo, United States, April 2006, ed. J. T. Clark and E. M. Hagemeister (Budapest 2006) 49-59; L. Isaksen, 'The application of network analysis to ancient transport geography: a case study of Roman Baetica', *Digitial Medievalist* 4 (2008).

[24] A. Collar, *Networks and religious innovation in the Roman Empire* (Unpublished PhD Thesis, University of Exeter, 2008).

[25] For a good discussion of its relevance to archaeology, see: T. Brughmans, 'Connecting the dots: towards archaeological network analysis', *Oxford Journal of Archaeology* 29.3 (2010) 277-303.

[26] See also: M. Zanin, 'Uncertainty in complex network,' *International Journal of Complex Systems in Science* 1 (2011) 78-82.

[27] P. Perlman, 'One hundred-citied Crete and the "Cretan *Politeia*"', *Classical Philology* 87.3 (1992) 193-205.

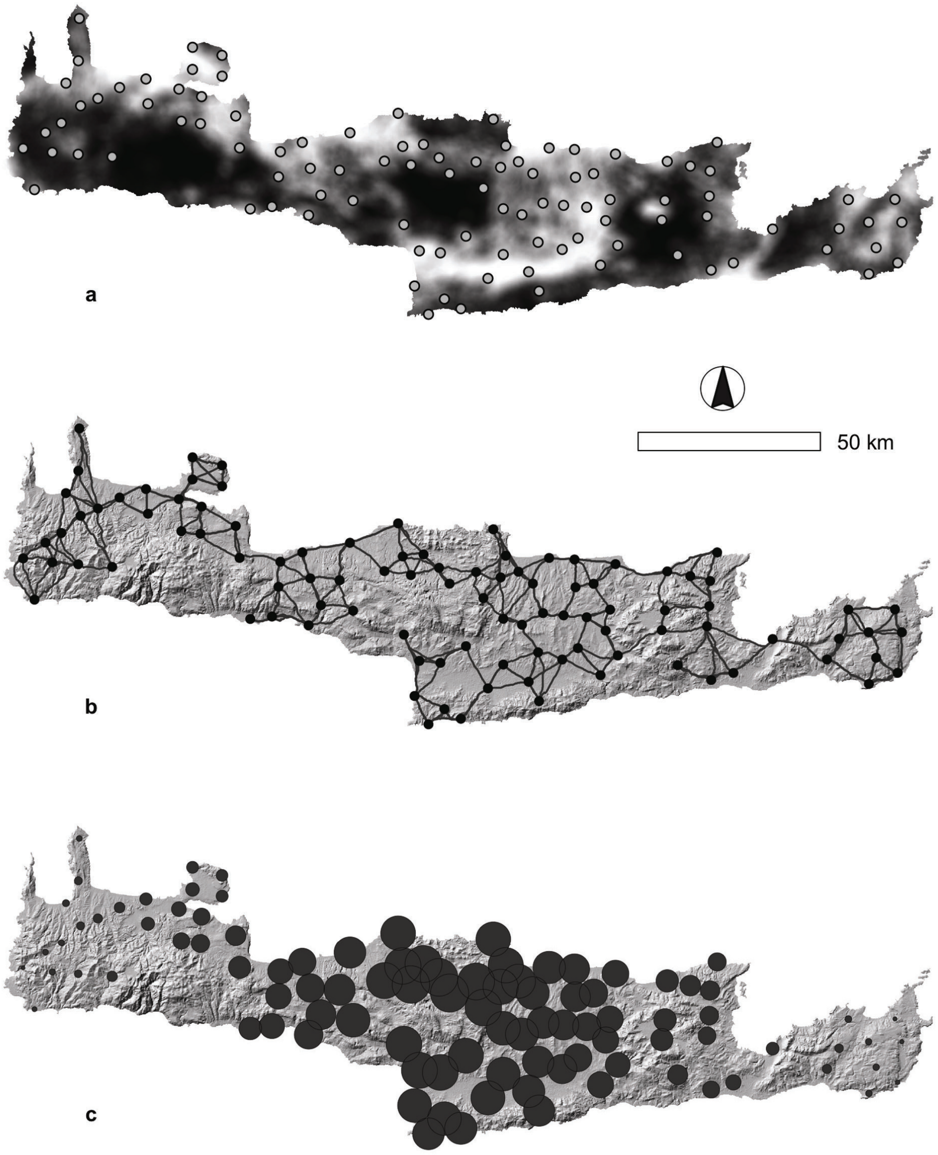


Figure 4 Cretan landscape affordances: (a) a set of 100 random points allocated preferentially on areas with better access to flat land (the latter shown as lighter shades of grey in the underlying intensity map), (b) an anisotropic path network linking each random point to its three nearest neighbours based on pedestrian travel time, (c) closeness centrality measures for the same sites, but based on a full weighted network (*i.e.* number of out-nodes = 99, weights are travel time).

found within a 2.5km radius (*i.e.* about an hour's roundtrip and typical of a wide cross-cultural range of human daily travel budgets).[28]

This map was then used as an intensity surface on which to simulate 100 points while maintaining a minimum distance of 5km between them.[29] We can then link up these hypothetical sites into a network via computed paths that seem optimal in terms of travel time by a single pedestrian (*i.e.* calculated anisotropically as above). Figure 4b shows an example where we have retained only the three nearest neighbouring links out from each site. Already with this kind of modelled connectivity, we can see a propensity for the island to break itself up into smaller regional network components. However, archaeologically and historically we have no reason to assume that three paths is a valid level of connection in any given time and place. A less judgemental approach might be simply to work with a complete set of connections among sites (*i.e.* site 1 is connected to each of sites 2-100 via 99 direct paths and so on), but one in which the connecting edges of the network are weighted by the varying estimated travel time along them. Although this network is full, weighted, and directional, we can nonetheless calculate traditional network metrics.[30]

As an example, figure 4c shows a network measure known as 'closeness centrality'[31] for each hypothetical settlement across the island: it is wholly unsurprising to note that sites in the middle of the island prove to be more central within the network (in this case shown as larger circles). Likewise, a fuller analysis would ideally consider: (a) the degree to which these results vary over multiple point simulation runs and under variable densities of points, as well as (b) a wider range of network metrics. However, an important insight is the fact that this measure of centrality shifts abruptly at certain points across the island due to a variety of topographical pinch-points. In other words, there is not a continuous gradient of change to the far west and the far east of the island, but a built-in propensity to generate regional spheres of interaction, some of which potentially exist in near isolation from the rest, at least in terms of their terrestrial connectivity. Were we to consider the movement of larger groups of travellers or heavier cargo explicitly, this situation is only likely to become more extreme. If such regions were therefore ever to be integrated politically and economically, then we can probably assume that it was maritime linkages that often most easily achieved it. Indeed, in historical times, the island has only been under a unified authority when it has been occupied by external powers with strong navies (*e.g.* Roman, Venetian, and Ottoman),[32] and prior to this, the classical sources

[28] J. H. Ausubel and C. Marchetti, 'The evolution of transport', *The Industrial Physicist* 7.2 (2001) 20-24.

[29] For intensity-based point simulations, see: J. Baddeley and R. Turner, 'spatstat: an R Package for analyzing spatial point patterns', *Journal of Statistical Software* 12.6 (2005) 33-35.

[30] For examples, see: T. Opsahl, F. Agneessens, and J. Skvoretz, 'Node centrality in weighted networks: generalizing degree and shortest paths', *Social Networks* 32.3 (2010) 245-51.

[31] For an introduction to this, see: M. Newman, *Networks. An introduction* (Oxford 2010) 181-85.

[32] J. Bennet, 'Knossos in context: comparative perspectives on the Linear B administration of LM II-III Crete', *American Journal of Archaeology* 94.2 (1990) 193-211.

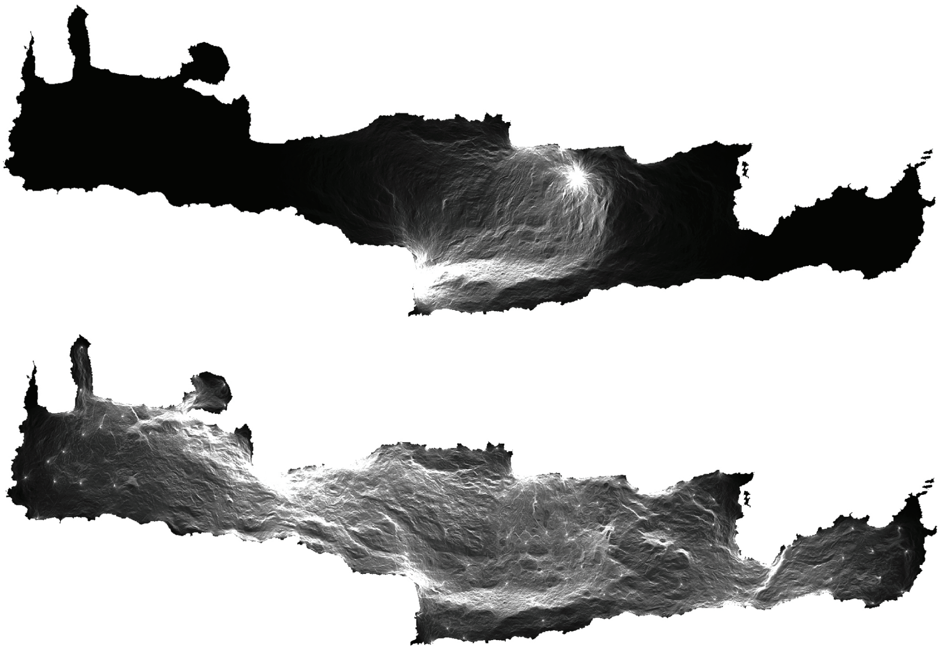


Figure 5 Current maps based on using terrain slope as a measure of electrical resistance: (a) between two random sites, and (b) aggregated for all sites.

claim a period of earlier political unification associated with the legendary figure of king Minos and underpinned by a Minoan 'thalassocracy'.[33]

3.3 Circuit theory and isolation by resistance

Further models of interaction can be generated via an approach inspired by the behaviour of electrical circuits.[34] In circuit theory terms, if we choose one of our random sites on Crete (as used above) and connect it to the ground, while connecting another site to the equivalent of a 1 amp electrical source, we can then model the behaviour of current flow over the intervening space with, for example, steepness of slope acting as a form of resistance. Figure 5a offers an example where current is allowed to flow from a site in central Crete (in electrical circuit terms, the 'source', and a stand-in for a major settlement such as Cnossus/Knossos in Bronze Age through to Roman periods) to one in the Mesara valley in central Crete (the 'ground', and perhaps a stand-in for a site such as Gortyn or Phaistos). Such a depiction is useful for three reasons: first, it suggests a broad corridor of likely movement between the two areas rather than a discrete single pathway or straight line. It also suggests that interaction via the Pediadha upland and the eastern end of the Mesara may

[33] For relevant archaeological discussion and caution over this later mythical tradition, see: C. Broodbank, 'Minoanisation: beyond the loss of innocence', *Proceedings of the Cambridge Philological Society* 50 (2004) 46-91; A. Bevan, 'Political geography and palatial Crete', *Journal of Mediterranean Archaeology* 23.1 (2010) 27-54.

[34] B. H. McRae, 'Isolation by resistance', *Evolution* 60 (2006) 1551-61.

have been just as important as a direct linkages northeast-southwest. Second, we might think of it as a more plausible, less deterministic model of possible connectivity than either an 'as-the-crow-flies', Euclidean distance or a single least cost path. For example, if we take three sites roughly equidistant apart, but with the connectivity of A and B benefitting from several possible alternative routes, then a least cost path approach will typically model just one of these in each case and assume interactions between A, B, and C are roughly the same, whereas multiple routes between A and B should in theory engender greater linkage. Third, the theoretical equivalence of a circuit theory approach to two-dimensional diffusion models and random walk models makes it an attractive 'null hypothesis' for interaction over a spatially heterogeneous space. At present it has mainly been used, with promising results, to model patterns of genetic variation with real-world distance.[35]

Figure 5b shows a cumulative current map produced by grounding one site at a time, then mapping the current from all ninety-nine other sites as individual sources, and then iterating for all sites. The map emphasizes certain short, sharp corridors (light grey filaments) of connection, for example between Chania and Rethymnon along the northwestern coast, north-south in the Ierapetera region, as well as broader zones of important linkage, for example between the Mesara plain and western Crete via the Amari valley.

Circuit theory offers an attractive way of exploring aggregate patterns of large-scale movement via multiple pathways. Resistances are likely to become an alternative model of distance effects to least cost paths or Euclidean measures and can be combined effectively with several of the other methods described below, such as network analysis or spatial interaction models (Section 3.4 below).

3.4 Evolving models

So far, we have considered methods for characterizing conditions for movement and interaction that are, perhaps slightly counter-intuitively, rather static in nature. While some network methods do also consider the dynamic properties of such configurations,[36] the last three types of model discussed briefly here are all promising ways to understand evolutionary patterns of movement, growth, decay, and interaction over time. To begin with, it is worth reconsidering an approach successfully introduced into archaeology and history some time ago by Tracey Rihll and Alan Wilson,[37] but rarely heard of since,[38]

[35] For examples: B. H. McRae and P. Beier, 'Circuit theory predicts gene flow in plant and animal populations', *Proceedings of the National Academy of Sciences of the USA* 104 (2007) 19885-90; J. van Etten, and R. J. Hijmans, 'A geospatial modelling approach integrating archaeobotany and genetics to trace the origin and dispersal of domesticated plants', *PLoS ONE* 5.8 (2010) e12060.

[36] For example: T. Evans, C. Knappett, and R. Rivers, 'Using statistical physics to understand relational space: a case study from Mediterranean prehistory', in *Complexity Perspectives on Innovation and Social Change*, ed. D. Lane, S. Van der Leeuw, D. Pumain, and G. West (New York 2009) 451-79.

[37] T. E. Rihll, and A. G. Wilson, 'Spatial interaction and structural models in historical analysis: some possibilities and an example', *Histoire et Mesure* 2.1 (1987) 5-32, and T. E. Rihll, and A. G. Wilson, 'Modelling settlement structures in ancient Greece: new approaches to the polis', in *City and country in the ancient world*, ed. J. Rich and A. Wallace-Hadrill (London 1991) 59-95.

despite its thoroughly established role in urban geography.[39] 'Spatial interaction models', as they are often known, are developed versions of gravity models and usually have at their heart an equation of the kind:

$$S_{ij} = \frac{O_i W_j^{\alpha} e^{-\beta c_{ij}}}{\sum_k W_k^{\alpha} e^{-\beta c_{ik}}}$$

where:

S_{ij} is a matrix recording the quantity of resources flowing from each site i to each other site j;

O_i is a measure of the size of flow originated at site i;

W_j is the attractiveness of site j;

C_{ij} is the distance from i to j;

α is a parameter used to model the advantages of concentrating resources in one place;

β is a parameter used to model the ease of communication over a distance;

e is an exponential function.

A model set up in this manner can then be run iteratively, updating the attractiveness (W_j) and size (O_i) of each site until collectively these all reach a point of equilibrium, with no remaining imbalances in modelled inflow to and outflow from sites ($\sum S_{ij}$). In an extreme case, and one particularly relevant to archaeological situations, the only necessary input for such a model is a set of point locations, some measure of the distances between them, and perhaps a rough idea of conceivable outcomes that might suggest a suitable range of α and β values to explore. Although further details of these methods will not be considered here, they certainly offer a rich set of analytical possibilities because: (a) they can support full, weighted, and directed spatial distance matrices in which very few assumptions are made *a priori* about the nature of the connectivity involved, and (b) they are at least potentially evolutionary in concept (*e.g.* the method can be used to explore the emergence of central places over different time-steps).

At present, spatial interaction models of this kind are useful for considering the hierarchy of population centres or of commercial outlets, but only include the effect of distance in relatively simple ways (typically as negative exponential isotropic decay). One fruitful future possibility would be to combine these with either least cost paths onto which the spatial interaction model flows can be loaded and mapped, separately from the centres (such as an example from Geometric Greece shown in Figure 6), or resistance distances (see above). Likewise, pathways are themselves subject to evolutionary forces: for example, without further guidance, pedestrians take the most direct route from A to B, B to C, and C to A, but once visible paths are established, they (a) act to encourage

[38] But now see: T. Evans, R. Rivers, and C. Knappett, 'Interactions in space for archaeological models', *Advances in Complex Systems* 85 (2011) 1150009.

[39] For more details on this see: A. G. Wilson, *Complex spatial systems. The modelling foundations of urban and regional analysis* (Harlow 2000).

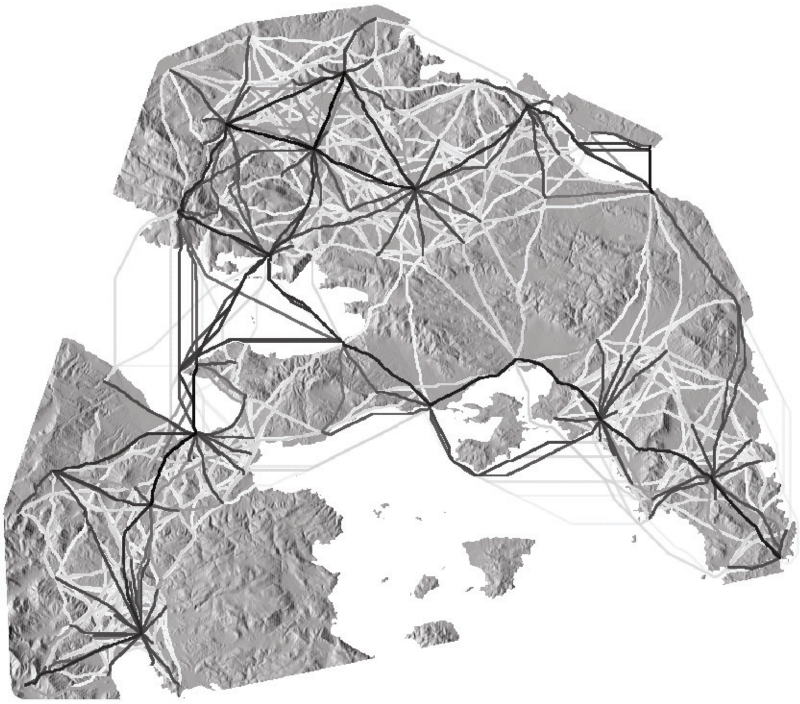


Figure 6 An example of a spatial interaction model for which minimum travel times over land and sea have been used as distances, and modelled flows have then been mapped back onto least cost routes, indicating major hubs of activity and preferred corridors of interaction (darker paths have higher flow). This is a preliminary reinvestigation (by Bevan, Dearden, and Wilson) of the Rihll-Wilson model of Greek geometric sites.

subsequent travel along roughly the same route, but (b) also can be shown gradually to adjust their respective course and 'bundle' into a more centralized set of paths over time.[40]

Two further approaches to movement modelling, arguably at opposite ends of the strategic spectrum, are diffusion-reaction equations and agent-based modelling. Diffusion-reaction models have not, as yet, received much consideration in classical archaeology but have been used to model the dispersal of farming practices into Europe,[41] or the southward spread of Palaeolithic hunter-gatherers into the Americas.[42] They are highly relevant to understanding how, for example, innovative ideas spread out across a population, where the latter has a clear spatial structure (*e.g.* is clumped into certain regions and into certain

[40] For example see: D. Helbing, J. Keltsch, and P. Molnar, 'Modelling the evolution of human trail systems', *Nature* 388 (1997) 47-50. See also: A. Bevan and A. Wilson, 'Models of settlement hierarchy based on partial evidence', *Journal of Archaeological Science* 40.5 (2013) 2415-27.

[41] A. J. Ammerman and L. L. Cavalli-Sforza, 'Measuring the rate of spread of early farming in Europe', *Man* 6 (1971) 674-88.

[42] For a full review of such models, see J. Steele, 'Human dispersals: mathematical models and the archaeological record', *Human Biology* 81.2-3 (2009) 121-40.

settlements). They are therefore also of potential relevance, for example, for understanding the spread of technological innovations or novel religious practices across the Graeco-Roman world. The differential equations that typically underpin them have much potential synergy with the spatial interaction models described above, and the opportunities for considering them all as a complementary family of models, of broadly Boltzmann-Lotka-Volterra type, has been recently emphasized.[43]

Agent-based modelling (ABM), on the other hand, attempts to model behaviour as a set of interactions among discrete agents who typically each have their own set of individual behaviours. ABM approaches potentially facilitate the exploration of the following in useful ways:[44] (a) emergence – whether higher-levels of structure in the model can be shown to emerge without being imposed in some way from the outset; (b) adaptation and fitness – which aspects of agent behaviour respond to changes in the overall environment and how does this affect their pursuit of particular objectives (*e.g*. successful reproduction, greater wealth, a particular destination, *etc*.); (c) sensory capacities – for example, the impact of an agent's local *versus* global knowledge about the surrounding environment; (d) interaction and collectivity – the degree to which agents interact in important ways with one another and/or form larger groups; and (e) stochasticity – the effect of inserting random chance into the model.

There have been several ABMs in archaeology that consider spatial interaction and/or movement explicitly. For example, Mark Lake has emphasized the degree to which we should consider how agents might share imperfect and partial spatial information about their surroundings (*e.g*. mental maps of resources in the landscape).[45] Likewise, it is possible for an ABM to consider the impact of localized knowledge on the navigational decisions made by sailors.[46] In contrast, Graham and Steiner demonstrate that it is also possible to consider some of the same issues of spatial interaction among human settlements in terms of real flows of travelling agents.[47] More precisely, they have developed an agent-based model that considers whether the random walks of individual agents from 'parent' sites to other sites in a region could approximate, stochastically and from the bottom up, the top-down spatial interaction models of settlement interaction suggested by Rihll and Wilson (see above). While both the exactness of their match with Rihll and Wilson's methods and the validity of the proposed model set-up might be argued about, their overall ABM-led approach to the issue is undeniably an attractive one.

[43] See A. G. Wilson, 'Boltzmann, Lotka and Volterra and spatial structural evolution: an integrated methodology for some dynamical systems', *Journal of the Royal Society. Interface* 5 (2008) 865-71.

[44] Following: V. Grimm, U. Berger, F. Bastiansen, S. Eliassen, V. Ginot, J. Giske, J. Goss-Custard, T. Grand, S. K. Heinz, G. Huse, A. Huth, J. U. Jepsen, C. Jørgensen, W. M. Mooij, B. Muller, G. Pe'er, C. Piou, S. F. Railsback, A. M. Robbins, M. M. Robbins, E. Rossmanith, N. Ruger, E. Strand, S. Souissi, R. A. Stillman, R. Vabø, U. Visser, and D. L. DeAngelis, 'A standard protocol for describing individual-based and agent-based models', *Ecological Modelling* 198 (2006) 115-26.

[45] M. W. Lake, 'The use of pedestrian modelling in archaeology, with an example from the study of cultural learning', *Environment and Planning B* 28.3 (2001) 385-403.

[46] G. Indruszewski and C. M. Barton, 'Simulating sea surfaces for modeling Viking Age seafaring in the Baltic Sea', in *Digital discovery*, ed. J. T. Clark and E. Hagemeister (n. 23 above) 616-30.

[47] S. Graham and J. Steiner, 'Travellersim' (n. 23 above).

All three of the above approaches – spatial interaction models, diffusion-reaction equations and ABM – have their advantages and disadvantages, as well as their conceptual overlaps, but all offer possibilities for modelling evolutionary trajectories and for factoring in patterns of geographic interaction in an explicit way, if necessary. The choice as to which one is most appropriate in a given modelling context is very much a strategic one, driven by the specific objectives in mind, and often some careful, combined exploration of several is wise.

4. Conclusion

This brief review has sought to consider the range of computational perspectives on travel and geographic interaction that are currently in use within archaeology, history, and geography, and to underscore some common problems that often still remain. Overall, it has emphasized the need (without demonstrating as yet the means by which) to develop models that are more aware of different priorities behind different kinds of travel, and different practicalities behind transitions from one transport mode to another, to name just two amongst a range of future objectives. Increasingly, the methodological and disciplinary boundaries between different kinds of spatial-analytical method are being eroded to the extent that we can integrate them in highly complementary ways. In any event, if we want to allow Mediterranean archaeology and history to contribute as effectively as it should to wider debates about the character of complex adaptive systems where humans play a role, then well-informed modelling will need to make an important contribution.

University College London a.bevan@ucl.ac.uk

Acknowledgements

I am grateful to Joel Dearden and Alan Wilson for discussion about spatial interaction models – figure 6 is part of a preliminary example of some research we are pursuing together. Thanks also to Todd Whitelaw for reading a chapter draft and discussing the modelling implications for Crete with me, as well as to Mark Lake for general discussion and three anonymous reviewers for a range of helpful, specific comments.

References

Adams, C., *Land transport in Roman Egypt* (Oxford 2007).

Ammerman, A. J., and L. L. Cavalli-Sforza, 'Measuring the rate of spread of early farming in Europe', *Man* 6 (1971) 674-88.

Ausubel, J. H., and C. Marchetti, 'The Evolution of Transport', *The Industrial Physicist* 7.2 (2001) 20-24.

Baddeley, A. J., and R. Turner, 'spatstat: an R Package for analyzing spatial point patterns', *Journal of Statistical Software* 12.6 (2005) 1-41.

Bell, T., and G. Lock, 'Topographic and cultural influences on walking the ridgeway in later prehistoric times', in *Beyond the map: archaeology and spatial technologies*, ed. G. Lock (Amsterdam 2000) 85–100.

Bennet, J., 'The structure of the Linear B administration at Knossos', *American Journal of Archaeology* 89.2 (1985) 231-49.

Bennet, J., 'Knossos in context: comparative perspectives on the Linear B administration of LM II-III Crete', *American Journal of Archaeology* 94.2 (1990) 193-211.

Bevan, A., C. Frederick, and N. Krahtopoulou, 'A digital Mediterranean countryside: GIS approaches to the spatial structure of the post-Medieval landscape on Kythera (Greece)', *Archeologia e Calcolatori* 14 (2003) 217–36.

Bevan, A., 'Political geography and palatial Crete', *Journal of Mediterranean Archaeology* 23.1 (2010) 27-54.

Bowman, A., and A. G. Wilson, ed., *Quantifying the Roman economy. Methods and problems* (Oxford 2009).

Braudel, F., *The Mediterranean and the Mediterranean world in the age of Philip II*, (London 1972).

Broodbank, C., 'Minoanisation: beyond the loss of innocence', *Proceedings of the Cambridge Philological Society* 50 (2004) 46-91.

Broodbank, C., 'Ships a-sail from over the rim of the sea: voyaging, sailing and the making of Mediterranean societies *c.* 3500–800 BC', in *The global origins and development of seafaring*, ed. A. A. Anderson, J. H. Barrett, and K.V. Boyle (Cambridge 2010) 249-64.

Brughmans, T., 'Connecting the dots: towards archaeological network analysis', *Oxford Journal of Archaeology* 29.3 (2010) 277–303.

Casson, L., 'Speed under sail of ancient ships', *Transactions and Proceedings of the American Philological Association* 82 (1951) 136–48.

Chrysoulakis, N., M. Abrams, H. Feidas, and D. Velianitis, 'Analysis of ASTER multispectral stereo imagery to produce DEM and land cover databases for the Greek islands: the REALDEMS project', in *e-Environment: Progress and Challenge*, ed. P. Prastacos, U. Cortés, J.-L. Díaz de León, and M. Murillo (Mexico City 2004) 404-17.

Collar, A., *Networks and religious innovation in the Roman Empire* (Unpublished PhD Thesis, University of Exeter, 2008).

Collischonn, W., and V. Pilar, 'A direction dependent least cost path algorithm for roads and canals', *International Journal of Geographical Information Science* 14 (2000) 397-406.

Conolly, J., and M. Lake, *Geographical Information Systems in archaeology* (Cambridge 2006).

Cotterell, B., and J. Kaminga, *Mechanics of pre-industrial technology* (Cambridge 1992).

Dijkman, J. T., 'A note on the influence of negative gradients on the energy expenditure of donkeys walking, carrying and pulling loads', *Animal Production* 54 (1992) 153–56.

Douglas, D. H., 'Least cost path in GIS using an accumulated cost surface and slope lines', *Cartographica* 31.3 (1994) 37–51.

Duncan-Jones, R., *Structure and scale in the Roman Economy* (Cambridge 1990).

Evans, T., C. Knappett, and R. Rivers, 'Using statistical physics to understand relational space: a case study from Mediterranean prehistory', in *Complexity Perspectives on Innovation and Social Change*, ed. D. Lane, S. Van der Leeuw, D. Pumain, and G. West (New York 2009) 451-79.

Evans, T., R. Rivers, and C. Knappett, 'Interactions in space for archaeological models', *Advances in Complex Systems* 85 (2011).

Fontenari, S., S. Franceschetti, D. Sorrentino, F. Mussi, M. Pasolli, M. Napolitano, and R. Flor (2005). r.walk. GRASS GIS.

Graham, S., 'Networks, agent-based models and the Antonine itineraries: implications for Roman archaeology', *Journal of Mediterranean Archaeology* 19.1 (2006) 45-64.

Graham, S., and J. Steiner, 'Travellersim: growing settlement structures and territories with agent-based modelling', in *Digital discovery: exploring new frontiers in human heritage. CAA 2006. Computer applications and quantitative methods in archaeology. Proceedings of the 34th conference, Fargo, United States, April 2006*, ed. J. T. Clark and E. M. Hagemeister (Budapest 2006) 49-59.

Grimm, V., U. Berger, F. Bastiansen, S. Eliassen, V. Ginot, J. Giske, J. Goss-Custard, T. Grand, S. K. Heinz, G. Huse, A. Huth, J. U. Jepsen, C. Jørgensen, W. M. Mooij, B. Muller, G. Pe'er, C. Piou, S. F. Railsback, A. M. Robbins, M. M. Robbins, E. Rossmanith, N. Ruger, E. Strand, S. Souissi, R. A. Stillman, R. Vabø, U. Visser, and D. L. DeAngelis, 'A standard protocol for describing individual-based and agent-based models', *Ecological Modelling* 198 (2006) 115-26.

Habermann, W., 'Statistiche Datenanalyse an den Zolldokumenten des Arsinoites aus römischer Zeit II', *Münstersche Beiträge zur Antiken Handelsgeschichte* 9 (1990) 50-94.

Helbing, D., J. Keltsch, and P. Molnar, 'Modelling the evolution of human trail systems', *Nature* 388 (1997) 47–50.

Herzog, I., 'Theory and practice of cost functions,' in *Computer Applications and Quantitative Methods in Archeology. Computing Applications in Archaeology,* ed. F. Javier Melero and P. Cano (Granada 2010) 375-82.

Horden, P., and N. Purcell, *The corrupting sea* (Oxford 2000).

Indruszewski, G., and C. M. Barton, 'Simulating sea surfaces for modeling Viking Age seafaring in the Baltic Sea', in *Digital discovery: exploring new frontiers in human heritage. CAA 2006. Computer applications and quantitative methods in archaeology. Proceedings of the 34th conference, Fargo, United States, April 2006*, ed. J. T. Clark and E. Hagemeister (Budapest 2008) 616-30.

Isaksen, L., 'The application of network analysis to ancient transport geography: a case study of Roman Baetica', *Digitial Medievalist* 4 (2008).

Kohler, T. A., J. Van West, C. Carr, and R. Wilshusen, 'Be there then: a modeling approach to settlement determinants and spatial efficiency among late ancestral Pueblo populations of the Mesa Verde region, U.S. Southwest', in *Dynamics in human and primate societies: agent-based modeling of social and spatial processes*, ed. T. A. Kohler and G. J. Gumerman (New York 2000) 145-78.

Lake, M. W., 'The use of pedestrian modelling in archaeology, with an example from the study of cultural learning', *Environment and Planning B* 28.3 (2001) 385-403.

Langmuir, E., *Mountaincraft and leadership* (London and Edinburgh 1995).

Llobera, M., 'Understanding movement: a pilot model towards the sociology of movement', in *Beyond the map: archaeology and spatial technologies*, ed. G. Lock (Amsterdam 2000) 65-84.

Llobera, M., and T. J. Sluckin, 'Zigzagging: theoretical insights on climbing strategies', *Journal of Theoretical Biology* 247 (2007) 206-17.

McRae, B. H., 'Isolation by resistance', *Evolution* 60 (2006) 1551-61.

McRae, B. H., and P. Beier, 'Circuit theory predicts gene flow in plant and animal populations', *Proceedings of the National Academy of Sciences of the USA* 104 (2007) 19885-90.

Minetti, A. E., 'Optimum gradient of mountain paths', *Journal of Applied Physiology* 79.5 (1995) 1698–1703.

Minetti, A. E., 'Efficiency of equine express postal systems', *Nature* 426 (2003) 785–86.

Newman, M., *Networks. An introduction* (Oxford 2010).

Opsahl, T., F. Agneessens, and J. Skvoretz, 'Node centrality in weighted networks: generalizing degree and shortest paths', *Social Networks* 32.3 (2010) 245-51.

Pendlebury, J. D. S., *The archaeology of Crete* (London 1939).

Perlman, P., 'One hundred-citied Crete and the "Cretan *Politeia*"', *Classical Philology* 87.3 (1992) 193-205.

Philpott, A., and A. Mason, 'Optimising yacht routes under uncertainty', *Proceedings of the 15th Chesapeake Sailing Yacht Symposium* (2001). URL:<http://www3.esc.auckland.ac.nz/people/staff/amas008/Papers/ChesapeakeStochasticRouting/Chesapeake.pdf>

Rennell, J., 'On the rate of travelling, as performed by camels; and its application, as a scale, to the purposes of geography', *Philosophical Transactions of the Royal Society of London* 81 (1791) 129–45.

Rihll, T. E., and A. G. Wilson, 'Spatial interaction and structural models in historical analysis: some possibilities and an example', *Histoire et Mesure* 2.1 (1987) 5-32.

Rihl, T. E., and A. G. Wilson, 'Modelling settlement structures in ancient Greece: new approaches to the polis', in *City and country in the ancient world*, ed. J. Rich and A. Wallace-Hadrill (London 1991) 59-95.

Sherratt, A., 'Portages: a simple but powerful idea in understanding human history', in *The significance of portages. Proceedings of the first international conference on the significance of portages, 29th Sept-2nd Oct 2004, in Lyngdal, Vest-Agder, Norway*, ed. C. Westerdahl (Oxford 2006) 1-13.

Sijpesteijn, P. J., *Customs duties in Graeco-Roman Egypt* (Zutphen 1987).

Steele, J., 'Human dispersals: mathematical models and the archaeological record', *Human Biology* 81.2-3 (2009) 121-40.

Tripcevich, N., 'Llama caravan transport. A study of mobility with a contemporary Andean salt caravan', in *73rd Annual Meeting of the Society of American Archaeology* (2008).

URL: <http://works.bepress.com/tripcevich/7>

Van Etten, J., and R. J. Hijmans, 'A geospatial modelling approach integrating archaeobotany and genetics to trace the origin and dispersal of domesticated plants', *PLoS ONE* 5.8 (2010) e12060.

Van Tilburg, C., *Traffic and congestion in the Roman Empire* (Oxford 2007).

Watts, D. J., and S. H. Strogatz, 'Collective dynamics of "small-world" networks', *Nature* 393 (1998) 440-42

Wilson, A. G., *Complex spatial systems. The modelling foundations of urban and regional analysis* (Harlow 2000).

Wilson, A. G., 'Boltzmann, Lotka and Volterra and spatial structural evolution: an integrated methodology for some dynamical systems', *Journal of the Royal Society. Interface* 5 (2008) 865-71.

Wood, J., 'Constructing weighted surface networks for the representation and analysis of surface topology', in *5th International Conference on GeoComputation (September 23rd-25th, 2000).*

URL: http://www.soi.city.ac.uk/~jwo/geoComp2000/

Zanin, M., 'Uncertainty in complex networks', *International Journal of Complex Systems in Science* 1 (2011) 78-82.

‘ONLY INDIVIDUALS’: MOVING THE BYZANTINE ARMY TO MANZIKERT

VINCE GAFFNEY, PHIL MURGATROYD, BART CRAENEN[1]
and GEORGIOS THEODOROPOULOS[2]

Introduction

In 1987 Margaret Thatcher contentiously asserted that there was no such thing as society and that there were only individuals.[3] Whilst not conceding the sentiments of the former Prime Minister, it is true that the role of the individual in society and history remains a point of contention. Individuals, of course, may achieve significance in a variety of manners. Traditionally, history has frequently emphasized the role of the ‘Great Man or Woman’, who may achieve greatness, or notoriety, through the consequences of their decision or whims. More problematic is the historical treatment of the mass of the population, or lumpen. The role of the majority of historic populations, in the absence of adequate written records, may often be glimpsed only through occasional written references, exceptional individual biographies, or, more likely, through gross statistical analyses which subsume the actions and intentions of vast numbers of people and provide a normative view of society.

There must be an argument, at least, that this situation is unsatisfactory. Historic analysis could benefit from research that sought to interpret the consequences of individual action at exponential scales and in contrast to the consequence of the actions of an eminent individual. Initiatives such as the *Prosopography of the Byzantine World* can expand the scope of historical analysis to such a degree that the patterns created by social structures can be seen, but it is still necessarily weighted towards the individuals whose names appear in the historical record.[4] Understanding the actions of hundreds, thousands, or millions of individuals would, of course, be a daunting task even if the historic data were available, but

[1] The VISTA Centre, Institute of Archaeology and Antiquity, University of Birmingham, Edgbaston, Birmingham, B15 2TT UK, (v.l.gaffney@bham.ac.uk; psm703@bham.ac.uk; bart.craenen@brunel.ac.uk).

[2] IBM Research, Dublin Research Lab, Ireland, (geortheo@ie.ibm.com).

3 The frequently cited quote ‘there is no such thing as society**.** There are individual men and women’ originates from an interview with Margaret Thatcher, published in *Woman’s Own*, 31 October 1987. However, the original transcript suggests that the phrase is an amalgam of several statements within the interview: ‘Who is society? There is no such thing! There are individual men and women’ and ‘There is no such thing as society. There is living tapestry of men and women and people’. Nothing suggests that the popular misquotation misrepresents Margaret Thatcher’s general view of the world (<http://www.margaretthatcher.org/document/106689>).

[4] *Prosopography of the Byzantine World*: <http://www.pbw.kcl.ac.uk/>.

there are methodologies that can attempt to approximate the consequences of individual action at a large scale. This chapter will outline one approach in relation to the logistical arrangements associated with the battle of Manzikert in AD 1071.

The general outline of the Manzikert campaign is relatively well known. In late February or early March AD 1071, the Byzantine Emperor Romanos IV Diogenes gathered an army, described by the Armenian monk Matthew of Edessa as 'more numerous than the sands of the sea',[5] and marched across Anatolia to the fortress at Manzikert. On August 26th he fought and lost a battle that would have far-reaching consequences for the Byzantine state as a whole, precipitating a civil war that allowed the Seljuk Turks to occupy much of Anatolia. Yet we know little of how the army transported itself over 700 miles to the site of the battle. We do not know the size of the army, the route it took or the effect it had on the communities it relied upon for supplies. There are many details regarding the organization and logistics of the Manzikert campaign that historical sources report either incompletely or not at all. Historians have effectively mined the historical records for relevant information, leaving no new sources to inform research. Direct archaeological sources for the campaign are non-existent due to the ephemeral nature of the physical evidence left by a medieval army on the move. New types of evidence are required to generate parameters within which the historical data can be analysed.

Recreating the movement of tens of thousands of men and animals, and thousands of tons of equipment is unfeasible in real life. Computer simulation represents our best tool to examine the practicalities involved. Due to the complexity of the processes involved, previous attempts at modelling the movement and provisioning of pre-modern armies have been restricted by the technology available to top-down, systemic approaches.[6] Recent advances in modelling techniques and hardware, however, give us new tools to add fresh insight to old arguments regarding the organization and supply of large bodies of soldiers marching across a pre-industrial landscape.

Agent-based modelling (ABM) is a computer-based simulation technique that has already been used to simulate the movement of large groups of individuals, from Craig Reynolds's work on flocking behaviours,[7] through to work in safety science and sociology.[8] There are also a number of archaeological ABM projects from the small and

[5] A. E. Dostourian, *The chronicle of Matthew of Edessa* (New Brunswick 1972) 231.

[6] D. W. Engels, *Alexander the Great and the logistics of the Macedonian army* (Berkeley and London 1978).

[7] C. W. Reynolds, 'Flocks, herds, and schools: a distributed behavioral model', *Computer Graphics*, 21.4 (1987) 25-34.

[8] For example: P. A. Thompson and E. W. Marchant, 'A computer model for the evacuation of large building populations', *Fire Safety Journal* 24 (1994) 131-48, and J. M. Epstein and R. Axtell, *Growing artificial societies: social science from the bottom up* (Cambridge MA 1996) respectively.

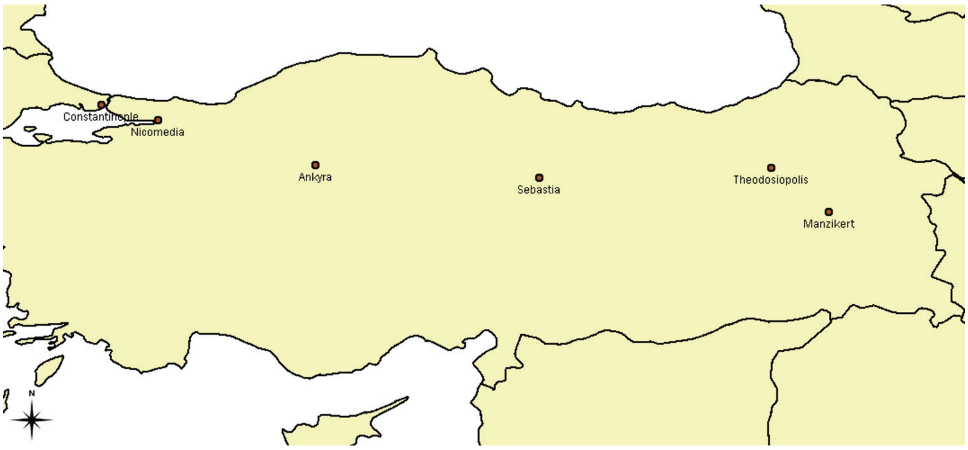


Figure 1. Anatolia

abstract[9] to the large and multidisciplinary.[10] It is mainly used to examine complex systems in which the behaviour of the whole emerges from the actions and interactions of its constituent individuals. Its architecture, in which autonomous software entities behave according to their own internal rules and inhabit a landscape that can act as a source of resources and a constraining factor to movement, is ideal for examining military organization and supply. The 'Medieval Warfare on the Grid' project[11] seeks to investigate the movement of the Byzantine army across Anatolia to the battle of Manzikert using an agent-based model to examine the organization of people and resources. By creating this kind of 'bottom up' model whereby the movement of the army as a whole is dependent on the interactions of the movement of individuals, it is possible to study the likely effects of crowding and the differences that altering army size or composition would make, something not possible with previous 'top down' models. As the army is an organization with limited levels of individual agency, soldiers not being simply left to make their own way across Anatolia, each agent can be relatively simple.

Creating an ABM for the Manzikert campaign involving tens of thousands of agents moving across a distance of over 700 miles presents many challenges, both technical and historical. Due to the incomplete nature of the historical record there is no single source for agent organization, behaviours, or attributes. The environment around which each will move is also based on incomplete and inadequate data. All historical ABMs share similar problems but, by modelling various hypothetical scenarios and noting where and how they differ, we can rule out the more impractical scenarios and establish a set of parameters within which we can re-evaluate the historical record.

[9] E. A. Smith, and J.-K. Choi, 'The emergence of inequality in small-scale societies: simple scenarios and agent-based simulations', in *The model-based archaeology of socionatural systems,* ed. T. Kohler and S. E. van der Leeuw (Santa Fe 2007).

[10] The VILLAGE Ecodynamics project: <http://village.anth.wsu.edu>.

[11] Medieval Warfare on the Grid: <http://www.cs.bham.ac.uk/research/projects/mwgrid/>.

Modern ABMs are widely used in the simulation of crowd movement[12] and this is an important part of this research into Byzantine army logistics. How the many members of the army are organized while on the march affects the speed of the army as a whole, which in turn affects consumption of resources and the subsequent impact on the communities providing those resources. Efficient movement was acknowledged by Byzantine military writers as being essential to the success of a military expedition.[13] Unfortunately there are few sources regarding the organization involved and the procedures followed, contemporary histories being more concerned with the battles and personalities of the time. It is this subject's very mundaneness that results in a lack of adequate descriptions of the daily routine of an army on the march and no detailed analysis of how this affected the territories passed through.

This paucity of evidence for such an important aspect of military life makes it an attractive subject for modelling. Although the 'Medieval Warfare on the Grid' project will model many aspects of the logistics of the Byzantine army on the march to Manzikert, it is the historical and technical problems involved in organizing the movement of the agents that comprise the army that will be dealt with here. The documentation of the design and operation of the model as a whole will require far more space than is available here and will be covered in future publications. Aspects such as the modelling of the environment, weather, energy expenditure, and the theoretical aspects of modelling are beyond the scope of this article. At present the model implements the parts of the movement system described here which deal with a full day's march. It is currently possible to model the march of an army of any reasonable size on a single march between any two points in Northern and Central Anatolia and examine energy expenditure and the results of using different criteria for deciding how this movement will affect travel time and supply requirements.

Our intention is not to model all actions involved in the march of the Byzantine army. If such a model were even possible, it would require a massive amount of effort, expertise, and computing resources. The model instead forms a null hypothesis against which to test the historical record. It is to a certain extent deterministic, dealing mainly with movement rates, energy expended, and supplies consumed. This can, however, be used to produce parameters within which historical hypotheses can be re-evaluated. This approach, along with the largely simple goal of the army, to get to Manzikert in a good enough physical condition to win a battle, means that some of the pitfalls of sociological ABMs[14] can be avoided or at least mitigated.

Historical problems

What little is known about how the late-eleventh-century Byzantine army moved, mainly comes from the nearly contemporary military treatises that proliferated in the late tenth century. These give certain practical details about many aspects of leading a military

[12] D. Thalmann and S. R. Musse, *Crowd simulation* (Vienna 2007).

[13] G. T. Dennis, *Maurice's* Strategikon*: handbook of Byzantine military strategy*. (Philadelphia 1984) 20, and G. T. Dennis, *Three Byzantine military treatises* (Washington DC 1985) 245-328.

[14] Neatly summarized by such as J. S. Lansing, '"Artificial societies" and the social sciences', *Artificial Life* 8 (2002) 279-92, and K. A. Richardson, 'On the limits of bottom-up computer simulation: towards a nonlinear modeling culture', in *Proceedings of the 36th Annual Hawaii International Conference on System Sciences* (2003)

expedition, including how to organize an army on the march. One particular treatise, *Campaign Organization and Tactics*, dated by George Dennis to the last decade of the tenth century, is an excellent source of Byzantine advice for moving large bodies of troops around.[15] Written by an experienced general, probably as a handbook for an inexperienced emperor, this treatise mainly concerns itself with logistical matters, unlike other contemporary treatises that focus on tactical matters such as skirmishing, battlefield tactics, or other aspects of warfare. Despite being written around eighty years prior to the Manzikert campaign, it still provides a lot of organizational detail written by someone who it seems had practical knowledge of leading an army across both friendly and enemy territory. The treatise not only advises on how to set up camps and move between them, but also gives us an idea of the kind of things which generals found important. It does not specify all the organization and behaviours that are needed to create an ABM, but it can give us a framework around which to test hypotheses. Other relevant, non-contradictory information can be taken from similar historical sources such as Constantine Porphyrogenitus’ treatises on Imperial campaigns.[16] There is at least one account of the campaign from someone who was actually part of the army, written by Michael Attaleiates.[17] This is not written to give specific advice about military organization but can give us useful information specific to the Manzikert campaign.

Working from the sources, a model can be built of the army’s logistical organization and movement that can be used in the ABM. Romanos IV Diogenes, though he undoubtedly made mistakes during the Manzikert campaign, had previously enjoyed a successful career as a general. He also appreciated that some of his men needed more training and would have looked to the daily routine as described in the treatises to instil order and fitness in his troops, so it is reasonable to assume that the procedures described in the sources are similar to those actually enacted on the road to Manzikert.

The sources allow certain assumptions to be made for the purposes of modelling, around which our movement model can be constructed. These are:

- The army is likely to have picked up troops along the way;
- Supplies would also have been picked up along the way;
- There is likely to have been a route planned in advance;
- The army is likely to have had a regular routine for marching;
- The army is likely to have had a set order of march;
- The army is likely to have had a consistent and organized method for setting up camp;
- The emperor makes all the strategic decisions;
- The length of a daily march must be flexible and practical.

[15] Dennis, ‘Three Byzantine military treatises’ (n. 13 above).

[16] J. F. Haldon, *Constantine Porphyrogenitus: three treatises on Imperial military expeditions* (Vienna 1990).

[17] Attaleiates, *Historia*, ed. I. Bekker (CSHB) (Bonn 1853).

The army is likely to have picked up troops along the way

Some elements of the army, particularly the provincial levies, would have been scattered across various parts of Anatolia in February 1071. Requiring these troops to march all the way to Constantinople only to have to march all the way back through their own territories on their way to Manzikert seems highly impractical. The Anatolian *thematic* troops in particular, being levies that spent most of the year as farmers, would have been locally mustered at towns near where they lived and would have waited for the army at the nearest large settlement on the route.[18] These settlements would have had the resources and infrastructure to feed and shelter hundreds or thousands of troops as they would also have been stockpiling supplies from the surrounding area for the main body of the army.[19]

Supplies would also have been picked up along the way

There are plenty of contemporary historical accounts of the army requiring settlements to provide supplies of both food and equipment in return for the commutation of taxes.[20] Engels's work on the army of Alexander the Great illustrates that there are many problems involved with carrying supplies for too many days, with twenty-five days being given as an absolute maximum.[21] Given the total length of the journey, around six months for those that started at Constantinople with the Emperor, the army would have needed resupplying many times along the route.

There is likely to have been a route planned in advance

Picking up supplies and personnel on the way necessitates a knowledge of the route in advance, at least in general terms. If Constantine Porphyrogenitus' treatises on military organization are indicative of campaign planning, a great deal of effort went into ensuring the route of march was suitable for the army and its objectives.[22] The entirety of the route would have been through lands at least nominally under Imperial control and as such would have been well known to the Empire's administrative machinery. Officials would have been dispatched to warn settlements along the proposed route that an army would be passing through and that supplies would be needed. This enabled the *themes*, the administrative districts of the Empire, to muster their men and resources in settlements convenient for the army. Michael Attaleiates' eyewitness account places the army in Ankyra, at Krya Pege (thought to be on the River Halys between Ankyra and Charsianon), and at Theodosiopolis, giving us some set points for the route. One of the advantages of ABMs, however, is their ability to create 'what if?' scenarios in order to examine the circumstances behind the decisions made. Based on the information, alternative models can be run in which the overall route is not pre-determined but is based on certain criteria. We can then investigate

[18] Haldon, *Constantine Porphyrogenitus* (n. 16 above) 83.

[19] J. F. Haldon, *Warfare, state and society in the Byzantine world 565-1204* (London 1999) 182.

[20] Haldon, *Warfare, state and society* (n. 19 above) 140.

[21] Engels, *Alexander the Great* (n. 6 above).

[22] Haldon, *Constantine Porphyrogenitus* (n. 16 above), particularly text B.

the implications of choosing each route, which in turn will enable us to form hypotheses regarding the decision-making process of the army on campaign.

The army is likely to have had a regular routine for marching

The treatise on *Campaign Organization and Tactics* gives a series of steps for organizing movement for the day.[23] It specifies that the Imperial tent is the first to be set up at a new camp and the first to come down when leaving the camp. Each morning a set routine is followed:

- The trumpets sound to get everybody up;
- Officers in command go to the Emperor to get their orders;
- The trumpets sound again and some units are sent out of the camp to cover the movement of the rest of the army as it exits the camp;
- Everyone ensures that equipment and baggage is loaded onto pack animals;
- The trumpets sound for a third time and the Emperor rides out of the camp with everyone else following in turn.

Although this treatise is not exactly contemporaneous with the Manzikert campaign, this series of actions forms a plausible starting point for our own modelling. It does not require a computationally intensive model to tell us that the movement of tens of thousands of troops cannot be organized as an uncoordinated free-for-all, even if circumstances may cause it to end up that way.

The army is likely to have had a set order of march

Historical sources contain a sample order of march[24] which can be used as a template for the way the army moves in the ABM (fig. 2). A set order of march enables each unit to know its place in the army and creates a framework for movement. New units can quickly determine their place within this framework. In some medieval armies, the order of movement is varied to ensure different units get first access to clean water and have fewer problems with the dust kicked up by preceding units.[25] It is not certain to what size of army this applies; certainly a smaller army with fewer units would find this easier to organize. For the purposes of this simulation, the army on the Manzikert campaign is assumed to be too large for this to be feasible.

Having a set order of march also helps with setting up and breaking camp. Units arriving first can head to the far side of the camp area and be out of the way when later units arrive. Similarly, when setting off, the first units will already be at the side of the camp nearest the direction of travel and will not have to manoeuvre round units who will be setting off later.

[23] Dennis, *Three Byzantine military treatises* (n. 13 above) 277.

[24] For example: Haldon, *Constantine Porphyrogenitus* (n. 16 above), although this varied depending on circumstance.

[25] C. J. Rogers, *The middle ages* (Santa Barbara 2007) 76.

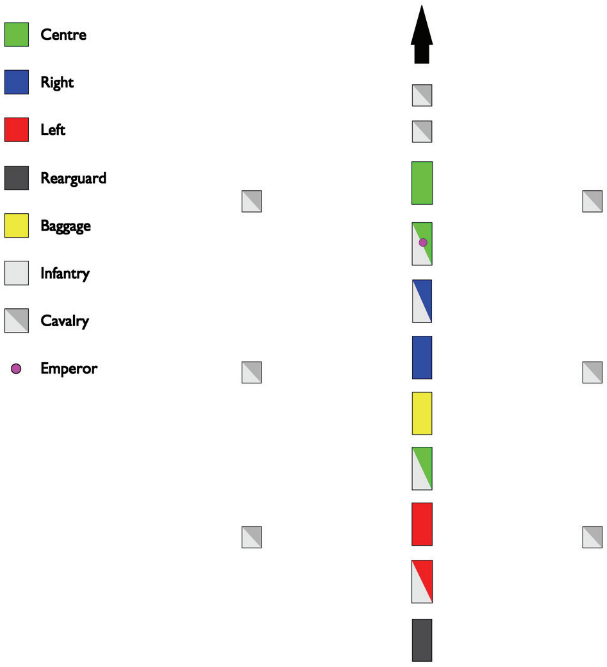


Figure 2. March formation in friendly territory (after Haldon, 1999)

The order of movement in friendly territory as described by John Haldon is:[26]

- Advance scouts;
- Vanguard;
- Infantry centre;
- Cavalry centre + Emperor;
- Cavalry right wing;
- Infantry right wing;
- Baggage and siege train;
- Cavalry centre second line;
- Infantry left wing;
- Cavalry left wing;
- Infantry rearguard;
- Rearguard.

There were units of outriders on either side of the column. The linear nature of this marching formation, as opposed to the semi-deployed formation that was used in enemy territory, ensured that roads could be more closely followed, resulting in less damage to

[26] Haldon, *Warfare, state and society* (n. 19 above) 162.

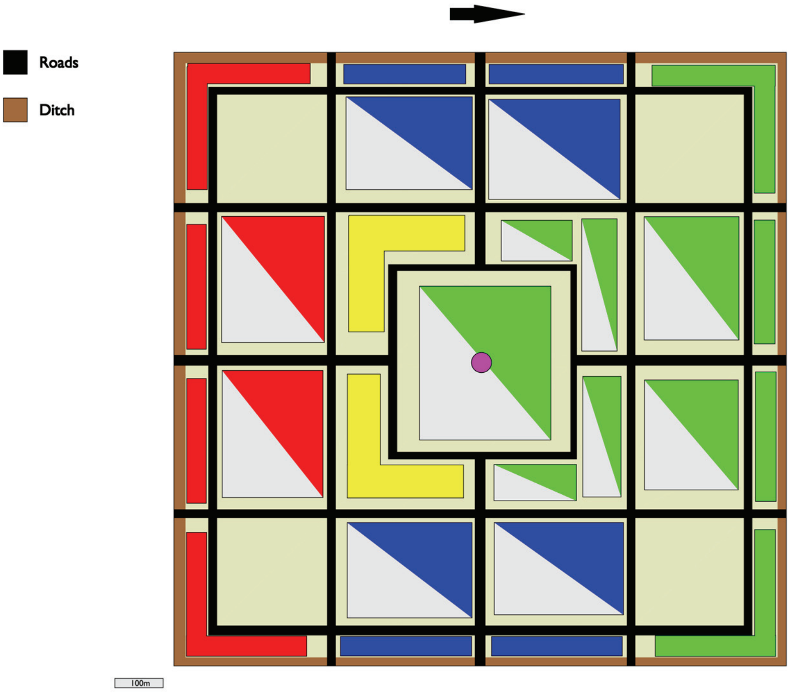


Figure 3. The camp, populated by the units from fig. 2.

properties on the path of march. Within the model, each unit needs to be assigned to one of these categories when joining the main body of the army. The Emperor and his household troops can be considered the cavalry centre. Other troops can be added on an arbitrary basis to ensure all other categories are of roughly equal quantity, with the advance scouts and vanguard as cavalry and the rearguard as infantry.

The army is likely to have had a consistent and organized method for setting up camp

It is obvious from the treatise on campaign organization that there were well-established rules for setting up camp:

> The best generals and those who have acquired a good deal of experience over a long period can study the size of the body of troops drawn up within the fortifications and determine well in advance the precise circumference of the site in which the whole army, horse and foot, is going to encamp.[27]

Although there is no consistent and reliable camp plan illustrated in the treatise for campaign organization, George Dennis managed to piece together the details from the text to create a hypothetical plan.[28] This plan has been used as the basis for camping locations for each of the units (Figure 3).

[27] Dennis, *Three Byzantine military treatises* (n. 13 above) 247.

[28] Dennis, *Three Byzantine military treatises* (n. 13 above) 335.

The camp will be positioned so that the units at the head of the column are camped nearest to the side of the camp facing 'forwards' (usually east). This will ensure they do not have to manoeuvre round everyone else's tents on the way into or out of camp. Taking the order of march from Haldon, it is clear that it is possible to fill in the camp plan from Dennis in an orderly way, ensuring the furthest reaches of the camp are set up first. Although no location is specified for the baggage train in the treatise, a space has been allocated for it on the plan as it will form a discrete unit on the march and therefore will simplify organization if it is separate. Dennis points out that there are no specific instructions on how the corner areas are used, other than for light infantry if they are too numerous to fit into their allocated areas.[29] This then can be used as valuable overspill space in case the camp becomes too tightly packed or extra space for baggage is required.

The Emperor makes all the strategic decisions

Strategic decisions regarding the route the army will take or whether the army will split into separate columns will all be made by the Emperor within the ABM. In reality the strategic situation will have been discussed within the upper ranks of the army. The extent to which this happened on the Manzikert campaign is unknown and this level of detail is unnecessary from the point of view of the model. Some historical sources mention that the Emperor detached his baggage train from the main body of the army at some point on the route and went a separate way. If this event is modeled, a substitute 'Emperor', who is the highest ranking officer remaining with the main body of the army, will adopt all the Emperor's decision-making functionality.

The length of a daily march must be flexible and practical

Past work has assumed either a certain distance for the army to move in a day[30] or has tried to calculate this based on the size of the army.[31] Calculating a daily movement distance runs contrary to our goals in building an ABM; the distance an army can move is determined by its size and organization and not vice versa. The model should also be able to help examine the relationship between army size, organization, composition, and speed of movement, along with other factors, such as weather and terrain, to be dealt with in future publications. Therefore the model requires a dynamic system, where the length of a day's march is dependent on how far an army of a given size can comfortably move. If an army struggles to reach the day's camp in time to get set up in the daylight, then the length of march must be shortened. If the army finds itself arriving in plenty of time, its length of march can be increased.

Our system cannot assume that the army simply moves until the end of the day and then stops. Historical sources state that it was standard Byzantine practice to send a group

[29] Dennis, *Three Byzantine military treatises* (n. 13 above) 329.

[30] Engels, *Alexander the Great* (n. 6 above).

[31] J. H. Pryor, 'Introduction: modeling Bohemond's march to Thessalonike', in *Logistics of warfare in the age of the Crusades*, ed. J. H. Pryor (London 2006).

of surveyors a day’s march ahead to set out the next camp site.[32] This camp was set out to the same layout each time so each unit would know upon arrival where they were supposed to pitch their tent. Our model replicates this system, yet still allows some flexibility in the case of movement problems; it is dynamic but always a day behind. By the time the army has all reached the day’s camp the surveyors have already planned out the following day’s camp. If the length of a daily march is too long and the troops are arriving late in the camp, the surveyors will have to make allowances for this when setting up subsequent camps.

The sources give us a framework around which to base our movement rules. They indicate locations along the route that can allow us to construct hypothetical routes of march. The usefulness of ABMs, however, lies in the creation of ‘what if?’ scenarios that let us examine the implications of changing parts of the system. In the initial ABM each agent will follow its rules based on the information it has. This creates an ideal example of behaviour that can be used as a null hypothesis to compare with information derived from other sources, such as archaeological or historical research. If, under optimum conditions, an army of a certain size takes longer than six months to reach Manzikert, then it is highly unlikely that in real life it would fare any better. But the model’s performance can also be deliberately degraded to introduce negative aspects, such as incompetence, into the system. We can investigate what happens when the camp is not set up in advance and the location has to be chosen on an *ad hoc* basis by the units at the front of the column. This enables the ABM to model what happens when water needs are ignored or streams dry up. Historical sources give little information about the organizational problems that *did* happen but the model can provide new evidence for what *could have* happened.

Creating a historically plausible framework is only the start of the process of modelling. It must also be translated into a working computer model. The mechanisms by which the agents plan and execute their movement are vitally important to the model as a whole. The whole point of the campaign was to get the army across Anatolia to Manzikert; if the movement aspect of the ABM does not work accurately, efficiently, and reliably this will not happen, regardless of the historically derived parameters.

Technical problems

The method of modelling the Byzantine army’s march to Manzikert has to take into account all of the above factors. It has to be reliable enough to ensure that if agents leave the marching army it is because of legitimate behaviour of the agents and not an error of the model. It also has to be robust enough to cope with any problems that may occur on the way. Troops that encounter problems should attempt to resolve them in plausible ways. The movement system must be able to cope with the macro-level decisions that the Emperor makes regarding the overall route, as well as the micro-level decisions made by a single soldier making his way around the camp. The behaviour emerges from how the movements of individual agents on a day-to-day basis affect the army’s performance and speed as a whole. The speed of the army cannot be extrapolated from the speed of its individual components and we do not know how increasing the number of men and

[32] Dennis, *Three Byzantine military treatises* (n. 13 above) 249.

animals in the army will reduce the overall speed. Modelling the movement of the army in a detailed and plausible manner can help us understand these factors in a way impossible through other methods.

The core of the movement system is route planning. A route plan is a series of individual moves that each agent will have to make in order to reach its destination. The army needed to have a planned route in order to ensure that the settlements on the route could have supplies ready. Supplies sufficient to feed and equip the army would be inconvenient and expensive to move over land and changes in this macro route, while possible, would be difficult to effect. But it is not just the army as a whole that needs a route plan; each individual agent will need one in order to move anywhere. A route plan must be worked out in advance, but can be changed at any time if circumstances render the initial route invalid. Any time an agent needs to move, it takes its current location and intended destination and passes these to a route-planning algorithm which formulates a series of individual moves that will get the agent from point A to point B.

Route planning is a very well-documented branch of computer science.[33] Several methods exist to allow an agent to select between different possible routes and manoeuvre itself across a landscape. The movement system needs to cope with the planning of the route of the army as a whole as well as the everyday movement of each agent. Any solution that treats both strategic route planning and micro-level movement in the same way risks either overloading itself when planning macro routes or missing out on the detail required to plan plausible micro routes.

A route planning*

One of the most popular algorithms used for route planning is the A* algorithm.[34] A* is a graph search algorithm that is used for route planning by representing each possible destination as nodes on a graph. The algorithm then searches through the nodes to find the route to the destination with the least 'cost': cost here can represent anything, including distance, energy expended via established models from medical science,[35] time taken, or any other method of differentiating between routes. The presence of terrain and weather within the environment means we can model the energy expenditure of each individual in the army, the mechanisms for which will be dealt with in future publications. At each step of its search the algorithm combines the cost to get to this node with an estimate of the cost to get from this node to the destination. If the cost of each move is represented by the energy expended to make the move, then the search algorithm will attempt to find the route that uses the least energy.

The estimated part of the equation, or heuristic, is based on the distance to the destination, and its presence allows us to prioritize routes that get us closer to our goal. Therefore the nodes first searched are the ones that cost less energy to reach and that reduce

[33] From E. W. Dijkstra, 'A note on two problems in connexion with graphs', *Numerische Mathematik* I (1959), 269-71 onwards.

[34] S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach* (New Jersey 2003) 97.

[35] D. Balado, *ACSM's guidelines for exercise testing and prescription* (Baltimore 1995) 276.

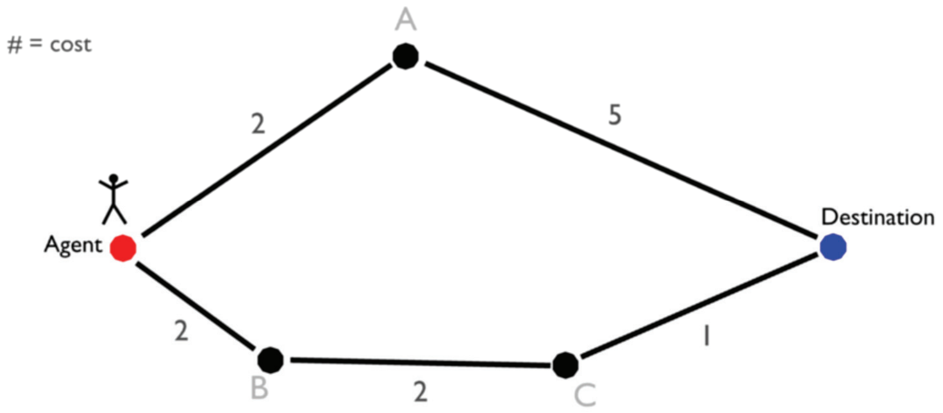


Figure 4. An example of A* in action

the distance to the destination. This ensures the search procedure prioritizes more likely routes in order to speed up the process.

In the example in figure 4, the agent has two possible routes to its destination. In order to prioritize its search towards the route likely to be the most efficient, it examines each node based on the cost to reach it, plus an estimate of the cost to get from that node to the destination. For ease of calculation, this estimate is the distance between the node and the destination, equivalent to assuming the cost of each move will be 1 from there onwards (Table 1).

Node	Cost to reach	Estimated cost to destination	Total
A	2	1	3
B	2	2	4

Table 1: A first planning move*

In this case the cost to reach nodes A and B is the same – 2. However, when estimating the cost to get from each of these nodes to the destination, the estimate for A is 1, whereas the estimate for B is 2, giving a total estimated cost of 3 for node A *versus* 4 for node B. This means that from where the agent starts, the move to node A seems the most attractive. Now the route planner expands node A and sees that the cost to move to its destination is an extra 5, making a total of 7 (table 2).

Node	Cost to reach	Estimated cost to destination	Total
A	2 + 5 = 7	-	7
B	2	2	4

Table 2: A second planning move.*

This exceeds the estimated total of moving via node B; so, with node B now looking the most attractive option, the route planner backtracks to node B and goes from there. Following the same procedure, the cost via this route will never exceed the cost via node A, so the route planner will complete its plan and return to the agent a route plan of 3 moves: starting location to B, B to C, C to destination. In this example the route planner ended up expanding all possible nodes to return the most efficient plan, but if the cost of moving from A to the destination had been more in line with the estimate, then the best route would have been planned in 2 moves, with no need to examine further nodes B or C, saving much processing time.

In order to use A*, which is a very simple, efficient, and effective search algorithm when properly used, the environment must be represented in a way that can be represented in graph form. During the development of the model, two main approaches were tried.

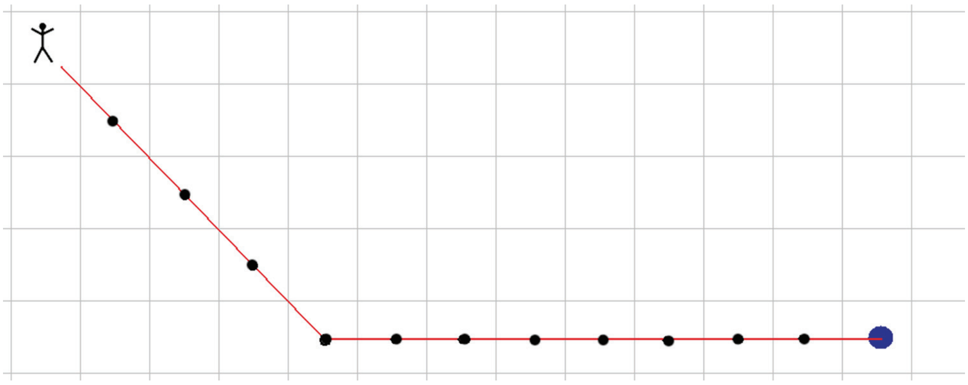


Figure 5. Grid movement

Grid movement

Grid movement (Figure 5), moving each agent from one cell of the environment to another, has the advantage of being easy to process conceptually and to programme. Each agent occupies a square of the environment, the number of agents in each cell is limited based on the size of the cell and the size of the agents (cavalry taking up more space than infantry, for example). Agents move from their cell to an adjacent cell, with each move having a cost associated with it. This cost can be used to plan routes based on the A* planning algorithm. Disadvantages with this approach arise when the route being planned results in a large number of cells of diverse costs being visited. Short distances are resolved quickly but long distances face an ever-increasing trade-off between lengthy processing time and sub-optimal routes. A key factor in A* planning performance is tree depth, a measure of the minimum number of nodes needed to reach any given destination. Each cell further away from the start that the destination is, the greater the tree depth. Unless the heuristic involved is very accurate, each increase in tree depth also increases the tree width, the number of nodes visited per step closer to the destination. This rapidly increases the number of nodes to be visited by the planner.

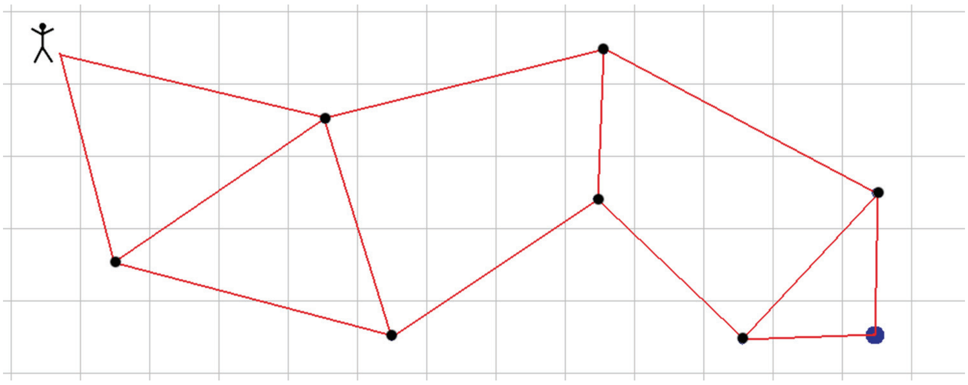


Figure 6. PRM movement

Probabilistic RoadMap movement

Probabilistic RoadMap (PRM) movement (Figure 6) relies on a series of nodes to be created over the environment. These nodes are linked by edges, which are the paths that an agent can move between nodes. Therefore an agent can move from node to node instead of from cell to cell, aggregating a whole series of movement costs into a single cost of moving from one node to another. This decreases the processing time of A* route planning because the number of steps required to traverse large numbers of cells is reduced. Disadvantages with this method arise when the nodes or edges are not created in places that would enable agents to access certain resources. This could render an agent unable to perform tasks that it would be able to do in real life, because of the ABM’s design, a situation that is clearly to be avoided.

Why not just A?*

A* graph search algorithms are both admissible and complete when properly implemented. The term ‘complete’ means that a solution will always be found if one exists; if an algorithm is ‘admissible’ then it returns the optimal solution as long as the heuristic does not overestimate the cost of reaching the goal. The closer the heuristic is to the actual cost of movement, the quicker and more optimally the algorithm will run, as seen in table 3.

Comparison of heuristic to actual cost	**Result**
Heuristic overestimates cost of remaining moves.	Algorithm runs fast but result may be sub-optimal.
Heuristic estimates cost accurately.	Algorithm runs fast, result is optimal.
Heuristic underestimates cost of remaining moves.	Result is optimal but algorithm runs inefficiently, expanding more nodes than necessary.

Table 3: Effects of heuristic values on the running of the A algorithm.*

A* route planning over an environment consisting of discrete cells works most rapidly when the distance covered is small. When the distance covered is considerable, matching the heuristic to the actual movement, cost becomes more important. Using cells of 5m^2, the area covered by the ABM results in a grid of 280,700 x 88,900 cells. Planning a route across the whole of this area presents an insurmountable problem for the A* algorithm, unless the heuristic estimates the remaining cost precisely. Finding a plausible route, however, relies on a variety of movement costs, making any estimate inaccurate. The difference in desirability between a smooth, flat road and a hike over a hilltop is considerable. The specific movement values are not important, but the relationship between them is. For a steep movement uphill to be twice as undesirable as a smooth level movement, the movement cost must be twice as much. As the minimum and maximum movement values diverge, so the heuristic is more likely to be further from the actual cost of movement. So with straight A* we are stuck in a situation with two undesirable options:

- Ensure the distances are never long by having more preset waypoints;
- Ensure the movement costs are more predictable by making the costs differ by smaller amounts.

The first option is undesirable because it reduces the autonomy of the agents, which ideally should be able to choose their own route based on our defined rules, not have it preordained from the start. The second option is likewise undesirable because the agents should make sensible route-planning choices, not be more likely to select an unreasonable route because of a design decision.

What is the solution?

Our A* route planning can be set up to work well over either long or short distances. Thankfully these can be combined by using a mixture of grid-based and PRM movements. Supplies would have been concentrated at settlements which in turn tend to be linked by roads. Therefore the army would have tended to move from settlement to settlement along the road network. This makes the army's macro-level route planning ideally suited for PRM movement. Whereas true PRM creates a random series of nodes spread over the environment, our node network can be created using settlements as nodes. Edges can be automatically created between neighbouring nodes and costs assigned to each edge will be based on the likely supply level of each settlement and the presence of a road linking them. The A* route planner can be run on this node network to create a macro route, which is then converted to a series of waypoints over which the grid-based A* planner can work. If a specific route needs to be tested, then these waypoints could be specified in advance and this step skipped.

Even chopping the route into discrete sections, an unmodified A* route planner will not do everything needed in a reasonable timescale. One way in which performance can be improved is to have a dynamic method of calculating the heuristic modifier. The heuristic calculates how many steps it takes to reach our destination and assigns a cost of 1 per cell. If movement costs are 1 or higher per cell, this means the result will be the lowest cost route. It can be assumed that the average cost of movement will be greater than 1 per cell and increase the heuristic cost accordingly. This will speed up the route planning but if the heuristic cost exceeds the actual cost, then the route may not be the one with least cost.

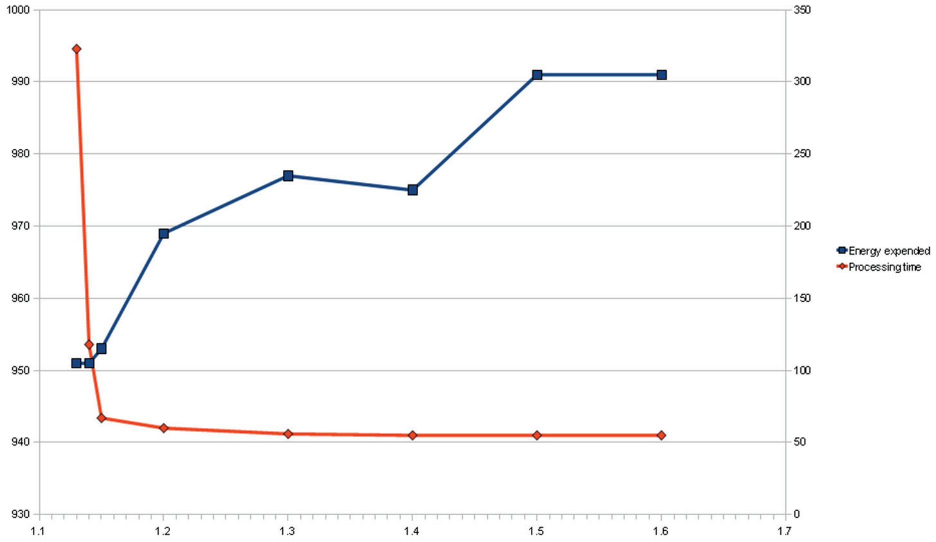


Figure 7. Graph showing relationship between heuristic modifier and performance

There is clearly a trade-off between performance and quality. In order to test this, an arbitrary measurement of energy expended during movement was created. An agent used more units of energy when moving uphill and along longer routes than level or downhill on shorter routes. This gave a coarse measure of a given route’s easiness: the lower the energy expended, the more the route planner avoided going uphill or on long detours. Tests reveal there will be a point of reasonable compromise where the route planned is near optimal and the processing time is acceptable (Figure 7).

As can be seen, when the heuristic cost per cell is between 1.14 and 1.15, the processing time is acceptable and the route is very close to being optimal. Our dynamic method of calculating this point assumes a deliberately high heuristic modifier and calculates the route. It repeats this process with decreasing values for the heuristic modifier until it exceeds a set time limit for the route-planning process. It then takes the route calculated by the last successful run and uses that as the route. This method ensures all routes fall within an acceptable area of accuracy and performance.

A separate step can be added which attempts to find a route along a road from the start location to the destination. If such a route exists, then this route will be taken, working on the assumption that a road route is preferable to an off-road route even if the road route is longer (Figure 8).

Conclusion

Just as there were a variety of levels on which campaign organization worked in the Byzantine army, so there are a variety of levels on which this simulation is modelled. The macro decisions regarding the overall route can be either pre-scripted or based on criteria built into our PRM network. This creates waypoints that are navigated between using our grid-based A* route planner. The route is then subdivided based on the length of the daily march and a series of camps specified. These can be altered at the end of each day based

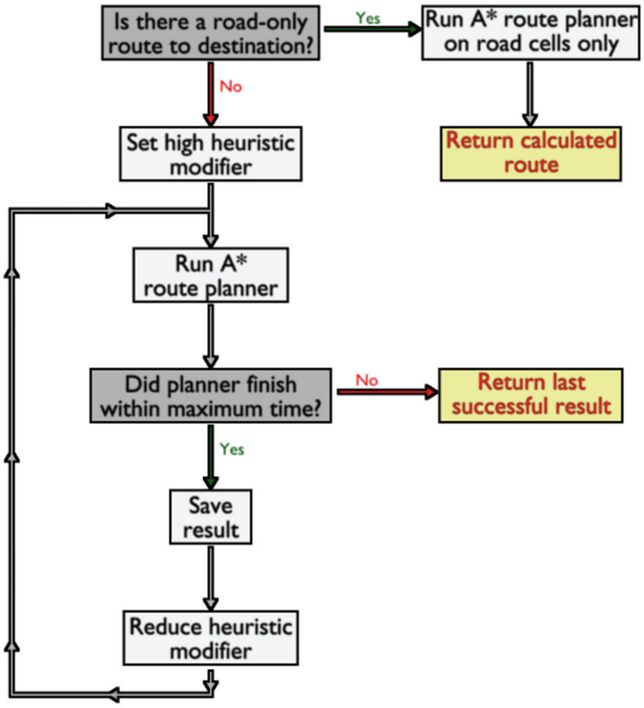

Is there a road-only route to destination?
Yes
Run A* route planner on road cells only
Return calculated route
No
Set high heuristic modifier
Run A* route planner
Did planner finish within maximum time?
No
Return last successful result
Yes
Save result
Reduce heuristic modifier

Figure 8. Flowchart detailing dynamic route-planning logic

on the experiences of the previous day's march. Each day the army makes and breaks camp according to a set routine.

The result of our technical efforts will be an ABM that incorporates the terrain and weather of Anatolia with hypothetical supply levels and distribution into an environment over which our agents can travel. By creating and running a series of 'what if?' scenarios with armies of different sizes and compositions travelling in a landscape with different levels of supply and water availability, the results of these scenarios will allow us to draw conclusions about how armies behave in ideal circumstances. We can then compare this with the historical record and attempt to explain any differences.

Finally, whilst appreciating that all models are wrong,[36] the project attempts to make a break with traditional historic analysis in the manner in which it incorporates individual action within an interpretative framework that can add to our understanding of a historic event or process. While small glimpses into individual behaviours can be found in the work of Attaleiates and modern research tools such as the online *Prosopography of the Byzantine World*, these do not allow us to examine the interactions common in complex systems. With the emergent behaviour modelled in the ABM, the project seeks to add new evidence to existing debates about the movement of large numbers of troops across a pre-industrial landscape. We can examine the relationship between human and animal stamina, unit organization, and how this impacts on army speed in ways previously

[36] G. E. P. Box and N. R. Draper, *Empirical model-building and response surfaces* (New Jersey 1987) 424.

unavailable. To the movement model will be added modelling of the use and transport of food and supplies. Levels of supplies can be varied to investigate the impact on the army as a whole. With their modular nature, ABMs can test different hypotheses regarding agricultural productivity and settlement by ensuring that the route-planning decision making remains constant, while varying the resource levels. The differences this creates between otherwise identical runs of the model can help inform our interpretations of how Byzantine military operations were planned and executed.

Vince Gaffney (*University of Birmingham*) v.l.gaffney@bham.ac.uk
Phil Murgatroyd (*University of Birmingham*) p.s.murgatroyd@bham.ac.uk
Bart Craenen (*Brunel University*) bart.craenen@brunel.ac.uk
Georgios Theodoropoulos (*University of Durham*) theogeorgios@gmail.com

ON USING DIGITAL RESOURCES FOR THE STUDY OF AN ANCIENT TEXT: THE CASE OF HERODOTUS' *HISTORIES*[*]

ELTON BARKER, LEIF ISAKSEN, NICK RABINOWITZ, STEFAN BOUZAROVSKI, and CHRIS PELLING

1. Approaching the world of Herodotus

About a generation after the unexpected and remarkable defeat of the Persians by a ragtag assembly of Greek city-states, Herodotus of Halicarnassus set out to investigate the great deeds of these two peoples and why they had come into conflict in the first place. His enquiry took him all round the Mediterranean, 'traversing in detail towns of men both small and great alike, for', as he explains, 'of the places that were once great, most have now become small, while those that were great in my time were small before' (ὁμοίως σμικρὰ καὶ μεγάλα ἄστεα ἀνθρώπων ἐπεξιών· τὰ γὰρ τὸ πάλαι μεγάλα ἦν, τὰ πολλὰ σμικρὰ αὐτῶν γέγονε· τὰ δὲ ἐπ' ἐμεῦ ἦν μεγάλα, πρότερον ἦν σμικρά: 1.5). Previous scholarship has typically depicted that world as being divided into three separate units, Europe, Asia, and Libya, and has highlighted the importance of water bodies, in particular rivers, for organizing that space.[1] Such 'mental maps' (see Figure 1)[2] are useful for drawing attention to the social, political, and cultural constructions of space,[3] but are

[*] We are extremely grateful to the Arts and Humanities Research Council for funding this research. The initial funding (for Barker's Early Career grant) ran from September 2008 to June 2010. In July 2013 the AHRC awarded Hestia 'Follow on' funding for one year in order to consolidate the technologies and promote knowledge exchange with groups within and outside academia. An earlier account of this work, entitled 'Mapping an ancient historian in a digital age: the Herodotus Encoded Space-Text-Image Archive (HESTIA)', is published by the *Leeds International Classical Journal* 9.1 (2010): (<http://www.leeds.ac.uk/classics/lics/>). All translations of Herodotus are our own.

[1] For the importance of natural boundaries to Herodotus' conception of history and, in particular, rivers as demarcation boundaries and markers of transgression, see esp. H. Immerwahr, *Form and thought in Herodotus* (Cleveland 1966), in the index under *river motif*; *cf.* D. Braund, 'River frontiers in the environmental psychology of the Roman world', in *The Roman army in the east*, ed. D. L. Kennedy, *JRA* Supp 17 (Ann Arbor 1996) 43-47.

[2] This map has been chosen because, published under a Wikimedia Commons licence, it can be reproduced without any infringement of copyright. But it also typifies other reproductions of Herodotus' world: see O. A. W. Dilke, *Greek and Roman maps* (London 1985) 58.

[3] See *e.g.* N. J. W. Thrower, *Maps and civilization: cartography in sulture and Society* (Chicago 1996); J. B. Harley, 'Deconstructing the map', *Cartographica: The International Journal for Geographic Information and Geovisualization* 26 (1989) 1-20; P. Jackson, *Maps of meaning: an*

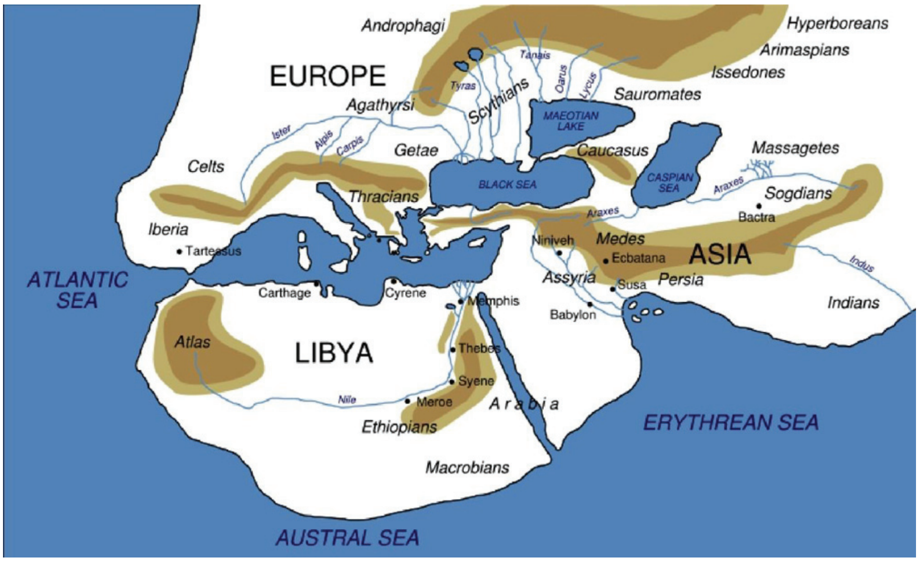


Figure 1: Herodotus' world (source: http://en.wikipedia.org/wiki/File:Herodotus_world_map-en.svg, Last accessed 15.11.2013)

inevitably hampered by the medium of representation. The printed map can offer only a compressed snapshot of how space is conceived across Herodotus' narrative as a whole, and tends to define it, moreover, in terms of separate territories, whose boundaries are rigidly policed by rivers, seas, and the odd desert. Herodotus himself appears to be deeply sceptical of the capacity for maps – a new technology at the time of the author – to represent adequately the kinds of engagement between peoples that he depicts in his narrative.[4] Our own technological advances, however, are now creating the possibilities not only of relating space more systematically and sensitively to Herodotus' text, but also of broadcasting the results to a much wider community. In short, the digital medium offers the potential to capture a world in flux, which changes over time (as Herodotus recognizes in the quotation above), and to represent places within that world in relation to each other.

Involving the collaboration of researchers from Classics, Geography, and Archaeological Computing, Hestia has been experimenting with various ways of extracting and analysing the different ways in which geographic space is organised in Herodotus' narrative.[5] Using the latest digital technologies in combination with close textual study, we examine the

introduction to cultural geography (London 1994); K. T. Jones, 'Scale as epistemology', *Political Geography* 17 (1998) 25-28.

[4] Hdt 4.37. Rather than talking in terms of abstract conceptions of space, Herodotus narrates the geography from the perspective of one travelling: *i.e.* the order of description does not necessarily map onto the nearest place but the one that comes next, as one travels to it. For 'hodological' approaches to space in Herodotus, see A. Purves, *Space and time in ancient Greek narrative* (Cambridge 2010) 144-49; *cf.* P. Janni, *La Mappa e il Periplo. Cartografia antica e spazio odologico* (Marcerata 1984); J. S. Romm, *The edges of the earth in ancient thought: geography, exploration, and fiction* (Princeton 1994) 34-44.

[5] For more information, please go to the project website: <http://hestia.open.ac.uk/>.

geographical concepts through which Herodotus describes the conflict between Greeks and Persians,[6] and explore the connections between places that Herodotus makes over the course of his narrative. Our research aims are, first, to question the received Greek-other dichotomy of the *Histories* by revealing ambiguities in the ways in which Greek and Persian worlds are mapped out and interrelated,[7] and, second, to use a network analysis for thinking about Herodotus' textualization of space – how, in other words, the narrator puts spatial ideas and concepts into words.[8] In both cases, we experiment with various kinds of network graph to investigate the topological relationships between places rather than try to reconstruct a topographical reality.

For the purposes of this volume we discuss three main digital aspects to the project: the data capture of place-names in Herodotus; their visualization and dissemination using digital technologies like Geographical Information Systems (GIS)[9] and web-mapping applications such as Google Earth and Timemap.js;[10] and the interrogation of the relationships that Herodotus draws between different geographical concepts using the digital resources at our disposal. Our concern here will be to set out in some detail the digital basis to our methodology and the technologies that we have been exploiting, as well as the problems that we have encountered, in the hope of contributing not only to offering a more complex picture of space in Herodotus but especially to setting out a basis for future digital projects across the humanities that deal with the spatial representation of large text-based corpora.[11] With

[6] In a recent article T. Harrison, 'The place of geography in Herodotus' *Histories*', in *Travel, geography and culture in ancient Greece, Egypt and the Near East*, ed. C. Adams and J. Roy (Oxford 2007) 44-65, makes the '*search of* geography and its place in the *Histories*' (44) the highest priority.

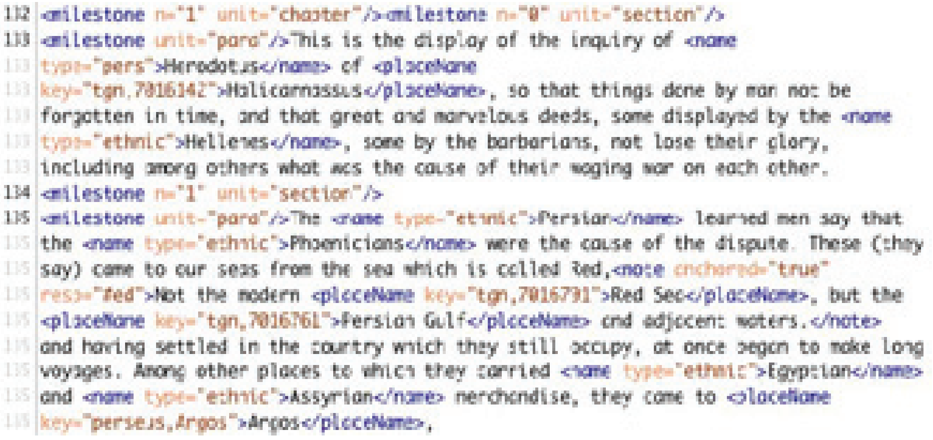
[7] *E.g.* E. Hall, *Inventing the barbarian* (Oxford 1989); P. A. Cartledge, *The Greeks and others* (Bristol 2002 [1993]). The 'East *vs*. West' dichotomy in Herodotus has recently been challenged, however: see for example, Chris Pelling, 'East is east and west is west – or are they?', *Histos* 1997 51-66: <http://research.ncl.ac.uk/histos/documents/1997.04PellingEastIsEast5166.pdf> (revised and updated with further bibliography, in *Herodotus volume 2: Oxford readings in Herodotus*, ed. R. V. Munson (Oxford 2013) 360-79).

[8] Network theory in the field of ancient Greek history: I. Malkin, *The returns of Odysseus* (Berkeley 1998); C. Constantakopolou, *The dance of the islands* (Oxford 2005). The two-dimensional 'Cartesian'-style map has dominated Western horizons since the Enlightenment: A. J. Gurevich, *Categories of Medieval culture* (London 1985). For a fuller articulation of the geographic theory underpinning our analysis of Herodotean space, see E. T. E. Barker, S. Bouzarovski, C. B. R. Pelling, and L. Isaksen, ed., *New worlds out of old texts: developing techniques for the spatial analysis of ancient narratives* (Oxford forthcoming).

[9] According to Wikipedia (<http://en.wikipedia.org/wiki/Geographic_information_system>), Geographic information system (GIS) is 'a system designed to capture, store, manipulate, analyze, manage and present all types of geographical data [last accessed on 15/11/2013]. In short, GIS allows the researcher to represent spatial data visually – to think with maps, as it were.

[10] 'Timemap.js is a Javascript library to help use online maps, including Google, OpenLayers, and Bing, with a SIMILE timeline': https://code.google.com/p/timemap/ [last accessed on 15/11/2013].

[11] Developments in technology are fast revolutionizing research into and visualization of geographical concepts, all of which has the potential to transform customary notions of mapping space in humanities more broadly: see, for example, R. J. A. Talbert, 'Greek and Roman mapping: twenty-first century



```
<milestone n="1" unit="chapter"/><milestone n="0" unit="section"/>
<milestone unit="para"/>This is the display of the inquiry of <name
type="pers">Herodotus</name> of <placeName
key="tgn,7016142">Halicarnassus</placeName>, so that things done by man not be
forgotten in time, and that great and marvelous deeds, some displayed by the <name
type="ethnic">Hellenes</name>, some by the barbarians, not lose their glory,
including among others what was the cause of their waging war on each other.
<milestone n="1" unit="section"/>
<milestone unit="para"/>The <name type="ethnic">Persian</name> learned men say that
the <name type="ethnic">Phoenicians</name> were the cause of the dispute. These (they
say) came to our seas from the sea which is called Red,<note anchored="true"
resp="ed">Not the modern <placeName key="tgn,7016791">Red Sea</placeName>, but the
<placeName key="tgn,7016761">Persian Gulf</placeName> and adjacent waters.</note>
and having settled in the country which they still occupy, at once began to make long
voyages. Among other places to which they carried <name type="ethnic">Egyptian</name>
and <name type="ethnic">Assyrian</name> merchandise, they came to <placeName
key="perseus,Argos">Argos</placeName>,
```

Figure 2: TEI P4 XML encoding of Herodotus – see note 12 (source: Perseus)

this in mind, we conclude with a brief discussion of some ways in which this study is being developed, with assistance from Research Grants from Google and JISC, which is taking this research (we hope) into exponentially larger data sets, such as Google Books.

2. Capturing the Data

Given our blended methodology of using digital tools to complement textual analysis, we first needed to acquire or generate a digital version of Herodotus' *Histories*. Fortunately help was at hand in the form of the *Perseus Digital Library*, which not only makes texts freely accessible, but also, by virtue of releasing them under a Creative Commons Attribution Non-Commercial Share-Alike licence, allows them to be used and adapted in whatever way users, such as ourselves, want.[12] We further benefited from the fact that the 'placenames' or *toponyms* – the data which we were interested in extracting from Herodotus – had already been marked up in the English edition of the text according to Text Encoding Initiative (TEI) compliant XML in a semi-automated process (see Figure 2).

While the acquisition of the digital text from *Perseus* gave our project an initial boost, a number of issues resulting from that inheritance were raised, which had to be overcome before the data could be properly stored in a database and used. First, the text had to be updated from the P4 TEI schema used by *Perseus* to the latest version, P5, which we accomplished with minimal information loss using an automated conversion tool

perspectives', in *Cartography in antiquity and the middle ages: fresh perspectives, new methods*, ed. R. J. A. Talbert and R. W. Unger (Leiden 2008) 9-27; and most recently, M. Dear, J. Ketchum, S. Luria and D. Richardson, ed., *GeoHumanities: art, history, text at the edge of place* (New York 2011). *Cf.* J. B. Harley, 'Deconstructing the map', *Cartographica: The International Journal for Geographic Information and Geovisualization* 26 (1989) 1-20.

[12] For more information about Perseus, see http://www.perseus.tufts.edu/hopper, adding /opensource for their licensing policy, /publications for related scholarship, and /dltext?doc=Perseus%3Atext%3A1999.01.0126 to access the text that we used—the 1920 Loeb translation of A. D. Godley.

developed by Sebastian Rahtz.[13] Second, we settled on using the English text of Herodotus' *Histories* for investigating spatial data, primarily on the basis that the alternative would have meant going through the text and tagging the *toponyms* 'by hand', when they were already widely available to us from *Perseus* using the English version, but also with a view to disseminating the results as widely as possible. Bearing in mind the scholarly importance of the Greek text, however, we assigned each section of the text in both documents (English and Greek) a common identifier in order to draw an association between them.[14] These sections, assigned by the original TEI encoders, equate more or less to a sentence of text, thereby giving a finer level of granularity than the chapter citations traditionally used. At the same time, each section was assigned to its canonical citation, in order to make identification at this level equally possible. Third, we had to strip out all footnotes, which came embedded in the *Perseus* Loeb edition, in order to prevent contamination by modern *toponyms* within them. These 'reference free' documents form the basis for the analysis.

Lastly, many of the 'placenames' inherited from *Perseus* came already tagged with certain kinds of data, such as geographical co-ordinates, modern-day location name, and place type, deriving from both the *Perseus Gazetteer* and the *Getty Thesaurus of Geographic Names*. The information was, however, automatically generated, and thus of a variable nature and sometimes wrong – different identifiers were used for the same location while sometimes places were misidentified altogether. In order to create a consistent basis for further work, inconsistencies and inaccuracies needed to be corrected by hand, specifically: removing duplicate entities for places, correcting references to false places (in particular those with homonyms), and updating co-ordinates. Furthermore, in order to filter and query places and references more effectively so that different place types could be handled separately, we also introduced categories to distinguish between toponyms in the database on the basis of their being:

1) identifiable communities, encompassing both Greek *poleis* and other kinds of habitations, both Greek and non-Greek (= 'settlements');

2) larger areas which may contain a variety of communities within an approximate region (= 'territories');

3) natural features of the environment (= 'physical features').

The whole process was labour intensive, but most efficient when the principal academic investigator (Barker) and digital specialist (Isaksen) worked together, allowing for targeted solutions to be developed for clearly defined problems. Additionally, we found the visualization technologies described below, particularly Google Earth, especially useful for editing problem locations, particularly when combined with more traditional reference materials such as the *Barrington Atlas*.[15]

[13] TEI P4 to P5 conversion tool: <http://www.tei-c.org/Guidelines/P5/p4top5.xsl>.

[14] That is to say, although we are using the English text to extract spatial data, at every point in the process the Greek text is available for scrutiny alongside it. We also transformed the Greek text from Betacode to Unicode using the Transcoder tool developed by Hugh Cayless, which is available for download via the Open Source directory SourceForge: <http://sourceforge.net/projects/epidoc/files/transcoder/>.

[15] R. J. A. Talbot, ed., *Barrington atlas of the Greek and Roman world* (Princeton 2000).

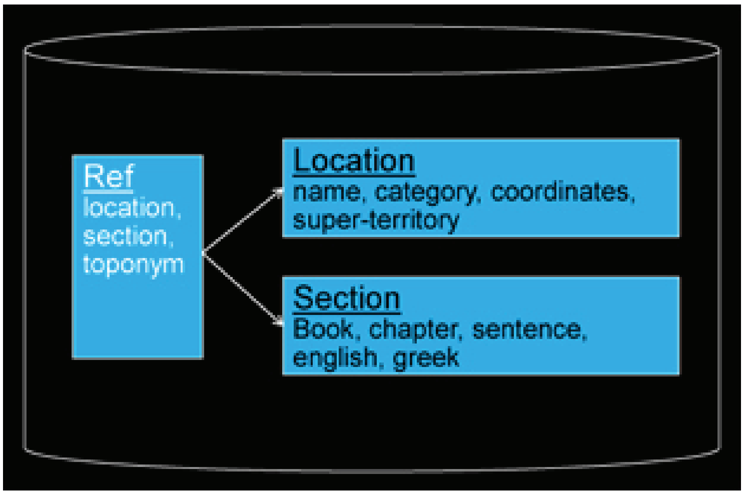


Figure 3: The Hestia database: PostgreSQL

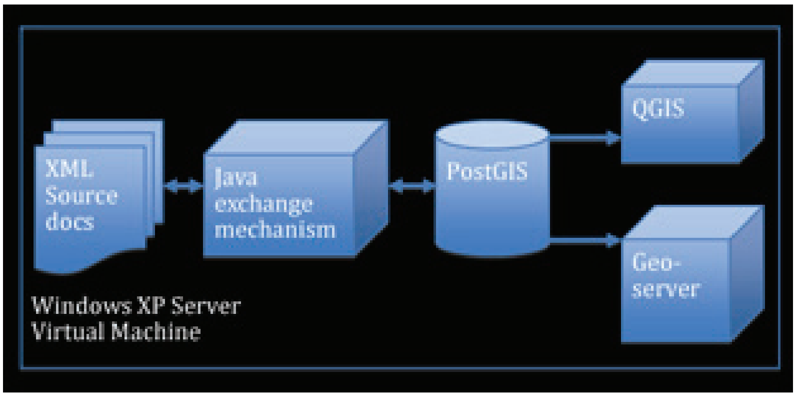


Figure 4: Hestia's architecture

As part of the data cleaning and categorization process, we exported the spatial references to a database, structured simply by (see Figure 3): a *Section* table containing information about the portion of Herodotus' text in which the locations occur, a *Location* table identifying each unique place, and a *Ref* table providing a unique ID for all references to spatial locations within the text.[16] We chose the database type PostgreSQL,[17] an industrial-strength open-source database, on the basis that its PostGIS[18] extension

[16] The advantage of using a structure of this nature is that updates to the Section or Location tables automatically filter through to their References. A virtual table (called a 'view') combining all these together can thereby be created which is accessible to the webmapping server and GIS application.

[17] PostgreSQL is a powerful, open source object-relational database system that has standard compliance: <http://www.postgresql.org/about/>.

[18] According to the website <http://postgis.refractions.net/>, 'PostGIS adds support for geographic objects to the PostgreSQL object-relational database. In effect, PostGIS 'spatially enables' the

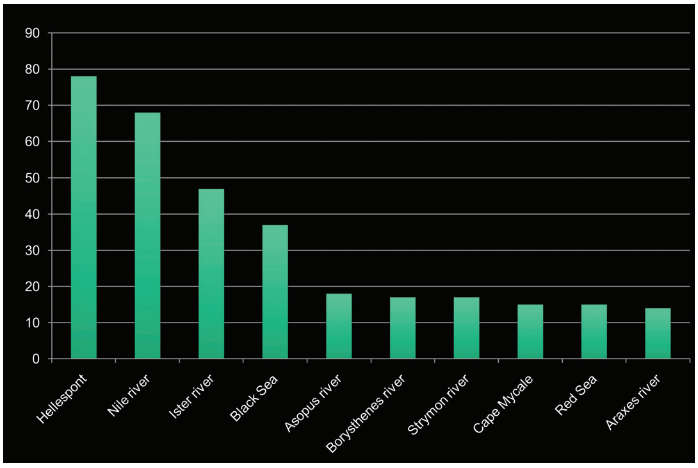


Figure 5: Bar chart showing the ten physical places mentioned most often in the *Histories*

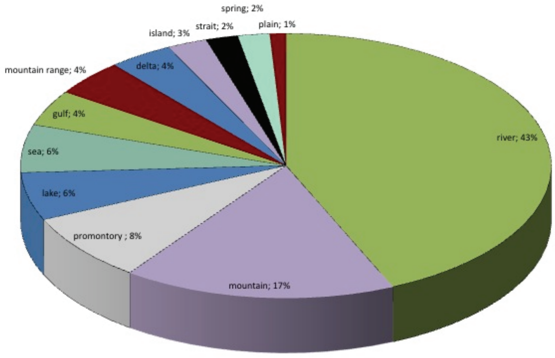


Figure 6: Pie chart showing the proportions according to physical type mentioned in the *Histories*

provides excellent functionality for spatial data and is widely supported by other applications. In particular, we had in mind that, by storing information about references, locations, and the text, it would be possible to link the database to both a Desktop GIS system and Webmapping server simultaneously (see Figure 4).

Just armed with this database, one can now aggregate, query, and analyze Herodotus' spatial data quickly and easily in numerous ways, as demonstrated by Figures 5 and 6. Figure 5 depicts the results of an enquiry into the physical places most frequently mentioned, represented in a tabular form, while Figure 6 shows a pie-chart of the distribution of physical types in the *Histories*. But it is with GIS technology and webmapping applications that we wished to experiment, primarily because of their potential to bring together different kinds of data in a visual representation of narrative. Yet, as we shall see, there is a marked difference in the *usability* of these resources, which, we believe, will prove a significant factor in their

PostgreSQL server, allowing it to be used as a backend spatial database for geographic information systems (GIS)' [last accessed on 15/11/2013].

adoption in the humanities more broadly. As well as outlining the new uses to which Herodotus' spatial data can be put, the rest of this chapter will also highlight limitations in these resources for humanities researchers, not least the considerable investment of time needed for non-IT specialists to become accustomed to the technology. Indeed, it is important to remember that all digital approaches come with their limitations as well as opportunities: we strongly believe that is only through a blended methodology – merging digital techniques with more traditional scholarship – that interesting new insights may emerge.

3. Thinking about the Histories *with GIS*

With this dataset now pre-processed, we were able to apply a variety of technologies to it in order to examine spatial phenomena in Herodotus' *Histories*, primary among which was using GIS. Already well established in archaeology and the social sciences, Geographic Information System (GIS) technology is increasingly being applied to historical studies,[19] though it is widely acknowledged that 'the humanities raise fundamental epistemological and ontological issues for GIS applications', in particular with regard to the contingent and qualitative nature of much of humanities research.[20] To the best of our knowledge, applying GIS to the study of an ancient historical narrative marks a distinctive new direction in the use of this technology, and suggests fruitful new lines of enquiry, as well as raising a number of additional challenges, which we set out in sections 4-6 below.[21]

As far as *Hestia* was concerned, GIS enabled us to approach Herodotus' narrative in an innovative and thought-provoking way. While the database may be used to filter records and even create new data, its information can only be delivered internally in table form; GIS, on the other hand, allows us to visualize Herodotus' spatial references more memorably, on a map. Two points about these maps are worth emphasizing from the outset. First, all the maps that we have produced on this project can be reproduced by anyone else: in fact, users are free to generate the maps of *their* choice, a possibility that we have sought to facilitate by using open source GIS applications ourselves.[22] Second, all maps generated should be regarded *not* as products demonstrating the world according to

[19] See especially I. N. Gregory and P. S. Ell, *Historical GIS: technologies, methodologies and scholarship* (Cambridge 2007); I. N. Gregory and R. G. Healey, 'Historical GIS: structuring, mapping and analyzing geographies of the past', *Progress in Human Geography* 31 (2007) 638-53; A. K. Knowles, ed., *Placing history: how maps, spatial data, and GIS are changing historical scholarship* (Redlands 2008).

[20] See *e.g.* T. M. Harris, S. Bergeron and L. J. Rouse, 'Humanities GIS: place, spatial storytelling, and immersive visualization in the humanities', in *GeoHumanities*, ed. Dear *et al.* (n. 11 above) 227-40 (quotation on 228).

[21] Combining GIS with narrative analysis is discussed by Mei-Po Kwan and Guoxiang Ding, 'Geo-narrative: extending Geographic Information Systems for narrative analysis in qualitative and mixed-method research', *The Professional Geographer* 60.4 (2008) 443-65.

[22] We use the open source application QGIS (<http://www.qgis.org/>), on the basis that anyone else wishing to query our data can use it. It also provides functionality for importing NASA's free Web Mapping Service (WMS) called the Blue Pearl Mosaic, thereby providing a helpful raster backdrop of the Earth's surface, on which Herodotus' spatial data can be projected.

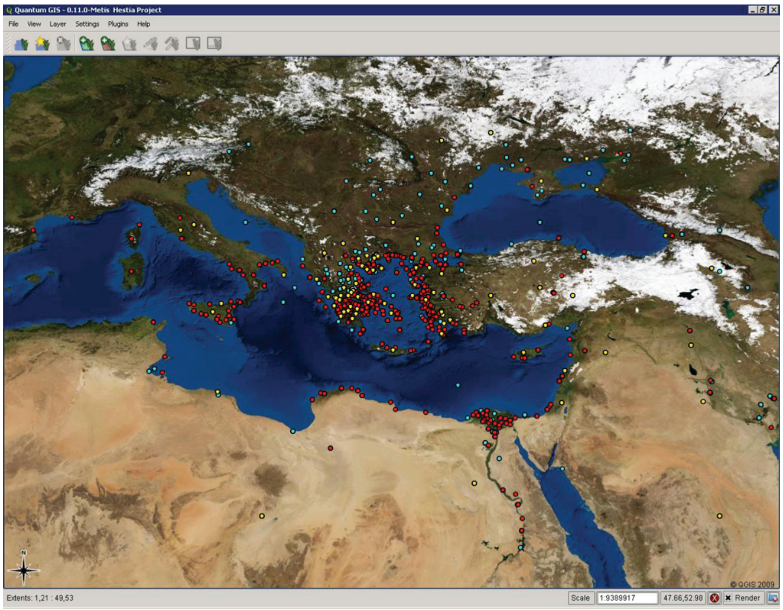


Figure 7: Map (in QGIS) showing all settlements (red), territories (yellow) and physical features (blue) in Herodotus' Histories

Herodotus *but as tools of the enquiry*: in other words, unlike the customary use of maps as ideological representations of various kinds of spatial data, all *Hestia* maps are interrogative, flagging up in visually arresting ways potential areas of interest that demand analysis and further study.[23]

This fundamental point should be clear from the most basic maps that we can generate. Figure 7 simply represents a 'flat' image of the spatial data: it marks all the places that Herodotus mentions over the course of his work with a single dot, thereby providing a snapshot of the huge scope of his enquiry. Yet, such a view gives an idea only of the spatial intensity of the distribution (around the Aegean), not of the frequency of any given place. However, since each place is given a unique identifier in the database that ties it to a particular point in the narrative, it is also possible to generate a count of the references and visualize the results according to a graded symbology. For example, Figure 8 depicts those places categorized as 'settlements', scaled according to the number of references each receives over the course of the *Histories*: the largest circles show those settlements mentioned most often. Of these, Athens and Sparta hardly come as a surprise, being the two most important Greek city-states. Sardis is the place from which Herodotus launches his narrative, and represents an important signifier of growing Persian influence. Arguably Delphi is equally important as a cultural and political centre in the Hellenic world;

[23] As Herodotus himself was acutely aware, the medium in which space is represented can fundamentally alter the way that it is conceived and understood, which is no less true of our own latest satellite imaging technologies in spite (or because) of their apparent claim to accuracy – hence the importance of conceiving these maps as part of the investigative process, not as its end.

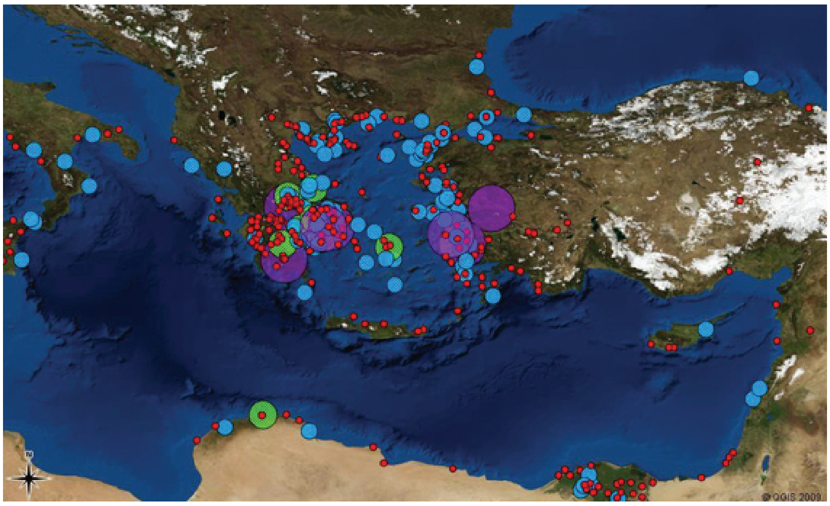


Figure 8: Map (in QGIS) showing settlements according to their number of citations in the *Histories*. The places most frequently mentioned (represented by the large purple circles) are: Sparta, Delphi, Athens, Salamis, Samos, Miletus and Sardis

anyway, there is little doubt of its importance in Herodotus' narrative.[24] The same is obviously true of the island of Salamis, the place of the decisive defeat of the Persian fleet by the Greek coalition. The value of this map arguably, then, lies in the attention that it draws to Miletus and Samos: Miletus, as the centre of the Ionian revolt, is, in Herodotus' words, 'the beginning of all evils for Greeks and barbarians alike' (ἀρχὴ κακῶν Ἕλλησί τε καὶ βαρβάροισι: 5.97.3);[25] Samos, on the other hand, appears to be important because of the text's post-history and the growth of the Athenian empire, which had resulted in – known to both Herodotus and his audience – the brutal subjugation of Samos.[26]

[24] See: H. W. Parke and D. E. W. Wormell, *The Delphic oracle* (Oxford 1956); R. Crahay, *La literature oraculaire chez Hérodote* (Paris 1956); H. Flower, 'Herodotus and Delphic traditions about Croesus', in *Georgica: Greek studies in honour of George Cawkwell*, ed. M. A. Flower and M. Toher, *BICS* Supp. 58 (London 1991) 57-77; L. Maurizio, 'Delphic oracles as oral performance: authenticity and historical evidence', *ClassAnt* 16 (1997) 308-34; E. T. E. Barker, 'Paging the oracle: interpretation, identity and performance in Herodotus' *History*', *Greece & Rome* 53 (2006) 1-28; J. Kindt, 'Delphic oracle stories and the beginning of historiography: Herodotus' *Croesus logos*', *Classical Philology* 101 (2006) 34-51. See also Purves, *Space and time* (n. 4 above), 150-58.

[25] Herodotus is referring to the ships that the Athenians send to Miletus in support of the revolt. On the significance of this phrase, which comes from Homer (*Il*. 5.62-3), see the contributions of R. V. Munson, 'The trouble with the Ionians: Herodotus and the beginning of the Ionian revolt (5.28-38.1)', 146-67 (esp. 149-59); C. B. R. Pelling, 'Aristagoras (5.49-55, 97)', 179-201 (esp. 182, 186); and J. Henderson, '"The fourth Dorian invasion" and "the Ionian revolt"', 289-310 (esp. 305), all of which are in: *Reading Herodotus. A study of the* logoi *in Book 5 of Herodotus'* Histories, ed. E. Irwin and E. Greenwood (Cambridge 2007).

[26] See E. Irwin, 'Herodotus and Samos: personal or political?', *Classical World* 102 (2009) 395-416; *cf*. P. Stadter, 'Herodotus and the Athenian *Arche*', *Annali della Scuola Normale Superiore di Pisa* 22 (1992) 781-809; J. Moles, 'Herodotus warns the Athenians', *Papers of the Leeds International*

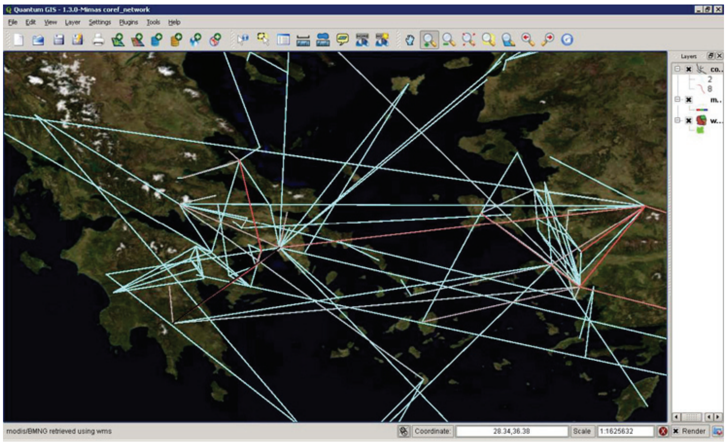


Figure 9: Close-up of map (in QGIS) showing the 'settlement network' across the *Histories*: lines indicate places mentioned together in the text on more than one occasion; those in red indicate the most intense relationships based on count alone

4. Networking places

Both of these kinds of map depict places in isolation from each other: as we outlined above, however, we have been especially keen to approach the question of Herodotus' narrativization of space through the lens of '*topological*' connections between places – those links that the narrator draws between different spatial concepts as he tells his story. With this in mind, we undertook a close textual and qualitative-based analysis of spatial concepts across one stretch of narrative (Book 5).[27] Since, however, this in-depth study was extremely time consuming and labour intensive, we also trialled the extent to which it was possible to use the *Hestia* database and GIS technology as a short-cut to highlighting potential patterns of interest.[28]

These chiefly included the connections between places that Herodotus draws in his narrative, for which we used a SQL query of the database based on simple co-presence of terms within the same section of text, and visualized the results in QGIS. Since these links were fully automated and we had no idea of telling what the connections were, beyond the fact that two or more places were being mentioned in 'the same breath', we tried to minimise arbitrariness by counting only those links that occurred more than once. Still, the resulting picture can be rather chaotic, as seen in Figure 9, which depicts the rough

Latin Seminar 9 (1996) 259-84. On the importance of Samos more generally, including the use of data from *Hestia*, see C. B. R. Pelling's 2010 Barron Memorial Lecture, 'Herodotus and Samos', *BICS* 54 (2011) 1-19.

[27] For a discussion of the methodology and results, see: E. T. E. Barker and S. Bouzarovski, 'Developing a qualitative network analysis: tools, methodological issues and preliminary results', in *New worlds*, ed. Barker *et al*. (n.8 above).

[28] For an example of using GIS in network analysis, see: M. L. Berman, 'Boundaries or networks in historical GIS: concepts of measuring space and administrative geography in Chinese history', *Historical Geography* 33 (2005) 118-33.

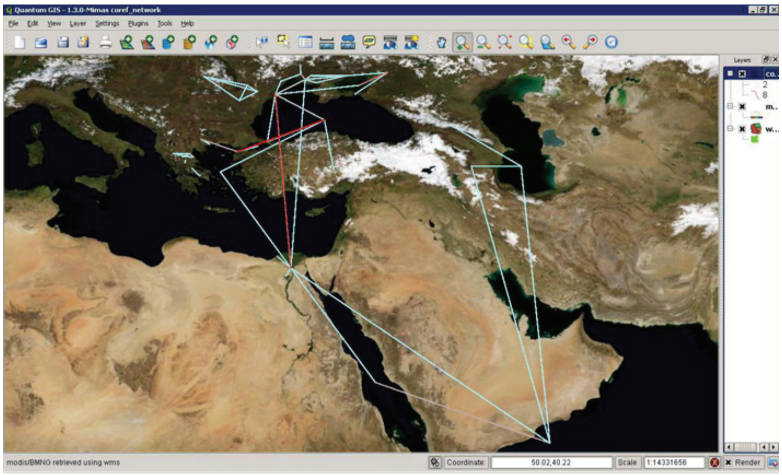

Quantum GIS - 1.3.0-Mimas coref_network
File Edit View Layer Settings Plugins Tools Help
Layers
co..
2
8
m..
w...
modis/BMNG retrieved using wms
Coordinate: 50.02,40.22
Scale 1:14331656
Render

Figure 10: Map (in QGIS) showing the ‘physical features’ network across the *Histories* – those features mentioned together in the text on more than one occasion. A red cross identifies the two strongest links, one between the Ister and Nile, the other between the Hellespont and the Pontus.

‘network’ picture of settlements across the entire *Histories* (focused on the Aegean Sea). It should be remembered, however, that in the GIS application the network can be tested: the user can click on any one these links and bring up the relevant passages in which the relationship is mentioned. Such a map is best thought of as a stimulus for the analysis and closer reading of the text, rather than a representation of (or replacement for reading) it.[29]

To give a brief example, consider figure 10, which depicts the physical features mentioned together in the text on more than one occasion. Two relationships stand out in red. The first is the connection between the river Ister (the modern-day Danube) in Scythia and the Nile. The strength of this relationship relates, not to a ‘real-life’ network of trade exchange, but rather to a conceptual network utilized by Herodotus to organize the world he narrates. Scythia and Egypt, on the margins of the known world, function as places of comparison through which Herodotus brings to light important aspects of cultural and political identity, in particular in relation to the Greeks and Persians.[30] The other strong connection marked in this map is between the Pontus and the Hellespont. This axis is important because it relates to the growing stretch of the Persian Empire, as Darius casts his eyes westwards from the Pontus towards the Hellespont, the bridging point between east and west, in a move that anticipates his son’s literal bridging of the Hellespont in his invasion of Greece.[31] The map also demonstrates the extent to which

[29] SQL or ‘Structured Query Language’ is ‘a special-purpose programming language designed for managing data held in a relational database’: http://en.wikipedia.org/wiki/SQL [last accessed on 15/11/2013].

[30] F. Hartog, *The mirror of Herodotus* (Berkeley 1988 [1980]). At 2.33-34 Herodotus explicitly draws the comparison between the Ister and Nile.

[31] On the importance of the Black Sea in Herodotus, using *Hestia* data, see: E. T. E. Barker, S. Bouzarovski, C. B. R. Pelling, and L. Isaksen, ‘Extracting, investigating and representing

physical features envelop anthropogenic constructs like territories, which act as anchors for the geography of linkages between social formations constructed by Herodotus.

A whole range of other basic network maps can be quickly and easily generated that vary in nature according to the place type one chooses to focus on (settlement, territory, or natural feature), the general scene in a particular book, the connections enjoyed by any one specific place, those places that enjoy the most number of connections (the central 'hubs' in a network), or those networks that have the strongest connections (the thickest 'edges'). Thus the database can be used not only to gain a general sense of a broad network culture represented in Herodotus, but also to generate a series of different types of network representations depending on the queries that one asks of the data. It is worth reiterating, however, that none of these maps are sufficient in and of themselves: rather, they are there to prompt new questions and provoke further investigation. In this way, all these GIS-generated maps, both the simple 'places in isolation' and the more complex networks, should be regarded as *complementing and preceding* rather than replacing close textual analysis.

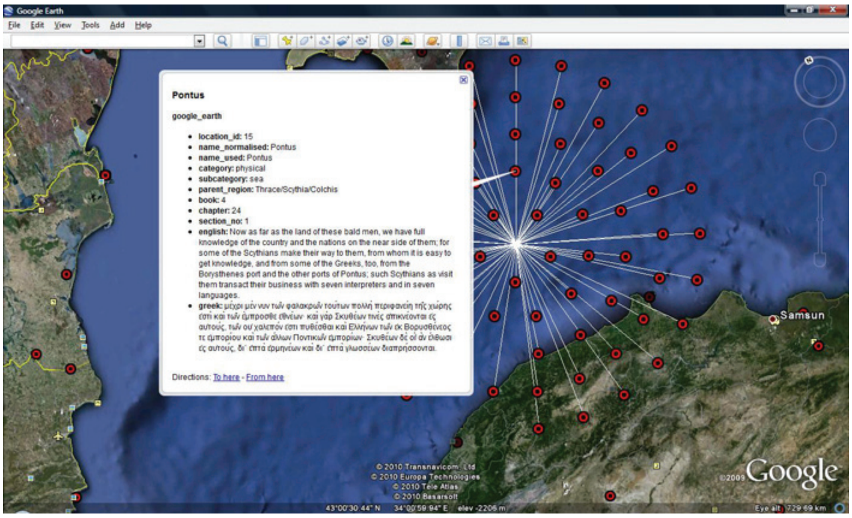
5. Bringing Herodotus' world into everybody's home

The foregoing sections have discussed the possibilities for GIS spatial and topological analysis of places within Herodotus' *Histories*. At the same time, our experiences have found that using GIS has demanded a level of expertise higher than that enjoyed by most of the team – a point not often acknowledged in the literature on humanities GIS. Limitations in its usability, however, were for us an important issue, since we have been determined from the outset to make our data, methodologies, technologies, and outcomes open for general use and reuse.[32] GIS, in our minds, did not have the widespread applicability that we were looking for.

In order, then, to start experimenting with public dissemination, we decided to expose the PostGIS data as Keyhole Markup Language (KML), a markup format that can be read by a variety of webmapping applications including the hugely popular Google Earth. This was achieved by installing the Open Source GeoServer, which serves spatial data in a variety of web-friendly formats simultaneously, including KML, SVG, WMS, WFS, and PDF. Significantly, it is automatically readable over the web on any machine that has Google Earth installed, and, since the link is a network link rather than a static KML file, any changes that we made to the *Hestia* database would result in automatic (and more-or-less instantaneous) updating of the data in the viewing application without any need on the user's part to do anything.

geographical concepts in Herodotus: the case of the Black Sea', in *The Bosporus: Gateway between the Ancient West and East*, eds. G. R. Tsetskhladze *et al.* (Oxford 2013) 7-17.

[32] In part this was in response to the concern expressed by our original sponsor, the AHRC, that their funded projects be disseminated as widely as possible. But we also hold that practising openness has proven overwhelmingly beneficial for our research by encouraging potential collaborators to get in touch. We note too that our own open approach has been largely possible because of *Perseus*' relatively liberal Creative Commons licence, which directly encourages further use and wider dissemination. Within the digital world, overly restrictive licencing (such as No-Derivatives clauses) can be almost as pernicious to scholarship as the failure to make the data available in the first place.



Pontus

google_earth

- **location_id:** 15
- **name_normalised:** Pontus
- **name_used:** Pontus
- **category:** physical
- **subcategory:** sea
- **parent_region:** Thrace/Scythia/Colchis
- **book:** 4
- **chapter:** 24
- **section_no:** 1
- **english:** Now as far as the land of these bald men, we have full knowledge of the country and the nations on the near side of them; for some of the Scythians make their way to them, from whom it is easy to get knowledge, and from some of the Greeks, too, from the Borysthenes port and the other ports of Pontus; such Scythians as visit them transact their business with seven interpreters and in seven languages.
- **greek:** μέχρι μέν νυν τών φαλακρών τούτων πολλή περιφανείη τής χώρης έστι καί τών έμπροσθε έθνέων· καί γάρ Σκυθέων τινές άπικνέονται ές αυτούς, τών ού χαλεπόν έστι πυθέσθαι καί Ἑλλήνων τών έκ Βορυσθένεός τε έμπορίου καί τών άλλων Ποντικών έμπορίων· Σκυθέων δέ οί άν έλθωσι ές αυτούς, δι᾽ έπτά έρμηνέων καί δι᾽ έπτά γλωσσέων διαπρήσσονται.

Directions: To here - From here

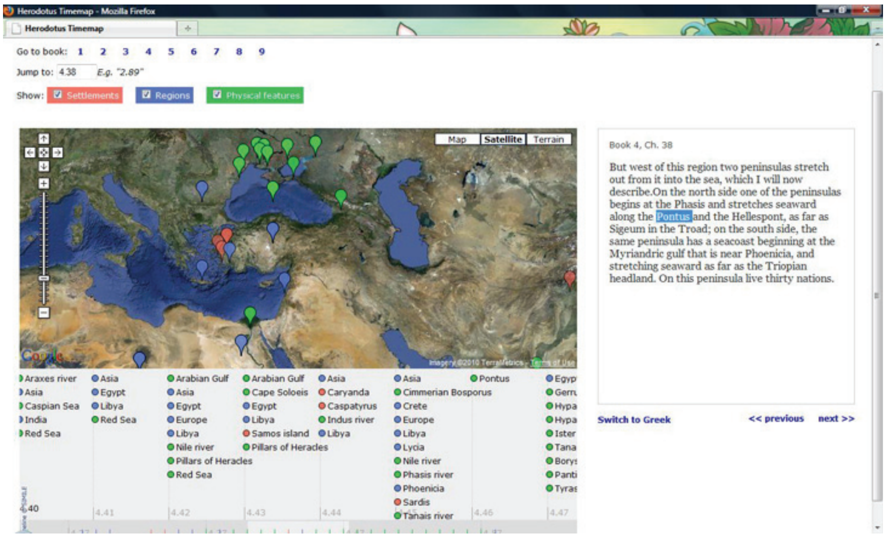
Figure 11: The *Histories* in Google Earth, highlighting one of the passages (each indicated by a red circle) in which Herodotus talks about the Pontus
http://hestia-geo.open.ac.uk:8080/geoserver/wms/kml?layers=hestia:google_earth

With this 'Herodotus geodata', users are able to construct 'mashups' of visual and textual data.[33] For example, since all places are linked to entries in the database, by simply clicking on a particular location in Google Earth, users are able to bring up a balloon box containing Herodotus' text (in both English and Greek) for that particular location for every occasion when it is mentioned in the narrative (Figure 11). It is this part of the project that potentially has the greatest impact, with its capacity to reach beyond an academic environment and bring the world of an ancient historian into peoples' homes. In the case of Herodotus' *Histories* this might involve linking different places that are mentioned with each other in the same stretch of narrative in chronological order, so that it would be possible to follow the 'journey', say, of historical agents within the text. One could, for example, imagine following Xerxes' passage into Greece from the 'down-on-the-earth' perspective of one of his myriad of troops.

While Google Earth is an excellent vehicle for enabling users of all kinds to explore the geography of Herodotus' *Histories*, like GIS it is less effective at displaying the changing points of geographical reference over the course of the narrative. The inherent danger in this kind of technology is that one forgets that there is a text to be read or, in Herodotus' day, to be heard – and reading or listening is an experience that happens in time. Therefore in collaboration with Nick Rabinowitz, developer of Timemap.js,[34] we

[33] As the account in Wikipedia puts it, 'a mashup in web development is a web page, or web application, that uses content from more than one source to create a single new service displayed in a single graphical interface'. For more information, see: http://en.wikipedia.org/wiki/Mashup_(web_application_hybrid) [last accessed on 15/11/2013].

[34] For Timemap.js, see n.10 above. The Herodotus TimeMap has since been superseded by GapVis. See section 6 below.



Herodotus Timemap - Mozilla Firefox

Herodotus Timemap

Go to book: 1 2 3 4 5 6 7 8 9

Jump to: 4.38 *E.g. "2.89"*

Show: ☑ Settlements ☑ Regions ☑ Physical features

Book 4, Ch. 38

But west of this region two peninsulas stretch out from it into the sea, which I will now describe.On the north side one of the peninsulas begins at the Phasis and stretches seaward along the Pontus and the Hellespont, as far as Sigeum in the Troad; on the south side, the same peninsula has a seacoast beginning at the Myriandric gulf that is near Phoenicia, and stretching seaward as far as the Triopian headland. On this peninsula live thirty nations.

Switch to Greek << previous next >>

Figure 12: Herodotus TimeMap. Reading text, map, and (narrative) time alongside each other. http://www.nickrabinowitz.com/projects/timemap/herodotus/basic.html

have created a web-based interface combining a map, a 'narrative timeline', and a reading pane, allowing the reader to view locations as they are mentioned in the text. For example, Figure 12 shows all the places mentioned by Herodotus in or near 4.38, divided according to settlement, territory or physical feature. The central Google Map displays geographical concepts roughly collocated in the narrative; the timeline below depicts 'narrative time' represented as the sequence of book sections, each listing geographical references cited within them; to the right, a reading window presents the text of the current chapter, which can be toggled between English and Greek. Navigational elements allow the user to page through the text or jump to a specific section, while clicking on a location in the map or timeline will move the text to the relevant chapter, highlighting the selected *toponym*. Markers on the map fade to transparency as the narrative progresses beyond their mention, mimicking their gradual disappearance from the scope of the reader's attention.

This 'timemap' of Herodotus' world represents a novel use of the Timemap.js library, which was designed to combine maps with historical, rather than narrative, timelines. The JavaScript library, which integrates the Google Maps API with the SIMILE Timeline API,[35] facilitates the loading of data simultaneously onto both a map and timeline, with a separate data structure to manage cross-references and interactions between the two visual representations of a given record. The result is an online interface in which user interaction with either the map or the timeline can be reflected in both elements; for example, clicking an item on the timeline may bring up an informational window on the map, and markers on the map may be hidden when their corresponding timeline item is out of view. To customize Timemap.js to the Herodotus Timemap application, a custom set-up script was required to convert the date-based chronological timeline into one displaying the narrative sequence of book chapters and subsections. To avoid the need for

[35] SIMILE project Timeline web widget: <http://www.simile-widgets.org/timeline>.

revisions to the core architecture of the SIMILE Timeline, this was accomplished through the relatively simple expedient of mapping chapters to an arbitrary sequence of dates, then using that mapping to determine the timeline labels and item positions.[36] The other technical challenge was to find a method for loading a large number of data items without overloading the browser-based interface. Under normal circumstances, a Google Map cannot display more than a few hundred items without significant performance issues – and our dataset included over 4,000 place references. Our solution, which addressed both performance and load-time concerns, was to segment both chapter text and location data into static JSON[37] files of manageable size, which could then be loaded 'on demand' as the user progressed through the narrative.

The Timemap is less an analytical tool than an enhanced interface for engaging with Herodotus' text as a reader. Unlike several of the data visualization approaches discussed above, it displays comparatively little information at any given time, focusing attention instead on the reading experience of the *Histories* and the relationship of geographical concepts to their place in the narrative. Our hope is that the rich visual detail that the Timemap adds to the text can offer value at the level of both the novice reader, who may benefit from a clear depiction of the locations under discussion, and the researcher, for whom the interface may help to identify quickly places co-occurring in the narrative and highlight chapters where Herodotus' use of geographical reference is particularly focused or wide-ranging. In fact, by trying to map as closely as possible the reading experience, we hope that the Herodotus Timemap may have the additional utility of facilitating research into the ways in which geographical concepts in the *Histories* undergo change over time.

All-in-all, these different technologies have allowed us to develop unprecedented means of disseminating the world view of an ancient historian as well as providing a glimpse of the potential for digital resources to help expose the spatial data for examination and ask new questions of it.

6. Towards a digital analysis of ancient places?

Applying new technologies to a dataset such as that of an ancient Greek historian inevitably raises (at least) as many questions as it resolves. Issues that have arisen during this project relate to data collection, management and aggregation, accuracy and precision, expertise required to use the tools, above all, the 'so what?' question. Indeed, our experience of using digital resources and developing an automated process for the querying and visualization of spatial data in Herodotus has demonstrated beyond doubt the need to adopt plural approaches that make use of, but are not dependent on, the technology. In short, we conclude that the employment of the digital medium can greatly enhance our understanding of spatial concepts in Herodotus *provided that* it is accompanied by a close engagement with the text at all stages of the analysis. Ultimately, it is this need to marry 'quantitative' and 'qualitative' approaches that we regard as one of the most interesting implications of the initial steps that we have taken.

[36] Besides, many of the dates associated with events or characters within the *Histories* are known only within very rough limits, particularly those that occur before the fifth century BC.

[37] JSON (JavaScript Object Notation) is a 'lightweight data-exchange format':
<http://www.json.org/>.

Of course *Hestia* is not alone in pioneering approaches to adopting digital technologies in classical studies. A similar project in the field of archaeology is *Open Context*, an Open Repository of archaeological excavation data developed and maintained by the Alexandria Archive and UC Berkeley.[38] In 2010, a funding call from Google, under its Digital Humanities Award program, sought to enable researchers to 'apply quantitative research techniques for answering questions that require examining thousands or millions of books'. This presented an opportunity to researchers on both *Open Context* and *Hestia* to pursue the common goal of tying classical resources together through referencing places.[39] The central principle behind the GAP (Google Ancient Places) project, identified while developing the *Hestia* Timemap, is that places referenced in narrative texts generally cluster together to maintain narrative coherency. In other words, given a set of *toponyms* with multiple possible identifications, the set of identifications with the shortest overall path between them is likely to be correct. We can further weight the influence of each *toponym* on our decision by the number of possible locations to which it could refer. Somewhat counter-intuitively, this means that small, obscure places with unusual names are much better guides to location than well-known places with many namesakes.[40]

Since 2011, the GAP team have been working in conjunction with the *Pleiades* project[41] to associate the local identifiers used by the *Hestia* project with *Pleiades* Uniform Resource Identifiers (URIs) for each place to which multiple names (*toponyms*), locations (such as spatial co-ordinates), and categories (like 'settlement') can be assigned. Not only does this make it much easier to handle the problem of synonymy, but it also means that, once an identification is made, it can be permanently fixed with a non-ambiguous identifier.[42] At a stroke this process renders digital data much more valuable, for now various kinds of enquiry can be asked of the data and various kinds of information can be brought together: it will be possible to find out about ancient places not only in texts but in other resources, including tables, databases, maps and images.[43] In addition, GAP is experimenting with ways

[38] For more on *Open Context* see: <http://opencontext.org/>.

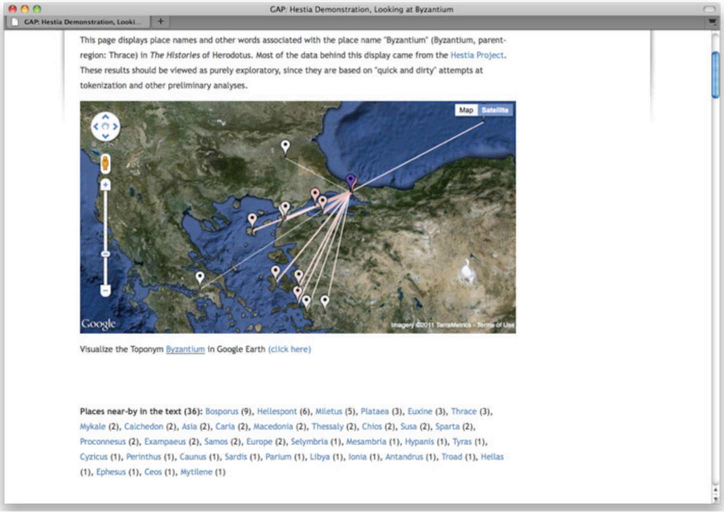
[39] Google Research Blog: 'Our commitment to the digital humanities': http://googleresearch.blogspot.co.uk/2010/07/our-commitment-to-digital-humanities.html [last accessed on 15/11/2013].

[40] For more information about the GAP project, including regular posts describing the work that has been carried out, see: <http://googleancientplaces.wordpress.com/>. The GAP team have also produced a new reading interface called GapVis (http://gap.alexandriaarchive.org/gapvis/index.html). This platform provides an initial landing page showing all the places in a text visualised on a map and as a histogram; a 'reading page' for viewing the places alongside the narrative; and a 'place page' for focusing on information about a particular place, such as those places mentioned most frequently with it (see figure 13).

[41] The *Pleiades* project (<http://pleiades.stoa.org/>) is in the process of digitizing the *Barrington atlas of the classical world*. The fact that Pleiades' coverage for Herodotus' place names had not been available at the time of the initial Hestia project demonstrates the rapidly changing digital environment.

[42] Or URIs (Uniform Resource Identifiers): <http://en.wikipedia.org/wiki/Uniform_Resource_Identifier>.

[43] Led by Elton Barker and Leif Isaksen of *Hestia*, Pelagios aims to make sense of the sea of data related to ancient world places. Funded by JISC in two consecutive programmes (Geospatial


GAP: Hestia Demonstration, Looking at Byzantium
GAP: Hestia Demonstration, Looki...
Map
Satellite
Google
Imagery ©2011 TerraMetrics - Terms of Use

This page displays place names and other words associated with the place name "Byzantium" (Byzantium, parent-region: Thrace) in *The Histories* of Herodotus. Most of the data behind this display came from the Hestia Project. These results should be viewed as purely exploratory, since they are based on "quick and dirty" attempts at tokenization and other preliminary analyses.

Visualize the Toponym Byzantium in Google Earth (click here)

Places near-by in the text (36): Bosporus (9), Hellespont (6), Miletus (5), Plataea (3), Euxine (3), Thrace (3), Mykale (2), Calchedon (2), Asia (2), Caria (2), Macedonia (2), Thessaly (2), Chios (2), Susa (2), Sparta (2), Proconnesus (2), Exampaeus (2), Samos (2), Europe (2), Selymbria (1), Mesambria (1), Hypanis (1), Tyras (1), Cyzicus (1), Perinthus (1), Caunus (1), Sardis (1), Parium (1), Libya (1), Ionia (1), Antandrus (1), Troad (1), Hellas (1), Ephesus (1), Ceos (1), Mytilene (1)

Figure 13: 'Place' page view of the GapVis platform, showing the places most frequently mentioned with 'Byzantium' in the *Histories*

of visualizing and analyzing spatial data that does not demand specific expertise in using specialist applications, such as GIS. Figure 13, for example, depicts a network analysis map of Byzantium's network in Herodotus (based on the co-reference of places in the text) in KML, which is the kind of map that any users will be able to generate and explore for themselves.

In conclusion, *Hestia* is not an end product, but just the beginnings of what promises to be a radical reconfiguration of the research (and teaching) process. We believe that the technologies described here, when used together and in conjunction with traditional scholarship, will help not only facilitate innovative lines of enquiry and lead to new insights, but also bring Classics to a whole a new audience. Digital Classics is already starting to move beyond the era of stand-alone projects, and the tools and data described here are already openly available. It is now time for the community to build upon this material, opening it up both to classicists and to the world at large.

Elton Barker (*Open University*) Elton.Barker@open.ac.uk
Leif Isaksen (*University of Southampton*) L.Isaksen@soton.ac.uk
Nick Rabinowitz nick.rabinowitz@gmail.com
Stefan Bouzarovski (*University of Manchester & University of Gdansk*)
stefan.bouzarovski@manchester.ac.uk
Chris Pelling (*Christ Church, Oxford*) chris.pelling@chch.ox.ac.uk

Engagement and Community Outreach programme; Resource and Discovery programme) Pelagios (meaning 'of the sea') is establishing generic processes and practices for referencing ancient places in online documents. For more information, please see the project blog: <http://pelagios-project.blogspot.com/>. As this publication goes to press, Pelagios is entering its third phase of development. Now with the support of the Andrew W. Mellon Foundation, and led by Isaksen, Pelagios will annotate, link, and index place references in digitized Early Geospatial Documents that use written or visual representation to describe geographic space prior to 1492.

MEASURING THE INFLUENCE OF A WORK BY TEXT RE-USE

MARCO BÜCHLER, ANNETTE GEẞNER,
MONICA BERTI, THOMAS ECKART

1. Introduction

Before the invention of modern book printing it was a very difficult, onerous, and expensive task to copy a text: the text itself had to be obtained, writing material was expensive, and even good scribes took a long time to make good copies by hand. It seems evident that everybody would think quite carefully about which books to copy and which not, and we thus have to assume that only certain kinds of books have been preserved until today. So what could have been the criteria behind these decisions?[1]

There are different reasons why a work was copied often enough to survive until today. One of them was good fortune: for example, it was conserved on a papyrus in the sands of Egypt or was 'accidentally' passed on as a palimpsest and was not destroyed in a fire, as in the Library of Alexandria. Another reason could be that this work had not been forbidden for ideological reasons (for instance in times of iconoclasm) and many other reasons could be added here as well.[2]

But we also have to consider the possibility that the significant influence of a work was one of the main reasons for its uninterrupted tradition. A work must have held the genuine interest of enough people who considered it to be worth copying and could also afford to do so. While it cannot always be determined why one work has been transmitted and another one has not, we have to assume that a work passed down until today had a certain impact that made people copy it again and again throughout the centuries. Those works must have been important, valuable, and/or useful to those who decided to copy them, or requested a copy. So, if the tradition of a work is long enough that it still exists in some form today, this is in itself quite significant and presents a question worth researching.

Classicists have long tried to determine the influence of a work by measuring its reception by citation indices throughout the centuries. The most common method used is to examine the texts of authors whose works have survived to look for traces of quotations and text re-use in their works. The major problem with this approach is that the manual collection of every passage that bears evidence of text re-use (especially of lost works) is a very demanding and time-consuming task. In order to obtain faster and hopefully more

[1] For an interesting discussion of this question see: H. A. Cayless, *Ktêma es aei: digital permanence from ancient perspective*, in *Digital research in the study of classical antiquity*, ed. G. Bodard and S. Mahony (London 2010) 139-50.

[2] See: H. Hunger, *Handschriftliche Überlieferung in Mittelalter und früher Neuzeit, Paläographie*, in *Einleitung in die griechische Philologie,* ed. H. G. Nesselrath (Stuttgart and Leipzig 1997).

complete results, classicists can now use a variety of tools at least partially to automate this kind of research. The approach described in this chapter is to provide an automatic search for textual re-use[3] and then to visualize the results, taking different aspects of re-use into consideration in order to make them easier to interpret.

Any discussion of textual re-use must also address the larger question of so-called 'fragments' of lost authors and works. The term 'fragment' is applicable to a wide range of ancient evidence, which includes archaeological ruins, epigraphical and papyrological documents, and many other pieces of the material record. By 'fragments', however, we mean not only the material remains of ancient writings, but also quotations of lost texts preserved through other texts. A huge number of quotations of lost texts have been gathered together in print collections, enabling scholars to reconstruct lost works and depict the personalities of 'fragmentary authors'.[4]

The importance of gathering quotations (fragments) of lost works is due to the fact that a significant majority of ancient texts have been lost. Nonetheless we can reconstruct this inestimable cultural patrimony thanks to traces of text re-use preserved in later works. At the same time, collecting fragments of lost authors also permits us to provide a useful measure of the shifting boundaries of canon formation over time.[5] Moreover, working with quotations of lost works serves as an extraordinary methodological exercise in attempting to discover patterns that could be useful within the fields of allusion discovery, plagiarism detection, and text re-use. Finally, gathering fragments of ancient works and representing them digitally is a fundamental exercise: model is built that can also be very useful for tracking modern quotations and, in particular, for use in multi-million book libraries such as Google Books or the Internet Archive.[6]

2. Related work and state of the art

In the field of text re-use, much research has already been conducted and, while it is impossible to address all relevant work here, some important aspects are summarized in this section. Scientifically, the linking of two text passages is formalized as a graph $G=(V,E)$ consisting of a set V of vertices and a set $E= VxV$ of edges between elements of V. The set V represents a non-overlapping corpus that is segmented into large linguistic units such as a sentence or a paragraph. This task can typically be done with a linear cost of $O(n)$. The set of

[3] By 'text re-use' we mean a textual congruence, which can be a good indicator for finding quotations. Nevertheless until proven it remains uncertain if this re-use is indeed a (direct or even indirect) quotation or a text passage, which has been quoted by later authors.

[4] For more on this see: M. Berti, M. Romanello, A. Babeu and G. Crane, 'Collecting fragmentary authors in a digital library (Greek fragmentary historians)', in *Proceedings of the 9th ACM/IEEE-CS joint conference on digital libraries* (Austin, TX 2009) 259-62, and M. Berti, 'Fragmentary texts and digital libraries', in *Philology in the age of Corpus and Computational Linguistics,* ed. G. Crane, A. Lüdeling, and M. Berti, CHS Publication (forthcoming).

[5] See: G. Most, ed., *Collecting fragments - Fragmente sammeln,* (Göttingen 1997).

[6] O. Kolak and B. N. Schilit, 'Generating links by mining quotations', in *Proceedings of the nineteenth ACM conference on hypertext and hypermedia (HT 2008). Pittsburgh, Pennsylvania*, (New York, NY 2008) 117-26.

edges *E* between two elements $v_i, v_j \in V$ represents *pairwise* links between two text passages. Computing those links is much more complex than defining the set *V*.

Trying to compute textual re-use by pairwise comparison is quite time-expensive due to the squared complexity of $O(n^2)$. This approach is useful for comparing smaller corpora such as the Dead Sea Scrolls with the Hebrew Bible.[7] But using it with an ancient Greek corpus like the CD-ROM Version E of the *Thesaurus Linguae Graecae* called *TLG*-E,[8] which has about 5.5 million sentences, would require *3.025e13* comparisons. Assuming that about 1000 comparisons can be done in a second, this process would approximately require a run time of almost 1000 years. Even if only all the sentences of a single author, such as Plato, were compared with a corpus like the *TLG*, the processing time would still require about one year.

Reviewing more complex algorithms, most of them can be summarized as a four-step process:

- *Fingerprinting*: Every re-use unit such as a paragraph or a sentence first needs to be fingerprinted. In detail, this means that the re-use unit can be quantified by a set of syntactical features such as any *n-gram approach* and semantic methods like *semantic clustering*. Within this chapter we decided for reasons of reliability to utilize a syntactical approach. Following this, two questions must be answered: first, does an overlapping or a non-overlapping approach make the most sense for this kind of problem, and, second, does a static or a non-static n-gram size fit best for the investigated question? To explain in greater detail, overlapping features means that a sentence or a paragraph is quantified by pairwise overlapping n-grams (shingling). Furthermore, this implies that every word is part of at least two n-gram features. In contrast to this, non-overlapping fingerprinting means that every word is exclusively part of one feature. For this research, we decided to choose an overlapping and non-static fingerprinting approach that is named *Longest Common Consecutive Words* (*LCCW*). Depending on both the research question and the problem, syntactical features are sometimes preferred, while in other cases fingerprints that are more semantic are required (see section *Measuring influence by hypertextuality*).
- *Selection*: Quoting a text passage always implies purposeful re-use and this means that both the original text and the subsequent quotation have similar patterns as far as the fingerprints are concerned. For this reason, not all possible fingerprint values are necessary. Imagine that two identical re-use units of twenty words exist that are then compared by a bi-gram fingerprinting. Without any selection all of the nineteen possible bi-gram features would link these two sentences with each other. It would be necessary, however, to have just one link between these sentences. In order to avoid any missing links by too strong a

[7] R. Hose, *CS490 Final report: investigation of sentence level text reuse algorithms,* Boom 2004 Bits On Our Minds: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.90.9835> [accessed on 2nd July 2011].

[8] *TLG* Consortium, *Thesaurus linguae Graecae,* CD-ROM Disk E, University of California,Irvine, released in February 2000.

selection process, four or five significant features would represent the re-use units perfectly.

- *Linking*: This step links two passages as either directed or undirected. In a historical context it is often useful to highlight who has used texts from whom, but without metadata it is quite difficult to make a directed link. Typical approaches reduce the complexity in comparison to the above-mentioned naive method of $O(n^2)$ to $O(n*log(n))$, which decreases the computation time dramatically.
- *Scoring*: After two text passages are linked, the next step is to score the similarity of the two linked passages.

In both of these last two steps, links of some passages are rejected. Depending on the text and the degree of textual re-use, there is often a strong selection in the linking step. Several experiments on different corpora and languages have shown in the past that only one in 100 million possible linking candidates is considered as an actual case of textual re-use.[9] The scoring itself can be seen more as a fine-tuning that removes less similar sentences.

The *fingerprinting* step is divided into two strongly correlated sub-tasks, first a window size and then an algorithm need to be selected. While it depends on both the selected corpus and the research question, typically used observation windows include *sentences*,[10] *paragraphs*,[11] and a *fixed word number window.*[12] For applications in the humanities, however, the choice of the window size will strongly depend on the following question: ‘How was an author quoted?’ If there is a strong literal re-use, then approaches using sentence segmentation or a fixed window are good choices. However, if a given piece of textual content is paraphrased or strongly mixed in with the referring author’s own words, then a larger context like a paragraph is necessary, otherwise the probability of a match decreases.

In the second step of the fingerprinting process, the link features are defined. Generally, there are three different clusters of approaches:

- *Words as features*: after all the function words, such as articles and conjunctions, are removed, passages that have the same words are linked. The general idea of these approaches is to identify those passages of a text that have a significant common semantic density.[13]

[9] See: M. Büchler, *Informationstechnische Aspekte des historischen Wissenstransfers*. (Engl. *Computational aspects of historical knowledge transfer*). (PhD thesis submitted Leipzig University 2013).

[10] Hose, *Final report* (n. 7 above)

[11] John Lee, ‘A computational model of text reuse in ancient literary texts’, in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, (Prague 2007) 472-79.

[12] B. Mittler, J. May, P. Gietz and A. Frank, *QuotationFinder - Cluster Asia and Europe - Uni Heidelberg* (2009): <http://www.asia-europe.uni-heidelberg.de/de/forschung/heidelberg-research-architecture/hra-projects/quotationfinder> [accessed on 11th January 2010].

[13] Mittler *et al*., *QuotationFinder* (n. 12 above), Lee, *Computational model* (n. 11 above).

- *N-grams as features*: to extract textual re-use syntactically, several n-gram approaches for bi-grams and tri-grams exist. The key idea is to link units having a significant large overlap of n-grams.[14]
- *Sub-graphs as features*: graph-based approaches, as shown in this chapter, deal with semantic relations between words. In the *Lexical Chaining* approach[15] that is often used for text summarization,[16] a *semantic construct* or a *semantic representation* of linguistic units is generated. When applying these approaches to a huge amount of text, an implicit feature-expansion of paradigmatic word relations in terms of language evolution or different dialects is often observed. This is caused by the fact that these words are connected with the other words of a unit as well.

While the cluster of n-gram approaches is strongly focused on syntactical features, the approaches of both other clusters can also deal with textual re-use in a free word order.

To score a found link, a measure is used to compute the similarity of both linked units. Therefore the features themselves or the words of both units are taken to compute any kind of similarity. Measures like the *Dice coefficient* compute the *similarity* of two pairwise linked passages by commonly used (and overlapping) words, while other measures like the *city block metric*, *Euclidean distance,* or the *Jenson-Shannon divergence* calculate the *semantic* distance between two units.[17] The main difference between these two types of measures is that a *similarity* measure scores relevant links with a high score, whereas *distance* measures score a relevant link of two units as close as possible to zero.

Given a corpus *C,* a re-use graph $G=(V,E)$ can be described by the following generalized algorithm:

```
1. V = segment_corpus(C) with $v_1, v_2, ..., v_n \in V$, $\cup v_i = C$ and $v_i \neq v_j$
2. for each $v_i \in V$
3. $F_i$=train_features($v_i$);
4. for each $v_i \in V$
5. for each $f_k \in F_i$
```

[14] Hose, *Final report* (n. 7 above); M. Büchler, G. Heyer, and S. Gründer, *Bringing modern text mining approaches to two thousand year old ancient texts, e-Humanities – an emerging discipline*. Workshop in the 4th IEEE International Conference on e-Science (2008); M. Büchler, *Medusa release homepage – a statistical engine for natural language processing matters*. <http://mbuechler.e-humanities.net/medusa/, 2005-2011> (2011).

[15] U. Waltinger, A. Mehler, G. Heyer, 'Towards automatic content tagging: enhanced web services in digital libraries using lexical chaining', in *4th Int. Conf. on Web Information Systems and Technologies (WEBIST '08), 4-7 May, Funchal, Portugal*, ed. J. Cordeiro, J. Filipe and S. Hammoudi (Barcelona 2008) 231-36.

[16] L. Yu, J. Ma, F. Ren, S. Kuroiwa, 'Automatic text summarization based on lexical chains and structural features', in *snpd, vol. 2, Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, (Qingdao 2007) 574-78.

[17] For all these see: S. Bordag, *Elements of knowledge-free and unsupervised lexical acquisition,* (Unpubl. PhD thesis, Leipzig University 2007).

```
6. $e_i=(v_i,v_j)\in E=$select all $v_j$ containing feature $f_k$
7. for each $e_i\in E$
8. $s_i=$scoring$(e_i=(v_i,v_j)\in E; F_i; F_j)$;
9. if($s_i<threshold$)$\{E=E\backslash\{e_i\}\}$
```

Listing 1: Generalization of a textual re-use algorithm consisting of 4 steps: Line 1: Segmentation of a corpus to linguistic units v_i (builds set V of a graph =(V,E)), lines 2-3: Training of features set F_i for every unit v_i, lines 4-6: Linking process of units (builds initial set E of a graph =(V,E)) and lines 7-9: Scoring and removing of less significant edges (cleans set E of a graph =(V,E))

Within the humanities, text re-use has always been an important field of research, especially in classical philology, and there has often been a strong focus on fragmentary works. The importance of providing technical tools for conducting research in this area has been a subject of increasing importance recently as millions of texts have been digitized. Typically, classicists had to conduct manual searches of textual concordances in databases such as the *TLG*, and the main drawback is that this work takes a lot of time and can be incomplete.

Regarding the fragments of lost works, we presently have many printed collections of Classical fragmentary authors. In particular, we have many collections of the fragments of Greek historians.[18] These collections have allowed scholars to reconstruct the characteristics and personalities of otherwise-lost authors. One of the major limits of printed collections of fragmentary authors is that one can only see the quotation *extracted* from the context in which it has been preserved. New digital models for collecting and representing fragments permit us to see the text of the fragments inside their *contexts* of transmission and according to different editions. In this multilevel structure, we can reconstruct the whole tradition of a text from its 'original' form to its reception and transmission across the centuries.[19]

Digital libraries and hypertextual models allow us to rethink the fundamental question of the relation between the fragment and its context of transmission, including representing and expressing every element of print conventions in a more dynamic and interconnected way. A fragment is in itself a perfect model of hypertext and in a digital library the fragment can be linked to the whole text of the source in which it is preserved. In this way it is possible to see the excerpt directly inside its context of transmission, avoiding the misleading idea of an independent material existence of fragmentary texts.[20]

[18] See: Berti *et al*., 'Collecting fragmentary authors' (n. 4 above) and M. Berti, 'Fragmentary texts' (n. 4 above).

[19] See: Berti *et al*., 'Collecting fragmentary authors' (n .4 above) and M. Berti, 'Fragmentary texts' (n. 4 above).

[20] See: Berti *et al*., 'Collecting fragmentary authors' (n. 4 above) and M. Berti, 'Fragmentary texts' (n. 4 above).

Author	Work	Century	Genre
HELLANICUS	Fragmenta	5 BC	Hist.
XENOPHON	Memorabilia	5 BC	Hist.
ZENO	Testimonia et fragmenta	4 BC	Hist.
PLATO	Timaeus	5 BC	Phil.
EPHORUS	Fragmenta	4 BC	Phil.
ARISTOTELES et CORPUS ARISTOTELICUM	De anima	4 BC	Phil.
Flavius JOSEPHUS	Contra Apionem (= De Judaeorum vetustate)	1 AD	Hist.
PLUTARCHUS	Pompeius	1 AD	Hist.
APPIANUS	Mithridatica	1 AD	Hist.
GALENUS	Ad Glauconem de medendi methodo libri ii	2 AD	Phil.
CELSUS	Ἀληθὴς λόγος	2 AD	Phil.
ALEXANDER	De anima	2 AD	Phil.

Table 1: An overview of the selected authors and works in this paper (names as in the *TLG-E*).

3. Investigated works

In order to test our approach, we have chosen twelve authors from two different time periods and two different genres (see Table 1). The most important criterion for choosing these works was comparability. One of the requirements was that the works had to have a similar length (*i.e.* number of tokens). Then they needed to belong to similar genres and time-spans and also had to include fragmentary authors. We have chosen philosophy and historiography as genres not only because of our genuine interest in these two literary fields, but also because we had observed some differences between the quotation of philosophical and historical texts.[21] For time-spans we have chosen the fifth and fourth centuries BC and the first and second centuries AD since these centuries seemed to contain enough interesting authors with works in the two selected genres as well as with almost the same text-length (*ca*. 15,000 to 30,000 tokens). The works chosen are not supposed to be considered the most important or influential ones of their time and genre, but to offer some variety in order to compare the results.

4. Data used and pre-processing

4.1 Data used

All illustrated methods and results are based on the *Thesaurus linguae Graecae* Version E,[22] a comprehensive collection of Greek writers, including many well-known authors like Plato and Sophocles and coverage from Homer's time to the fall of Constantinople in the

[21] See Büchler, *Informationstechnische Aspekte* (n. 9 above).

[22] *TLG* Consortium, *Thesaurus Linguae Graecae,* CD-ROM Disk E, University of California,Irvine, released in February 2000.

fifteenth century. This corpus has been created and provided by the *TLG* research centre at the University of California, Irvine, and today it is one of the most important digital resources when dealing with ancient Greek texts. The version used for this study contains around 7200 works written by more than 1800 different authors over a time period of more than 1800 years. Since the origin of this digital corpus goes back to the 1970s, all textual data and metadata are encoded in a binary format (*TLG*-E) that is not a suitable basis for efficient Text Mining applications. A rather comprehensive tool chain of pre-processing steps therefore had to be built.

4.2 Pre-processing

Several specific tools were either developed or adapted, including an extractor for the binary input data, a Beta Code to Unicode converter as well as different tools for dealing with problems concerning a strongly inflected language such as ancient Greek and its various changes over a long period of time.

> Step 1: Sentence segmentation
> As a first pre-processing step, a newly developed, rule-based sentence boundary detection algorithm splits the text. To deal with various extraneous information that is unimportant to the detection of textual re-use (such as the marking of speaker roles), different lists of boundary marks are used in combination with abbreviation lists to enhance the sentence boundary detection rate.
>
> Step 2: Tokenization
> As the next step, all tokens were extracted from the segmented sentences. In comparison with modern languages as modern English or German, a more active *tokenization* process was needed for dealing with ancient Greek. Specifically, this means that more irrelevant parts of the input material had to be removed to gain usable text. In addition to punctuation marks, all brackets of the Leiden Conventions were also removed.

As a result of these first two steps, all *TLG* works are segmented into about 4.96 million sentences with an average length of 13.51 words. Table 2 shows the resulting cumulative sentence length distribution.

Sentence length	<=5	<=10	<=15	<=20	<=25	<=30	<=35	<=40	>40
Cumulative distribution in %	29.63	51.82	68.40	79.39	86.48	90.99	93.92	95.64	100

Table 2: Cumulative distribution of sentence length in words

> Step 3: Normalization
> Since the ancient Greek language contains a large number of diacritical marks and many words also exist in a variety of upper/lower-case letter combinations, many different forms of the same word type can be found in the corpus. As an example, the

conjunction καί exists in the *TLG*-E in more than fifteen different versions.[23] Since many of these variants exist due to changes in writing or modern modifications of the original text (such as the usage of lower case letters), a re-use detection process based on these variants might ignore a huge portion of relevant text passages. Therefore, a normalization process is executed that reduces all words internally to a lower-case representation and removes any diacritics. Table 3 shows the number of different spelling variants for some high frequency words of the *TLG*.

Word	*τοῦ*	*πρός*	*τοῖς*	*κατὰ*	*τοῦτο*	*εἶναι*	*βασιλεία*
Number of variants	15	8	8	21	10	15	14

Table 3: Number of word variants with identical normalized word form

Step 4: Lemmatization
Another class of variations of the same word are due to morphology. Therefore, all words have been analysed and internally reduced to their base form by using the morphological analyser Morpheus, which was developed by the *Perseus Digital Library* (*Perseus*).[24] As Morpheus can also identify dialects, even dialectical variants are reduced to the same base form.

4.3 Pre-processing as an ongoing process

The pre-processing of highly structured text, especially on a corpus that covers a very long period of time, is not a task that can simply be done once and then be considered as complete. Because of various differences between ancient Greek and modern languages, many standard Natural Language Processing (NLP) tools failed when used on ancient Greek text: tokenization and sentence segmentation proved particularly difficult. For this reason, existing tools had to be substituted by specialized replacements, various parameters had to be adjusted, and dedicated tools had to be created to deal with special phenomena. Thus the quality of the whole pre-processing system had to be evaluated regularly by classicists and the task turned out to be an ongoing process.

5. Re-use methodology

Text re-use is represented in a formal re-use graph *G=(V,E)* having *V* as the set of re-use units (such as the sentences of a text corpus) and *E* as the set of pairwise edges between elements of *V*. Within this study, we decided to use the *Longest Common Consecutive Words* fingerprinting (*LCCW*) with a selection of at least 5 words in a row and with feature frequency of at least 2. We have chosen to use the dice coefficient as the similarity measure with a threshold of 0.4 that provides good results. This threshold is low enough not to ignore embedded quotations and is high enough typically to ignore phrases such as ‘in the Name of our Lord Jesus Christ’.

[23] *καὶ, Καὶ, καί, Καί, και, Καῖ, Και, καῖ, καἴ, καἰ, καἱ, Καϊ, Καΐ, Κἆι, Κάι.*

[24] See the Digital Classicist wiki entry for Morpheus: <http://wiki.digitalclassicist.org/Morpheus>.

Starting with an n-gram of size 5, in every iteration all n-grams of length *l* of the previous iteration are taken to compute new, statistically significant n-grams of size *l+1*. 'Statistically significant' means that the new n-gram must have a log-likelihood score not smaller than 6.63 and a minimum n-gram frequency of 2. This step is iterated until no more n-grams can be computed.

Expanding significant n-grams in such a way has one benefit and one consequence. The benefit is that the longest common match between a text re-use and the original text can be found. With this information available, visual access for philologists can be provided quite simply since the boundaries of an n-gram are determined by one of the following three causes:

- the beginning of a sentence,
- the end of a sentence, or,
- any kind of a differing word due to causes such as language evolution, dialect change, an inserted word or the boundaries of an embedded re-use within a larger sentence.

A negative consequence of this approach is that all common prefixes of the longest match that consist of at least five words are produced.[25] Consequently a post-processing step removes these prefixes. In addition, finding the prefix properties of those n-grams requires a frequency heuristic such as:

$$eps = \log_2\left(\frac{Frequency(x_1 x_2 \ldots x_n)}{Frequency(x_1 x_2 \ldots x_n x_{n+1})}\right)$$

Empirically, an epsilon between 0.1 and 0.2 yields the best results and only prefixes with a smaller score than *eps* are removed. A larger score indicates that there is at least one more unit referring to the same original text. However, this text passage may just have a less common longest n-gram match. Given a set of those longest matching n-grams, all sentences containing the same n-gram are pairwise compared for similarity. To compute the similarity of both linked units, the *dice coefficient* is used. Words of both sentences are then compared for a common overlap in relation to the words that could be overlapped.

The reasons for deciding to use *LCCW* as the text re-use fingerprinting method for this study are twofold. First, by way of using this most restrictive fingerprinting, we can assume a higher level of precision. Since it is impossible to validate all detected hypertextual edges in *E* manually, it made sense to us to choose this very restrictive technique. By way of the similarity threshold of 0.4, however, the re-use detection remains aware of embedded quotations. Secondly, using *LCCW* keeps the amount of data at an acceptable level.

6. Measuring influence by hypertextuality

Within this study we introduce two parameters of measuring hypertextuality: *re-use coverage* C_{LCCW} and *re-use temperature* T_{LCCW}, whereas *LCCW* indicates the fingerprinting methodology.

[25] The minimum threshold of five is chosen by statistical properties of n-grams. With respect to the audience, however, we skipped a detailed description.

Given a set V of re-use units such as a sentence, *re-use coverage* C_{LCCW} measures the ratio between quoted re-use units of a work in relation to the total amount of re-use units of this work. It is, however, not of interest if a re-use unit is quoted more than once or not.

The *re-use temperature* T_{LCCW}, however, measures the frequency f for a dedicated re-use unit and scales it as described by the following formula:

$$\mathcal{T}_{LCCW} = log_{10}(f) + 1$$

Table 5 illustrates the behaviour of this formula in a bit more detail. The logarithmic scaling of f down-scores peaks significantly. If T_{LCCW} increases by 1, ten times more quotations can be observed (see Table 5).

Re-use temperature $\mathcal{T}_{\mathrm{LCCW}}$	0,00	1,00	1,50	2,00	2,50	3,00	3,50
frequency of re-use f	0,00	1,00	3,16	10,00	31,62	100,00	316,23
color	black	blue	violett	red	dark orange	orange	yellow

Table 5: T_{LCCW} vs. f. Translation table of T_{LCCW}(f) for some significant re-use temperatures. Furthermore, T_{LCCW}(f) can be mapped to colours that are used in Figure 1.

Whereas C_{LCCW} measures the 'breadth' of re-use, *i.e.* how many parts of the work have been re-used at least once, T_{LCCW} indicates the 'depth' of re-use, called 'temperature' by way of measuring the *re-use frequency* of a dedicated re-use unit. By way of ordering all re-use units as they occur in the text, a temperature map as shown in figure 1 can be generated. At the x-axis these ordered re-use units are shown. They are normalized to relative positions. A text position of 0.4 represents a text re-use unit at 40% of this work. The y-axis is solely to add one dimension, so that the plots in figure 1 are a rectangle instead of a single line. The colours represent the *re-use temperature* T_{LCCW} as indicated in Table 5.

Before interpreting these results, it has to be taken into consideration that the chosen text re-use discovery algorithms find passages that use *exactly* the same words as the chosen text, *i.e.* they are very literal textual concordances. The majority of located passages would likely be considered to be intentional references. Due to no limitations of the time span in which the re-use is found, this method not only shows by whom the chosen authors (see table 1) have been re-used, but also what older text-passages these authors re-used themselves. Nevertheless, by demonstrating intertextuality, the results are of great help for classicists trying to answer our research question.

The results shown in figure 1 have to be interpreted by a classicist. Works like the *De Anima* of Aristotle and Celsus' *Ἀληθὴς λόγος* seem to have been referenced very broadly, due to the fact that at least one re-use has been found for almost every part of these works. But this high amount of referencing could be for a variety of reasons: for instance, it could

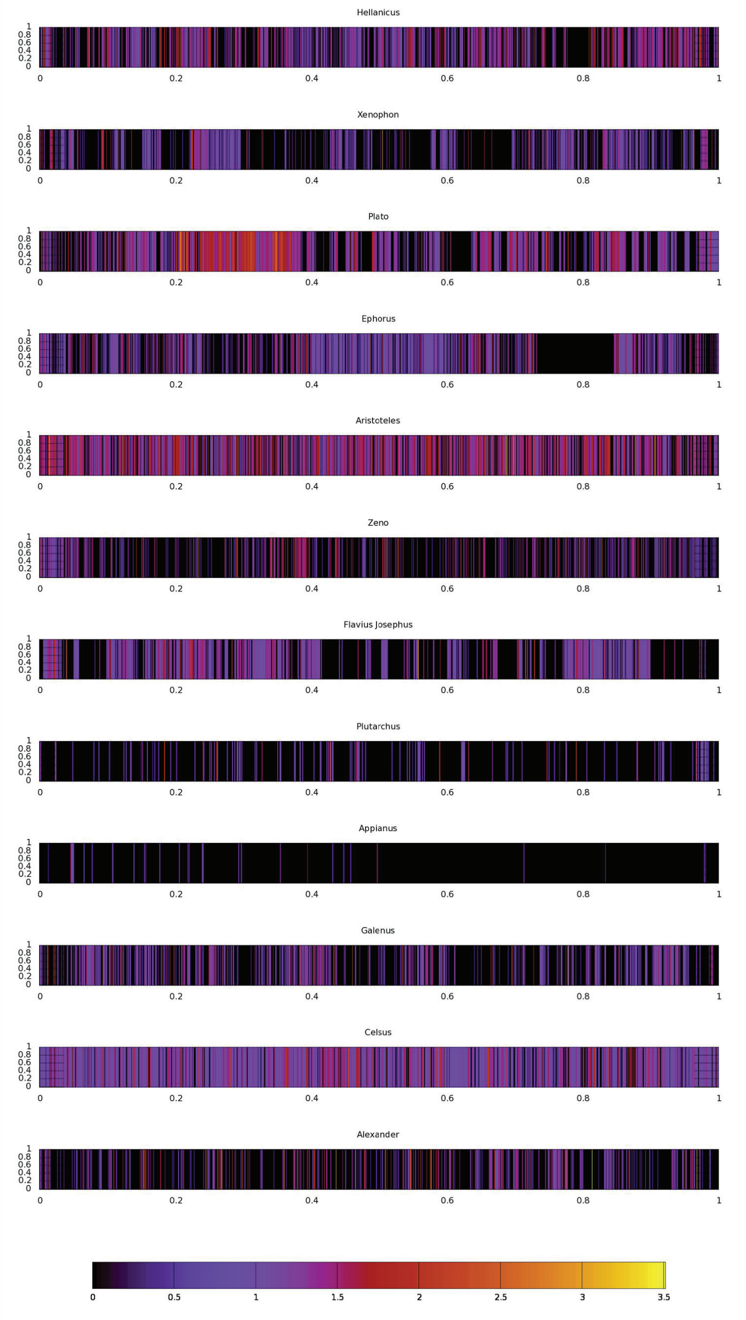


Figure 1: Re-use temperature T_{LCCW} for all works from table 1. The x-axis represents the re-use units (sentences). A score of 0.2 represents a text position of 20% within this work. The colours represent the re-use temperature T_{LCCW}.

be the result of anthologists, who decided to add some parts of these texts into their collection. While the fact that a work is chosen to be part of an anthology itself seems to indicate the relative importance of a work, is it really enough to measure its larger influence? We think that, in order to measure the impact of a work, we should consider factors beyond how *many* parts of this text have been re-used and should also explore if there are some passages of this text that became so 'popular' that they were used much more *often* than others (*i.e*., they had a (very) high temperature). As we know from famous sayings such as 'I know that I know nothing' and 'To be or not to be', the popularity of an author or work can frequently be demonstrated by only a few words that are cited very often. Table 6 demonstrates this concept, showing the top three works with the highest numbers of passages found for every temperature.

As this table demonstrates, it is quite possible to state that Aristotle's work can indeed be considered very influential, because it has not only been re-used *broadly* but also because it ranks in the top three in every category of temperature. Celsus has some high temperatures too, but passages of his work are not as frequently quoted as Aristotle. It is comparable with the *Timaeus* of Plato, which is not so much *broadly* referenced, but has lots of high temperatures for distinctive parts of the text.

By comparing Plato with Hellanicus in figure 1, the difference between a work of Plato that still exists and a fragmentary work can be shown. Whereas for Hellanicus only smaller clusters of re-use can be observed (which depend significantly on the order of the fragments), one strongly quoted cluster exists in Plato's *Timaeus* between the text positions 0.2 and 0.4.

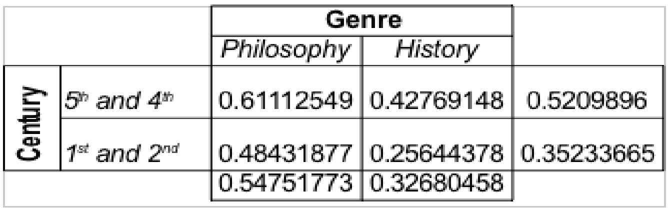
Perhaps the most interesting result observed is that the work of Alexander of Aphrodisias, which is not broadly cited at all, has one of the highest temperatures of all chosen works. This could mean that single-text passages of this work have had a significant influence on other authors.

Furthermore, philosophical texts tend to be quoted more broadly, whereas historical texts are quoted in a more focused and frequent fashion. Table 6 supports this point. While a high text re-use temperature of 2.5 finds historians as part of the top three, they disappear significantly from the top three as the text re-use temperature is gradually decreased. On the other hand, a philosopher such as Alexander of Aphrodisias is, significantly, within the top three. When decreasing the text re-use temperature, however, his work decreases in terms of its score. In detail, this means that Alexander seems to have been quoted in some text passages very frequently wherever his work has an atypical quotation usage for less frequent re-used text passages. A possible reason for these few highly quoted passages could be due to the fact that this work of Alexander of Aphrodisias is a commentary on the work *De Anima* written by the famous philosopher Aristotle, which includes quoting passages from this work, which can be very often re-used phrases themselves.

One significant result of the temperature maps of Figure 1 and the data from table 6 is that philosophical texts tend to be re-used more literally than historical texts. This is likely due to the fact that in philosophical texts certain ideas and thoughts of famous philosophers are discussed and thereby repeated much more often and more accurately than in historiography. In table 7 this fact is highlighted in detail. Table 7 aggregates all the authors from table 1 within the genres of philosophy or history on the one hand. On the other hand, all authors are grouped into two clusters based on their living time. Comparing both columns representing the genre, a significant difference of 0.22 (0.5475-0.3268) between philosophy and history is observed. This indicates a more literal quotation style of philosophical rather than historical texts. On the other hand, by comparing both rows of dating clusters, a similar quotation style is observable. Whereas, in the fifth and fourth centuries BC coverage of 0.52 is possible it is only 0.35 for the first and second centuries AD.

		Re-use temperature					
		3,5	*3*	*2,5*	*2*	*1,5*	*1*
Authors	*HELLANICUS*	0,00000	0,00000	0,00000	0,01004	0,08233	0,53614
	XENOPHON	0,00000	0,00045	0,00223	0,00848	0,03527	0,35536
	PLATO	0,00000	0,00000	0,00426	0,04691	0,16311	0,61301
	EPHORUS	0,00000	0,00000	0,00000	0,00179	0,03461	0,50776
	ARISTOTELES et CORPUS ARISTOTELICUM	0,00160	0,00480	0,01279	0,03437	0,24860	0,74820
	ZENO	0,00000	0,00000	0,00561	0,01402	0,05329	0,30208
	Flavius JOSEPHUS	0,00000	0,00000	0,00102	0,00917	0,04383	0,47299
	PLUTARCHUS	0,00000	0,00000	0,00000	0,00444	0,01110	0,13762
	APPIANUS	0,00000	0,00000	0,00000	0,00000	0,00102	0,03176
	GALENUS	0,00000	0,00075	0,00373	0,01343	0,03731	0,37537
	CELSUS	0,00000	0,00000	0,00166	0,02244	0,08728	0,86367
	ALEXANDER	0,00074	0,00297	0,00742	0,01782	0,05197	0,25390

Table 6: Selected re-use temperature T_{LCCW} for all works of table 1. Those cells that are marked with grey background colour are part of the top three authors at this temperature.



		Genre		
		Philosophy	*History*	
Century	*5th and 4th*	0.61112549	0.42769148	0.5209896
	1st and 2nd	0.48431877	0.25644378	0.35233665
		0.54751773	0.32680458	

Table 7: A contingency table of re-use coverage of dating by century and genre.

7. Further work

This chapter has examined two text re-use properties out of the much larger set that could be investigated. Both *re-use coverage* and *re-use temperature* are properties that can be extracted easily. An ever more detailed consideration of this topic, however, would necessitate including additional data, such as the dating of authors and texts. In this chapter we did not make a distinction if one of our selected authors quoted another one or if the author himself was quoted. The key issue is that the necessary date information is of too poor a quality and this makes this type of work almost impossible. By improving this information, however, we will separate both *re-use coverage* and *re-use temperature* by the additional dimension of the degree of an own contribution or a quotation.

Furthermore, we could show that the same algorithm – such as the *Longest Common Consecutive Words* – works differently on two genres of the same language. On one hand, we can apply different algorithms. On the other hand, it seems to be obvious that in different genres the re-use style tends to be significantly different. For this reason one further dimension such as the degree of closeness to the original makes sense. More generally, this dimension corresponds to the *Kolmogorov complexity* – that is an algorithmic distance between an input and an output. In the context of reception importance, the degree of change is of interest, since, if an author is quoted more literally, this quotation can be weighted higher than less literal ones, since the degree of re-using on purpose is much more significant.

Since classical scholars especially require more than just the *LCCW* algorithm, it is part of our current research plan to include further algorithms. In detail, there are currently active developments on the *TRACER* Java library that provide much more functionality.[26] The software library is not designed as a monolithic algorithm, but as a *6 level architecture* where, for each level, at least one implementation has to be selected to build the complete algorithm. Currently, there already exist about 120,000 possible combinations that have recently been evaluated in detail. This includes not only text re-use algorithms but also the extraction of *canonical references*[27] in order to apply the metrics *re-use temperature* and *re-use coverage* to those kinds of quotation traces as well.

8. Conclusion

Classicists have always considered trying to determine the influence of an ancient work by measuring its hypertextuality as an effective approach. Undertaking this work by means of text-mining methods thus presents the next logical step to take. And indeed it has led to very interesting results and the visualization proved to be very useful.

Measuring influence has made it clear that we have to consider different aspects of a text's impact on other authors. It seems best to divide the found results into different categories:

- Works referenced broadly with high temperatures (Aristotle, Plato – both philosophers);
- Works referenced broadly, but not with high temperatures (Celsus);
- Works with single passages of high temperatures, but not referenced broadly (Alexander, Xenophon, Hellanicus, Zeno);
- Works which meet none of these criteria (Appianus, Plutarch – both writing about history);
- Works somewhere in between.

This research question, however, cannot be answered simply by looking at the figures and tables. Research and interpretation by scholars will always play a large role in deciding

[26] The *TRACER* library (http://etraces.e-humanities.net/TRACER) is a text re-use engine that will be available by a Creative Commons licence in summer 2013. Due to the complexity of TRACER, it is planned to initiate in summer 2013 a series of teaching courses that will be announced on the aforementioned link.

[27] For more on this see: M. Romanello, M. Berti, A. Babeu, G. Crane, 'When printed hypertexts go digital: information extraction from the parsing of indices', in *HT 09. Proceedings of the 20th ACM Conference on Hypertext and Hypermedia,* (Turin and New York 2009) 357-58.

how influential an author or work could have been, as will the data selected and used for analysis. For example, we cannot say that Plutarch was an author without any influence on other authors, but we can affirm that he was not referenced quite literally and/or those referencing works got lost. Therefore, it is necessary to search to see if he had an impact on other authors in different ways. One of the reasons for the few re-uses of Plutarch could be the fact that historiographical works do not seem to be quoted as close to the original texts as, for instance, philosophical texts, which is an interesting fact in itself.

Natural Language Processing Group, Institute of Mathematics and Computer Science, University of Leipzig, Germany [mbuechler|teckart]@e-humanities.net

Ancient Greek Philology Group, Institute of Classical Philology and Comparative Studies, University of Leipzig, Germany agessner@e-humanities.net
Dipartimento di Studi Umanistici, Università di Roma Tor Vergata Department of Classics, Tufts University monica.berti@uniroma2.it monica.berti@tufts.edu

References

Berti, M., M. Romanello, A. Babeu, and G. Crane, 'Collecting fragmentary authors in a digital library (Greek fragmentary historians)', in *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries (JCDL 2009)* (Austin, TX 2009) 259-62.

Berti, M., 'Fragmentary texts and digital libraries', in *Philology in the age of Corpus and computational linguistics*, ed. G. Crane, A. Lüdeling, and M. Berti, CHS Publication (forthcoming).

Beta Code *Thesaurus Linguae Graecae – the beta code manual*, online publication [accessed: 25th October 2010]: <http://www.tlg.uci.edu/encoding>.

Bordag, S., *Elements of knowledge-free and unsupervised lexical acquisition,* (Unpubl. PhD thesis, Leipzig University 2007).

Büchler, M., *Performante Textstatistiken auf großen Textmengen: Kookkurrenzanalyse in Theorie und Anwendung* (Saarbrücken 2008).

Büchler, M., G. Heyer, and S. Gründer, *Bringing modern text mining approaches to two thousand year old ancient texts, e-Humanities – an emerging discipline*. Workshop in the 4th IEEE International Conference on e-Science (2008).

Büchler, M., *Medusa release homepage – a statistical engine for natural language processing matters*: http://mbuechler.e-humanities.net/medusa/, 2005-11 (2011).

Büchler, M., *Informationstechnische Aspekte des historischen Wissenstransfers*. (Engl. *Computational aspects of historical knowledge transfer*). (Unpubl. PhD thesis, to be submitted at Leipzig University 2013).

Cayless, H. A., *Ktêma es aei: digital permanence from ancient perspective*, in *Digital research in the study of classical antiquity*, ed. G. Bodard and S. Mahony (London 2010) 139-50.

Hose, R., *CS490 Final Report: investigation of sentence level text reuse algorithms,* Boom 2004 Bits On Our Minds [accessed: 2nd July 2011]: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.90.9835>.

Hunger, H., *Handschriftliche Überlieferung in Mittelalter und früher Neuzeit, Paläographie*, in *Einleitung in die griechische Philologie,* ed. H.G. Nesselrath (Stuttgart and Leipzig 1997).

Kolak, O., and B. N. Schilit, 'Generating links by mining quotations', in *Proceedings of the nineteenth ACM conference on hypertext and hypermedia (HT 2008). Pittsburgh, Pennsylvania*, (New York, NY 2008) 117-26.

Lee, J., *A computational model of text reuse in ancient literary texts*, in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, June 2007*, Association for Computational Linguistics 2007) 472–79.

Mittler, B., J. May, P. Gietz, and A. Frank, *QuotationFinder - Cluster Asia and Europe - Uni Heidelberg* [accessed: 11th January 2010]: <http://www.asia-europe.uni-heidelberg.de/de/forschung/heidelberg-research-architecture/hra-projects/quotationfinder>.

Most, G., ed., *Collecting fragments - Fragmente sammeln* (Göttingen 1997).

Perseus Digital Library, online publication [accessed: 23rd October 2010]: <http://www.perseus.tufts.edu/hopper>.

Romanello, M., M. Berti, A. Babeu, G. Crane, 'When printed hypertexts go digital: information extraction from the parsing of indices', in *HT 09. Proceedings of the 20th ACM Conference on Hypertext and Hypermedia,* (Turin and New York 2009) 357-58.

TLG Consortium, *Thesaurus Linguae Graecae,* CD-ROM Disk E, University of California,Irvine, released in February 2000.

Waltinger, U., A. Mehler, G. Heyer, 'Towards automatic content tagging: enhanced web services in digital libraries using lexical chaining', in *4th Int. Conf. on Web Information Systems and Technologies (WEBIST '08), 4-7 May, Funchal, Portugal*, ed. J. Cordeiro, J. Filipe and S. Hammoudi (Barcelona 2008) 231-36.

Yu, L., J. Ma, F. Ren, S. Kuroiwa, 'Automatic text summarization based on lexical chains and structural features', in *snpd, vol. 2, Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, (Qingdao 2007) 574-78.

TOWARDS A VIRTUAL DATA CENTRE FOR CLASSICS

TOBIAS BLANKE, MARK HEDGES,
AND SHRIJA RAJBHANDARI

1. Introduction

For many years Classics researchers have been producing a variety of digital outputs, whether in the form of relational databases, corpora of texts marked up in XML (Extensible Markup Language), or other formats. Naturally, these resources tend to focus on specific research topics that reflect the interests of their creators, whether in terms of the nature of the source material, or the time periods, communities, and geographical areas addressed; nevertheless, they are reusable resources that support research beyond that intended, or even envisaged, by their creators. Moreover, the content of these various digital resources is often conceptually related, each representing a small part of the increasingly rich data landscape available for the ancient world, and they would be of much greater utility to researchers if they could be linked up in a way that allowed them to be explored as a unity. However, while the resources may be reusable, the variety of data representations and formats used militates against such an integrated view. This is the challenge of interoperability – without interoperability, each resource remains an island, which can be combined with other resources only with a great deal of effort on the part of the researcher.

One way of approaching the interoperability challenge has been standardization. Many discussions have taken place (and not just in the humanities) about establishing standards for the creation of digital resources, with the aim (among others) of facilitating the creation of highly interlinked corpora. An important example of this from Classics is EpiDoc,[1] which provides standards and guidelines for the mark-up and interchange of inscriptions and other ancient documents which conforms to the TEI (Text Encoding Initiative) XML guidelines.

However, although the development of standards such as EpiDoc is an important step forward, standardization is unlikely to solve all issues around linking up heterogeneous data in the humanities, for a number of reasons. Firstly, there exists already a great deal of legacy data in diverse, non-standard, and often obsolete formats.[2] Secondly, users have first to be trained in the correct application of a standard, which requires potentially a large investment in terms of time and money that not all projects may be able to accommodate. Thirdly, even when standards are used, the sheer variety of the data means that there is significant flexibility in how the standards are applied (a selection of TEI documents, for example).

[1] EpiDoc website: <http://epidoc.sourceforge.net>.

[2] T. Blanke, M. Hedges, and S. Dunn, 'Arts and humanities e-science: current practices and future challenges', *Future Generation Computer Systems* 25. 4 (2009) 474-80.

Finally, standards are generally developed within particular disciplines or domains, such as (in the example of EpiDoc) epigraphy, whereas research is often inter-disciplinary, making use of varied source material and incorporating data conforming to different standards. There will inevitably be diversity of representation when information is gathered together from different domains and for different purposes, and consequently there will always be a need to integrate this diversity.

The approaches that we describe are based on the principle of respecting the integrity of existing representations of data, while virtualizing data and services over the web – that is to say, creating a virtual version of a resource by means of an abstraction layer that is independent of the underlying data structures and storage systems, allowing heterogeneous resources to be treated in a common fashion. It is outside the scope of this work to address the very broad range of digital resources available in Classics; rather, we investigated the issues raised by integrating structured datasets relating to ancient documents, albeit structured datasets that contained a significant quantity of unstructured text and structured data. Specifically, we used datasets relating to epigraphy and papyrology, although the issues raised are of relevance to other datasets of analogous form. We present two case studies representing different approaches, the first using data-grid technologies to provide integrated views of resources, the second enabling integrated content-based retrieval of resources. We consider the two approaches to be complementary, each providing in different ways dynamic integrated views over the data, and reducing uncertainties about the information by linking it to related information in other sources. We elaborate on this further elsewhere.[3]

The structure of the chapter is as follows. First, we present the background to the work, in Section 2 describing the specific collections used for our experiments and examining the interoperability issues that they raise, and in Section 3 introducing the technologies and approaches that we used to virtualize the collections. These approaches are addressed in two case studies presented in Sections 4 and 5. Finally, in Section 6, we deliver the vision of a Classics virtual data centre that emerged from our investigations. In Section 6, we also discuss in which related disciplines like archaeology such a service already exists. We suggest a virtual data centre for Classics, as the data resources are smaller and more spread out. Also, a virtual data centre could be more easily embedded in larger initiatives and would be therefore cheaper to maintain. Finally, all the experiments presented here will be integrated into the architecture discussion and realization of the emerging European infrastructure for digital arts and humanities, *DARIAH*.[4]

2. The data and its challenges

The digital resources used in our experiments primarily concern ancient documents, in particular classical epigraphy and papyrology, and in terms of formats included relational databases with different schemas and implemented using different data technologies, as well as a corpus of XML data. Specifically, the resources used were:

[3] T. Blanke and M. Hedges, 'Humanities e-science: from systematic investigations to institutional infrastructures', in *E-SCIENCE 10: Proceedings of the Sixth IEEE International Conference on e-Science* (Washington, DC 2010) 25-32.

[4] *DARIAH* website: <http://www.dariah.eu>.

- The *Heidelberg Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens* (*HGV*),[5] a Filemaker Pro database containing metadata on some 55,000 papyri, mostly from Roman Egypt and its environs, including bibliography, dates, and places (*e.g.* findspots and provenances);
- *Projet Volterra*,[6] a Microsoft Access database of Roman legal pronouncements and associated metadata, from various sources, whether epigraphic, papyrological, juristic, or literary;
- The *Inscriptions of Aphrodisias* (*IAph*),[7] an XML corpus of 1500 inscriptions from the ancient city of Aphrodisias in Asia Minor, including transcribed texts and metadata marked up using EpiDoc TEI, as well as images.

These three datasets vary significantly, both in terms of their information content and the formats used for implementing them; however, there is sufficient overlap in content to make integrating them profitable to researchers. For instance, although the *Volterra* collection is specific to legal texts, it contains some papyri and therefore some references to find-spots that also occur in the *HGV* metadata. Similarly, although none of the *Volterra* texts are inscriptions from Aphrodisias, there are attestations of persons that appear in both the *Volterra* and *IAph* texts (especially in the late antique period, for which the Aphrodisias material is most richly annotated). The *IAph* and *HGV* collections do not have any content in common, but the categories that are used to organize the texts within these resources have a certain overlap, for example letters, decrees, honours, contracts. All three datasets overlap fairly closely in date, and have similar (but not identical) mechanisms for recording dates, date ranges, periods, and uncertainty of dating.[8] Cross-corpus searches based on these areas of overlap should provide realistic tests of our approaches, as well as yielding potentially useful scholarly results. Note, however, that these three resources are just three examples from a much larger pool of related datasets that might have been included in a larger-scale integration project, and were selected in order to investigate the feasibility of our approaches and to identify issues that might arise.

We may make the following observations about these three datasets and the researcher's broader data environment that have consequences for data interoperability:

- Data formats are very diverse, and involve multiple media and standards;
- Databases rarely follow standard database schemas. The use of mark-up can vary significantly, particularly in resources developed before much effort had been made towards standards (such as EpiDoc), but natural variation occurs even in applying these standards;
- The material may be highly complex, with many structural and semantic relationships both internal – for example within a TEI document – and

[5] *HGV* website: < http://www.rzuser.uni-heidelberg.de/~gv0/>.

[6] *Projet Volterra* website: <http://www.ucl.ac.uk/history2/volterra>.

[7] *Inscriptions of Aphrodisias* website: <http://insaph.kcl.ac.uk>.

[8] A particularly challenging issue being investigated is that of handling different levels of uncertainty in temporal data: some dates are extremely precise – even to the day – whereas many others are very vague – perhaps to a span of 50 or 100 years.

contextual. The interpretation of an object (*e.g.* an inscription) may depend on its relationships to other resources and collections (*e.g.* other inscriptions, literary texts, archaeological surveys, concordances), which are moreover not necessarily digital;

- Data may be incomplete, indeed incompletable – the capture of the data cannot be repeated nor the data enhanced to fill in the gaps. For example, an inscription may be damaged, a papyrus's provenance not recorded, a corpus of texts fragmentary, the date of an event unknown;
- Data may be fuzzy or uncertain, or even contradictory. For example, there may be several sources for the date of an event, with various degrees of precision (to the year, to the decade) and various degrees of reliability;
- The resources are not easily available for use. In many cases, they are locked away on departmental machines; in other cases they are 'published' on a web site but not in a way that makes the resources particularly usable by a researcher;
- Even when a resource is available it is often available only in isolation. Many of these resources may be regarded as fragments of a larger picture, and would have vastly more value if researchers could have access to this larger picture, rather than just the parts;
- The resources may be owned by different communities and subject to different rights; the scholars who created them may be unwilling to accept anything that affects the integrity of the original resources. Consequently, any integration initiative must respect this autonomy and integrity, if it is to be successful;

We would not argue that such issues arise with respect to data only in Classics, nor that all Classics data can be characterized in this way. These issues will, however, be recognized by a significant number of researchers in many humanities disciplines.[9]

3. Virtualization approaches

Our general aim is to enable sharing of heterogeneous data resources in Classics in an integrated fashion, rather than as a number of isolated resources. Our broad approach may be described as one of virtualization of data and of access to data, by means of abstracted and standardized interfaces and protocols. Virtualization describes in computing all approaches that create a virtual version of a physical resource and goes back to the early days of computing with hardware virtualization strategies. In our case, instead of the actual data resources, the users can directly interact with virtual combination of them. Data can generally be *virtualized* in relation to several aspects:

- Location. Access is provided independently of where the datasets reside.
- Autonomy. Data may be governed by independent management regimes, owned by different communities and subject to different rights. Access is made more uniform while respecting the integrity of the original data and the environments

[9] For an extensive discussion and systematic analysis of the characteristics of humanities data see also M. Doerr, 'The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata', *AI Magazine* 24.3 (2003) 75-92.

in which it is managed, so that access to the data is in accordance with the terms of the data holders.

- Heterogeneity, both the infrastructural heterogeneity of the storage, and the structural heterogeneity of the data. Virtualization means that datasets do not need to be accessed in possibly idiosyncratic ways.

Although all three are relevant for our work, the third aspect is the key one for integrating diverse resources. Virtualization can hide 'irrelevant' (for whatever purpose we have in mind) differences between data resources, giving the user more seamless access to them. Distributed, autonomous, and heterogeneous datasets can be federated and regarded as a single resource, enhancing the visibility of the data and multiplying the uses to which it can be put. Virtualization offers, therefore, new ways of defining interfaces between datasets, where irrelevant aspects are ignored and the common information across the datasets retained. In this paper we discuss two approaches to virtualizing the data resources. The first publishes the resources as data services that expose datasets in a standard, relational, database-like way, while the second allows virtual representations of resources to be constructed by building common indexes from existing datasets.

4. Virtualization case study 1: linking and querying ancient texts

The *LaQuAT* (*Linking and Querying Ancient Texts*) project[10] investigated the use of the OGSA-DAI (Open Grid Service Architecture-Data Access and Integration)[11] middleware. OGSA-DAI is widely used for supporting virtual integration of diverse, distributed data resources, providing 'on-the-fly' common interfaces to data. Its primary focus was on relational databases, and it supports integrated views across many different database management systems, with a particular view to querying, transforming, and delivering data in different ways via a simple toolkit for developing client applications. It was used in the first instance to provide an integrated view across relational databases with different schemas, namely the *HGV* and *Projet Volterra* data resources. OGSA-DAI is designed to be extensible, and subsequently our work was extended to integrate the *IAph* XML corpus, providing an integrated view over the three structured data resources.

Figure 1 shows how the *LaQuAT* architecture integrates different database resources. OGSA-DQP (Distributed Query Processing) was the main abstraction mechanism, hiding the details of the database implementations from the user. Our approach to virtual data integration is thus to specify the local data sources as views over the global schema. Out of the separate databases we create a large, virtual one.

In the case of the integration of the *Volterra* and *HGV* data resources, two principal alternatives were discussed and evaluated. On the left hand side of figure 1, the architecture contains a single abstraction layer using OGSA-DQP, which hides the implementation details of the underlying databases. In addition, we bridge the language divide between the German *HGV* and the English *Volterra* data resources by using a *join* table to map between German and English keywords. On the right hand side, the

[10] *LaQuAT* project website: <http://www.laquat.cerch.kcl.ac.uk/>. *LaQuAT* was funded by JISC (Joint Information Systems Committee).

[11] OGSA-DAI website: <http://www.ogsadai.org.uk>.

architecture uses an additional OGSA-DQP abstraction layer to hide the fact that *HGV* is a German database. We decided that the former was the preferable solution, as access to the *join* table may be beneficial for other data resources.

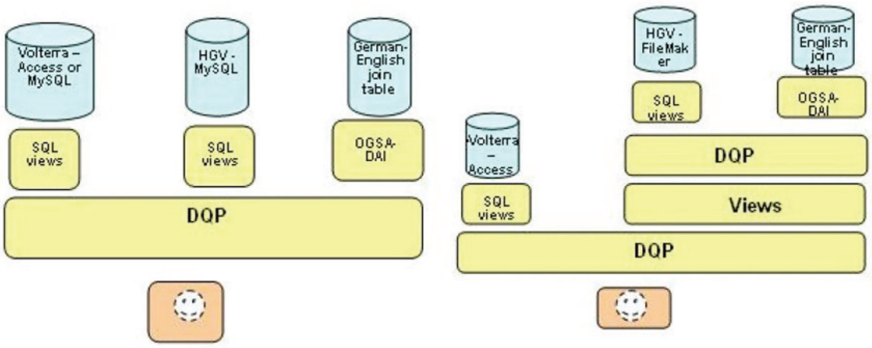


Figure 1: LaQuAT Architecture

OGSA-DAI uses SQL (Structured Query Language) views to hide the details of a data resource. In OGSA-DAI, everything from a standard database, to an XML file, to an indexed text resource will look to the user as if they were interacting with a single, large SQL data resource. To this end, OGSA-DAI generalizes the concept of an SQL view and virtualizes it. For *LaQuAT*, the following combination of traditional database technologies and OGSA-DAI technology will realize the virtualization of the data resources.

SQL views can handle the following requirements:

- Expose TEXT date column types as DATE date column types. In *Volterra*, for example, all date-related fields are defined as text fields in MS Access.
- Form a union of N tables so they are treated as a single table. This is standard-view functionality. However, some of the data resources like MS Access have very specific ways of realizing them.
- Expose German column and table names as English, handling any spaces and German characters.

OGSA-DAI DQP can additionally handle the following requirements:

- Expose multi-lingual column contents as English. This is done using the already-mentioned join table.
- Perform text searches over the contents of individual fields.
- Perform a join across databases.

LaQuAT's results were promising from a Classicist's point of view, as new lines of enquiry by combining existing data resources could be explored, for example by discovering references to homonymous (and possibly identical) persons in different texts that could be dated to within a small number of years of one another. Nevertheless, *LaQuAT* also identified limitations to this approach to data integration in the case of humanities resources.

In general terms, OGSA-DAI is optimized for working with *data-centric* rather than *text-centric* resources. The distinction here is between resources that contain significant quantities of unstructured text (text-centric), and those that consist primarily of structured data such as numerical data, dates, or very short text fields. In the humanities, however, researchers work more commonly with text-centric resources, such as text documents, within which they want to find relevant information so that standard document retrieval techniques can be applied and adapted for dealing with the specifics of handling additional structural constraints.[12] Indeed, the limitations of our first approach became particularly apparent when it came to working with XML files of inscriptions rather than with databases. Jackson *et al.* discussed *LaQuAT* in more detail and also presents the results of our user acceptance testing.[13] The next section presents an approach that addresses these issues.

5. Virtualization case study 2

The JISC-funded Virtual Research Environment (VRE) project *gMan* investigated how to build a research environment for everyday data-driven research in Classics.[14] Specifically, it showed how to provide a variety of integrated views over heterogeneous archives that correspond to specific research interests and reflect the actual day-to-day working practices of the researchers that work with the resources. These practices range from highly specialized semantic annotations using community standards such as EpiDoc, to the use of standard, online search tools such as integrated library catalogues and Google Scholar.[15] In the *gMan* experiment, we wanted to support those communities of humanities researchers that would like to work with a specialized set of digital collections but are not satisfied with standard search and retrieval tools. These researchers may want to search across resources based on looser criteria of relevance – for example by searching for all Roman legal texts in one resource containing information on punishments that are also mentioned in papyri from another resource – and where their needs are served neither by the sort of search functionality investigated in Section 4 – which is too highly structured – nor by such very general-purpose search tools as Google, which will fail to deliver the specific functionality required.

gMan addressed services that would enable more general-purpose Classics research activities, such as integrating and organizing the heterogeneous and often unstructured digital resources through advanced discovery facilities. We investigated how the UK and European research infrastructure can be exploited to support data-driven, collaborative

[12] T. Blanke, M. Hedges, and S. Dunn, 'E-science in the arts and humanities – from *ad hoc* experimentation to systematic investigation', in *E-SCIENCE '07: Proceedings of the Third IEEE International Conference on e-Science and Grid Computing* (Washington, DC 2006) 103-10; M. Nentwich, *Cyberscience. Research in the Age of the Internet* (Vienna 2003).

[13] M. Jackson, M. Antonioletti, T. Blanke, G. Bodard, M. Hedges, A. Hume, and S. Rajbhandari, 'Building bridges between islands of data – an investigation into distributed data management in the humanities', in *Proceedings of the Fifth IEEE International Conference on e-Science* (Washington, DC 2009) 33-39.

[14] *gMan* project website: <http://gman.cerch.kcl.ac.uk>.

[15] T. Blanke, L. Candela, M. Hedges, M. Priddy, and F. Simeoni, 'Deploying general-purpose virtual research environments for humanities research', *Phil. Trans. R. Soc. A* 368, 1925 (2010) 3813-28.

research in Classics by using the gCube environment,[16] which was developed by the EU-funded *D4Science* project.[17] gCube allows virtual research communities to deploy VREs on demand by making use of the shared resources of the European research infrastructure, and provides services that match closely the sort of information organization and retrieval activities that we identified as being typical in humanities research. It enables this use by virtualizing these resources in full-text indexes, which can be interrogated using various standard search and browse tools. This way, content can be delivered to Classics researchers more effectively, independently of the location and implementation of that content, and with special facilities provided for customizing the retrieval, management, and manipulation of the content.

In our experiment, researcher communities were able to ingest the three data resources described in Section 2 into the gCube environment, which involves mapping the resources and their metadata into the generic data model used by gCube. Researchers were supported in this task by an import service that provides standard workflows for importing data, workflows that can be customized by using a simple scripting language. The environment also allows a variety of text-based indexes to be created for the collections, thus generating a number of different views onto the collections.

Using the imported collections, researchers could then deploy specific VREs to work on specific research questions by combining the data resources to which their virtual organization has access with tools and services that support interaction with the underlying data. Various search and browse tools offer access to the collections using keywords or geo-locations as entry points. Finds can be brought together in so-called virtual collections, which assemble references to items in existing collections. These virtual collections and the items in them can in turn be shared among the group of researchers that come together in the VRE. Other tools and services include a report-writing tool, as well as several annotation services.

6. A Classics data service

Our final aim is to integrate disparate, heterogeneous data sources for Classics using virtualization technologies. Our current design, based on the experiences and issues outlined in the *LaQuAT* and *gMan* experiments, is outlined in figure 2. Using OGSA-DAI, we integrate database resources with connectors, which allow users to query multiple remote databases as if they were a single virtual database. Using gCube, we join together remote document collections using a single, joint index of the textual sources. Again, remote, heterogeneous datasets can be queried. The final aim would be a network of integrating servers, *e.g.* for disciplines in humanities, maintained by a trusted arts and humanities data service, similar to existing services such as the Archaeology Data Service[18] in the UK or the Dutch DANS.[19] Each of these is a member of the European

[16] L. Candela, D. Castelli, and P. Pagano, 'gCube: a service-oriented application framework on the grid', *ERCIM News* 72 (2008) 48–49.

[17] *D4Science* website: <http://www.d4science.eu>.

[18] Archaeology Data Service website: <http://archaeologydataservice.ac.uk>.

[19] DANS website: <http://www.dans.knaw.nl>.

DARIAH project,[20] a European infrastructure for digital arts and humanities, which will take up some of the ideas expressed here.

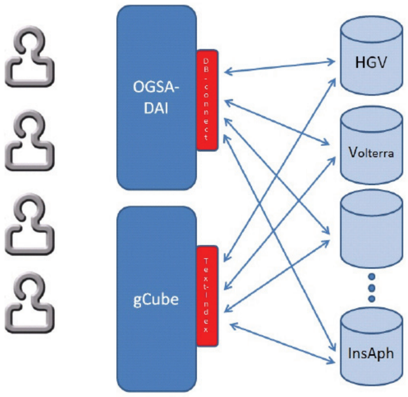


Figure 2: Classics virtual data centre

Thus, in our attempts to set up a virtual data centre for integrating Classics resources, to date we have two ways of connecting, developed by these two projects. Each way requires a trusted intermediary that would allow those remote sources to expose their data. This connection could be established in several ways – for example, the data could be transmitted and kept in a trusted vault, or the data centre could be allowed to query the remote data source directly. This will depend on the requirements of the remote data source and agreements made. Using gCube as an integration platform, the remote data source could also allow only the publication of the index of its textual resources.

7. Conclusions and future work

The main output for each case study was a demonstrator that provided integrated views over the three datasets used in the experiments (the same datasets were used in each case), although with quite different results in each. The conclusions from *LaQuAT* concerned limitations to the approach rather than solutions. The relational model followed by OGSA-DAI was more effective for resources that consist primarily of *structured* data (which we call *data-centric*) rather than for largely *unstructured* text (which we call *text-centric*), which makes up a significant component of the datasets we were using. This approach was, moreover, insufficiently flexible to deal with the semantic issues described in Section 2. The *gMan* project, on the other hand, addressed these problems by virtualizing data resources using full-text indexes, which can then be used to provide different views onto the collections and services that more closely match the sort of information organization and retrieval activities found in the humanities, in an environment that is more interactive, researcher-focused, and researcher-driven.

[20] *DARIAH* website: <http://www.dariah.eu>.

Subsequent to the projects described here, we were funded to experiment with a linked data approach to the issues described, as part of the *SPQR* project.[21] The primary aim of *SPQR* was to link and integrate datasets related to classical antiquity using RDF (Resource Description Framework) or equivalent formalisms, taking particular account to address the semantic issues described above. We followed core standards, in particular the Europeana Data Model (EDM),[22] which has been developed by the EU-funded *Europeana* project for modelling cultural heritage data, as well as OAI-ORE (Open Archives Initiative Object Reuse and Exchange)[23] and emerging domain-specific ontologies and vocabularies. Ontologies form the centrepiece of the data integration project here, acting as semantic mediators for heterogeneous databases, which are mapped onto ontologies to provide semantic views over the datasets.[24]

It should be noted, however, that the resources used in the experiments described in this chapter were just three examples from among numerous others to which these various approaches could be applied. There are many small, scattered, yet related resources that would be much more useful to researchers if they were linked along these lines. Their utility would increase greatly once a certain critical mass is reached and together they would form a whole much greater than the sum of the parts, enabling researchers to ask questions that would not otherwise have been possible. An analogy might be a map, where each dataset represents a small area, say a few houses within a street; integrating a few of them is of limited utility, but after a certain point is reached there will be sufficient information to navigate from one place to another. A further output of the work is the definition of a Classics interoperability service for data resources, as defined in the previous section.

The benefits for the researchers include the ability to ask new questions by integrating the data resources. A great quantity of digital resources has been produced by humanities researchers in recent years, and a significant proportion of these are in the form of databases and of corpora of texts marked up in XML (usually some form of TEI). Although there are a number of initiatives to create standards in particular areas, such as EpiDoc, there will inevitably be a certain degree of variety in the representation of information gathered in different circumstances and for different purposes. A striking example is provided by museum databases, developed for the purposes of object cataloguing, and ill-suited to interact with other treatments of the objects within the museum. In any case, there is also a considerable body of legacy resources, especially databases, that exist in a variety of forms. However, there is a definite and pressing need to enable researchers to link up disparate data resources, but to do so using virtualization without affecting the original resources, which may be owned by different communities and subject to different rights.

Tobias Blanke (*King's College London*) tobias.blanke@kcl.ac.uk
Mark Hedges (*King's College London*) mark.hedges@kcl.ac.uk

[21] *SPQR* project website: <http://spqr.cerch.kcl.ac.uk>.

[22] EDM documentation: <http://pro.europeana.eu/edm-documentation>.

[23] OAI-ORE website: <http://www.openarchives.org/ore>.

[24] T. Blanke, G. Bodard, M. Bryant, S. Dunn, M. Hedges, M. Jackson, and D. Scott, 'Linked data for Humanities research – the *SPQR* experiment', in *Proceedings of the Sixth IEEE International Conference on Digital Ecosystems Technologies* (Washington DC 2012) 1-6.

THE *SON OF SUDA ON-LINE*

RYAN BAUMANN

Introduction

Integrating Digital Papyrology (*IDP*) is a multi-institutional project aimed at establishing and improving relationships between three digital papyrological resources: the *Duke Databank of Documentary Papyri* (*DDbDP*), the *Heidelberger Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens* (*HGV*), and the *Advanced Papyrological Information System* (*APIS*). Started in 1983, the *DDbDP* collects a number of digital transcriptions of ancient documentary papyri from print editions. *HGV* and *APIS*, meanwhile, collect metadata (place of origin, date, keywords, bibliography, *etc*.) and images of much of the same material. A unification of these data sources would allow linking the digital transcriptions of texts with the images, dates, and other metadata, and in 2007 the Andrew W. Mellon Foundation funded a project, *Integrating Digital Papyrology*, to begin this process. Over the years the *DDbDP* has undergone a number of transitions, and this project also supported its transition from an idiosyncratic SGML encoding to standards-based EpiDoc XML markup.[1] In addition, the grant provided funding to improve and finish the first generation of a tool for searching and browsing the unified collection of materials, called the Papyrological Navigator (or PN). At the conclusion of the *IDP* grant, Mellon funded a second phase of the project (called *IDP*2) with the following goals:[2]

1. Improve operability of the PN search interface on the merged and mapped data from the *DDbDP*, *HGV*, and *APIS*.
2. Facilitate third-party use of the data and tools.
3. Create a version-controlled, transparent and fully audited, multi-author, web-based, real-time, tagless, editing environment, which – in tandem with a new editorial infrastructure – will allow the entire community of papyrologists to take control of the process of populating these communal assets with data.

The environment described in the last item, inspired by the *Suda On-Line* (*SOL*),[3] was named the *Son of Suda On-Line* (*SoSOL*). Though it takes its name and inspiration from *SOL*, *SoSOL* was written from the ground up to incorporate new technologies, address

[1] SGML (Standard Generalized Markup Language), first formalized as a standard in 1986, is the generalized document markup language of which XML (Extensible Markup Language) is a descendant. EpiDoc <http://epidoc.sourceforge.net> is a set of community guidelines for marking up digital editions of ancient texts.

[2] J. D. Sosin *et al.*, 'Integrating Digital Papyrology 2' Mellon Foundation (New York 2008): <http://www.duke.edu/~jds15/IDP2-FinalProposalRedacted.pdf>

[3] The *Suda On-Line*: <http://www.stoa.org/sol/> is a project aimed at collaborative translation of the massive tenth century Byzantine encyclopedia known as the *Suda*.

project-specific problems, and move toward more open data and tooling.[4] This chapter aims at not only a description of the resulting software, but also of the challenges encountered and solutions chosen in its formulation to encourage broader adoption or discussion of both.

The Son of Suda On-Line

Though collaborative online editing environments, such as Wikipedia, have the advantage of allowing anyone to contribute, many question the scholarly integrity of resources which can be edited by anyone, unvetted. The *Suda On-Line*, which actually predated the existence of Wikipedia by two years, addressed this problem by marking submitted translations with their level of editorial vetting.[5] This combination of openness to contribution with strong editorial control was the guiding principle in the design of the *Son of Suda On-Line*.

However, even more than *SOL*, the papyrological projects encompassed in *Integrating Digital Papyrology* value the scholarly integrity of data published under their aegis. Thus, *SoSOL* attempts to digitally replicate the scholarly mechanisms of the peer review these projects would normally enforce. This results in somewhat of an inversion of where and how editorial control is exerted in comparison with *SOL*. While *SOL* users are authorized by editors during registration and are assigned work or must request a specific entry,[6] in *SoSOL* users are not screened and at present work on whatever they feel needs emendation or inclusion in the corpus. However, this distinction in the assignment of work may just arise naturally from the differing natures of the texts involved; whereas the *Suda* – while large – is a bounded unit of work, the papyri do not yet show signs of halting their expanding numbers in transcriptions and publications.

Standing in starker contrast is how submissions that have not received editorial oversight are handled: in *SOL*, they are immediately publicly searchable and accessible but marked as ‘draft’; in *SoSOL*, submissions undergo review and voting by an editorial board before publication as ‘canonical’ and being made available for searching in the Papyrological Navigator. This may seem restrictive, or even contradictory to claims of openness. It is indeed the former, but only inasmuch as the editorial boards are controlling the quality of what they are willing to put their names to in the tradition of peer review. The latter requires some discussion.

Data and openness

We no longer see *IDP* as representing at any given moment a synthesis of fixed data sources, directed by a central management; rather, we see it as a constantly changing set of fully open data sources, governed by the scholarly community and maintained by all active scholars who care to participate. One might go so far as to say that we see this

[4] T. Elliott, *Background and funding: integrating digital papyrology* (2008): <http://idp.atlantides.org/trac/idp/wiki/BackgroundAndFunding>; J. D. Sosin, ‘Digital papyrology’, in *26th Congress of the International Association of Papyrologists* (2010): <http://www.stoa.org/archives/1263>.

[5] R. Finkel, W. Hutton, P. Rourke, R. Scaife, and E. Vandiver, ‘The *Suda On Line*’, *Syllecta Classica* 11 (2000) 178-90: <http://www.stoa.org/sol/about.shtml>.

[6] A. Mahoney, ‘Tachypaedia Byzantina: the *Suda On Line* as collaborative encyclopedia’, *Digital Humanities Quarterly* 3.1 (2009).

nexus of papyrological resources as ceasing to be 'projects' and turning instead into a community.[7]

What do we mean here by 'fully open'?

On one level, it is the terms under which data is published. Though asserting any sort of copyright on the 2,000-year-old texts themselves is perhaps nonsensical (at least in the USA),[8] the complete set of *IDP* XML files are published with a Creative Commons Attribution 3.0 Licence,[9] explicitly permitting the typical varieties of scholarly reuse and citation anticipated for the data, in line with other recent calls for open access in the humanities.[10] (Atypical and unanticipated forms of reuse would be even more exciting.)

On another level, it is the manner in which data is published. For collaborative online projects, this is usually a challenge. If the data is constantly changing, how do you publish it in any traditional sense? Perhaps even more challenging is this: how do you publish the changes themselves, both retroactively and proactively?

By retroactively, we mean that the revision history of the data up to the present may itself be important; by proactively, we mean that if a user has already obtained the complete revision history at some point in time, it is better to allow them to simply download the changes since that point. Many online collaborative environments, such as MediaWiki and the original *SOL*, store all changes in a database system. This usually makes distribution of the complete revision history, particularly proactive distribution, extremely difficult. As an example, the English-language Wikipedia was unable to distribute its complete dataset for several years, and was only most recently able to dump its revision history in January 2010, with no successful exports since.[11] Even if they were able to do so, the only mechanism for updating such a data dump is to download the entire several-hundred-gigabyte file each time.

Next-generation version control

We felt that the best way to approach this problem was to use a Revision Control System as the backend for the data itself, instead of a traditional database. Though there are many

[7] R. Bagnall, 'Integrating digital papyrology' in *Online humanities scholarship: the shape of things to come*, ed. F. Moody and B. Allen, Mellon Foundation (New York 2010): http://hdl.handle.net/2451/29592;

[8] See *e.g.* Bridgeman Art Library *v*. Corel Corp.: <http://www.law.cornell.edu/copyright/cases/36_FSupp2d_191.htm>, which rules that 'slavish copies' of public domain works are not copyrightable. For scholarly transcriptions and images of ancient texts, much of the goal is to produce as faithfully slavish a copy as possible.

[9] Creative Commons Attribution 3.0 License: <http://creativecommons.org/licenses/by/3.0/>.

[10] G. Crane, 'Give us editors! Re-inventing the edition and re-thinking the humanities', in *Online humanities scholarship: the shape of things to come*, ed. F. Moody and B. Allen, Mellon Foundation (New York 2010): <http://cnx.org/content/m34316/latest/>.

[11] Wikipedia: 'Database download – latest complete dump of English Wikipedia': <http://en.wikipedia.org/w/index.php?title=Wikipedia:Database_download&oldid=393163797\#Latest_complete_dump_of_English_Wikipedia>.

well-established centralized systems such as CVS and Subversion,[12] in recent years there has been an explosion in the popularity of Distributed Version Control Systems (DVCSs). Typically this means there is no 'central' server except by social convention; all copies of the repository are, in a sense, equal and can share changes with one another. This allows for a variety of workflow styles, and has a number of other important impacts.

One of the most popular distributed version control systems is Git, initially developed by Linus Torvalds for managing the Linux kernel software project. Due to its broad acceptance, design choices, and proven performance on a number of large projects, it is the backend we selected for data in *SoSOL*. The *SoSOL* codebase itself was also managed with Git from the beginning, and is available online.[13]

As a result of using a DVCS for the data backend, it is possible to use Git not only to retrieve the complete revision history of the *IDP* data as managed by *SoSOL*[14] (retroactive publication), but also to easily update your copy of the repository as changes are published (proactive publication). Due to the distributed nature of Git, the concepts of branching development and merging changes have been integrated into its design, making it easy to keep your copy of the data up-to-date even if you have made your own modifications. (After all, if anyone can pull changes from any other copy of the repository, merging needs to be fast and easy.) The long-running version of this behaviour of splitting off your own modifications is known in the open source world as 'forking'. Git reduces the overhead of both forking a project, as well as of contributing your forked changes back.

That a DVCS makes these things trivial also represents a significant decision in the design of *SoSOL*: for the 'canonical' data repository it interacts with, it does not need to care about any external mechanisms or workflows used to introduce changes. *SoSOL* only needs to keep track of changes within its domain; it is merely a front-end and social infrastructure for easing and managing contributions. When a user edits data in *SoSOL*, it is forked from the main repository to allow them to do their work without interruption. They then submit their changes for editorial review, and when they pass muster they are merged back into the canonical repository. However, the repository may be updated by any external process in the interim, typically without drastically impacting the work that must be done to perform the merge.

Thus, the fact that *IDP* now uses Git for its public data repository, in combination with the licence the data is distributed under, represents the complete realization of 'a constantly changing set of fully open data sources governed by the scholarly community and maintained by all active scholars who care to participate'.[15] For us, 'participate' in fact has two senses: participating within our system (that is, participating in our editorial

[12] CVS, or the 'Concurrent Versions System,' was started in 1986 as a version control system built atop the even-earlier 'Revision Control System', which only operated on a single file. 'Subversion' was started as a later project for version control similar to CVS, but with various fixes and improvements.

[13] The *Son of Suda On-Line*: <https://github.com/papyri/sosol>.

[14] *IDP* Data: <http://github.com/papyri/idp.data>. See also *IDP* Data available on GitHub: <http://digitalpapyrology.blogspot.com/2011/01/idp-data-available-on-github.html>.

[15] Bagnall, 'Integrating digital papyrology' (n. 7 above).

review process), or participating in any enterprise you choose with the complete dataset which we make freely available.

Implementation

The *Son of Suda On-Line* environment itself is written in Ruby using the popular Rails web framework.[16] Instead of the mainline Ruby interpreter written in C (usually referred to as Matz's Ruby Interpreter, or MRI, after the language's creator), we use a Java implementation called JRuby. Though this was initially done to enable deployment of *SoSOL* in any Java Servlet Container such as Tomcat, *SoSOL* has come to use a number of Java libraries (particularly for interacting with XML data), facilitated by JRuby's tight Java integration.

Git internals

Some discussion of how Git works and internally represents version history is perhaps necessary to illustrate how its design enables and informs other design decisions in *SoSOL*. In Git, the version history of a project is encoded as a directed graph of three kinds of internal objects, all of which are identified by the unique SHA-1 hash of the object's content.[17] These objects form the 'nodes' in Git's graph, while their contents contain the directional arrows linking them together.

The simplest instance of version history (in Git, and conceptually) is a single piece of content with one version. Bare content is the 'blob' object in Git, and has no additional metadata associated with it – these can be thought of as leaf nodes in the graph (that is, nodes which do not point to other nodes).

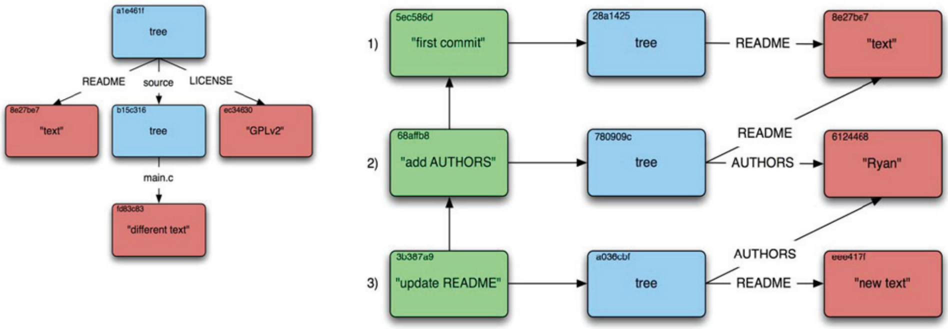
However, having just one file is not very useful for most projects. Git organizes and collects multiple blob objects into a file structure using tree objects. These tree objects are simple, plaintext files, which list identifying hashes with their filenames. Each tree object represents a single directory; for subdirectories, trees can point to other tree objects in addition to blobs, as in figure 1a.

Revisions in Git are stored as commit objects. These point to a single tree, and contain metadata about the commit (such as author, time, commit message) as well as pointers to one or more parent commit objects. Thus, a simple merge would have two parent commits, while a branch or fork would be two different commits pointing to the same commit. The fact that all of this history is stored as a connected graph is what allows the Git system itself to examine things such as where a fork occurred when attempting to merge concurrent changes, and intelligently make the right decision based on the chosen merge strategy.

Let us walk through a simple, linear commit history to illustrate things, winding up with the graph shown in figure 1b.

[16] 'Ruby' is a dynamic, object-oriented, interpreted programming language. 'Rails' is a web framework written for and in Ruby which uses the Model-View-Controller design pattern to organize web applications.

[17] Think of the SHA-1 hash as a 160-bit number that uniquely identifies any given input string. Even a small change in the string results in a very different hash, and collisions are designed to be rare.



(a) A tree graph in Git (b) A series of commits in Git

Figure 1: Visualizations of Git's internal graph structure. Red squares are blobs, blue are trees, green are commits. Text in the top-left corner is the object's truncated hash.

1. We start our project with just a README file containing the string 'text'. Since we want to record this momentous occasion, we commit the state of our repository with the commit message 'first commit'. This commit points to the hash of the tree, which contains the hash of our README.
2. Since this project is sure to be our *magnum opus*, we quickly decide we want to immortalize contributors by crediting them in an AUTHORS file. We write this out and hastily make a new commit to add it. Since the README is unchanged, the same object from before is reused. However, because we have added a new file, the tree has changed, so a new tree object is made for the commit to point to.
3. Later, the mood strikes us to update the README, so we do so, and make a new commit. Again, a new blob object is constructed for the new content, although this time the AUTHORS blob is able to be reused because we did not modify it. Because changing the README changes the blob object's hash, the tree object must store the new hash and receive a new hash itself.

Though there are many other facets and features, this is the core of Git.

The consequence of all objects being identified by a unique hash of their contents means these objects can easily be shared between copies of the repository – knowing both that there will not be conflicts between objects with different content having the same hash, and that the same object will have the same hash no matter where it is. Since all links between objects are part of the hashed content, any change in the graph would result in a cascade of changing hashes (as in step three of our example). Because Git verifies and uses these hashes for its own operation, the integrity of the repository is incredibly robust; if you have a copy of the repository, and someone tries to rewrite history by removing or altering an object which is already referenced, you will notice when you try to pull their changes because all descendant objects will have different hashes from yours. If an object file is corrupted, it can easily be restored from any copy of the repository; likewise, any copy of the repository is a complete copy of the repository.

Data sources, publications, and workflow

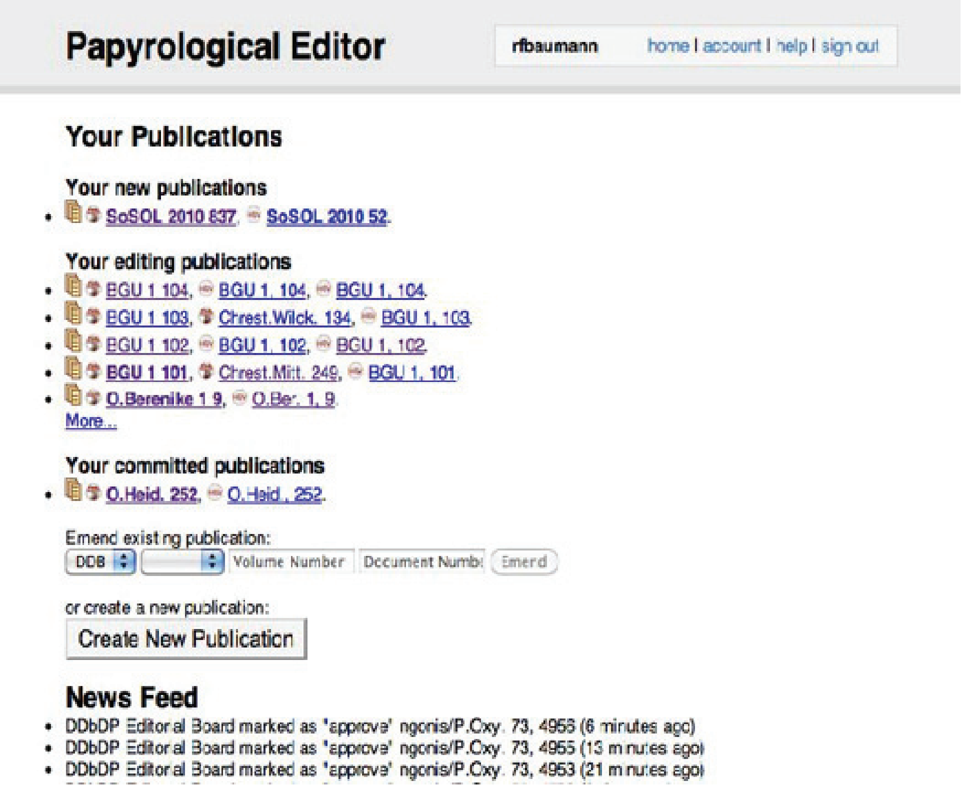
As outlined in the introduction, *IDP* represents the synthesis of a variety of papyrological projects managed by different institutions. This has informed *SoSOL*'s design in how it interacts with and models these disparate data sources.

Because these papyrological resources evolved separately, the concept of a 'publication' and what defines an object or text may in some cases be slightly different. As an example, two distinct hands may have written two different texts on a single piece of papyrus; *HGV* keeps two metadata records, while the *DDbDP* keeps a single transcription (but still indicating the distinct hands inside it). These relationships can become quite complicated with things such as reprints of texts, or ancient, military receipt records containing hundreds of texts. In addition, the data itself is different enough that different methods may be preferable for editing or interacting with data from a given resource. Although *IDP* has standardized on EpiDoc XML encoding, it still collects a variety of different kinds of information about papyri, including transcriptions, translations, and complex metadata.

SoSOL deals with this internally by representing each 'publication' as a collection of one or more resources, which we call 'identifiers'. Because there are a variety of types of resources, we use an identifier base class which implements common methods for all types of resources (such as getting or setting the identifier's content in the Git repository), while using subclasses of this to implement behaviour specific to a given type of resource. Each identifier is actually called such because it is named with a string which it is assumed corresponds to exactly one resource – for example, *HGV*'s or *DDbDP*'s name for an object. Because these resources are stored as separate XML files with their own particular directory structure, each type of identifier has its own method for turning its name into a file path.

SoSOL uses the 'publication' as the unit of work for the editorial workflow – each publication corresponds to a development branch in the Git backend. When the identifiers belonging to a publication have been modified, it can be submitted to the editorial boards for review. Because we have had to deal with disparate types of resources from the beginning, the review process is able to have different editorial boards for each type of identifier. Currently this is implemented as a sequential workflow; if a user submits a publication with changes to *HGV* metadata, *DDbDP* transcription, and *HGV* translation, it will be reviewed in that order, requiring approval from each board before going to the next. Each editorial board has their own membership and voting rules. If a submission is rejected, the user is able to see the reasons given during voting and revise and resubmit their work. Users can also see a list of other users working on the same 'publication', as well as their contact information. If they desire to coordinate amongst themselves, they can share links to their publications, which are viewable (but not editable) by any logged-in user who knows the link.

Because of Git's design, *SoSOL* is actually able to maintain a separate Git repository for each user and editorial board in its backend (see figure 2). Despite the fact that *IDP*'s canonical data repository is around 1GB in size, each copy can be on the order of kilobytes because it can simply reference objects already stored in the repository it has been forked from. This means that when a user begins editing a publication, the branch for that publication is made on their copy of the repository, and only new objects which they create in the course of making updates must be stored in it. When they submit the publication, this branch and its related objects are copied to the board's repository, and, when they approve the publication, they then merge this branch back into the canonical data repository. Eventually, this design could be integrated with a Git server (in the style of GitHub),[18] allowing each user to have direct access to their own Git repository to make changes easily, using any process they choose, before submission.



Your Publications

Your new publications
- SoSOL 2010 637, SoSOL 2010 52.

Your editing publications
- BGU 1 104, BGU 1, 104, BGU 1, 104.
- BGU 1 103, Chrest.Wilck. 134, BGU 1, 103.
- BGU 1 102, BGU 1, 102, BGU 1, 102.
- BGU 1 101, Chrest.Mitt. 249, BGU 1, 101.
- O.Berenike 1 9, O.Ber. 1, 9.

More...

Your committed publications
- O.Heid. 252, O.Heid., 252.

Emend existing publication:

or create a new publication:

News Feed
- DDbDP Editorial Board marked as 'approve' ngonis/P.Oxy. 73, 4956 (6 minutes ago)
- DDbDP Editorial Board marked as 'approve' ngonis/P.Oxy. 73, 4955 (13 minutes ago)
- DDbDP Editorial Board marked as 'approve' ngonis/P.Oxy. 73, 4953 (21 minutes ago)

Figure 2: A user’s dashboard in SoSOL, with publications being worked on.

Another advantage of Git’s design is that accurate, transparent attribution is easily maintained in the history of each piece of data. Interventions made by the editorial board after submission are preserved as being authored by them, rather than by the submitter. Because Git allows a distinction between ‘author’ (who wrote a commit) and ‘committer’ (who put a given commit in the repository), we can record which editor made the merge to the canonical repository without losing information about who actually authored the underlying changes. We also record the members of the editorial board at the time a submission is accepted, by adding that they have signed off to the commit message. All of this is done as part of a process we call ‘finalization’ – after a board approves something, it is assigned to a random member of the board to undergo any final revisions and manual oversight of the merge into the canonical repository. As part of this, we flatten multiple commits made before submission into a single commit (as each time the user saves it introduces a commit, which was deemed more revision granularity than necessary for our

[18] Secure source code hosting and collaborative development – GitHub: <http://github.com/>.

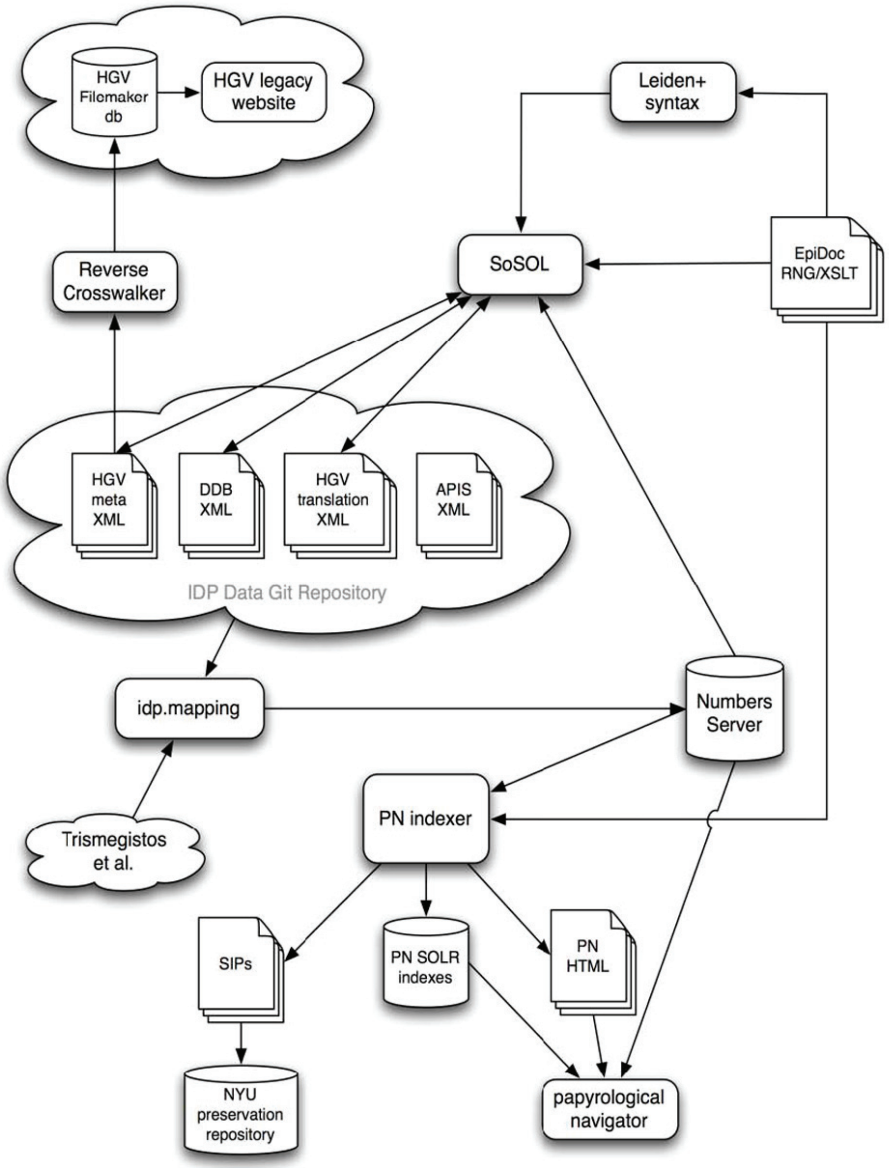


Figure 3: Software components and data flow at the conclusion of IDP2

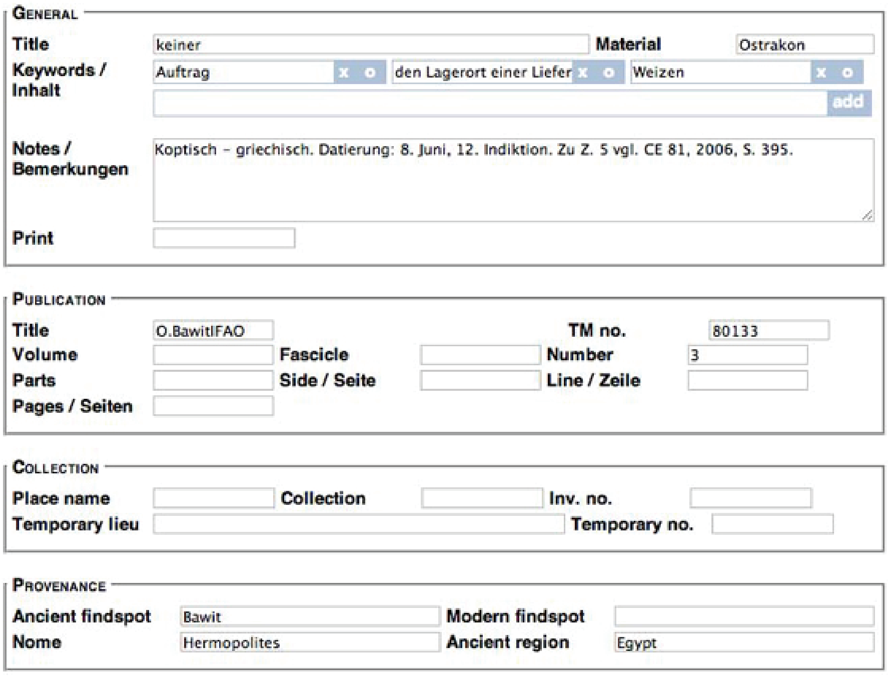
use), which is rewritten to have the content of any individual commit messages as well as the submission reason and editorial sign-off messages.

The advantage of all this is that the core design of *SoSOL* deals mainly with this identifier/publication model, and providing functionality for using this abstraction to have an editorial workflow on top of a Git repository. The goal was to make these core components reusable, while providing your own implementations for how your own identifiers are edited and aggregated into publications. One can imagine the simplest implementation as being an identifier whose name is the file path, which just presents the plaintext contents of the file for editing, and which has no relationships with other identifiers, so that each publication is a single identifier.

Under *IDP*2, *SoSOL* manages the complex relationships between identifiers by interfacing with a piece of software developed for the project, which we call the Numbers Server. This is implemented as an RDF triplestore,[19] which is built up by processing the entire canonical *IDP* dataset and looking for associations between resources. *SoSOL* can then simply query any single identifier in the Numbers Server to find all other identifiers related to it, in order to aggregate them into a logical publication. The relationships and data flow between *SoSOL*, the Papyrological Navigator, and the Numbers Server are illustrated in Figure 3.

Alternative syntax for XML editing

One of the proposed items for the *SoSOL* environment was to provide a 'tagless' editing environment for the EpiDoc XML data used by *IDP*. For metadata, where there is some fixed number of possible elements, this can be achieved by simply presenting the user with a form specific to the needed data types which translates to and from XML. This is what we have done for *HGV* metadata, as shown in Figure 4.



GENERAL

Title	keiner	Material	Ostrakon
Keywords / Inhalt	Auftrag x o den Lagerort einer Liefer x o Weizen x o add		
Notes / Bemerkungen	Koptisch – griechisch. Datierung: 8. Juni, 12. Indiktion. Zu Z. 5 vgl. CE 81, 2006, S. 395.		
Print			

PUBLICATION

Title	O.BawitIFAO			TM no.	80133
Volume		Fascicle		Number	3
Parts		Side / Seite		Line / Zeile	
Pages / Seiten					

COLLECTION

Place name		Collection		Inv. no.	
Temporary lieu				Temporary no.	

PROVENANCE

Ancient findspot	Bawit	Modern findspot	
Nome	Hermopolites	Ancient region	Egypt

Figure 4: HGV metadata entry form in SoSOL

[19] An RDF (Resource Description Framework) triplestore provides a way of storing and querying the relationships between resources as 'triples' in subject-predicate-object form, such as 'dogs are animals' or, in this case, '*DDbDP*'s P.Oxy. 1 53 relates to *HGV* number 20715'.

However, for freeform text transcriptions like those recorded by the *DDbDP*, the concept of a ‘tagless’ environment is more challenging. Because many papyrological materials are damaged and difficult to read, scholarly transcriptions record a number of things about the reading of the text itself. If letters of a word cannot be made out on the object, but you can interpolate from context what they likely were, you should indicate that your restoration is a result of that process. This happens so often when editing papyrological texts that a shorthand for indicating them in the text itself was developed and, in a 1931 meeting at the University of Leiden, standardized as a set of rules called the Leiden conventions. Epigraphers also adopted these conventions, as they faced many of the same challenges and, along with papyrologists, have used them for publishing printed transcriptions ever since (as in Figure 5).

[ἔτους α (?) Αὐτοκράτορος] . . [.]. . του
[- ca.12 -] Σεβαστοῦ
[εἴργ(ασται) ὑ(πὲρ) χω(ματικῶν) ἔ]ργ(ων) τοῦ αὐτοῦ̣ πρώτου (ἔτους)
[-ca.?-] κ κϛ ἐ̇[ν] τῆ̣ Ἐπα -
[γαθιαν]ῆ̣ διώ(ρυγι) Βακχιά(δος)
[-ca.?-] Πατκ(όννεως) τοῦ Θεαγένους
[. . . . . .] μη(τρὸς) Ταύρεως
[-ca.?-] (hand 2) σεση(μείωμαι)

Figure 5: Typical print transcription following Leiden conventions (P.Sijp., 41a)

1. [ἔτους] [<#α=1#> (?)] [Αὐτοκράτορος] .2[.1].2του
2. [ca.12] Σεβαστοῦ
3. [(εἴργ(ασται)) (ὑ(πὲρ) χω(ματικῶν))] ([ἔ]ργ(ων)) τοῦ αὐτοῦ̣ πρώτου ((ἔτους))
4. [.?] <#κ=20#> <#κϛ=26#> ἐ̇[ν] τῆ̣ Ἐπα
5.- [γαθιαν]ῆ̣ (διώ(ρυγι)) (Βακχιά(δος))
6. [.?] (Πατκ(όννεως)) τοῦ Θεαγένους
7. [ca.6] (μη(τρὸς)) Ταύρεως
8. [.?] $m2 (σεση(μείωμαι))

Figure 6: Leiden+ representation of the P.Sijp., 41a text

EpiDoc is a TEI-based XML encoding standard for marking up the same sort of textual semantics represented in Leiden, with additional standardized markup for other features typically needed when encoding ancient texts. For example, numbers which are written as Greek text can be marked semantically as numbers with their value, orthographic corrections can have both the normalized and original word linked, and so forth. Thus, the key advantage of EpiDoc is that it acts as a superset of Leiden with explicit, computationally actionable semantics.

Because we have our transcriptions already encoded in EpiDoc, we wanted to surface these facets of it to users, without burdening them with the full verbosity of XML markup. We also wanted them to be able to explicitly mark up new texts in the same environment.

We contemplated trying to use the contentEditable HTML attribute to provide a sort of 'rich text', what-you-see-is-what-you-get (WYSIWYG), text entry form. However, browser implementations of this feature vary wildly in behaviour and often confound user expectations, and what exactly the meaning of 'WYSIWYG' is when semantically marking up things such as numbers is debatable. As a result, we decided to use a simple, plaintext, form element, utilizing a transformation of the XML to make the text more legible and easier to edit quickly. Of course, to update the modified plaintext would require a transformation back to XML to save it in our system. One way to do this would be to write two separate transform processes, one from XML to plaintext, and one from plaintext to XML. However, verifying and maintaining such a process would be difficult.

Instead, we use a tool called XSugar to perform both directions of the XML transformation.[20] This utility allows you to define a single, context-free grammar, where each rule has both an XML representation and a non-XML representation. Thus it can parse either representation into an intermediate form, and then use the same ruleset to output the opposite representation. Additionally, the tool can check that this transformation is reversible – that is, round-trips of a given input do not alter it.[21] Due to the immense size of the *DDbDP* corpus (over 55,000 transcriptions), we use automated nightly runs of transformations on the entire corpus to both verify and improve our definition of the Leiden+ grammar, as well as reduce encoding errors in the source XML (many being difficult, edge cases left over from the transitioning of the *DDbDP* from SGML to EpiDoc). We also use an XML normalization process to reduce the amount of 'thrashing' in the version history – small changes to the text should not alter unrelated parts of the XML and make it hard to spot the actual change when looking through the file's history.

We call our non-XML representation of EpiDoc markup 'Leiden+', as it attempts to use the same symbols as Leiden where possible, but is also able unambiguously to represent the additional markup enabled by EpiDoc encoding. Figures 5 and 7 illustrate how a traditional print transcription might be marked up in EpiDoc XML, with figure 6 being the Leiden+ transformation of that XML. As you can see, things like the Greek letter 'κ' on line four being the number '20' are implicit in print, but explicit in both EpiDoc and Leiden+.

Though Leiden+ must in some cases be more verbose than traditional Leiden, this is due to the fact that Leiden+ must be able to be transformed into unambiguous, valid EpiDoc XML. For example, on line 3, abbreviations expanded by an editor use nested parentheses instead of a single pair around just the expansion, because the unit of text which is being expanded must be marked up as well. Because multiple, standalone Unicode combining underdots (indicating vestiges of illegible characters, as in line 1 of the example text) can be confusing to type and count by themselves, Leiden+ simply uses a period followed by the number of characters. However, for characters which are unclear but can be inferred from context (as in 'ῦ' of 'αὐτοῦ' at the end of line 3), Leiden+ preserves the combining underdot for readability, and we provide a JavaScript helper for inserting the character (a screenshot

[20] XSugar – Dual Syntax for XML Languages: <http://www.brics.dk/xsugar/>.

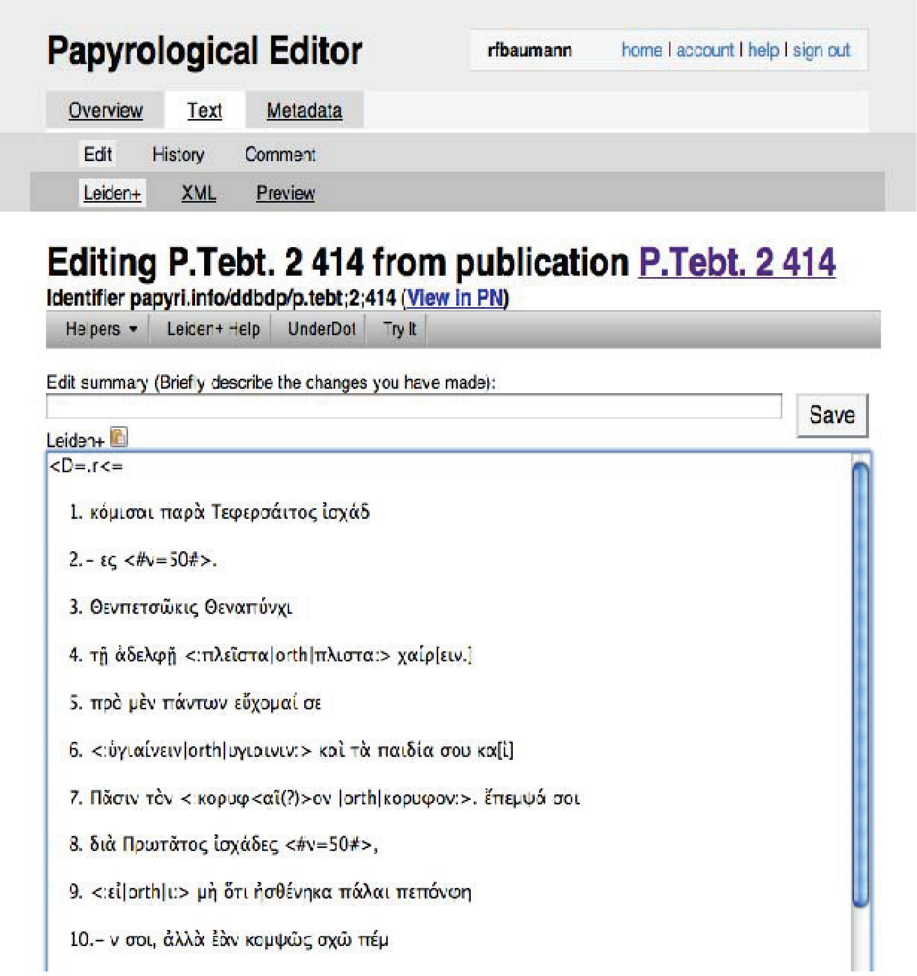
[21] C. Brabrand, A. Møller, and M. I. Schwartzbach, 'Dual syntax for XML languages', *Information Systems* 33.4-5 (2008) 385-406.

of Leiden+ as it appears with helpers in the editing environment is shown in Figure 8). Users can, of course, still edit the XML directly, with a button provided to copy the entire content of each text area to their clipboard so they can paste it into their own editor. In either case, submissions are validated against the EpiDoc RELAX NG schema before saving in order to ensure that invalid XML does not make its way into the system.[22]

```
<div xml:lang="grc" type="edition" xml:space="preserve">
 <ab>
  <lb n="1"/><supplied reason="lost">ἔτους</supplied> <supplied
   reason="lost" cert="low"><num value="1">α</num> </supplied> <supplied
   reason="lost">Αὐτοκράτορος</supplied> <gap reason="illegible" quantity="2"
   unit="character"/><gap reason="lost" quantity="1" unit="character"/><gap
   reason="illegible" quantity="2" unit="character"/>του
  <lb n="2"/><gap reason="lost" quantity="12" unit="character"
   precision="low"/> Σεβαστοῦ
  <lb n="3"/><supplied reason="lost"><expan>εἴργ<ex>ασται</ex></expan>
   <expan>ὑ<ex>πὲρ</ex> χω<ex>ματικῶν</ex></expan></supplied>
   <expan><supplied reason="lost">ἔ</supplied>ργ<ex>ων</ex></expan> τοῦ
   αὐτο<unclear>ῦ</unclear> πρώτου <expan><ex>ἔτους</ex></expan>
  <lb n="4"/><gap reason="lost" extent="unknown" unit="character"/> <num
   value="20">κ</num> <num value="26">κ ς </num> ἐ<supplied
   reason="lost">ν</supplied> τῇ ’Επα
  <lb n="5" type="inWord"/><supplied reason="lost">γαθιαν</supplied>ῇ
   <expan>διώ<ex>ρυγι</ex></expan> <expan>Βακχιά<ex>δος</ex></expan>
  <lb n="6"/><gap reason="lost" extent="unknown" unit="character"/>
   <expan>Πατκ<ex>όννεως</ex></expan> τοῦ Θεαγένους
  <lb n="7"/><gap reason="lost" quantity="6" unit="character"
   precision="low"/> <expan>μη<ex>τρὸς</ex></expan> Ταύρεως
  <lb n="8"/><gap reason="lost" extent="unknown" unit="character"/>
   <handShift new="m2"/><expan>σεση<ex>μείωμαι</ex></expan>
 </ab>
</div>
```

Figure 7: EpiDoc XML fragment equivalent to the preceding Leiden+

[22] RELAX NG (REgular LAnguage for XML Next Generation) allows the creation of complex XML validation rules not possible with traditional XML DTDs (Document Type Definitions).



Editing P.Tebt. 2 414 from publication P.Tebt. 2 414

Identifier papyri.info/ddbdp/p.tebt;2;414 (View in PN)

Edit summary (Briefly describe the changes you have made):

Leiden+

```
<D=.r<=

  1. κόμισαι παρὰ Τεφερσάιτος ἰσχάδ

  2.- ες <#ν=50#>.

  3. Θενπετσῶκις Θεναπύνχι

  4. τῇ ἀδελφῇ <:πλεῖστα|orth|πλιστα:> χαίρ[ειν.]

  5. πρὸ μὲν πάντων εὔχομαί σε

  6. <:ὑγιαίνειν|orth|υγιαινιν:> καὶ τὰ παιδία σου κα[ὶ]

  7. Πᾶσιν τὸν <:κορυφ<αῖ(?)>ον |orth|κορυφον:>. ἔπεμψά σοι

  8. διὰ Πρωτᾶτος ἰσχάδες <#ν=50#>,

  9. <:εἰ|orth|ι:> μὴ ὅτι ἠσθένηκα πάλαι πεπόνωη

  10.- ν σοι, ἀλλὰ ἐὰν κομψῶς σχῶ πέμ
```

Figure 8: Editing Leiden+ in SoSOL

Conclusions

While Leiden+ does take some experience to get used to, in EpiDoc training seminars where we have introduced students and papyrologists to using it as an alternative for editing XML, the response thus far has been very positive. Users have entered and submitted thousands of new texts for inclusion in the *DDbDP*, using the production version of *SoSOL* running on papyri.info, dubbed the Papyrological Editor.[23] As of this writing, almost 25,000 commits have been made by over 200 different authors since the transition of *IDP* to using Git for its data backend, the majority of these through *SoSOL*. New, electronic editions of texts can now be made available much more quickly, and the potential for born-digital editions with vetting and version control has been enabled by the system. In addition, the *Perseus Digital*

[23] Papyrological Editor: < http://www.papyri.info/editor/> the public, running instance of *SoSOL* for *IDP*.

Library has recently announced that they plan to use *SoSOL* 'to decentralize the curation, annotation, and general editing of the TEI XML texts that it hosts'.[24]

The model of loosely-coupled tools operating on their interests over standard interfaces (for example: *SoSOL*'s interactions with Git, *SoSOL*, and PN using standard RDF to interact with the Numbers Server, *etc*.) has allowed for flexible and unexpected uses with very little additional work. For example, because the Leiden+ grammar definition and transformation code is completely separate from *SoSOL* and only included externally, the same code should be easily reusable by any project wishing to use it in conjunction with their standard EpiDoc XML texts. The Numbers Server being implemented as an RDF triplestore has also had great utility, allowing arbitrary SPARQL queries to be written, which reveal useful information about complex relationships in the data.[25]

We believe that exposing our complete dataset and its history to the community is the best approach to enabling true community ownership of the data. Using a DVCS for our data backend in the editing environment, instead of a traditional relational database backend, is what allows us to do this with very little friction, and facilitates direct interaction with the data repository without necessitating going through our project-specific editing environment. This is an approach which could be adopted by other Digital Humanities projects, even if not using *SoSOL* itself. That the system transparently preserves attribution at every step will, we hope, foster a sense that users do have some investment and ownership in their contributions. Additionally, it is hoped that this will enable academic institutions to recognize individuals' work in the system as scholarly activity, equivalent to work with traditional print publications.

This is also a way of moving away from the rigidity and implied authority of print publications. The *DDbDP* is not a fixed resource, finished at some date, unwavering and confident that it knows all; rather, it is a collection of conjectures, now easily capable of being revisited, revised, and improved. The technological framework now in place aims to reduce the overhead of these activities, to speed the expansion of knowledge and detection of error. It invites in the Popperian spirit: 'if you are interested in the problem which I tried to solve by my tentative assertion, you may help me by criticizing it as severely as you can'.[26] By publishing our data as a resource which is easily capable of decentralization and reuse, we hope also to apply this ideal to the system itself. While submissions which pass our system of editorial review derive their authority from the composition of the editorial board and the submitter, others may independently modify and publish our data under their own authority.

Ryan Baumann (*Harvard Centre for Hellenic Studies*) rfbaumann@gmail.com

[24] *Perseus Digital Library* news item, March 20th 2012: <http://www.perseus.tufts.edu/hopper/>.

[25] SPARQL (SPARQL Protocol and RDF Query Language) is the language used to query and discover relationships in RDF triplestores. (See also n. 19 above).

[26] K. Popper, *Conjectures and refutations* (London 1963) 30–36.

Bibliography

Bagnall, R., 'Integrating digital papyrology' in *Online humanities scholarship: the shape of things to come*, ed. F. Moody and B. Allen, Mellon Foundation (New York 2010): <http://hdl.handle.net/2451/29592>.

Brabrand, C., A. Møller, and M. I. Schwartzbach, 'Dual syntax for XML languages', *Information Systems* 33.4-5 (2008) 385-406.

Crane, G., 'Give us editors! Re-inventing the edition and re-thinking the humanities', in *Online humanities scholarship: the shape of things to come*. ed. F. Moody and B. Allen, Mellon Foundation (New York 2010): <http://cnx.org/content/m34316/latest/>.

Elliott, T., *Background and funding: integrating digital papyrology* (2008). <http://idp.atlantides.org/trac/idp/wiki/BackgroundAndFunding>.

Finkel, R., W. Hutton, P. Rourke, R. Scaife, and E. Vandiver, 'The *Suda On Line*', *Syllecta Classica* 11 (2000) 178-90: <http://www.stoa.org/sol/about.shtml>.

Mahoney, A., 'Tachypaedia Byzantina: the *Suda On Line* as collaborative encyclopedia', *Digital Humanities Quarterly* 3.1 (2009).

Popper, K., *Conjectures and refutations* (London 1963).

Sirks, A. J. B., and K. A. Worp, ed., 'Papyri in memory of P. J. Sijpesteijn', *American Studies in Papyrology* 40 (2007) 275-76.

Sosin, J. D., 'Digital papyrology', in *26th Congress of the International Association of Papyrologists* (Lexington KY 2010): <http://www.stoa.org/archives/1263>.

THE *LEXICON OF GREEK PERSONAL NAMES* AND CLASSICAL WEB SERVICES

ELAINE MATTHEWS[1] AND SEBASTIAN RAHTZ

Introduction

The *Lexicon of Greek Personal Names* was established in 1972 as a Major Research Project of the British Academy. The overall objective of the *LGPN* project is to create a comprehensive and authoritative record of the names of all individuals attested in Greek (or with Greek names attested in Latin) in the ancient Greek-speaking world, and so provide the classical research community worldwide with a unique and fundamental resource for the study of all aspects of the ancient Greek world.

Research publications about the *Lexicon* provide a pointer to the range of research fields which the *LGPN* can illuminate, and, in some instances, makes possible for the first time, including linguistics, the history of religion, historiography and literary history, demographic studies, and above all, cultural interaction.[2] In practice, of course, there is no limit, nor should there be, to the uses researchers will make of the *LGPN*'s material.

LGPN is internationally recognized as a resource which has transformed the basis on which names may be studied and used. It has done so to date primarily through its publications; so far, over a quarter of a million individuals sharing over 35,000 names have been published in six regional volumes:

1. I, Aegean Islands, Cyprus, Cyrenaica (1987);
2. II, Attica (1994);
3. IIIA, Peloponnese, W. Greece, Sicily, Magna Graecia (1997);
4. IIIB, Central Greece (2000);
5. IV, Macedonia, Thrace, Northern Regions of the Black Sea (2005);
6. VA, Coastal Asia Minor: Pontos to Ionia (2010);

with at least two more in preparation:

- VB, Coastal Asia Minor from Caria to Cilicia (due 2012);
- VC, Inland Asia Minor;

[1] Elaine Matthews was unfortunately unable to complete her planned enhancements to this paper before her untimely death in June 2011, so there is less critical engagement with other prosopography and person databases than had been planned.

[2] See: S. Hornblower and E. Matthews, ed., *Greek personal names: their value as evidence*, Proceedings of the British Academy 104 (Oxford 2000); E. Matthews, ed. *Old and new worlds in Greek onomastics*, Proceedings of the British Academy 148; R. W. V. Catling, F. Marchand and M. Sasanow, ed., *Onomatologos: studies in Greek personal names presented to Elaine Matthews*, (Oxford 2010); P. M. Fraser, *Greek ethnic terminology* (Oxford 2009).

(and, if the project funding continues, extending to a second tranche of work on Palestine, Syria, and the Trans-Euphratic Regions, as well as possible work on Egypt, for which some material has been collected).[3]

The regional basis on which the *Lexicon* research and publication has been undertaken has meant that use of the collection as a whole has been relatively limited. The project offers summary data online, which provides the numbers of hits per name, and allows the reader to establish, for example, that the name ’Αβάσκαντος is attested 147 times.[4] However, the other data about those 147 uses of the name have remained on paper. This chapter attempts to unlock some of that information, and show how it can be accessed in a variety of ways.

Lexicon *data categories*

The key to understanding the *LGPN* records is the set of *data categories* established at the start of the project. It was decided then to record the following pieces of information:

1. Normalized primary name form;
2. Sex of person named;
3. Place where the person belonged;
4. Date of the attestation (which can vary wildly in precision);
5. Bibliographical references;
6. Assorted other data. This may include placename variations *e.g.* alternative places of citizenship; name variants (orthography, dialect), corrections *etc*; parent/child relationships to other *people*; status or profession; and editorial corrections/alterations to the record.

The initial data collection was made on paper slips (an example is shown in Figure 1) on the basis of scholars familiar with the region reading primary and secondary sources.

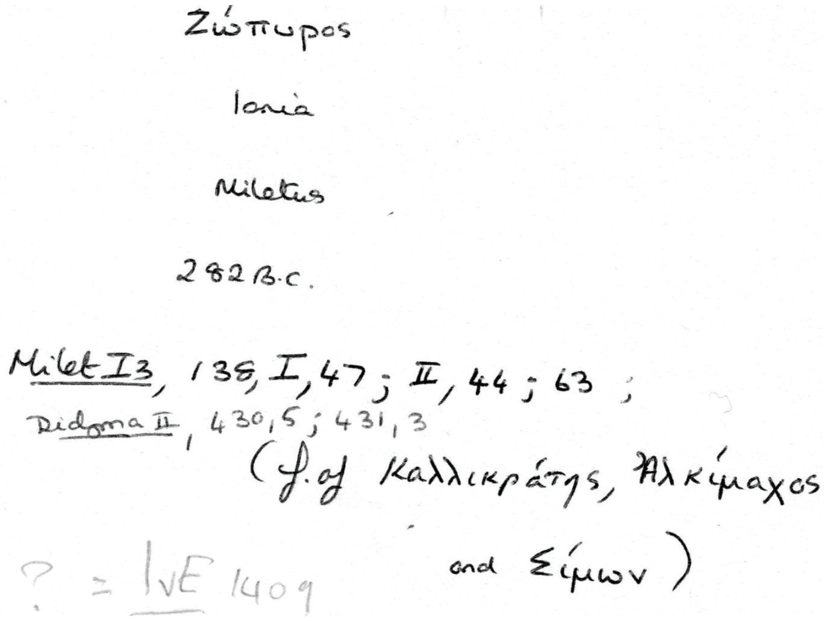
The initial phase of work consisted solely of data collection, but by the late 1970s it was beginning to be recognized that conventional typesetting of the desired publication was likely to be prohibitively expensive, and that managing the entries on a computer should allow for delivery of camera-ready copy to the publisher. Given the technology of the time, it was not clear how this would be achieved, but a text format was agreed which provided the minimal distinctions. In this compact text form of the *Lexicon* data, a record looks like this, with six fields of information separated by @ characters:

```
Nani1s @ f @ Athens? @ F4B @ +IG II<2> 12229 @ (%_Na!nei1s!)
```

This asserts that there is a record of a person called Νανίς, a woman, probably from Athens, in the first half of the fourth century BC, with a bibliographical reference (publication of an inscription). The transliteration is the *Lexicon*’s own internal system (Table 1): numbers are used to indicate accents and breathings, except for 6 and 7 (used to

[3] For more details see the *LGPN* state of preparation page: <http://www.lgpn.ox.ac.uk/publications/stateprep.html#4prep>.

[4] The search interface for *LGPN*: <http://www.lgpn.ox.ac.uk/database/lgpn.php>.



Ζώπυρος

Ionià

Miletus

282 B.C.

Milet I3, 138, I, 47; II, 44; 63;
Didyma II, 430, 5; 431, 3
(f. of Καλλικράτης, Ἀλκίμαχος
and Σίμων)

? = IvE 1409

Figure 1: Lexicon input slip

indicate archaic characters); the number 4 is used to indicate the editorial dot under an unclear reading). The name is normalized from the attested form which has the syllable νείς. We should note that this is not guaranteed to be different from the person in another record with the name Νανίς, but the *Lexicon* believes it is a distinct individual. Any other inscriptions which mention the same person will be conflated with this record. The bibliographical reference is usually, but not always, exhaustive. If the person is very well attested, another reference work, such as an encyclopaedia, will be cited, where the full references can be found.

The relationship to the place name is almost always place of birth, usually an ancient city or region, though the name of the modern find-spot may be given where the ancient site cannot be identified.

6 and 7	6=F 7=h
A	*A=῾A *A1=῞A *A3=῟A A =᾿A A'e=᾿A'ε A'i=Aἰ A'i1=Aἴ A'i3=Aἶ A'o=᾿A'o A'u=Aὐ A'u1=Aὔ A'u3=Aὖ A1=῎A A14=῎Ạ A3=῏A A4=᾿Ạ
B	B=B
C	C=X C4=Χ̣
D	D=Δ D4=Δ̣
E	*E=῾E *E1=῞E E=E E'i=Eἰ E'i1=Eἴ E'i3=Eἶ E'o=᾿E'o E'o1=᾿E'ό E'u=Eὐ E'u1=Eὔ E'u3=E ὖ E1=῎E E14=῎Ẹ E18=᾿Έ E4=᾿Ẹ
F	F=Φ F4=Φ̣
G	G=Γ G4=Γ̣
H	*H=῾H *H1=῞H *H3=῟H H=H H1=῎H H3=῏H H4=᾿Ḥ
I	*I=῾I *I1=῞I I=I I1=῎I I3=῏I
K	K=K K4=Ḳ
L	L=Λ L4=Λ̣

M	M=M M4=Μ̣
N	N=N N4=Ν̣
O	*O=῾Ο *O1=῝Ο O=Ο O1=῎Ο O4=᾿Ο̣
P	P=Π P4=Π̣
Q	Q=Θ Q4=Θ̣
R	*R=῾Ρ R=Ρ
S	S=Σ S4=Σ̣
T	T=T T4=Τ̣
U	*U=῾Υ *U1=῝Υ U=Υ U1=῎Υ U14=῎Υ̣
W	*W=῾Ω *W1=῝Ω *W3=῏Ω W=Ω W1=῎Ω W3=῟Ω
X	X=Ξ
Y	Y=Ψ
Z	Z=Z Z4=Ζ̣
a	a=α a1=ἄ a14=ἄ̣ a3=ἆ a34=ἆ̣ a4=ἀ̣
b	b=β b4=β̣
c	c=χ c4=χ̣
d	d=δ d4=δ̣
e	e=ε e1=ἔ e14=ἔ̣ e18=ἔ e3=□ e34=□̣ e38=ἐ͂ e4=ἐ̣ e48=ἐ̣ e8=ἐ e81=ἔ
f	f=φ f4=φ̣
g	g=γ g4=γ̣
h	h=η h1=ἤ h14=ἤ̣ h3=ἦ h34=ἦ̣ h4=ἠ̣
i	i=ι i1=ἴ i14=ἴ̣ i145=>ΐ̣ i15=>ΐ i3=ἶ i34=ἶ̣ i4=ἰ̣ i5=>ϊ
j	j4=᾿7̣
k	k=κ k4=κ̣
l	l=λ l4=λ̣
m	m=μ m4=μ̣
n	n=ν n4=ν̣
o	o=ο o1=ὄ o11=ὄ o14=ὄ̣ o148=ὄ̣ o18=ὄ o3=□ o38=ὀ͂ o4=ὀ̣ o48=ὀ̣ o8=ὀ o81=ὄ
p	p=π p4=π̣
q	q=θ q4=θ̣
r	r=ρ r4=ρ̣
s	s=ς s4=σ̣
t	t=τ t4=τ̣
u	u=υ u1=ὔ u14=ὔ̣ u15=>ΰ u3=ὖ u34=ὖ̣ u4=ὐ̣ u5=>ϋ
w	w=ω w1=ὤ w14=ὤ̣ w3=ὦ w34=ὦ̣ w4=ὠ̣
x	x=ξ x4=ξ̣
y	y=ψ y4=ψ̣
z	z=ζ z4=ζ̣

Table 1: *LGPN* transliteration

Some more complex examples of the *Lexicon* text markup format:

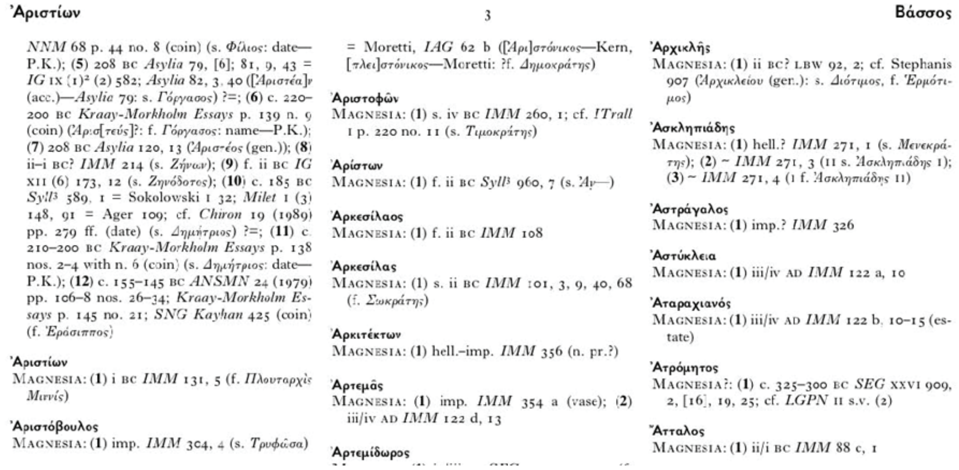
```
Qeo1frastos @ m @ Hagnous @ F1B @ Paus. i 37. 1; Plu., +Mor. 843c ==
 +PA 7169; +IG II<2> 1961, 19; 3510;= +IEleusis 301, 8 f., 14; +IG
 II<2> 3511;= +IEleusis 302?; 300, 32, 39, 45; Thompson, +New +Style
 +Coinage 1230 & +Chiron 21 (1991) pp. 13 f. @ (II s. %Qemistoklh3s I,
 %Ake1stion, f. %Qemistoklh3s II, %Sofoklh3s)

E'utuci1s @ f @ Amphissa @ 139-122BC @ +FD III (2) 122, [3-4], 7, 9-10 @ (%E'utu1cios (gen.) - l. 7:
endogen. slave/freed.)

Ga1gios @ m @ Apollonia-Sozopolis @ 4B @ +BMNBurgas 4 (2002) p. 124 no. 18 @ (%GAGIW (gen.) - ed.,
%Gagi1hs?: f. %Da3os)
```

show how the last field is very considerably overloaded with its own fields of information, especially in the 'final bracket' (beloved of generations of *Lexicon* staff), with subfields separated by ':' characters, and within those by commas and many other editorial conventions (*e.g.* marking of italics with +, and Greek with %). This is markup rather characteristic of its time, managed within the project over 30 years of work.

Critically, the records shown above provide no unique identifier for a *Lexicon* record, a problem which may have a considerable impact on future work.



'Αριστίων 3 Βάσσος

NNM 68 p. 44 no. 8 (coin) (s. *Φίλιος*: date—P.K.); (**5**) 208 BC *Asylia* 79, [6]; 81, 9, 43 = *IG* IX (1)² (2) 582; *Asylia* 82, 3. 40 ([*'Αριστέα*]ν (acc.)—*Asylia* 79: s. *Γόργασος*) ?=; (**6**) c. 220–200 BC *Kraay-Morkholm Essays* p. 139 n. 9 (coin) (*'Αρισ[τεύς]*?: f. *Γόργασος*: name—P.K.); (**7**) 208 BC *Asylia* 120, 13 (*'Αριστέος* (gen.)); (**8**) ii–i BC? *IMM* 214 (s. *Ζήνων*); (**9**) f. ii BC *IG* XII (6) 173, 12 (s. *Ζηνόδοτος*); (**10**) c. 185 BC *Syll*³ 589, 1 = Sokolowski I 32; *Milet* I (3) 148, 91 = Ager 109; cf. *Chiron* 19 (1989) pp. 279 ff. (date) (s. *Δημήτριος*) ?=; (**11**) c. 210–200 BC *Kraay-Morkholm Essays* p. 138 nos. 2–4 with n. 6 (coin) (s. *Δημήτριος*: date—P.K.); (**12**) c. 155–145 BC *ANSMN* 24 (1979) pp. 106–8 nos. 26–34; *Kraay-Morkholm Essays* p. 145 no. 21; *SNG Kayhan* 425 (coin) (f. *'Εράσιππος*)

'Αριστίων
MAGNESIA: (**1**) i BC *IMM* 131, 5 (f. *Πλουταρχὶς Μινίς*)

'Αριστόβουλος
MAGNESIA: (**1**) imp. *IMM* 304, 4 (s. *Τρυφῶσα*)

= Moretti, *IAG* 62 b ([*'Αρι*]*στόνικος*—Kern, [*τλει*]*στόνικος*—Moretti: ?f. *Δημοκράτης*)

'Αριστοφῶν
MAGNESIA: (**1**) s. iv BC *IMM* 260, 1; cf. *ITrall* I p. 220 no. 11 (s. *Τιμοκράτης*)

'Αρίστων
MAGNESIA: (**1**) f. ii BC *Syll*³ 960, 7 (s. *'Αρ—*)

'Αρκεσίλαος
MAGNESIA: (**1**) f. ii BC *IMM* 108

'Αρκεσίλας
MAGNESIA: (**1**) s. ii BC *IMM* 101, 3, 9, 40, 68 (f. *Σωκράτης*)

'Αρκιτέκτων
MAGNESIA: (**1**) hell.–imp. *IMM* 356 (n. pr.?)

'Αρτεμᾶς
MAGNESIA: (**1**) imp. *IMM* 354 a (vase); (**2**) iii/iv AD *IMM* 122 d, 13

'Αρτεμίδωρος

'Αρχικλῆς
MAGNESIA: (**1**) ii BC? LBW 92, 2; cf. Stephanis 907 (*'Αρχικλείου* (gen.): s. *Διότιμος*, f. *'Ερμότιμος*)

'Ασκληπιάδης
MAGNESIA: (**1**) hell.? *IMM* 271, 1 (s. *Μενεκράτης*); (**2**) ~ *IMM* 271, 3 (II s. *'Ασκληπιάδης* I); (**3**) ~ *IMM* 271, 4 (I f. *'Ασκληπιάδης* II)

'Αστράγαλος
MAGNESIA: (**1**) imp.? *IMM* 326

'Αστύκλεια
MAGNESIA: (**1**) iii/iv AD *IMM* 122 a, 10

'Αταραχιανός
MAGNESIA: (**1**) iii/iv AD *IMM* 122 b, 10–15 (estate)

'Ατρόμητος
MAGNESIA?: (**1**) c. 325–300 BC *SEG* XXVI 909, 2, [16], 19, 25; cf. *LGPN* II s.v. (2)

"Ατταλος
MAGNESIA: (**1**) ii/i BC *IMM* 88 c, 1

Figure 2: Lexicon typeset output

Figure 2 shows the three-column typeset output which was designed for the *Lexicon* in the early 1980s and has remained more or less consistent for all the printed volumes to date.

The Lexicon*'s IT history and current status*

The *Lexicon* has lived through all the generations of humanities computing, in each period espousing the technology of the moment where possible.

During the 1970s, the project had embraced digital storage, and had started transferring data on cards to files using a locally-written flat database (Famulus); there was no method of retrieval, and only an outline plan for producing camera-ready copy by using a pen plotter. By the start of the 1980s, however, the project had to confront the problem of selective retrieval, output, and checking the integrity of data. A period of intensive examination of the data so far input meant that by the mid-1980s, the *Lexicon* was loading material into a network database (IDMS), had retrieval programs written in FORTRAN, and was able to typeset pages using procedural markup on a Monotype Lasercomp. The database was subsequently converted to a relational model (using Ingres), retrieval programs were rewritten in Pascal and C, and the typesetting switched to producing PDF using TeX. The most significant landmark in the project's IT history was the design and implementation of the database structure to reflect and provide access to all the research components of an *LGPN* record, which in the publication books are simply presented as text, for example chronological and topographical data, and socially relevant data such as statuses and relationships.[5]

The database design (an early draft is shown in Figure 3) was crucial in imposing consistency of format on complex evidence, and the published volumes have all been generated from it in camera-ready form,[6] but it has not been opened up for general research. In the present century, staff continue to use text files with markup as their main editorial tool, backed by a relational database.

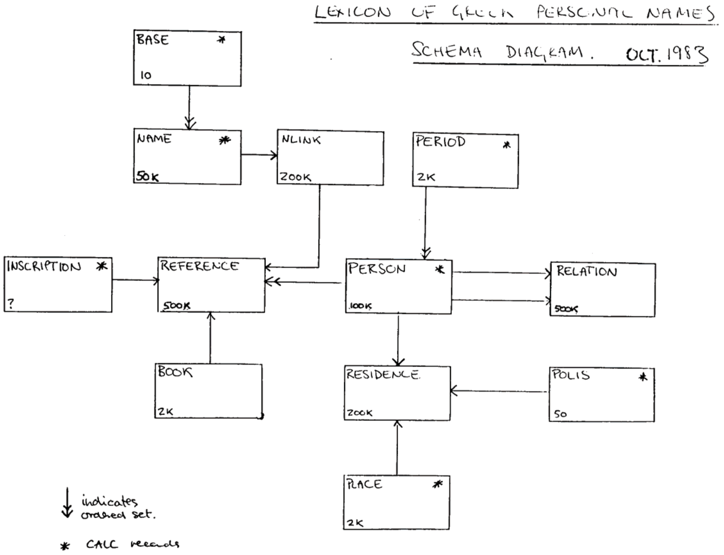


Figure 3 Database original schematic

The increasing requirement for data models which emphasize collaboration with other projects, and concerns over sustainability, caused the *Lexicon* to initiate a project in 2005 to remodel the data so that it could be represented in XML conformant with the guidelines of the Text Encoding Initiative (TEI).[7] The aim was to set up an IT infrastructure to support the future maintenance and preservation of irreplaceable research data, and provide direct online access to all the data, thus opening up the full potential of the *LGPN* data to researchers. This included the integration of data from all published (and to be published) volumes into one resource. The intention was that the *LGPN* play as significant a part in the e-research environment as it had played in traditional scholarship, and take a lead in setting standards for encoding names in documents and achieving interoperability with online material worldwide.

[5] E. Matthews and S. Rahtz, 'Designing and using a database of Greek personal names', *Proceedings of the VIII International Symposium of the Association of Literary and Linguistic Computing*, (Nice 1984).

[6] The first volume was typeset by database routines generating code for a Monotype Lasercomp; subsequent volumes utilized a complex relational database retrieval which extracted and transformed all the fields into TeX markup.

[7] TEI Consortium, ed., *TEI P5: guidelines for electronic text encoding and interchange* (Charlottesville VA 2007): <http://www.tei-c.org/release/doc/tei-p5-doc/en/html>.

The technical components of the XML phase of the *LGPN* which helped the *Lexicon* enhance its publishing and interchange capability consisted of several phases. First, there was the definition of an XML schema, as a customization of the TEI, for an archival form of the *Lexicon* data suitable for repositories. Existing database retrieval routines were then adapted to output XML conformant to the agreed schema.[8] This allowed the project to provide a simple forms-based interface for searching the database, and delivery of results in XML against the TEI schema; with the option of transforming that to other modern web delivery formats such as HTML for web pages, JSON (data optimized for consumption by Javascript in web pages) or RDF (for use in semantic web applications).

The *LGPN* XML work coincided with, and stimulated, a major revision of the TEI module relating to names and dates in 2006/07. The work done then to model persons, places, and organizations as first class objects, allowed the *Lexicon* schema to be a conformant pure subset of the TEI.

The major data categories present in the *Lexicon* map cleanly to TEI elements as follows:

Lexicon	**TEI**
Name	<nym>
Person	<person>
Name form	<persName>
Sex	<sex>
Date	Formally, *notBefore* and *notAfter* attributes on <birth>; informally in <floruit>
Status	<socecStatus>
Reference	<bibl>

The TEI format designed for the *Lexicon* makes use mainly of the <person> element to contain a *Lexicon* record.

After transformation, the new long-form XML record for the first *Lexicon* data example given above looks like this:

```
<person xml:id="V2-60057">
 <sex value="2"/>
 <persName type="full" nymRef="#Nani1s">
  <forename>Νανίς</forename>
 </persName>
 <birth notAfter="-0350" notBefore="-0399">
  <placeName key="Athens" cert="?">Athens</placeName>
 </birth>
 <floruit>f.iv BC</floruit>
 <persName type="namevariant" xml:lang="grc">Να<seg type="orth">νείς</seg>
 </persName>
 <bibl>
  <title>IG</title> II<hi rend="sup">2</hi> 12229</bibl>
</person>
```

The issue of 'fuzzy' dates has been dealt with by storing the human-readable string provided by the editorial compiler (in this case *f.iv BC*, meaning 'first half of the fourth century BC') as the content of the <floruit> element, but mapping it onto absolute year

[8] Via a set of *ad hoc* Perl cleaning scripts, and XSLT transformations.

range using the notAfter and notBefore attributes of <birth>. This allows for sorting and date range searching of the records.

Names themselves are stored in a set of <nym> records, providing for the name as a first class object distinct from the <person>. For example, the name in our example above looks like this:

```
<listNym>
 <nym xml:id="nNani1s">
  <form xml:lang="el-grc">Νανίς</form>
  <form xml:lang="el-grc-x-noaccents">Νανις</form>
  <form xml:lang="el-grc-x-lgpn">Nani1s</form>
  <form xml:lang="el-grc-x-lgpnnoaccents">Nanis</form>
  <form xml:lang="el-grc-x-perseus">*nani/s</form>
 </nym>
</listNym>
```

Variants of the main name in different encodings, with and without accents, are stored as well in order to make implementation of searching easier.

One of the more complicated examples shown earlier (lexical data categories Θεόφραστος) in short form exposes some of the less ideal uses of the TEI markup:

```
<person n="2-23" xml:id="V2-33229">
 <sex value="1"/>
 <persName type="main" nymRef="#nQeo1frastos">Θεόφραστος</persName>
 <birth notAfter="-0050" notBefore="-0099">
  <placeName key="LGPN_20500">Hagnous</placeName>
 </birth>
 <floruit>f.i BC</floruit>
 <state key="#relationship">
  <label>s. <persName type="relationship" xml:lang="el-
grc" nymRef="#nQemistoklh3s">Θεμιστοκλῆς I</persName>
  </label>
 </state>
 <state key="#relationship">
  <label>
   <persName type="relationship" xml:lang="el-
grc" nymRef="#nAke1stion">Ἀκέστιον</persName>
  </label>
 </state>
 <state key="#relationship">
  <label>f. <persName type="relationship" xml:lang="el-
grc" nymRef="#nQemistoklh3s">Θεμιστοκλῆς II</persName>
  </label>
 </state>
 <state key="#relationship">
  <label>
   <persName type="relationship" xml:lang="el-
grc" nymRef="#nSofoklh3s">Σοφοκλῆς</persName>
  </label>
 </state>
 <bibl>Paus. i 37. 1</bibl>
 <bibl>Plu., <title>Mor.</title> 843c == <title>PA</title> 7169</bibl>
 <bibl>
  <title>IG</title> II<hi rend="sup">2</hi> 1961, 19</bibl>
 <bibl>3510</bibl>
 <bibl>= <title>IEleusis</title> 301, 8 f., 14</bibl>
 <bibl>
  <title>IG</title> II<hi rend="sup">2</hi> 3511</bibl>
 <bibl>= <title>IEleusis</title> 302?</bibl>
 <bibl>300, 32, 39, 45</bibl>
 <bibl>Thompson, <title>New Style Coinage</title> 1230 &amp; <title>Chiron</title> 21 (1991)
pp. 13 f.</bibl>
</person>
```

It will be clear to the experienced TEI user that the <bibl> records could usefully be properly structured, and that the use of <state> to model relationship claims is not as good as a proper <relation> element would be. Unfortunately, the assertion in the *Lexicon* that ‘Θεόφραστος is the father of Θεμιστοκλῆς II’ does not permit us to automatically identify which record ‘Θεμιστοκλῆς II’ applies to (although this sort of record works well in print where a human can extrapolate).

Places are modelled using the TEI <place> element, pointed to by the *key* attribute on <placeName>. This allows us to maintain a single hierarchical <listPlace> containing all place names used by the *Lexicon*. Thus:

```
<listPlace>
 <place type="region" xml:id="LGPN_33014">
  <placeName>
   <region>Achaia</region>
  </placeName>
  <place type="settlement" xml:id="LGPN_33915">
   <placeName>
    <settlement>Aiga</settlement>
   </placeName>
  </place>
  <place type="settlement" xml:id="LGPN_33003">
   <placeName>
    <settlement>Aigeira</settlement>
   </placeName>
  </place>
  <place type="settlement" xml:id="LGPN_33917">
   <placeName>
    <settlement>Aigeira (Hyperesia)</settlement>
   </placeName>
  </place>
  <place type="settlement" xml:id="LGPN_33004">
   <placeName>
    <settlement>Aigion</settlement>
   </placeName>
  </place>
 </place>
</listPlace>
```

The *Lexicon* has defined five levels of data interchange as a result of the XML work:

1. Character interchange: ASCII text version of data separately from the binary format used by any database system. This was the minimal form of interchange supported in the initial stages of the project.

2. Character encoding: The *Lexicon* defined its own transliteration for Greek, independently of, for example, *TLG* betacode, and continues to use it for internal purposes. It is a happy accident that the characters used in the transliteration allow the names to be used in human-readable URLs.

3. Standardized structural markup: Data relationships follow the schema defined in 1983. The hierarchical and network structure used in project databases are maintained in the XML records.

4. Standardized semantic markup: The XML representation of the *Lexicon* is aligned with the vocabulary and semantics for XML elements of the Text Encoding Initiative. The TEI elements are themselves in the process of alignment with the CIDOC CRM,[9] allowing even wider understanding and a serious ontology within this field.

5. Information linking: For most categories of data (name, sex, data, bibliography), the *Lexicon* can be fully linked to comparable data. Places information is more complex and will be addressed later in this chapter.

One of the important additions to the new *LGPN* XML representation is the exposure of an ID for each record, to allow the project to offer a permanent URL. The IDs are of the form *volume number-person number*, for example V2-1030. Another form of identifier available is the sequence number for each name, as shown in the published volumes, for example 'Ἀρχίτιμος 10'. For users of the books, this is the only way they can refer to an entry, but such usage raises the very considerable problem of providing updates and additions for published material. The *Lexicon* has not yet resolved this issue.

Dealing with place names

There are approximately 3000 place names referred to in the *Lexicon*, in data which was collected long before it occurred to anyone to plot name occurrences on maps. The places are managed in a three-level geographical/political hierarchy (region, settlement, and *deme*) with occasional granularity down to the quasi-geographical tribe. In order to provide the map display described in the next section, we need to establish, at a minimum, a latitude and longitude for each place. This leaves aside, for the present, the issue of what point to use for a large place like Athens (the geographic centre? the Parthenon?) and the problem of variable size of settlements over time. To locate all 3000 places in the *Lexicon* from scratch is a considerable task, bearing in mind that:

1. the common geo-gazetteers (*e.g.* GeoNames)[10] do not include tiny villages in northern Greece where the name given in the secondary literature from which the *Lexicon* derives may be an idiosyncratic transcription, and an older name;
2. the precision of the location may only be regional (*e.g.* Crete);
3. the recorded name may be ambiguous when checked against modern atlases.

[9] Described at the CIDOC CRM Home page: <http://www.cidoc-crm.org>; compare this with the TEI Ontologies SIG: <http://www.tei-c.org/Activities/SIG/Ontologies>.

[10] GeoNames geographical database: <http://www.geonames.org>.

Desirable though it would be to revisit all the sites with a GPS, the *Lexicon* method is, in practice, three-fold. Firstly, place names which are unambiguous, and do appear in the modern gazetteers (*e.g.* Athens), are matched with a latitude and longitude quickly. Secondly, places which can be located in the *Barrington atlas*[11] can be given an intermediate record of name, page number and grid reference, in the knowledge that we will be able to use the work of the Pleiades project,[12] which is gradually digitizing all the material from *Barrington*, to resolve an identifier like akraiphiai-55-e4 (this place is in fact http://pleiades.stoa.org/places/540617/).[13] Finally, the *Lexicon* is a partner in the larger CLAROS[14] project, which brings together a set of classical art resources, including a large set of common place names. As each partner geolocates places, the *Lexicon* can share the data.

A more complete place record may now be shown, enhanced with latitude and longitude, modern place names, and reference to the *Barrington atlas*.

```
<place type="settlement" xml:id="LGPN_11230">
 <placeName>
  <settlement>Tan Solluch</settlement>
 </placeName>
 <placeName type="modern">Daryanah</placeName>
 <location type="batlas">
  <label>tansoluch-38-b1</label>
 </location>
 <location cert="medium">
  <geo>20.3533 32.3744</geo>
 </location>
</place>
```

At the time of writing, only about half the *Lexicon* placenames have been fully geolocated and/or linked to other gazetteers.

The Lexicon *online*

With the conversion to XML available, it is now possible to deliver 250,000 published records in a single new interface. The service offers a fairly conventional form-based interface to allow users to search by any of the available data fields: name (in various transliterations), date, place, *floruit*, and status.[15] An initial form showing name (with on-screen keyboard for Greek) and date (Figure 4) can be expanded to cover the other fields (Figure 5). Names and places can also be picked from pre-built summary lists (Figures 6 and 7). The fields are simply additive; the results must satisfy all criteria at once, thus not allowing for disjoint queries such as 'names from Cyprus or Messenia', or 'names ending in *ιμος* from the fourth century AD or fourth century BC'.

[11] R. Talbert, ed., *Barrington atlas of the Greek and Roman world* (Princeton 2000).

[12] Pleiades, a gazetter of ancient places: <http://pleiades.stoa.org/>.

[13] Thanks to help from Tom Elliott, we were able to make a trial digitization of *c*.60 places ourselves, ahead of Pleiades schedule, which helped us refine the *Lexicon* workings.

[14] Claros: < http://www.clarosnet.org>.

[15] *LGPN* search interface: <http://www.lgpn.ox.ac.uk/database/lgpn.php>.

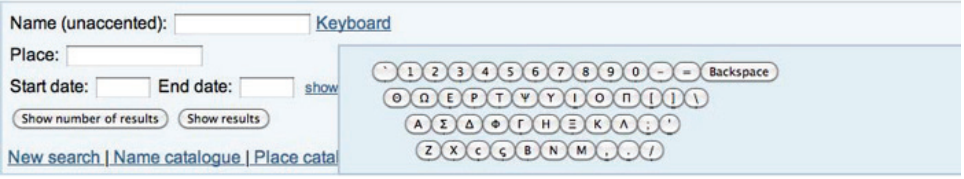
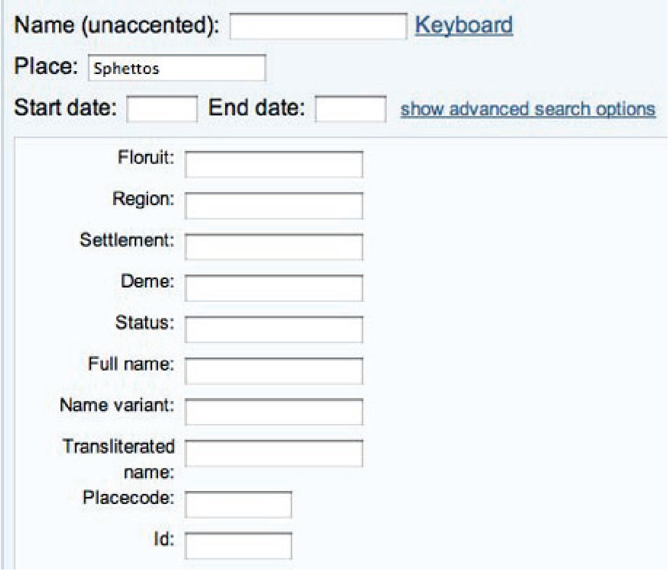
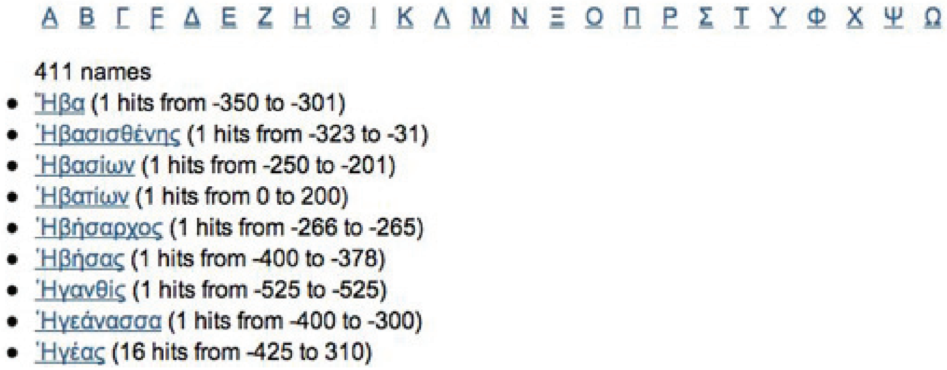


Figure 4: LGPN online simple search form



Place: Sphettos

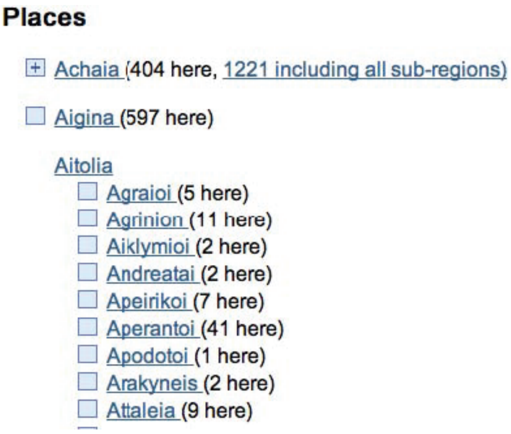
Figure 5: LGPN online advanced search form



411 names

- Ἡβα (1 hits from -350 to -301)
- Ἡβασισθένης (1 hits from -323 to -31)
- Ἡβασίων (1 hits from -250 to -201)
- Ἡβατίων (1 hits from 0 to 200)
- Ἡβήσαρχος (1 hits from -266 to -265)
- Ἡβήσας (1 hits from -400 to -378)
- Ἡγανθίς (1 hits from -525 to -525)
- Ἡγεάνασσα (1 hits from -400 to -300)
- Ἡγέας (16 hits from -425 to 310)

Figure 6: LGPN online, pick lists for names



Places

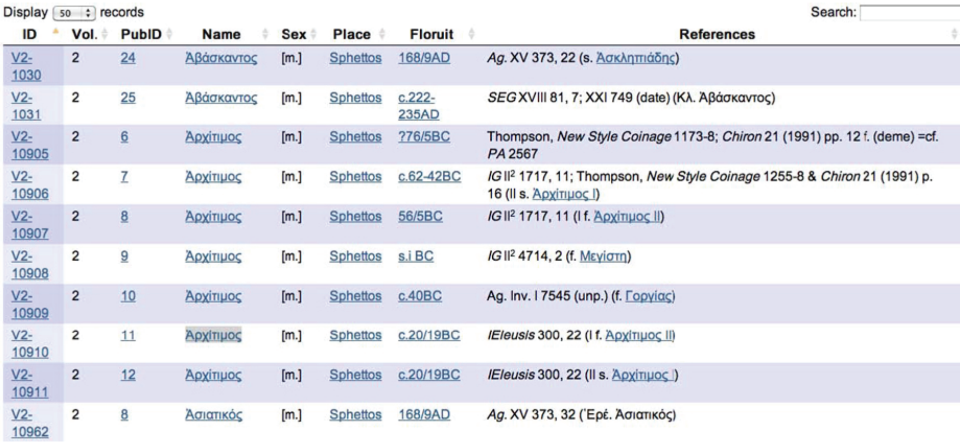
⊞ Achaia (404 here, 1221 including all sub-regions)

□ Aigina (597 here)

Aitolia
- □ Agraioi (5 here)
- □ Agrinion (11 here)
- □ Aiklymioi (2 here)
- □ Andreatai (2 here)
- □ Apeirikoi (7 here)
- □ Aperantoi (41 here)
- □ Apodotoi (1 here)
- □ Arakyneis (2 here)
- □ Attaleia (9 here)

Figure 7: LGPN online, pick lists for places

The results can be returned in a variety of ways. For normal browsing, the default is to return simply the number of hits, in order to avoid unnecessary data transfer. Alternatively, a tabular display is provided, with sortable columns and narrowing *via* a search box (Figure 7). Results are batched, but queries which would result in more than 10,000 records to display are not permitted.[16] For those interested in the geographical spread of results, the *Map* tab utilizes Google Maps to offer simple point display (Figure 8). It should be noted, however, that not all of the places known to the *Lexicon* have been geolocated, as discussed in the previous section.



Display 50 records Search:

ID	Vol.	PubID	Name	Sex	Place	Floruit	References
V2-1030	2	24	Ἀβάσκαντος	[m.]	Sphettos	168/9AD	*Ag.* XV 373, 22 (s. Ἀσκληπιάδης)
V2-1031	2	25	Ἀβάσκαντος	[m.]	Sphettos	c.222-235AD	*SEG* XVIII 81, 7; XXI 749 (date) (Κλ. Ἀβάσκαντος)
V2-10905	2	6	Ἀρχίτιμος	[m.]	Sphettos	?76/5BC	Thompson, *New Style Coinage* 1173-8; *Chiron* 21 (1991) pp. 12 f. (deme) =cf. *PA* 2567
V2-10906	2	7	Ἀρχίτιμος	[m.]	Sphettos	c.62-42BC	*IG* II^2 1717, 11; Thompson, *New Style Coinage* 1255-8 & *Chiron* 21 (1991) p. 16 (II s. Ἀρχίτιμος I)
V2-10907	2	8	Ἀρχίτιμος	[m.]	Sphettos	56/5BC	*IG* II^2 1717, 11 (I f. Ἀρχίτιμος II)
V2-10908	2	9	Ἀρχίτιμος	[m.]	Sphettos	s.i BC	*IG* II^2 4714, 2 (f. Μεγίστη)
V2-10909	2	10	Ἀρχίτιμος	[m.]	Sphettos	c.40BC	Ag. Inv. I 7545 (unp.) (f. Γοργίας)
V2-10910	2	11	Ἀρχίτιμος	[m.]	Sphettos	c.20/19BC	*IEleusis* 300, 22 (I f. Ἀρχίτιμος II)
V2-10911	2	12	Ἀρχίτιμος	[m.]	Sphettos	c.20/19BC	*IEleusis* 300, 22 (II s. Ἀρχίτιμος I)
V2-10962	2	8	Ἀσιατικός	[m.]	Sphettos	168/9AD	*Ag.* XV 373, 32 (Ἐρέ. Ἀσιατικός)

Figure 8: LGPN online result table

[16] Requests for data formats, rather than web pages for human consumption, are not limited by size.

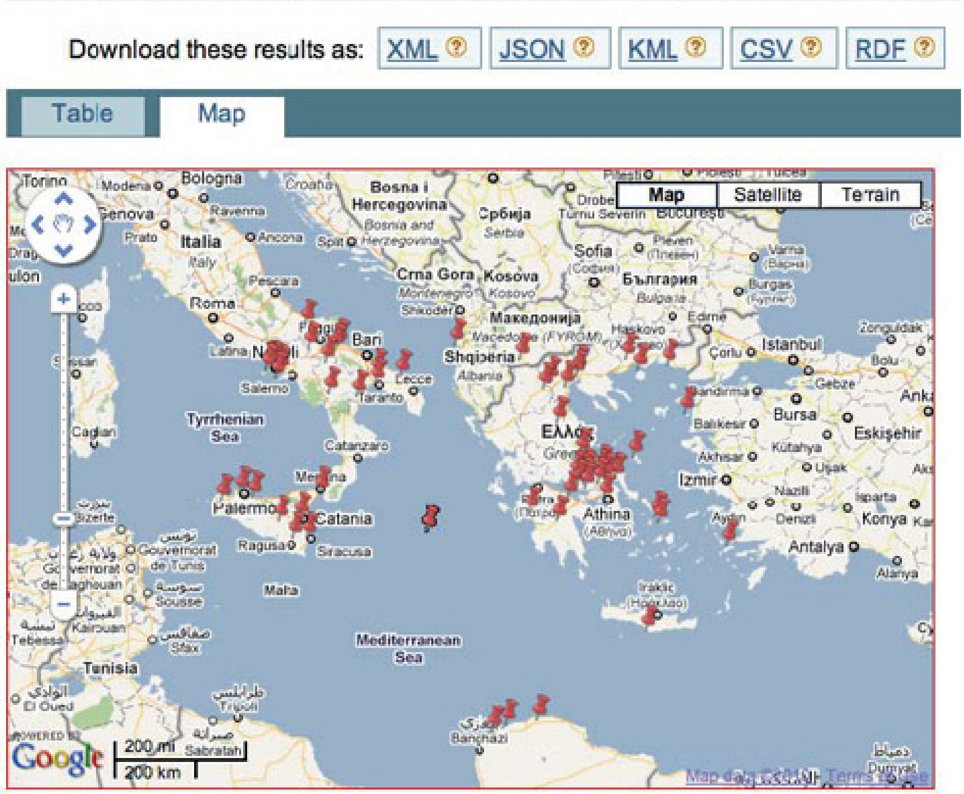


Figure 9: LGPN online result map display

The results of this web form-based searching can be exploited further using a variety of other data formats. The underlying TEI XML can be returned for those wishing to perform their own transformations, JSON data can be used for web-based visualization, the KML (Keyhole Markup Language) format of XML for Google Earth and Google Maps can be downloaded, and a CSV (comma-separated data fields) summary can be imported into a spreadsheet for data exploration.

The XML RDF format is a conversion of the TEI markup to the CIDOC CRM ontology for interchange and cross-searching with other data sets. Making *Lexicon* data available against this formal ontology, mapping *Lexicon* data categories onto well-defined concepts, allows it to be ingested immediately into semantic web databases, and start to participate in the universe of open-linked data. This is intended to be used in computer-to-computer interaction, using a standardized query language (SPARQL).[17]

The CLAROS project is an example of using the *Lexicon* RDF feed;[18] it provides an aggregating searchable cache across large classical art history databases, but also includes the *Lexicon* data. This can allow, for example, the formulation of queries which combine

[17] SPARQL is a W3C recommended query language for RDF.

[18] Claros: < http://www.clarosnet.org>.

questions about vase forms with names of people, linked by place name. Europeana is another example of a large cache based on aggregating RDF data against the CRM.[19]

The online search is implemented using a simple read-only relational database which contains chunks of TEI XML (*i.e.* a <person>), and a set of extracted index terms. This allows for efficient sorting and searching. The TEI XML fragments are taken from the database and converted to the appropriate output format (HTML, JSON, KML, *etc.*) using XSLT transformations. An initial implementation using an XML database (eXist)[20] was unable at that time to support very large result sets and complex additive queries with a sufficiently good response time. Many other systems could be used to provide the same service in future.

The web pages for use by classical scholars are only one way in which the data can be accessed. They are also available through *persistent and predictable URLs*. We try to follow here some of the modern guidance about providing 'cool URIs'.[21] *Lexicon* URLs take the form: http://www.lgpn.ox.ac.uk/*type*/*query*/*format*, where *type* is one of:

1. batlas: grid square, page and name in *Barrington atlas*;
2. date: date in the form *year*to*year;*
3. floruit: date range, using *Lexicon* conventions;
4. id: *Lexicon* person ID;
5. lexname: transliterated name (lexnamenoaccents: without accents);
6. n: a combination of *volume-name* in accented Greek-publication number;
7. name: name in UTF-8 Greek (namenoaccents: without accents);
8. place: place name from *Lexicon* authority list;[22]
9. placecode: *Lexicon* internal code for place ;
10. region: geographic region, using *Lexicon* names;
11. status: status from *Lexicon* list;[23]

…and *format* is one of:

1. csv: comma-separated values in table;
2. exhibitdata: JSON code suitable for consuming by Simile Exhibit;[24]
3. html: human-readable web page;
4. json: Javascript JSON data format;
5. kml: KML for display in Google Earth or Maps;
6. rdf: RDF XML;

[19] Europeana: <http://www.europeana.eu/portal/>.

[20] eXist, an Open Source database: <http://exist.sourceforge.net>.

[21] For more discussion of this topic see W3C, *Cool URIs don't change*: <http://www.w3.org/Provider/Style/URI>.

[22] *LGPN* place name authority list: <http://www.lgpn.ox.ac.uk/place>.

[23] *LGPN* status and occupations list: <http://www.lgpn.ox.ac.uk/status>.

[24] Simile: <http://simile.mit.edu>.

7. summaryjson: Javascript JSON showing summary detail of date range and count for name;
8. xml: TEI XML.

A selection of examples is given in Table 2.

Search by name, return HTML	/lexnamenoaccents/Paramonos
Search by name, return XML	/lexname/Para1monos/xml
Search by name in accented Greek, return KML	/name/Παράμονος/kml
Search by *Lexicon* ID, return TEI XML	/id/V1-2697/xml
Search by status, return JSON	/status/potter/json
Search by date, return RDF	/date/250to265/rdf
Search by place name, return comma-separated data	/place/Aloros/csv
Search for 1st Ἀργόφιλος in volume 3a	/n/3a-Ἀργόφιλος-1

Table 2: Lexicon persistent URL patterns

With *Lexicon* data available in a standardized way, what sort of services can be built on top of it? One example is a 'name decorator service'. If we have web pages showing Greek inscriptions, and the names are identified in some way, we can run over the page, look up each name in the *LGPN*, and enhance the page. Thus the HTML may have markup like this:[25]

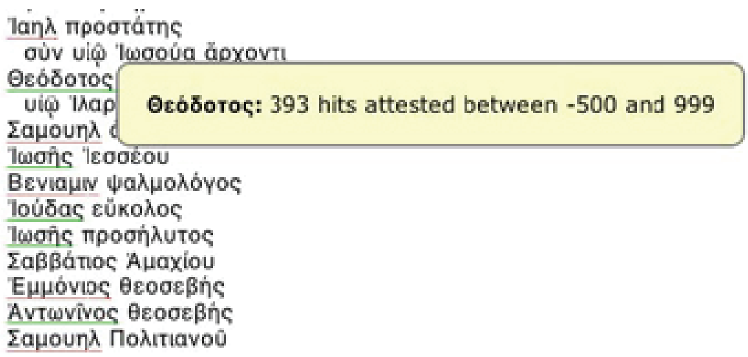
```
<div class="ab">...κὲ παντευλόγ [...]ων<br/>εἰς ἀπενθησίαν<br/>τῷ πλήθι ἔκτισαν<br/>ἐξ ἰδίων
μνῆμα<br/>
 <span class="lgpn">’Ιαηλ</span> προστάτης <br/>  σὺν υἱῷ
<abbr class="lgpn" title="’Ιωσούα">’Ιωσούᾳ</abbr>
 ἄρχοντι <br/>
 <span class="lgpn">Θεόδοτος</span>
 <abbr class="lgpn" title="Παλατῖνος">Παλατῖνος</abbr>σὺν<br/>υἱῷ
 ...</div>
<script type="text/javascript" src="greeknames.js"/>
```

in which names are identified using an HTML class attribute. The Javascript in greeknames.js will take care of the lookup by making a series of requests to, for example, http://www.lgpn.ox.ac.uk/name/Θεόδοτος/summaryjson which returns a record like this:

```
[{"query": "Θεόδοτος","id": "Qeodotos", "name": "Θεόδοτος",
 "notBefore": "-500", "notAfter": "999", "number": "393", "firstChar": "Θ"}]
```

containing the information about the date range and number of occurrences of the name. This can then be used to add a popup on the name, in which unknown names have a red underline and known names have a green underline. This complete example (facing page) is available at <http://clas-lgpn2.classics.ox.ac.uk/Demo/iAph110055.html>, which includes a link to the XML source.

[25] We are grateful to Gabriel Bodard at King's College London for this example from the Aphrodisias inscriptions. The HTML was created by transforming the project's TEI XML markup.



Ἰαηλ προστάτης
σὺν υἱῷ Ἰωσούα ἄρχοντι
Θεόδοτος
υἱῷ Ἱλαρ
Σαμουηλ
Ἰωσῆς Ἰεσσέου
Βενιαμιν ψαλμολόγος
Ἰούδας εὔκολος
Ἰωσῆς προσήλυτος
Σαββάτιος Ἀμαχίου
Ἐμμόνιος θεοσεβής
Ἀντωνῖνος θεοσεβής
Σαμουηλ Πολιτιανοῦ

Θεόδοτος: 393 hits attested between -500 and 999

Conclusions

We believe that this story of the *Lexicon of Greek Personal Names* illustrates four points. Firstly, the conceptual data model of the 1970s has survived the test of time; it has gone through many completely unforeseen changes and challenges, but has required no serious rethinking. Secondly, the *Lexicon* experience shows that the modern web techniques of machine/machine interchange, and rich exploratory tools, can be retrofitted effectively to older projects. Thirdly, we believe that the extra abilities added to interoperability by the adoption of open standards and linked data are crucial to the future of research data like the *Lexicon*.

Finally, we hope that there are entirely new academic questions waiting to be answered by this version of the *Lexicon* data, as well as uses we have not yet imagined.

Acknowledgements

We are grateful to Tom Elliott at the Institute for the Study of the Ancient World, NYU, for assistance during the work that led to this chapter and for essential help with locating places and ideas about URLs; Janet McKnight, for Javascript programming and geolocating; Gabriel Bodard, for forcing us to complete this work by giving us a deadline for the seminar paper on which this chapter is based; and Donna Kurtz and the Beazley Archive at Oxford for stimulating ideas about interchange.

Elaine Matthews (late of *All Souls College, Oxford*)
Sebastian Rahtz (*University of Oxford*) sebastian.rahtz@it.ox.ac.uk

Bibliography

Catling, R. W. V., and F. Marchand and M. Sasanow, ed., *Onomatologos: studies in Greek personal names presented to Elaine Matthews* (Oxford 2010).

Fraser, P. M., *Greek ethnic terminology* (Oxford 2009).

Hornblower, S. and E. Matthews, ed., *Greek personal names: their value as evidence,* Proceedings of the British Academy 104 (Oxford 2000).

Matthews, E., ed., *Old and new worlds in Greek onomastics*, Proceedings of the British Academy 148 (Oxford 2007).

Matthews, E., and S. Rahtz, 'Designing and using a database of Greek personal names', in *Proceedings of the VIII International Symposium of the Association of Literary and Linguistic Computing,* (Nice 1985).

Talbert, R., ed., *Barrington atlas of the Greek and Roman world* (Princeton 2000).

TEI Consortium, ed., *TEI P5: guidelines for electronic text encoding and interchange*. (Charlottesville VA 2007): <http://www.tei-c.org/Guidelines/P5/>.

HUMSLIDES ON FLICKR: USING AN ONLINE COMMUNITY PLATFORM TO HOST AND ENHANCE AN IMAGE COLLECTION

SIMON MAHONY

Introduction

The teaching of Classics is heavily dependent on the examination of artefacts, with images having an important pedagogical impact – they 'enhance learning, by illustrating concepts and providing visual memory cues'.[1] Academics and departments often have large collections of 35mm slides which have been the traditional medium for teaching and research. Since this format, along with associated projection equipment, is no longer manufactured, educators and researchers have had to seek alternatives. In response to this the School of Humanities at King's College London set up a pilot project (*HumSlides*) using the holdings of the Classics and the Byzantine and Modern Greek Departments to create a digital image resource for online delivery available to teachers, students, and researchers. This chapter builds on what was learned from that pilot project and suggests a way forward. It also questions the broader implications for collaborative environments and user interaction within an image-based pedagogical framework.

Digital media have almost completely taken over from traditional slide-based delivery systems and offer new potential opportunities and benefits. Students no longer have to be content with images only being briefly displayed in the lecture hall and in fought-over, specialist library books which are beyond their budgets. Images are now available over the web and via institutional networks to be downloaded and saved into personal collections on laptops and other portable devices. High-quality digital images can be enlarged to examine detail not possible in a print publication and the growth of online resources enables and indeed encourages the incorporation of images in numbers simply not feasible in the print medium.[2] Many other disciplines such as palaeography, manuscript studies, and library studies use images to support teaching and research and so the model that develops should be of benefit in many other areas.

[1] See JISC Digital Media guide, 'Using images in learning, teaching and research materials' <http://www.jiscdigitalmedia.ac.uk/stillimages/advice/using-images-in-learning-teaching-and-research-materials/>.

[2] An online publication is not limited in the number of images it includes in the way that a print edition is. For an example and a discussion of this and the possibilities for widening access to online resources for Classics see: G. Bodard, 'The inscriptions of Aphrodisias as electronic publication', *Digital Medievalist* vol. 4 (2008): <http://www. digitalmedievalist.org/journal/4/bodard>.

With the advent of the web, digitization – particularly that of large image collections and libraries – was looked to as the way forward and a solution to problems of distribution, access, and archiving.[3] The JISC Image Digitization Initiative (JDI) was set up specifically to manage digitized images for use in teaching and learning, making it 'the first large-scale multi-site and multi-foci digital imaging project undertaken in the United Kingdom'.[4] However, an awareness of the potential new threats and problems posed by digital technology alongside the new opportunities it offered, particularly in the face of the rapid technical changes, had been recognized early.[5] Much discussion of digital preservation centred around the possibilities afforded by emulation and migration as alternative approaches.[6] The new possibilities opened up by digital media brought new problems such as cost, storage, changing technologies, and copyright, which all need to be addressed (these are discussed below in relation to the *HumSlides* case study).[7] Image hosting sites such as Flickr offer more affordable alternatives to costly server infrastructure and many libraries and archives have joined *The Commons* there.[8] In doing so they now also have the added advantage of possibilities to engage and build a community of users and to harness them to enhance the collection.[9]

The HumSlides *pilot project*

Funding was secured from KCL's College Teaching Fund to create a resource for teaching and learning, with the digitization of slides taking place over the summer of

[3] For example see: S. Lee, 'Scoping the future of the University of Oxford's digital library collections, final report', (Oxford 1999) [accessed 2nd September 2011]: <www.bodley.ox.ac.uk/scoping/final.doc>. This report notes the many image and document archives set up at Oxford: Beazley Archive, Bodleian Broadside Ballads Project, Celtic and Medieval Manuscripts, Centre for the Study of Ancient Documents, *et al*. See also M. Deegan and S. Tanner, *Digital futures: strategies for the information age* (London 2002) and S. Lee, *Digital imaging, a practical handbook* (London 2002).

[4] S. Ross, *Image digitisation management models: an assessment of the JIDI programme* (Glasgow 2000) [accessed 2nd September 2011]: <http://eprints.erpanet.org/96/>.

[5] For a full exposition on this and the need for safeguards: 'Preserving digital information: report of the task force on archiving of digital information', (OCLC Research 1996) [accessed 2nd September 2011]:
<http://www.oclc.org/content/dam/research/activities/digpresstudy/final-report.pdf>.

[6] For discussion contemporary to this project see for example B. Lavoie, *The incentives to preserve digital materials: roles, scenarios, and economic decision-making* (OCLC online 2003) [accessed 2nd October 2012]. Available at:
<https://www.oclc.org/resources/research/activities/digipres/incentives-dp.pdf>

[7] For example JISClegal has been set up to provide guidance as the complexities of copyright are often a deterrent to setting up a digitization project:
<http://www.jisclegal.ac.uk/LegalAreas/CopyrightIPR.aspx>.

[8] See the section: *The Commons* and n. 31 below.

[9] For more on non-professional contributions and their potential see: M. Terras, 'Digital curiosities: resource creation via amateur digitization', *Literary and Linguistic Computing* 25.4 (2010) 425-38.

2005. This involved the bulk-scanning of over 6,000 35mm slide transparencies using 4,000 DPI resolution and 24 bit RGB colour into archive quality uncompressed Tiff files; after archiving, these were further processed into working copies as JPGs and then into upload copies at a more manageable size (most between 1 and 1.5 MB).[10] Also necessary was the collection of data taken from the slides and their containers, along with spreadsheets completed by the contributors.[11] Images were typically accompanied by a caption, some description (depending on data sheets completed by the contributors), tags (keywords), and other metadata elements all of which were searchable.[12] The amount of descriptive data varied considerably but to be included images needed to have at least a caption. It is not possible to include screen shots of the *HumSlides* images (as they are password protected) but, as well as a high resolution digital image with zoom function, all contained the following metadata:

- Slide Number (unique ID for that slide);
- Caption (a simple title for the image);
- Location (physical geographical location of the subject of the image at the time that it was taken; additional data about findspot and provenance could be included in the 'description' field);
- Century (to allow broad searching; a more precise date could be added to the description field);
- Course related (code for the course(s) for which the slide would be used);
- Description (descriptive detail about the subject and content of the image);
- Keywords (main keywords to describe the content);
- Creator (person that took the slide photo, if known);
- Source Provenance (where the image came from, in as much detail as possible);
- File ID (unique ID for that image file, which allows a link between spreadsheet and image file)

The project was implemented using a proprietary software package (ContentDM) supplied freely for a limited time by OCLC PICA.[13] All fields were hyperlinked and so any word clicked would bring every other occurrence of that word in the collection.

One major problem with this slide collection was that of uncertain copyright and the consequent need to restrict access. The majority of slides in the departmental holdings

[10] For good discussion about appropriate file sizes and resolution see: M. Terras *Digital images for the information professional* (Farnham 2008) chapter 4.

[11] Enriching the data and funding details are covered below.

[12] The Dublin Core was chosen as a widely used and accessible metadata standard: <http://dublincore.org>.

[13] A cooperative union between the Online Computer Library Center and the PICA Foundation: <http://www.oclc.org/news/releases/200677.htm>.
For more on OCLC as it is now see: <http://www.oclc.org/uk/en/default.htm>.
ContentDM was at that time the OCLC content management system for libraries and collections online: <http://www.oclcpica.org/?id=1101&ln=uk>.

were legacy material with no clear provenance, as well as many taken by the College Audio-Visual service from popular teaching books. This raised the issue of user rights and the need to restrict access, which was not possible with this management system other than by limiting users by IP address and restricting access to the College network.[14] The insistence on having the collection password protected, with a staff or student login required for access, was ultimately one of risk management rather than any legal requirement as, once images are online, there is nothing except for the copyright notice to prevent any user downloading and distributing these files as they wish.[15] Arguably this restriction greatly reduced the uptake and usefulness of this new digital image collection.

The difficulty here, as is often the case, is in attempting to apply copyright management retrospectively. In the transition from an analogue to a digital medium, what is acceptable as 'fair use for academic and educational purposes' for a 35mm slide transparency or the photocopied page of a teaching book is not acceptable in the new electronic medium. The potential for the use of these images changes when they are converted from analogue to digital and, unlike the 35mm slide, once they are distributed all control over their future use is lost. Also notable was the dual attitude of some of the slide contributors where they wished to see copyright-covered material made freely available to them and yet were unwilling to include material for which they themselves held the copyright.[16]

User study

To evaluate the usefulness of the resource, academic staff and students were sent a questionnaire after the project had been live for a complete academic year. Four lecturers replied and each was followed up with a semi-structured interview. No response from any student was received.[17] The next task is to consider these user studies from the pilot project: what we can learn from those and how might those findings be fed into a new resource? Another pertinent point to consider here is the distinction between the different needs of the various user groups: the teachers and the students/researchers.[18] The tutor would be looking for suitable images to support their pedagogical aims and may or may not have a specific image or subject of an image in mind. The student may be looking for suitable ones to include as evidence to support their essays, and the researcher for

[14] The decision to restrict access was necessary from the College's perspective to avoid any possibilities of copyright infringement. Although a selection process was built into the workflow to reject images that were clearly from published sources, it was possible that some images with uncertain provenance would slip through.

[15] For more on this see L. Hughes, *Digitizing collections: strategic issues for the information manager* (London 2004) chapter 2.

[16] A future project might insist on applying Creative Commons licences; see: <http://creativecommons.org/>.

[17] Note also that not all lectures are supported by images, particularly those that are language and literature based.

[18] These were the intended users, although opening up the collection would then make it more widely available and include schools and the public.

something new and perhaps unexpected. From the information collected it was possible to make several observations based on the user responses and these are broken down in the following sub-headings:

Metadata:[19] finding an image was often easy enough but, once found, they were often not suitable for use by the lecturers, as they lacked sufficient accompanying data. This raises additional issues around what might be termed 'user seeking behaviour' with a focus on the browsing habits of academics.[20] For the academic supplying the image, who already knows the importance of the subject and content, the return for the time spent filling out the datasheets which generate the descriptive data is minimal. The accompanying data that was available made things easier to find but accentuated the difficulties when insufficient. More and richer metadata was needed to make this resource more useful. In fact, its usefulness as a resource is wholly dependent on the metadata.[21] What was unclear from the respondents was the type of additional metadata that they thought would be needed to improve the resource and they clearly focussed on discovery metadata, although data for long-term preservation of the images themselves is equally important. It would seem that from their perspective, the most important addition would be more descriptive metadata; that would make appropriate images easier to find by using selective keyword searching, and, once found, users would have a better understanding of the context and importance of the subject of the image.

Usage: the pedagogy employed by lecturers (particularly in the study of material culture) required a wide mix of image materials and many of their own were often sourced from books. *HumSlides* provided a good source of high quality images (unlike most that they found on the web)[22] but lecturers needed teaching sets that covered their whole module and so needed to go beyond this collection.

Pedagogy: to be of use the images needed to be integrated into the module teaching materials and cover the whole module. In addition, students needed to be shown why they would need to use this resource. This might best be achieved by embedding the use of the image resource into their coursework. A workable solution would be to set formative tasks

[19] The term 'metadata' here refers to the information that accompanies each image such as, caption, description of the image, keywords, provenance, *etc*. See the Dublin Core for an example of a widely used metadata standard: <http://dublincore.org/>.

[20] See for example: D. Nicholas, P. Williams, I. Rowlands, and H. Jamali, 'Researchers' e-journal use and information seeking behaviour', *Journal of Information Science* vol. 36 no. 4 (London 2010) 494-516. A variety of case studies looking at the way in which humanities scholars search and make use of both analogue and digital resources are part of the research conducted by the Humanities Information Practices at the Oxford Internet Institute, University of Oxford: <http://www.oii.ox.ac.uk/research/projects/?id=58>.

[21] For a full discussion of the importance of attaching metadata to images see M. Terras, *Digital images* (n. 10 above) chapter 7. See also Visual Arts Data Service (VADS), Creating Digital Resources for the Arts: Standards and Good Practice, 4.3: 'Resource discovery metadata and the Dublin Core': <http://vads.ahds.ac.uk/guides/creating_guide/sect43.html>.

[22] This refers to the images that the lectures could generally source on the web rather that what is available there. Examples of good quality online images are below.

requiring the students to investigate chosen images and enrich the descriptive metadata themselves.

Accessibility:[23] as already noted, much of the collection consisted of inherited slides with the provenance long-since forgotten and so, because of potential copyright issues, access to *HumSlides* was restricted to the KCL IP address and thus only available onsite via the King's network. Users were unable to use images from this resource off-site without the images being downloaded in advance. This was particularly problematic for intercollegiate graduate modules which were often taught at the Institute of Classical Studies.[24]

Further conclusions can be drawn from observation and feedback rather than the survey. The content of the image collection is heavily influenced by staff as they hold the slides, know which ones there are, and will be biased towards their own requirements. It was notable that the coverage was limited with, surprisingly, history and topography being better covered than archaeology, despite the latter's focus on materiality.

It cannot be emphasized enough that it is the metadata that makes an image collection useful (and will ensure its long-term survival) as this data allows the user to search not only for what they know exists but with serendipity for that welcome and unexpected result that they were not expecting.[25] This only happens if the metadata is extensive and sufficiently rich such that it describes all the content of the image and not just the main focus and interest of the academic that submitted the slide. For example: an epigrapher would be interested in the inscription but may, in their description and keywords, ignore any detail of the object inscribed which might be of immense interest to other researchers such as architectural historians, archaeologists, and indeed, if the epigraph was a literary text, then also literary scholars. What needs to be considered is the possibility not only of categorizing the images and their content but also of categorizing the possible types of interest in that image.

With regard to copyright issues, there was a clear difference in response from staff who had submitted sets of slides and those who had not. Those that had submitted teaching sets (generally the more established staff members who had a better knowledge of the departmental collections) were aware of the images that had been excluded (for copyright concerns or for lack of accompanying metadata) and focussed on those. Newer members of staff did not have the same concerns and commented on this collection being a good source of quality images. They focussed on what was available rather than on what had been excluded (presumably they were not aware of the excluded slides).

As for the use of the collection in teaching: students are strategic learners and for them to use this type of material it needs somehow to be incorporated into teaching. Lecturers

[23] The term *accessibility* is used here to mean the barriers that prevent the images from being viewed and downloaded rather than issues from the point of view of disability and W3C compliance (for basic W3C guidelines see: <http://www.w3.org/TR/WCAG10/>) which relate to the image management and delivery system and are beyond the scope of this chapter.

[24] For example the many modules taught between King's College London, UCL, Royal Holloway, and the Institute of Archaeology.

[25] The image-management software allowed searching by any word in any of the text fields. Hence searching for 'Augustus' would return any image whose metadata included the word 'Augustus' as a keyword, in the caption, or description of the image.

were using a variety of materials: as well as images from *HumSlides* they had their own digital images, some taken from the web and also some which they had scanned or photographed from books. Staff saw the benefits of an online digital image collection as they all acknowledged the difficulties with slides. The projection equipment was problematic and they often could not get hold of the images they wanted, particularly if they were locked away in the office of a colleague. This project made images more accessible and reliable and hence encouraged their use in teaching. However, rather than changing the way in which images are used, this project changed the way in which users would like to employ images by making them more aware of the possibilities.

The question now is how we might learn from this experience and create a more useful resource. First, let us look at what else is out there to see what other pertinent initiatives there are to draw on.

Some other online image initiatives[26]

OxCLIC[27] is a HEFCE-funded,[28] image-management project at Oxford, using images from the department of the History of Art, and the faculties of Classics, Archaeology, and Oriental Studies, to store images and metadata in a distributed system using Open Source software. One of the stated aims was to set out to 'provide guidelines on how image material held by individual academics and material held in departmental collections might be combined and made available in a web-based environment'.[29]

OxCLIC uses MDID (Madison Digital Image Database).[30] The project organizers highlight the need for clear metadata conventions across the collections, particularly when the source material is held in different departments. In practice, the terms used to describe objects and their attributes varies across disciplines (and sometimes even within the same one) making some type of standardization necessary. These ideas and the use of taxonomy or controlled vocabulary are familiar to librarians, cataloguers, and anyone having to apply a structure to their data and consistently define their terms. It was found essential to develop a standardized set of useful contextual data for the digital image.

[26] This list is not intended to be definitive and there are other popular online image resources such as *ARTStor* <http://www.artstor.org/index.shtml> which are not included here as they are not directly relevant.

[27] For more on *OxCLIC* see their Public Wiki at: <http://wiki.oucs.ox.ac.uk/ltg-public/OxCLIC>.

[28] Higher Education Funding Council for England: <http://www.hefce.ac.uk/>.

[29] *OxCLIC* Annexe B Summary Report (10/05/07).

[30] MDID is an Open Source digital image database (available at SourceForge: <http://sourceforge.net/projects/mdid/>) developed at the James Madison University to host image collections primarily used for teaching and the study of art and history: <http://www.lib.jmu.edu/resources/more.aspx?id=1560>.
More details of MDID are on the JISC pages at:
<http://www.jiscdigitalmedia.ac.uk/stillimages/advice/image-management-madison-digital-image-database-mdid/>.

The Commons,[31] ('[y]our opportunity to contribute to describing the world's public photo collections') was a pilot project set up on Flickr in partnership with The Library of Congress (LoC) and launched in January 2008 with two main stated aims: to increase exposure to current content held in public institutions worldwide and to facilitate the collection of general knowledge about these collections. What is particularly interesting here is the extent to which this collaborative project has grown. Starting with the LoC and being quickly joined by The Powerhouse Museum (Sydney), the partner institutions grew in number almost monthly, such that a talk I gave on online image collections in the summer of 2008 listed eight and, at the time of writing, they now number sixty-four, including the National Archives UK, the National Maritime Museum, the Imperial War Museum Collections, and the National Library of Scotland.[32]

Another significant point is that for *all* these image collections, visitors are 'invited to help describe the photographs [they] discover […] either by adding comments or leaving tags'.[33] The user community is being encouraged to enrich the resource by adding additional data in the form of comments and tags as well as annotating the images themselves by adding 'notes' that can be attached to specific areas within the image. Further, all these images may be freely viewed and downloaded in a variety of resolutions up to and including the original. Images can be arranged and organized in 'favourites' or 'galleries' (personal sub-sets of images) and displayed individually or by using the built-in slideshows.

Any internet user is able to search, view, and download images in *The Commons* and to access the tags, description, and annotations. To be able to add data to the images and arrange personal sets of 'favourites', users need to be logged in to their Flickr account and thus identified by their Yahoo! details.[34] User accounts are either 'free', which allow users to up-load 100MB of images per month into a single collection, or 'Pro' ($24.95 per year at the time of writing), which allow unlimited image upload, storage for archiving, multiple collections, as well as aggregated statistics of page views, referrers, and user activity on the account.[35]

The Commons has a general umbrella 'Rights Statement' outlining copyright issues and how the images may be used.[36] As well as this there is a link to the individual 'Rights Statement' for each of the collections of the partner institutions. For example, the one for the LoC takes you to their own webpage for 'Copyright and Other Restrictions' for their

[31] *The Commons* <http://www.flickr.com/commons/>.

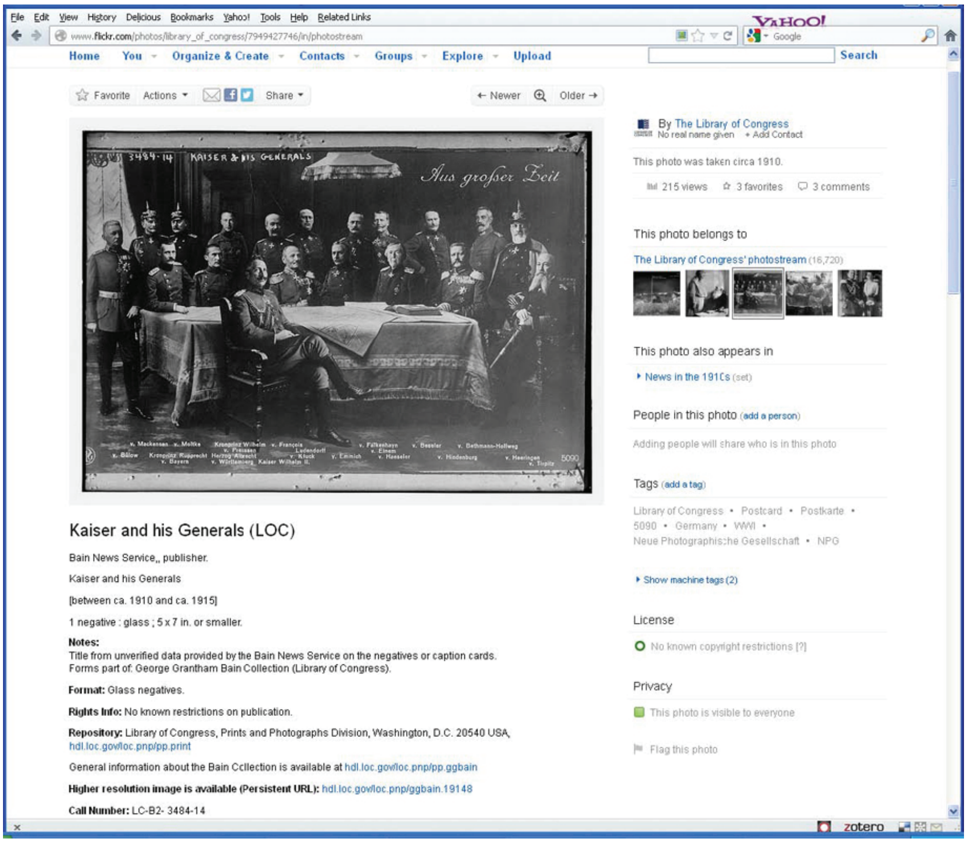
[32] For a full up-to-date list see 'Participating Institutions': <http://www.flickr.com/commons/institutions/>.

[33] 'Welcome to *The Commons*': <http://www.flickr.com/commons/>.
'Comments' may be added to a discussion box beneath the image similar to those used in a blog. 'Tags' are a collection of 'keywords' listed adjacent to the image and when clicked will return all other images that share the same tag. Random testing shows that when logged into a Flickr account it is possible to add tags and comments freely, although one might expect some type of moderation (after they have been added) to avoid abuse.

[34] Flickr is owned by Yahoo! and so this identifies users.

[35] Flickr account details are described at: <http://www.flickr.com/help/limits/>.

[36] Details of *The Commons*' 'Right Statement' and links to individual statements for each participating institution are at: <http://www.flickr.com/commons/usage/>.


File Edit View History Delicious Bookmarks Yahoo! Tools Help Related Links
www.flickr.com/photos/library_of_congress/7949427746/in/photostream
YAHOO!
Google
Home
You
Organize & Create
Contacts
Groups
Explore
Upload
Search
Favorite
Actions
Share
← Newer
Older →
3484-14 KAISER & HIS GENERALS
Aus großer Zeit
5090
zotero

Kaiser and his Generals (LOC)

Bain News Service,, publisher.

Kaiser and his Generals

[between ca. 1910 and ca. 1915]

1 negative : glass ; 5 x 7 in. or smaller.

Notes:
Title from unverified data provided by the Bain News Service on the negatives or caption cards.
Forms part of: George Grantham Bain Collection (Library of Congress).

Format: Glass negatives.

Rights Info: No known restrictions on publication.

Repository: Library of Congress, Prints and Photographs Division, Washington, D.C. 20540 USA,
hdl.loc.gov/loc.pnp/pp.print

General information about the Bain Ccllection is available at hdl.loc.gov/loc.pnp/pp.ggbain

Higher resolution image is available (Persistent URL): hdl.loc.gov/loc.pnp/ggbain.19148

Call Number: LC-B2- 3484-14

By The Library of Congress
No real name given + Add Contact

This photo was taken circa 1910.

215 views 3 favorites 3 comments

This photo belongs to

The Library of Congress' photostream (16,720)

This photo also appears in

News in the 191Cs (set)

People in this photo (add a person)

Adding people will share who is in this photo

Tags (add a tag)

Library of Congress • Postcard • Postkarte • 5090 • Germany • WWI • Neue Photographis:he Gesellschaft • NPG

Show machine tags (2)

License

○ No known copyright restrictions [?]

Privacy

🟩 This photo is visible to everyone

Flag this photo

Figure 1: an image from the LoC collection on The Commons showing the accompanying metadata below the image and the searchable tags and link to the licence statement to the right ('No known copyright restrictions' with links to both LC and Flickr statements). Image accessed 10th September 2012.
<http://www.flickr.com/photos/library_of_congress/7949427746/in/photostream>

Prints and Photographs Reading Room.[37] Further, within the collections, each individual image in Flickr has a 'Privacy' notice and is accompanied by a 'License' statement that again links to the institutions policy document.[38] The statement of rights and permissions for use as they apply to each collection are clearly stated for the user.

All these institutions are making use of their user communities to enrich these online collections by adding additional data in the form of tags, comments, and annotations. They also give users the right to download the full resolution image files and (in most cases) permission to re-use their material for non-commercial purposes providing it is appropriately attributed.

[37] Library of Congress Prints & Photographs Reading Room notice on copyright and restrictions: <http://www.loc.gov/rr/print/195_copr.html#noknown>.

[38] They can be restricted to 'friends' only as we will see later.

Issues raised by The Commons *on Flickr*

The intended audience for these collections is not the same. *HumSlides* and *OxCLIC* are aimed at students and academic staff and expect their collection to be used for teaching and research, whereas *The Commons* is aimed at the general public. However *The Commons* raises immediate questions about why these partner institutions are, firstly, allowing random users to interact with and add content to the data that accompanies their online image collection and, secondly, why they should be giving away their precious images freely. Judging by the rapid expansion of institutions joining, it would appear that there are definite benefits.

From the Library of Congress report on their Flickr pilot it seems that, although they were already a pioneer in digitizing their photo collection, there was little public awareness of that fact.[39] This initiative had significantly aided the discovery and use of their collections and subsequently raised the public profile of this already prestigious institution. They had increased the awareness of their image collections by reaching new audiences, which seems to have been their purpose. In addition, they had made use of and sourced the knowledge of the user community:

> Flickr members also have offered corrections and additions by identifying locations, events, individuals, and precise dates. This data is often supported by accompanying links to articles from the *New York Times* archive, Wikipedia, and subject-specialized websites. After verification by Library staff, information provided by the Flickr community is incorporated into our catalog records.[40]

Additional reasons given for the success of the project include the support they received from the online Flickr community. They asked for help, which clearly seems to have appealed to Flickr users, as did the fact that their image collection satisfied the desire for high-quality content without copyright restrictions.[41] Another important consideration is the additional weighting given to images in Flickr by the major search engines (Flickr is owned by Yahoo!), making them easier to find and at the same time helping to raise the online profile of the image supplier.

The report admits some initial reservations from sceptics, particularly in the area of the possibility of 'fake facts' and 'uncivil discourse', but notes that: '[i]ncreasing the ability to engage and connect with photos increases the sense of ownership and respect that people felt for these photos'.[42] As a result of the pilot they 'gained a deeper understanding of how users want to interact with […] the collection. The benefits appear to far outweigh the costs and risks.'[43] They had taken what had been a perceived risk and benefited through increased awareness and exposure of their collections as a result.

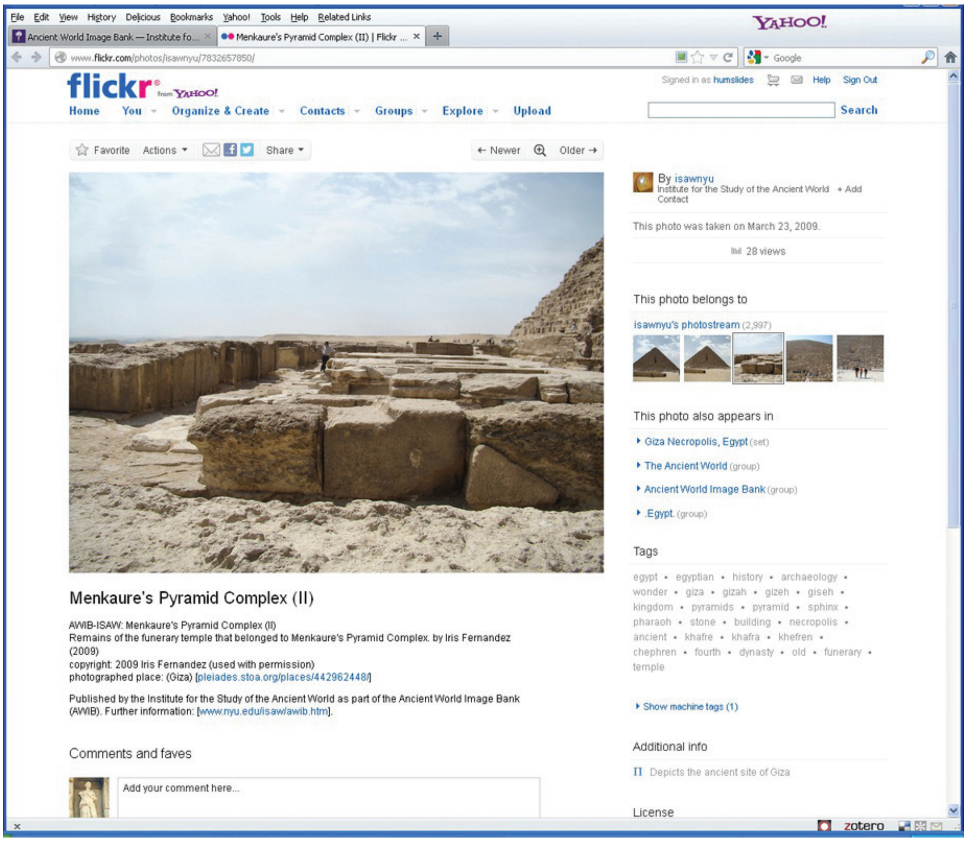
[39] Library of Congress blog on the release of the Flickr pilot report: <http://blogs.loc.gov/loc/2008/12/library-releases-report-on-flickr-pilot/>. The full report is at: <http://www.loc.gov/rr/print/flickr_report_final.pdf>, and the summary: <http://www.loc.gov/rr/print/flickr_report_final_summary.pdf>.

[40] 'For the Common Good: the Library of Congress Flickr Pilot Report' (summary), p.5: <http://www.loc.gov/rr/print/flickr_report_final_summary.pdf>.

[41] For extensive analysis of the tags and comments added by users see: 'For the Common Good: the Library of Congress Flickr Pilot Report' p.25: <http://www.loc.gov/rr/print/flickr_report_final.pdf>.

[42] 'For the Common Good' (n. 41 above) 25-26.

[43] 'For the Common Good' (n. 41 above) 36.



www.flickr.com/photos/isawnyu/7832657850/

By isawnyu
Institute for the Study of the Ancient World + Add Contact

This photo was taken on March 23, 2009.

28 views

This photo belongs to

isawnyu's photostream (2,997)

This photo also appears in

- Giza Necropolis, Egypt (set)
- The Ancient World (group)
- Ancient World Image Bank (group)
- .Egypt. (group)

Tags

egypt • egyptian • history • archaeology • wonder • giza • gizah • gizeh • giseh • kingdom • pyramids • pyramid • sphinx • pharaoh • stone • building • necropolis • ancient • khafre • khafra • khefren • chephren • fourth • dynasty • old • funerary • temple

Additional info

Π Depicts the ancient site of Giza

License

Menkaure's Pyramid Complex (II)

AWIB-ISAW: Menkaure's Pyramid Complex (II)
Remains of the funerary temple that belonged to Menkaure's Pyramid Complex. by Iris Fernandez (2009)
copyright: 2009 Iris Fernandez (used with permission)
photographed place: (Giza) [pleiades.stoa.org/places/442962448/]

Published by the Institute for the Study of the Ancient World as part of the Ancient World Image Bank (AWIB). Further information: [www.nyu.edu/isaw/awib.htm].

Comments and faves

Add your comment here...

Figure2: an image from the AWIB collection on Flickr with description, copyright (C.C. BY 2.0), and publication details below; and searchable tags to the right. Image accessed 10th September 2012. <http://www.flickr.com/photos/isawnyu/7832657850/>.

Ancient World Image Bank[44]

Another online image collection using Flickr is the *Ancient World Image Bank* (*AWIB*) at the Institute for the Study of the Ancient World (ISAW), New York University. Their opening statement describes this as:

> […] a collaborative effort to distribute and encourage the sharing of free digital imagery for the study of the ancient world. Beginning with the slide and digital photography collections of ISAW faculty, staff and affiliates, *AWIB* will expand to publish imagery donated by others as well. [45]

After testing in late 2009, this collection began to publish full image sets on Flickr in April 2010 under Creative Commons (CC) Attribution 2.0 Generic licence, which means that users are free to download, copy, share, and adapt the work, providing they attribute

[44] *Ancient World Image Bank*: <http://www.nyu.edu/isaw/awib.htm>.

[45] *AWIB*: <http://www.nyu.edu/isaw/awib.htm>.

the work in the manner specified.[46] Clicking an image on the *AWIB* webpage takes you straight to that image on their Flickr photostream.

All users can search, view, and download these images in a range of resolutions, including the original size, but to add comments you must be signed in to a user account (this is standard for Flickr images). However, unlike contributors to *The Commons*, *AWIB* has chosen not to allow Flickr account holders to be automatically able to add 'tags' to the existing list or 'notes' (annotations) to the images.

The *AWIB* occupies a different space in the sphere of publically available image collections in that it sits within a larger framework of online resources for ancient world studies. The images hosted there are currently from the collections of ISAW faculty, staff, and affiliates, although they do plan to extend this to publishing images that are donated, presumably by their user community, at a later stage. It is anticipated that it will soon participate in *Concordia*[47] which is an initiative to 'link up separately published online resources for ancient studies', as well as integrate with *Pleiades*[48] a digital gazetteer of historical geographic information and an electronic successor to the *Barrington atlas*. For this reason the options for user-generated content are more restricted, although it should be noted that the *Pleiades* project is developed by interaction with its community, using, creating, and sharing historical geographic data about the ancient world.[49]

Other relevant collections

An important initiative that draws on the user community to enrich its collection by tagging is the BBC's *YourPaintings* project[50] which was developed in collaboration with the Public Catalogues Foundation. This has gathered together and published online images of all publically owned paintings in the UK. As well as being given some information, after registering, viewers are invited to add tags relating to the content of the images, which then become searchable by keyword, greatly increasing their accessibility.

There are other popular online image resources such as *ARTStor*[51] which is a widely used subscription service now developing *Shared Shelf* image management system to allow users to upload and manage their own digital images. However, because of space limitations, further discussion here is restricted to those already mentioned above.

[46] For Creative Commons see: <http://creativecommons.org>.
For details of this specific CC licence see: <http://creativecommons.org/licences/by/2.0/deed.en>.

[47] *Concordia*: <http://concordia.atlantides.org>.

[48] *Pleiades*: <http://pleiades.stoa.org>.

[49] See 'The *Pleiades* Community':
<http://www.atlantides.org/trac/pleiades/wiki/PleiadesCommunity>.

[50] *YourPaintings*: <http://www.bbc.co.uk/arts/yourpaintings/>.

[51] *ARTStor*: <http://www.artstor.org/index.shtml>.

Cost implications: content management software versus *an online community platform*

Digitizing an image collection, whether that is 35mm slide transparencies, photographic prints, or whatever, has a cost element that can be calculated: equipment, staff hours, metadata collection, upload, *etc*.[52] The *HumSlides* pilot used the ContentDM management system which was supplied free of charge by OCLC for a limited time and the hosting was on existing departmental project servers.[53] *OxCLIC* made use of the Open Source package MDID which, although available freely, still has substantial costs involved in development and hosting. The hidden cost is that of the infrastructure necessary for hosting the resource and, regardless of whether content management software is proprietary or Open Source, there is still a considerable financial cost.

The HEFCE funding to *OxCLIC* in 2005 came to £56,270 and, although the largest expense was staff costs, approximately £16,000 was allocated for capital costs and software development, with an additional £12,000 for the Academic Computing Development Team (together making about half the total cost).[54]

The Library of Congress notes that their 'investments were relatively minor' and that 'no staff members were ever assigned to work full time on this project'.[55]

HumSlides *on Flickr*

When the licence for the ContentDM software expired in 2008, questions arose about how we could extend the life of this image collection, improve its usefulness, and take it forward as a resource. We had the digital images (more than 6,000), we had the metadata that needed to be enriched and we had the views of the existing user community; however, we had no funding. Using Flickr we had possibilities for a web-based interactive system for both delivery and collaborative annotation of these images. A Flickr Pro account was set up which allowed an unlimited number of images to be uploaded in an unlimited number of 'collections' and 'sets'. 'Collections' would equate to the academics who had originally submitted collections of their slide holdings for digitization and 'sets' would be arranged around individual 'teaching sets' as required.

The Pro account was an administrator account for uploading the images and data, and setting the permissions. A series of free accounts were also set up for submitters (academic accounts for those that had contributed slides for scanning) and for general

[52] *HumSlides* was originally funded in 2004 from the College Teaching Fund Competition (£24,900) as a project to digitize slides for teaching and learning.

[53] The server infrastructure required for any hosting service has a cost implication and should not be entered into lightly. *HumSlides* was hosted on existing servers used for funded projects at the Centre for Computing in the Humanities at King's College London.

[54] With thanks to Charles Crowther for details of the *OxCLIC* set-up costs.

[55] 'The Common Good' (n. 41 above) 9. They further identify 222 hours of technical programming over a six-month period (with a breakdown of details which includes testing) and 160 total staff hours on 'non-technical tasks'.

users.[56] These user accounts were necessary as the College was still unwilling, because of possible copyright issues, to allow the images to be open. Permissions were set to restrict access to 'friends' by adding the user accounts to the 'friends' list of the Pro account.[57] Thus, by limiting the login details (which would be changed routinely) of the user accounts for staff and students, access to the images and the ability to add additional information was likewise restricted to staff and students.

During this set-up time we were fortunate to have the help of three student interns. One researched the Flickr setup and support pages, while the other two, by desktop and book research, enriched the metadata (held in spreadsheets) by adding significantly to the keywords and descriptive information, where it existed, and supplying some where it was absent. This followed the findings of the user study and in addition allowed many more images to be included.[58]

The manual uploading of this quantity of images and accompanying data is clearly impractical and the LoC had surely automated the process. This is where a software developer is needed. Flickr has an open Application Programming Interface (API) which enables it to interact with other software by exposing its functionality and allowing it to be programmed against. We had to develop a bulk upload tool to associate the metadata with the appropriate images (by adding the image file name as a field in the data spreadsheet) for uploading. In addition, after creating a user interface, the tool needs to allow the Pro accountholder to authenticate the login through Flickr, locate the spreadsheet and associated images, and initiate the upload process.[59]

The *HumSlides* images in Flickr are displayed with the data (mapped to the Dublin Core) placed underneath and the keyword 'tags' listed alongside.[60] Permissions were set so that 'academic' account holders would be able to edit the descriptive data beneath each image and upload additional ones. Any user, once logged into a *HumSlides* user account, will (as in *The Commons* model) be able to add additional 'tags', comments', and annotations in the form of 'notes'.[61] This means that students as well as academics can add additional data. Participating academics are able to correct any errors found or add to the data by notifying the owner of the Pro account, who is automatically notified of activity on the *HumSlides* collection. These images are restricted to the user community. This expression is used purposefully as that is what we are doing here – building a user community, a community that will sustain and enrich this resource.

[56] Several 'user' accounts were set up as it was not known how the system would respond to multiple simultaneous logins. After testing we found that Flickr did allow several people to log in at the same time using one account.

[57] This is analogous to social software sites only allowing people from a fixed list of 'friends' to access certain information.

[58] With acknowledgement and thanks to Silvia Cinnella, Rebecca Collins, and particularly Greta Franzini.

[59] With many thanks to Payman Labbaf for the development of the bulk upload tool we used.

[60] This is a simplified set of data taken from the pilot to match the Flickr display: Slide Identifier, Caption, Description, Location, Century, Department, Submitter, Rights and Restrictions.

[61] Note that only after logging into a *HumSlides* user account are you able to access the *HumSlides* on Flickr images.

Allowing this interaction was intended to extend considerably the functionality of the original project. Users now had the ability to add their own thoughts and to comment on the content of the images, as well as adding additional data in the form of tags and annotations. We had the opportunity to build a user community that would be able to enrich this resource further. Lecturers could create their own teaching sets within *HumSlides* and students and researchers could gather together their own collections of 'favourite' images to support their studies. These images could now be embedded in teaching activities by setting students tasks surrounding specific images, or ones that they found for themselves.[62] In addition, the geographical features made available by Flickr open up additional possibilities by encouraging the creation of maps and bringing together images by location.

The essential user-research stressed as a requirement by JISC[63] had been drawn on, as had *OxCLIC* and *The Commons*. It is true that a login was still needed, but the collection was now available from any internet connection and not limited to the College network. The metadata had been considerably enriched and extended with more than 1,000 additional images now included. Why, then, had there been no reported activity and no corrections forwarded or further image sets uploaded?

The user statistics built into the *HumSlides* Pro account show that in twenty-four months after being set up only 261 of the images had ever been viewed (many would have been for testing and research for this chapter), none have been put into 'favourite' sets (with the exception of one used to test that feature) and none have been geotagged. No comments have been added, although eighteen images have been 'tagged' (again several of these would have been from testing the system). No corrections for any errors in the descriptive data (dates, location, content, *etc*.) have been sent to the author. Neither have any notifications of activity in the *HumSlides* collection been received at the Pro account. In short, two things are clear: firstly, that the images in the collection are not being viewed and certainly have not been incorporated into any task-based learning and, secondly, that the features that improve functionality and access through using Flickr as a platform are not sufficient to encourage its use.

If the online image collection, which is based on teaching sets from an academic department, is not being used, then how do we account for this? The answers are complex. One of the major issues raised in the pilot user study was that the collection was not available from anywhere other than the institution's network, whether password protected or not.[64] The second phase of the project hosted on Flickr *is* available from anywhere with an internet connection by using a generic user login, but it is still not *open*. The stumbling block is the large number of digital images from scanning legacy-collections of slides where the provenance is long forgotten. Attempting to apply copyright retrospectively is simply not possible and, to be open, the collection must start with source material for which copyright is held, or at least known, and with permissions secured. With this in

[62] For example, students could be asked to identify specific iconography used in mosaics or the attributes associated with heroes or gods.

[63] JISC, 'Using images in learning' (n. 1 above).

[64] Note that many areas of King's webpages are not publicly viewable but require an institutional login when off-site.

place there should be no need for the password requirement and, with the agreement of the copyright holders, the images can be freely distributed.[65] This collection (in suitable copyright-free form) would benefit greatly from being openly available on the web without the need for any login. In addition, there would seem to be a general lack of awareness of the image collection and it needs to become part of the routine toolkit of the teaching staff, acknowledged as such, and promoted to the students. Students, if not staff, are very familiar with so-called Web 2.0 applications and social media in general and there are many pedagogical possibilities, such as setting students tasks as already mentioned.[66] The students as well as teaching staff need to be encouraged to use these collections for their projects and research, and any barriers to their use need to be lifted.

HumSlides *2.0 on Flickr: a possible next phase*

The current *HumSlides* iteration is not being used, sitting as it does behind a password, and with no apparent promotion to the potential user community. It currently holds more than 4,000 slides digitized into high resolution image files stored on an online and low-cost community platform. Of those original slides, many were taken by the academic contributors on field and research trips, and are clearly identifiable as such from the content of the images, as well as from the original data taken at the time of collection. To ensure that all contributors received their slides back correctly, an inventory was kept of the boxes, trays, and sheets that the slides came in, along with a record of anything and everything written on the containers and the slides themselves.[67] For example, a box labelled ‘Rome 1992’ containing slide images of monuments in Rome would presumably be images taken on a trip to Rome by that contributor in 1992. It is generally clear to see from the content of the images and their arrangements in sets which ones had been taken by the academics themselves. Guidelines were given to contributors asking for slides for which either they or King’s held the copyright and prioritizing those used for teaching in the coming academic year. However, in practice, with the numbers involved it was often

[65] This is of course not only the case with digital images but with any copyright-protected material. A comparative example would be in re-purposing teaching materials from a class-based module to an e-learning programme distributed via a VLE (Virtual Learning Environment) such as Blackboard or Moodle. If the e-learning programme is conceived and constructed as such from the start then all copyright can be cleared and, where necessary, paid for in advance. Again the problems arise when attempting to apply copyright retrospectively. It is simply not possible and is a clear barrier to re-purposing teaching material for online delivery.

[66] By the term ‘social media’ I mean online communication platforms which allow the user to create and exchange content with the platform and other users. Flickr is an example of such a platform.

[67] Trust was an important issue here, persuading the holders of the slide collections to part with their precious and guarded material. Guidelines for collection, handling, and return were in the original documentation which also named this author, who had been both an undergraduate and research student in that department and so known to them (and presumably considered trustworthy). An inventory was kept of each slide container (each with a unique identifying code attached), when it was collected, from whom, and when returned.

not practicable to make a selection prior to scanning, but rather to exclude images at the QA (quality assurance) stage post-processing.[68]

The next phase then might be to solicit agreement from the contributors who submitted images that were clearly taken by them and so have no copyright implications (other than their consent) to have their images made publically available on Flickr.[69] There may also be the possibility of enrichment by machine tagging with geolocational data imported from *Pleiades* and perhaps also *Pelagios*.[70] This would then form the core on which we could build an open and freely available image collection, to which we would invite other institutions and individuals to contribute. This would be contingent on their agreement to a policy of open access under an appropriate Creative Commons licence and being prepared to supply the necessary metadata in an agreed format.

Further, adopting *The Commons* model pioneered by the LoC, users could be encouraged to enrich the data by adding tags, comments, and annotations to the images.[71] We then employ the user community in an activity that has become known as 'crowdsourcing'.[72] The LoC admitted (understandably) to having some initial reservations about this and,[73] in the introduction to their final report, asks: 'Could the Library tap the knowledge and energy of the user community to augment its own efforts?', and: 'What's the quality of the information gained through crowdsourcing?'[74] The answer must surely lie in the fact that, after the completion of the pilot project, the LoC collection on Flickr has grown considerably, as have the number of institutions that now participate in *The Commons*. To quote from the 'Recommendations and Conclusions' of their report:

> Ten months into the pilot, the question looms whether to move from pilot project to program. Performance measures documented in this report illustrate how the project has been successful in achieving the objectives and desired outcomes of the Library's strategic goals. The Flickr project increases awareness of the Library and its collections.[75]

[68] All images were surveyed for file size and optimization as well as image quality and it proved easier to exclude suspect images at that stage (on a monitor screen) rather than scrutinize a 35mm slide on a light-table.

[69] To an extent this is already the case as, once these images are on Flickr, there is nothing to prevent any user downloading and distributing these image files as they wish, except for the rights and restrictions notice attached to each one. The difference is that responsibility for copyright would now be removed from the institution and so there would be no insistence on password protection.

[70] Pelagios: <http://pelagios-project.blogspot.co.uk/>.

[71] Because of the way Flickr is set up the tags link to all tags with the same character string. What would be particularly interesting would be to see how the 'comments' and 'annotations' could be used as a means of building links between images and what kind of links they would turn out to be.

[72] Referred to by name, 'crowdsourcing' is what is taking place with the LoC and other image collections in *The Commons*. ('For the Common Good' [n. 40 above] 1). See also: n. 75 below.

[73] 'For the Common Good' (n. 41 above) 36.

[74] 'For the Common Good' (n. 41 above) 1.

[75] 'For the Common Good' (n. 41 above) 33.

And further:

> At the start of the pilot, critics pointed out several risks often expressed as questions. Experience so far has not borne out their concerns. The skeptics wondered: Would the public conversation contribute to a better understanding of the photos or would fan mail, false memories, fake facts, and uncivil discourse obscure knowledge? […] Since the Library first launched its account the public has allayed many of the misgivings by lauding the rapid access to interesting photographs that could be enjoyed and used without restriction. News media complimented the Library for making publicly held information widely and freely available and also praised our openness to participatory cataloging. Fellow cultural heritage organizations quickly began to join Flickr's Commons because 'taking the pictures to the people' resulted in reaching large new audiences.[76]

In addition they report that this increased 'ability to engage and connect with photos' gives an increased 'sense of ownership and respect […] for these photos'.[77] Importantly also, that they have 'gained a deeper understanding of how users want to interact with […] collections'.[78] Public engagement has also helped them to understand how they might manage interaction with their users:

> in ways that are less formal without diminishing the reputation of the institution; how to reconcile the inevitable loss of control over content with the recognition that we can significantly increase the reach of that content if people can access and interact with it in the communities in which they participate.[79]

It seems unsurprising, then, that their recommendation was to move from a pilot phase into an expansion of this activity and of engagement with their user community as 'the benefits appear to far outweigh the costs and risks'.[80] Having said this, we need to be mindful of the cost and staffing implications. Although this was minimal for an organization such as the LoC, it could still be a potential burden for an overstretched Classics department or Institute.

Another successful example of crowdsourcing is to be seen at the Victoria and Albert (V&A) Museum's 'Search the Collections' initiative.[81] Here (after setting up an account) users are improving the collection by selecting the best 'crop' of an image from a selection to maximize the user experience. At the time of writing, the website shows that more than 35,000 objects have been processed in that way. They also note an additional spin-off benefit from working with the public as gaining 'insight' into their 'users' views and preferences'.[82]

[76] 'For the Common Good' (n. 41 above) 35.

[77] 'For the Common Good' (n. 41 above) 35.

[78] 'For the Common Good' (n. 41 above) 36.

[79] 'For the Common Good' (n. 41 above) 35.

[80] 'For the Common Good' (n. 41 above) 36.

[81] V&A Crowdsourcing: Search the Collections: <http://collections.vam.ac.uk/crowdsourcing/>.

[82] V&A Crowdsourcing: Search the Collections (n. 79 above).

Crowdsourcing *per se* is a discussion for another publication.[83] However, one might consider the case of Wikipedia and, for a positive discussion, see Blackwell and Crane's evaluation of the unpaid labour contributed to that community platform.[84] A perhaps more established example of a successful crowdsourcing project would be the 'Reading Programme' of the Oxford English Dictionary, which has recruited both voluntary and paid readers since 1857 to supply the editors with quotations and examples of English language usage.[85] In the earlier Digital Classicist volume, Stuart Dunn notes the effectiveness of utilizing the accumulated knowledge of the human population in the field of neogeography.[86] To engage the help of the user community (as LoC asked), they must be able to see the benefits for themselves. The initial impetus can result from 'crowd-casting' (*i.e.* 'pushing' the need and incentivizing user participation) but ultimately the potential contributors of additional data need to be 'pulled' in the direction of the objective.[87] As noted in the LoC final report: giving users 'the ability to engage and connect […] increases the sense of ownership'.[88]

User comparisons of collections on Flickr

It has not been possible to gather user statistics from institutions participating in *The Commons* other than those published by LoC in their final report.[89] Launched in January 2008, these figures were collected on 23 October 2008 and so represent approximately ten months. The total images in their Flickr collection: 4,615.

- All time views: 10.4 million;
- 79% of the images had been made a 'favourite';
- More than 15,000 Flickr members had made LoC a 'contact';
- 7,166 comments were left on 2,873 photos by 2,562 unique Flickr accounts;
- 67,176 tags were added by 2,518 unique Flickr accounts;

[83] Crowdsourcing is the term often used when the content is generated or added to by users of a resource rather than the creators of that resource. An example of such a project is Transcribe Bentham: <http://www.ucl.ac.uk/transcribe-bentham>. See also the recent initiative by the AHRC: Crowd Sourcing Study: <http://crowds.cerch.kcl.ac.uk/>.

[84] See: 'The work of scholarship: new divisions of labor in the world of Google and Wikipedia', in 'Scaife digital library and Classics in a digital age', in Crane & Terras, *Changing the center of gravity: transforming Classical Studies through cyberinfrastructure*, 3:1 (2009) <http:// digitalhumanities.org/dhq/vol/003/1/000035/000035.html>.

[85] Oxford English Dictionary: Reading Programme: <http://www.oed.com/public/reading/reading-programme>.

[86] S. Dunn, 'Space as an artefact: a perspective on "neogeography"', in *Digital research in the study of classical antiquity*, ed. G. Bodard and S. Mahony (Farnham 2010) 53-69.

[87] For more discussion on this point see: A. Hudson-Smith, M. Batty, A. Crooks, and R. Milton, 'Mapping for the masses: accessing Web 2.0 through crowdsourcing', in *Social Science Computer Review* 27.4 (2009) 524-38.

[88] 'For the Common Good' (n. 41 above) 25.

[89] 'For the Common Good' (n. 41 above) iv.

- Fewer than 25 instances of user-generated content were removed as inappropriate;
- Average monthly visits to all their ‘Prints and Photographs Online Catalog’ web pages rose 20% over the five-month period of January-May 2008, compared to the same period in 2007.

Their move to Flickr had made a considerable positive impact on the traffic to their web pages.

The Flickr aggregated statistics for the *AWIB* collection on 22 November 2010 represent a period of approximately eight months:[90]

- View counts (total images in the collection: 2,098);
- Number of views on the previous day: 50;
- All time views: 19,492;
- Number of individual images viewed: 1,210;
- Images added as ‘favourites’: 24;
- Images that have had comments added: 12.

HumSlides on Flickr was launched to coincide with the start of the UK academic year in September 2009 and so statistics collected at 4 December 2010 represent approximately fourteen months:

- View counts (total images in the collection: 4,035);
- Number of views on the previous day: 0;
- All time views: 855;
- Number of individual images viewed: 261;
- Images added as ‘favourites’: 1;
- Images that have had comments added: 0.

These figures highlight the difference in use between these collections that share the same platform. *OxCLIC* is not included in the comparison here, as it is hosted on institutional infrastructure rather than Flickr and user statistics are not published.

Conclusion

Cost is an immediate reason for keeping this resource on Flickr. Setting up and developing the infrastructure, whether using proprietary software or an Open Source option, to host and maintain an online collection should not be taken on lightly. In addition, the major cost of the *HumSlides* pilot was the labour involved in the digitization of the slides and the processing of the images into usable content. With the ubiquitous nature of digital photography (one of the reasons for the demise of the slide medium in the first place), we have a readymade source of digital images from our community, many of whom already share their images on online platforms such as Flickr.

Using the *AWIB* model and releasing the images under the appropriate Creative Commons licence would mean that they may be freely downloaded, copied, shared, and

[90] With thanks to Tom Elliott at ISAW for providing this data.

adapted, providing that they are attributed correctly. Actively encouraging reuse will raise the profile of the collection (and that of the contributors) through citation and this raised profile will in turn encourage those in our community to contribute their images and improve the collection. With the bulk upload tool and metadata organized in a spreadsheet as per our standard, uploading the images and data is a trivial activity and no additional storage facility is required.[91]

This, then, seems to be the way forward and the only other consideration is whether or not to allow users to contribute additional data in the form of tags and annotations. The two models we have are *The Commons*, which allows crowdsourced data, and *AWIB*, which does not. The issue here may be one of the staffing implications involved in the management of any user data that is added. The parent institution of *AWIB* is ISAW and one of their partner projects is *Pleiades*, which is a community-based project where members create, modify, and share data,[92] and so there should not be any objection in principle. It is perhaps more an issue of staffing costs. This is confirmed by the director (Tom Elliott) who adds that their policy is to reciprocate contact requests and allow anyone who is serious about tagging images for the benefit of others the opportunity to do so.

With the possibilities for reuse, the images from an open *HumSlides* collection can be embedded in teaching modules. Further, they can by collected together, built into teaching objects, and incorporated in initiatives such as the Higher Education Academy (HEA) and JISC-funded Open Education Resources (OER) Programme to make teaching resources freely available to all.[93] With the appropriate attribution, this would again raise the profile of both the collection and its contributors. To increase awareness, the image collection needs to be promoted within the institution (and, if open, then also amongst other institutions) and wider community and perhaps packaged to be made available as OERs for embedding in teaching and student projects. The community could be greatly widened by establishing links with existing Classics collections such as those at *AWIB* and also The *Stoa* Consortium.[94]

The nature of internet is changing and so are the users. The so-called Web 2.0 is not a technological revolution, but a social revolution enabled by the new technology. It has reorganized the way we communicate and hence how we learn, with users becoming contributors and collaborators.[95] Here, in this suggested model, users become contributors of

[91] It would, of course, be prudent to retain an offline archive copy of the image files although this would have a trivial cost.

[92] See 'The *Pleiades* Community':
<http://www.atlantides.org/trac/pleiades/wiki/PleiadesCommunity>.

[93] See the HEA/JISC OER programme:
<http://www.heacademy.ac.uk/ourwork/teachingandlearning/oer>.

[94] The *Stoa* (<http://www.stoa.org/>) has been committed to Open Access since its inception and has images collected and made available at <http://www.stoa.org/gallery/> hosted on the institutional infrastructure at the University of Kentucky.

[95] For more on the changing relationship between 'production and consumption of content' see: D. Beer and R. Burrows, 'Sociology and, of, and in Web 2.0: some initial considerations', *Sociological Research Online* 12(5) (2007): <www.socresonline.org.uk/12/5/17.html> [accessed 9th February 2011].

images and the community as a whole become suppliers and improvers of the metadata. This is much more in keeping with Tim Berners-Lee's original conception of the Web, where, rather than being an online marketplace and entertainment centre, we should all 'be putting […] ideas in, as well as taking them out.'[96] Users, then, have a vested interest in and a sense of ownership of the collection, which will help to ensure its sustainability. Web based technologies now offer opportunities to develop user-driven models and collaborative environments for generative teaching, learning, and research image collections. The collective knowledge of the user community becomes a resource to enhance and drive its usefulness forward. In this case the more the resource is used, the more it will grow and improve.[97]

Simon Mahony (*University College London*) s.mahony@ucl.ac.uk

[96] T. Berners-Lee, 'Transcript of Tim Berners-Lee's talk to the LCS 35th Anniversary celebrations', Cambridge, MA (1999): <http://www.w3.org/1999/04/13-tbl> [accessed 10th November 2011].

[97] Since writing there has been an important new AHRC publication with a focus on engaging user communities to enrich humanities resources. S. Dunn and M. Hedges, 'Crowd-sourcing scoping survey: engaging the crowd with humanities research' (AHRC Project 2013). PDF available at: <http://crowds.cerch.kcl.ac.uk/>.

CONNECTING THE CLASSICS: A CASE STUDY OF COLLECTIVE INTELLIGENCE IN CLASSICAL STUDIES

VALENTINA ASCIUTTI AND STUART DUNN

Introduction

Between the mid-1990s and the late 2000s, there was much interest in the internet as an enabler of 'Collective Intelligence' (CI), which held that groups, or crowds of individuals, were better at decision making than even expert individuals.[1] This movement was defined by James Surowiecki in 2004.[2] In the context of collective intelligence, the 'Classics' as a collective of Classicists – and especially digital classicists – comprise what the anthropologist Clay Shirky has described as an 'undisciplined group': there is no single, shared objective, rather a mass of different interests, sometimes shared, sometimes not, sometimes conflicting, sometimes coinciding.[3] In order to support digital research in the Classics, therefore, research support infrastructures with a nuanced and critical approach to collective intelligence are needed. This reflects the view of Scott Rettberg, who notes that:

> Whenever people are collaborating on a project of knowledge sharing or creative production, they are collaborating not only with other people, but with a system which they, the other participants, and the communicative environment help to create.[4]

Classicists, however, are not simply those employed by universities, and as the internet breaks down the boundaries between the online spaces occupied by universities and the wider connected world, methods of enabling the two to collaborate will become ever more important. This paper will review how public participation in the digital sphere is changing Classics and Classical Archaeology and will examine the future of so-called crowd-sourcing in this area, and in particular what infrastructure is needed to support it. Crowd-sourcing is defined in a review as 'the process of leveraging public participation in

[1] See: D. Brabham, 'Crowdsourcing as a model for problem solving: an introduction and cases', *Convergence: the International Journal of Research into New Media Technologies* 1.14 (2008) 75-90.

[2] J. Surowiecki, *The wisdom of crowds: why the many are smarter than the few and how collective wisdom shapes business, economies, societies and Nations* (New York 2004).

[3] C. Shirky, 'The political power of social media: technology, the public sphere and political change', *Foreign Affairs* 90.1 (2011) 28-41.

[4] S. Rettberg, 'All together now: collective knowledge, collective narratives, and architectures of participation', in *Digital arts and culture conference*, (Copenhagen 2005) 1-2.

projects and activities',[5] and has been shown to have enabled the creation of complex knowledge in humanities research. In this paper, we expand on this to introduce the concept of *directed knowledge circulation*, which seeks to describe combining the power of crowd-sourced information in the humanities, registries of semantic knowledge derived from years of scholarly effort, and also vocabularies of reference information, especially those pertaining to geography and location.

Data creation versus *data linking*

Numerous infrastructures support the creation of data in the Classics by members of the public. Often, however, this data is unstructured or semi-structured, with contributor-generated semantic tagging. Flickr provides an excellent example: given that many locations of interest to classicists are also tourist destinations, much visual information about those features is available on Flickr and much of this is licenced using Creative Commons, rendering it available for academic use. However, this is a 'flat' dataset. For example, a simple search using the term 'Parthenon' returns (at the time of writing) 92,367 objects, many, but not all of which, depict the Parthenon at Athens. Many others are of Parthenon-like structures of varied ages.[6] Flickr is a classic example of the kind of information circulation which the internet privileges: flat, broad, and unable to focus around key subject areas or questions. Such foci in the content are generated if there is a shared interest in a particular subject or object. Wikipedia, of course, does this via crowd-generated subject areas in the form of page titles. Intuitively, one might believe that pages relevant to the Classics and archaeology would contain contributions from 'expert' authors. However, as Kittur *et al.* have shown,[7] whilst Wikipedia in its early days was driven by small groups of 'elite' users, contributing much of the new content and making most of the edits, this saw a decline in the mid-2000s towards larger numbers of users making smaller edits and contributions. Clearly, neither comprehensive subject coverage nor academic credibility is achievable by the distributed generation of very large amounts of very small units of (simple) information: the so-called 'long-tail'.

Much of the value of information generated or processed by 'the crowd' accrues to Classics when it is combined with *information infrastructures* – that is to say, Persistent Identifiers (PID), vocabularies, Uniform Resource Identifiers (URIs), and typologies which allow crowd-generated information to be tagged and organized automatically. Typically, such information infrastructures are based around notions of *what, where*, and *when*. Developing the Flickr example presented above, the best instance of this is the *Pleiades* project (www.pleiades.sioa.org) as an example of 'where'. The *Pleiades* approach is to provide stable and persistent unique identifiers for every place in the Greek and Roman world. This corpus has a facility which allows users to tag their own Flickr

[5] S. Dunn and M. Hedges, 'Engaging the crowd with humanities research', report for the Arts and Humanities Research Council *Connected Communities* programme (2012).
(Online at: <http://crowds.cerch.kcl.ac.uk>).

[6] For example: Parthenon, Centennial Park, Nashville, TN.:
<http://www.flickr.com/photos/purphoto/4080554853>.

[7] A. Kittur, E. Chi, B. A. Pendleton, B. Suh, and T. Mytkowicz, 'Power of the few *vs* wisdom of the crowd : Wikipedia and the rise of the bourgeoisie', *Algorithmica* 1.2. (2007) 1-9.

objects with machine-readable tags relating to *Pleiades* URIs. Thus, the ancient Roman bridgehead on the south bank of the Tyne at Chollerford in Northern Britain can have its own *Pleiades* URI created even though it does not have a name of its own,[8] and this can be automatically associated with a Flickr photo of the bridgehead itself. This is then associated, by the expert *Pleiades* editorial group, with the URI for Cilurnum (Chesters Fort), with which the bridgehead has both a direct geographical and chronological association.[9]

www.arts-humanities.net[10]

A similar editorial and vocabulary-based approach is taken by the www.arts-humanities.net project, an activity developed by the Centre for e-Research at King's College London and now being developed in collaboration with the CenterNet group. The www.arts-humanities.net structure aims to provide an integrated set of knowledge and services to scholars in all the humanities disciplines, but there are both requirements and resources that are distinct to the study of classical antiquity. As Crane *et al.* have noted, a digital library of classical sources typically deals with material that is far more fragmentary, sparse, and disconnected than a digital library of resources for, say, nineteenth-century London.[11] Whilst well-known digital library and infrastructure projects such as *Perseus*[12] seek to build solutions, and to *provide* material to meet the Classics-specific requirements, www.arts-humanities.net seeks to *describe* and *connect* existing material in a systematic and structured way.

The material in www.arts-humanities.net is described by a set of controlled vocabularies, expressed as tags. Tags are assigned by the user when they create a record. For example, at the time of writing, the resource holds a list of about five hundred project records, of which forty-eight are tagged with 'Classics and Ancient History'. However, there are other discipline tags, which allow more nuanced descriptions. Nine projects of the five hundred, for example, are tagged with both 'Classics and Ancient History' and 'Linguistics' (a figure which can be determined by searching using those terms in a free text search), indicating that these projects cross both areas. Although most of these projects indeed comprise the application of linguistic methods in the domain of Classics:

[8] Unnamed Roman Bridgehead at *Pleiades*: S. Dunn and T. Elliott, 'Places: 765102512 (Unnamed Roman bridgehead)', *Pleiades*. <http://pleiades.stoa.org/places/765102512> [accessed: 8th November 2013 12:08 pm]. Note that the place URI (765102512) assigned to this bridgehead is incorporated in the URL.

[9] Cilurnum: A. Esmonde Cleary, R. Warner, R. Talbert, T. Elliott, S. Gillies, and S. Vanderbilt. 'Places: 89144 (Cilurnum)', *Pleiades*. <http://pleiades.stoa.org/places/89144> [accessed: 8th November 2013 12:07 pm].

[10] arts-humanities.net: <http://www.arts-humanities.net/>.

[11] Crane *et. al.*, 'Building a hypertextual digital library in the humanities: a case study on London', *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital Libraries*, (Roanoke VA 2001) 426-34.

[12] *Perseus*: <http://www.perseus.tufts.edu/hopper/>.

for example, the *EpiDoc* project (of which more below),[13] but this combination (using all the search terms) also returns the 'Digital Research Infrastructure for the Arts and Humanities' (DARIAH)[14] project record, which is a generic European digital humanities infrastructure project and therefore tagged with every subject classification, including 'Linguistics', and 'Classics and Ancient History'. However, such generic projects are relatively rare, and therefore this does not affect the overall validity of the browsing model. Rather, filtered searches such as this allow the users interested in particular areas of the Classics to retrieve projects relevant to those areas.

Methods taxonomy

The principle of these vocabularies is a taxonomy of over one hundred digital research methods, classified into seven main categories and over one hundred sub-categories.[15] A method is defined here as the means by which digital technologies or tools are used to create new knowledge. Alongside the taxonomy, a catalogue of about 100 digital tools identified both by project staff and by the wider community is included. Using these descriptions of tools and methods, it is possible to identify workflows for specific research questions based on specific information from the knowledge base. For example, a researcher might be interested in extracting information from a digital corpus and turning it into a structured database. Of the forty-eight projects tagged 'Classics and Ancient History', three share method tags of 'Parsing' and 'Data Modelling'. From this combination, the user can identify projects of likely interest to them.

One key problem with this approach, of course, is that the terminology itself is not neutral, especially given the extremely interdisciplinary nature of classical studies (and digital humanities more generally). Each method term is accompanied by a detailed description, which at least seeks to acknowledge possible conflicts of meaning and semantics. In the example given above, 'parsing' is noted to refer to one thing in a linguistic context, and another in a computer science context.[16] www.arts-humanities.net seeks to deal with this by providing open descriptions of the methods which individuals can, if necessary, challenge by providing alternative reviews or initiating discussions via the forums.

Another problem with having such a large and detailed set of method definitions is that, typically, the number of methods associated with each project is high. For example, the fourteen projects tagged with both 'Classics and Ancient History' as a discipline and 'Data Modelling' as a method have an average of 7.4 other method tags associated with them. In many of these cases these other methods are extremely disparate, ranging in one case from '2D raster', to 'Geophysical Survey', and 'Image Enhancement'. In all cases the combinations are unique, the connections forming a 'methodological fingerprint' for each project. This highlights that, even with a robust well-documented taxonomy, its

[13] *EpiDoc*: < http://epidoc.sourceforge.net>.

[14] DARIAH: <http://www.dariah.eu>.

[15] See: www.arts-humanities.net methods: <http://www.arts-humanities.net/ictguides/methods>.

[16] See: www.arts-humanities.net methods *s.v.* parsing:
<http://www.arts-humanities.net/data_analysis/parsing>.

application in individual cases will reflect the idiosyncrasies of individual projects, even though the tags themselves are part of a controlled vocabulary.

Case study: Roman inscriptions of Britain

As noted above, the www.arts-humanities.net database contains numerous pieces of information on the study of epigraphy. Inscriptions are particularly interesting objects of study from both a cultural heritage and an information science point of view. Furthermore, in contrast to other kinds of humanities data, Roman inscriptions are relatively easy to categorize into certain groups: verse inscriptions, for example, form an easily definable group. Whether *in situ* or in museums, inscriptions are constantly viewed, received, and discussed by both expert and non-professional online communities. They therefore form a case of requiring what Copeland has described as 'constructivist' cultural heritage,[17] a set of data which lends itself to audiences participating in its interpretation and reception. This is against the background of the study of epigraphy, which has long regarded inscriptions primarily as texts, with relatively little attention given to their provenance as archaeological artefacts, although this view has changed in recent years.[18]

The *Roman Inscriptions of Britain* (*RIB*) corpus is, to date, the most authoritative and complete edition of Latin inscriptions from Roman Britain. The most recent edition was published in 2009;[19] however, the version employed in this study is based on the 1995 edition.[20] The inscriptions are listed in geographical order, starting from *Londinium* and 'walking' through the province up to the northern boundaries.

A key part of documenting and describing an inscription's context, as well as its text, lies in describing its location at different points in its history (as with any other kind of artefact) see Lock (2003),[21] as well as what kind of discourses particular inscriptions have provoked in the contemporary world. This inevitably means organizing and interrogating digital ephemera. Witness, for example, a tweet from @perlineamvalli of 23rd October 2012:

[17] T. Copeland, 'Presenting archaeology to the public: constructing insights on-site', in *Public archaeology*, ed. N. Merriman (London 2004) 132-44.

[18] See G. Bodard, 'Archaeology and epigraphic interchange and e-Science', *Stoa Consortium* (2009): <http://www.stoa.org/archives/857>, and also C. Tupman, 'Contextual epigraphy and XML: digital publication and its application to the study of inscribed funerary monuments', in *Digital research in the study of classical antiquity*, ed. G. Bodard and S. Mahony (Farnham 2010) 73-86.

[19] R. S. O. Tomlin, R. P. Wright, and M. W. C. Hassall*, Roman inscriptions of Britain, volume III: inscriptions on stone, found or notified between 1 January 1955 and 31 December 2006* (Oxford 2009).

[20] Our research on the Romano-British verse inscriptions was carried out before the new edition of *RIB* came out; however, to the best of our knowledge there are no new Latin verse texts in the new edition and there are no corrigenda to the readings and commentary on the previous ones. So the research is still valid.

[21] G. Lock, *Using computers in archaeology. Towards virtual pasts* (London 2003).

> #RIB3297 #centurialstone first seen 1986 at St Oswald's farm http://goo.gl/maps/fFRSu #hadrianswall #inscriptions[22]

The link provides georeferenced transcriptions from the *Roman inscriptions of Britain* corpus (numbers 1441, 3297 – hence the hashtag – and 1440), yet the only means of discovering this information, unless one happens to be following @perlineamvalli, is via the hashtags given, which are arbitrarily chosen by a single user, and do not conform to any agreed system.

On the other hand, www.arts-humanities.net provides an information structure for digitally representing tools, metadata, and standards, which can be used to document inscriptions as archaeological objects (see above). This ‘directed’ circulation of crowd knowledge contrasts with the *ad hoc* assignment of free text hashtags (analogous to the free-text descriptors on sites such as Flickr), which are neither unique nor persistent.

To illustrate this contrast, sample inscriptions from *RIB* were marked up using *EpiDoc* TEI[23] and associated with the online gazetteer *GeoNames*.[24] This is a straightforward object-level representation of several significant properties about the inscription, including its current location. While this does not provide the same level of access to the material as fully digitized corpora such as the *Inscriptions of Roman Tripolitania*[25] or the *Inscriptions of Roman Cyrenaica*,[26] it allows us to express and visualize various relationships within *RIB* using standard vocabulary and a pre-existing information structure.

As well as describing and representing such geographic meta-information about inscriptions in terms of points, lines, and polygons (the standard vector GIS approach), *GeoNames* can be used to link datasets such as the *RIB* with other, related sources of information (such as the locations of villas, Roman settlements, communication networks, and trading routes), using geography as the common factor,[27] much as *Pleiades* was used in the Cilurnum example above. In *EpiDoc*, the location reference can be embedded in the markup. As an example, we give *RIB* number 265, a metrical inscription from Lincoln**.** This may be marked up in *EpiDoc* as follows:

```
<provenance>
  <listEvent>
    <event type="found">
      <p><placeName ref="http://sws.geonames.org/2644487"
        >Lincoln</placeName>: context.</p>
    </event>
    <eve <provenance type="found">
```

[22] Twitter: <https://twitter.com/perlineamvalli/statuses/260651954668183552>.

[23] *EpiDoc* Training Workshop: <http://www.arts-humanities.net/user_tags/epidoc>.

[24] *GeoNames* geographical place names database: <http://www.geonames.org/>.

[25] *Inscriptions of Roman Tripolitania*: <http://irt.kcl.ac.uk/irt2009/>.

[26] *Inscriptions of Roman Cyrenaica*: <http://www.ircyr.kcl.ac.uk/>.

[27] P. S. Ell, ‘GIS, e-Science, and the humanities grid’, in *The Spatial Humanities*, ed. D. J. Bodenhamer, J. Corrigan, and T. M. Harris (Bloomington IN 2010) 143-66.

```
  <p><placeName ref="http://sws.geonames.org/>Lincoln</placeName>:
context.</p>
</provenance>
<provenance type="observed">
  <p>In the Lincoln Museum in <placeName
ref="http://sws.geonames.org/2644487">Lincoln</placeName>.</p>
</provenance>nt type="observed">
      <p>In the Lincoln Museum in <placeName
        ref="http://sws.geonames.org/2644487">Lincoln</placeName>.</p>
    </event>
  </listEvent>
</provenance>
```

This markup directly embeds the *GeoNames* reference (2644487) as a URI.[28] Making the original location explicit in the markup allows the original findspot to be associated with this record and mapped (see below, Figure 2). *RIB*, and most other epigraphic corpora, tend to provide *informal* georeferencing and in most cases this is an association of information based on place name. There are several problems with this: first of all, many place names are not unique, and it is possible that subsequent searches may return the wrong location, especially if the place in question has alternative spellings or is double-barrelled (*e.g.* Winterbourne Basset *versus* Winterbourne Steepleton). To avoid such problems, www.arts-humanities.net linked a selection of the *RIB* with the *GeoNames* database.

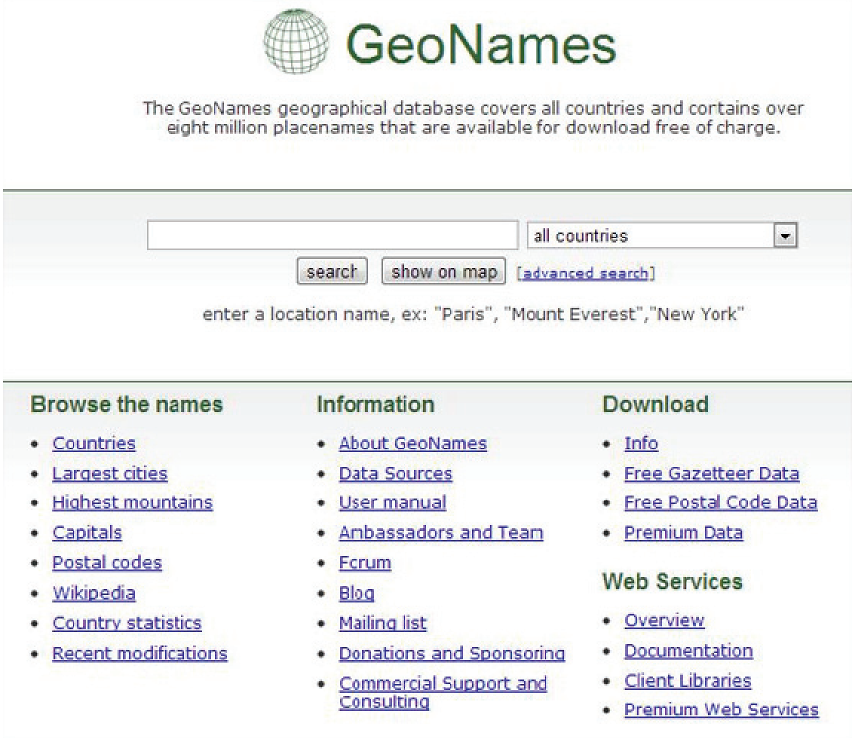
The example from Lincoln given above provides a case in point: *GeoNames* indicates that numerous entities, including 'lake' and 'administrative division', are described by the name 'Lincoln' and are located in Nebraska, Quebec, Ontario, South Dakota, Kansas, Florida, and Illinois, among others.[29] However, the URI ascribed to the entry for Lincoln in Lincolnshire, UK *is* unique. This makes *RIB* objects marked up in this way in Epidoc exposable to other methods in a-h.net, including (for example) Statistical Analysis, Spatial Data Analysis, 2D Scanning and Photography, Geophysical Survey, Manual Input and Transcription, Text Recognition, Collaborative Publishing, Resource Sharing, Geo-referencing and Projection, Image Enhancement, and Image Restoration.

Beyond GIS: georeferencing using shared standards

We term this method of dealing with geographic knowledge about epigraphic objects, using external taxonomies and reference lists to create aggregated knowledge sets (whose components could come from any online source) *directed circulation of knowledge*. It provides what we consider to be a useful distinction from collective intelligence or crowd-sourcing in that it combines scholarly direction with the diversity of crowd-created content. This may be seen as a more formal manifestation of the emergence of so-called

[28] *GeoNames*: <http://www.geonames.org/2644487/lincoln.html>.

[29] *GeoNames*: <http://www.geonames.org/search.html?q=lincoln&country>.



The GeoNames geographical database covers all countries and contains over eight million placenames that are available for download free of charge.

enter a location name, ex: "Paris", "Mount Everest","New York"

Figure 1. *GeoNames* front page

'neogeography'[30] and the question of how reliable volunteered geographic information can be said to be.[31]

We argue here that the process of associating information with latitude/longitude coordinates (georeferencing) is at the core of growing the 'geospatial web' for cultural heritage and that this process must be conditioned so as to allow the linking of resources, both those of established scholarship, such as *RIB*, and ephemera,[32] using geography.

As discussed above, georeferencing can be either formal, *i.e.* based on a mathematical string expressed in degrees, or some other decimal numeric projection, or informal, which usually means place names, or some other textual or semantic description. In a review of the area, Martyn Jessop notes that '[t]here is a requirement for more central archives and a metadata schema that would allow [humanities geodata] resources to be discovered and

[30] S. Dunn, 'Space as an artefact: a view of neogeography from the digital humanities', *Digital research*, ed. Bodard and Mahony (n. 18 above) 53-69.

[31] See M. Haklay, 'How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets', *Environment and Planning B: Planning and Design* 37.4 (2010) 682-703; M. Goodchild, 'Editorial: citizens as voluntary sensors: spatial data infrastructure in the world of Web 2.0', *International Journal of Spatial Data Infrastructures Research* 2 (2007) 24-32.

[32] Such as the Twitter example cited (n. 22 above).

shared across archives'.[33] Most kinds of data can be associated in some way with geographic information that can in turn constitute metadata. For example, a photograph can be associated with a specific point on the earth's surface by a caption specifying place, coordinates from a GPS-enabled camera, reference in a text, and so on. Texts, such as *RIB*, can refer to specific places or types of places in a great variety of ways. Finds on archaeological sites are often recorded using theodolite readings, or other spatial location techniques (which produce site-specific datasets that need to be related to a global coordinate system, or imagery with spatial distortions caused by perspective that need correcting, the process of orthorectification.

The ability to deploy shared standards such as KML[34] in epigraphy highlights how directed knowledge circulation can enable classicists to move beyond the constraints of 'conventional' analytical environments such as Geographic Information Systems (GIS).

Directed circulation of epigraphic knowledge

There are several diverse facets to documentable information about inscriptions and, in order to construct an epigraphic knowledge base around an epigraphic corpus, it is necessary to identify and list them. For example, out of roughly three thousand inscriptions known from Roman Britain, around thirty might be considered *metrical*. For these metrical texts the map (Figure 3) shows all the inscriptions that demonstrate some metrical features, those entirely metrical as well as those partially in verse, with metrical lines encapsulated in the body of the text (*e.g. RIB* 684, an epitaph in prose from York that incorporates an epigram in the middle of the text from *Secreti* to *finem*).

This is therefore a small corpus, yet a complex one. The analysis of metrical texts found around the Roman Empire has shown a certain level of homogeneity in treating verse inscriptions in the Roman provinces.[35] In particular, two things are evident in the wider corpus of metrical inscriptions. Firstly, the majority of the metrical inscriptions are funerary texts, and most of them are written in the dactylic metre. Secondly, a common feature that Roman provinces share is the concentration of verse inscriptions from major centres within the province.

[33] M. Jessop, 'The inhibition of geographical information in digital humanities scholarship', *Literary and Linguistic Computing* 23.1 (2008) 39-50.

[34] Keyhole Markup Language is the protocol used for encoding geographic information in Google Earth and Google Maps

[35] F. Buecheler, *Carmina Latina Epigraphica* (Leipzig 1897-1927); M. Buonocore, 'Carmina Latina Epigraphica regionis IV Augustae', *Avvio ad un censimento*, *Giornale di Filologia Italiana*, 49 (1997)1597-1628; P. Carletti-Colafrancesco, M. Maddaro, and M. L. Ricci, *Concordanze dei Carmina Latina Epigraphica* (Bari 1986); P. Cugusi, *Carmina Latina Epigraphica Provinciae Sardiniae* (Bologna 2003). P. Cugusi, *Carmina Latina Epigraphica Pannonica* (Bologna 2007); P. Cugusi, *Carmina Latina Epigraphica Moesica* (Bologna 2008). P. Cugusi, *Carmina Latina Epigraphica Thraciae* (Bologna 2008).

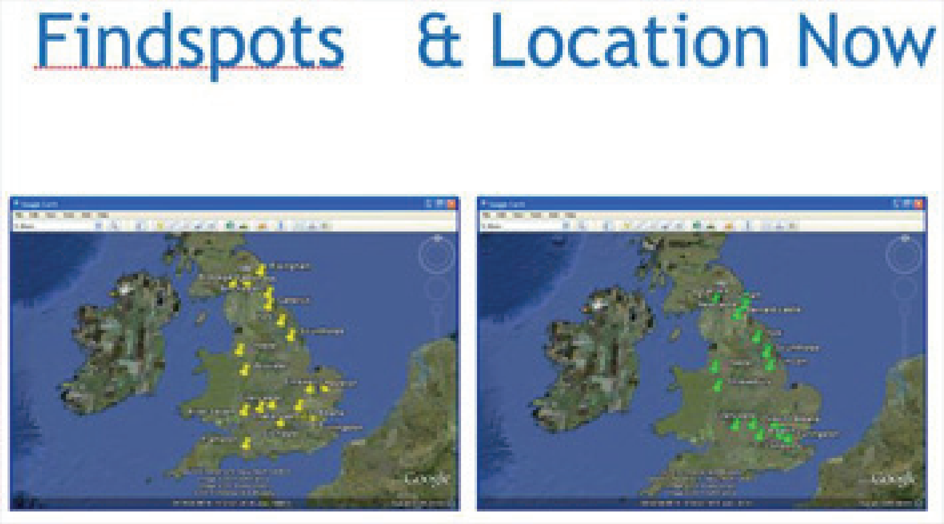


Figure 2.

The situation in Britannia, however, is different:

- There is no predominance of funerary texts;
- A wide range of metres were adopted;
- There are plenty of metrical texts inscribed on *instrumentum domesticum*, in particular mosaics, bowls, bricks, and tiles;
- There is no site, not even the largest and most important one, that produced more than one or two metrical texts.

In general, the vast majority of Romano-British inscriptions were found in the proximity of military settlements, particularly in the north of the province; however, the distribution of the metrical texts is instead more spread out throughout the country and only five texts out of thirty can be linked directly by their location and content to the military community.

Reasons for these anomalies can only be hypothesized, pending further comparative research of this kind, especially in areas with a greater concentration of material. It will be interesting to see if there are other provinces with a comparable proportion of verse inscriptions on *instrumentum domesticum*, or areas that produce a comparable range of metres. Perhaps the insular position of Britain beyond the limits of the continental empire contributed to this difference.

So far we have linked text and geographical location (*i.e.* findspot), and the map created (figure 2) helps to visualize the points made above, regarding the peculiarity of the metrical texts from Britannia.

We can mention at least four examples of iconographical references to Classical texts:

A unifying factor that underpins all these interpretations of metrical texts is of course the location of each inscription's findspot, and its context. Whilst the former is recorded systematically and authoritatively by the *RIB*, less attention is typically paid to the latter.

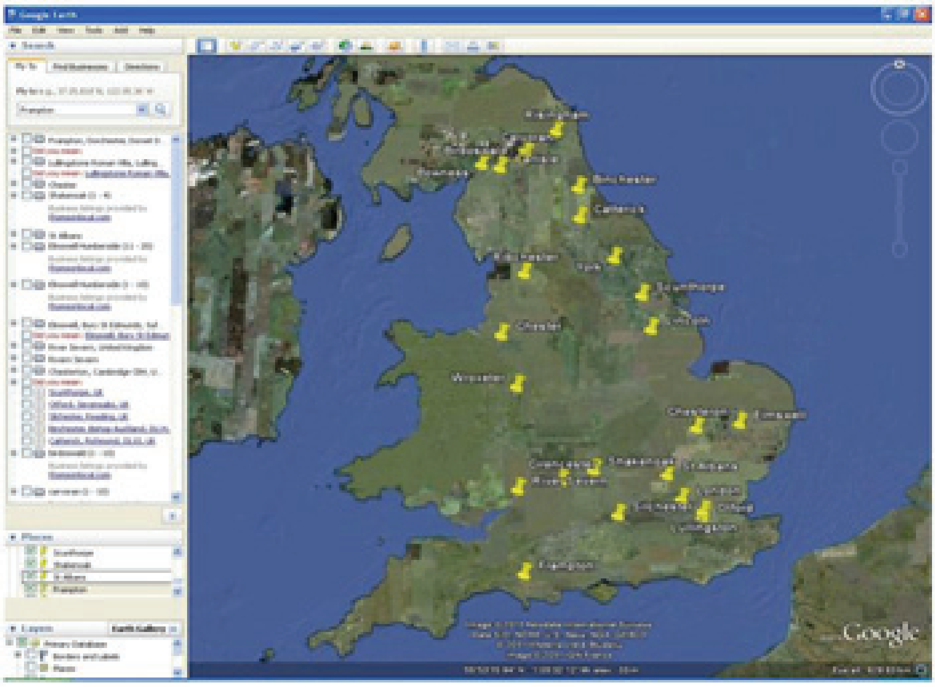


Figure 3.

Thus, by linking *RIB* using a KML file containing unique findspots, and documenting the process in the structured environment of www.arts-humanities.net, we create an integrated and multifaceted data object which combines content (inscriptions/objects), methods as defined in the www.arts-humanities.net schema,[36] standards (KML and *EpiDoc*), and a tool (Google Earth). Any other scholars, or interested members of the wider public, with relevant information to share can do so using this information structure.

The digital environment of www.arts-humanities.net allows us to present this research process as a linked set of objects, to which others can refer. The system provides the building blocks; we, the users, decide how to fit them together. In the past, epigraphic databases have naturally relied upon what might be termed *traditional* cartographic methods of representing the locations of their records, because their spatial component is based on point data. However, this is a relatively limited medium for effectively representing datasets which differ so significantly across space. We have made initial explorations into georeferencing the *RIB* information referred to here, and representing it in Google Earth, and our future plans include building a hierarchy of lines and polygons in Google Earth, which will seek to reflect the *RIB*'s complexities.

This work is one of numerous case studies from the digital humanities which can be cited to illustrate the 'push' that wider community involvement in analyzing data gives to the process of turning that data into information. As noted in the introduction, collaboration in

[36] Arts-humanities.net schema: <http://www.arts-humanities.net/ictguides/methods>.

the digital humanities is now as much about collaboration between people and digital systems, as collaboration between people and other people.

Conclusion

Much has been made in recent years of the power of crowd-sourcing and its potential as a means of generating academic digital resources.[37] In this chapter, we suggest that the creation of crowd-sourced digital resources in the Classics and Archaeology is most effective when it encourages the *circulation* of knowledge which otherwise would only be discoverable serendipitously as a result of internet protocols such as hashtags, open semantic searches, and hyperlinks. Such knowledge is directed using vocabularies which are either expert (such as *Pleiades*) or simply stable (such as *GeoNames*), or, of course, both. Our case of epigraphic data and the *RIB*, which is authoritative yet monolithic, shows that it is essential to express and visualize in evocative ways links that already exist but are not explicit across disciplines within Classics and the broader humanities field, in our case Epigraphy, Archaeology, and Geography. Using www.arts-humanities.net and the multifaceted building blocks it provides, we are able to construct a navigable digital information object which links all three. Our case-study also illustrates how important it is for a discipline like Classics to embrace technology and incorporate digital humanities infrastructures. Collaborations, sharing of expertise, and Collective Intelligence can only be possible if we bring to the surface and make transparent all the relevant interconnections that make research more comprehensive and meaningful.

In other words, directed knowledge circulation develops upon the idea of Collective Intelligence as a means of building information from disparate groups of people and combining and analyzing it. As in our example, a Digital Humanities Web would gather relevant information from web activities, providing a lively and continuously updated source of information.

By putting a project like this on www.arts-humanities.net, we allow these links to become alive. The project is tagged with all the relevant digital methods and tools, data formats and metadata standards used during the life of the project, so that it can be linked to other similar projects as well as relevant tools, methods, events, jobs in the field, and a library of papers, audio, video files, and images, too.

Valentina Asciutti (*King's College London*) valentina.asciutti@kcl.ac.uk
Stuart Dunn (*King's College London*) stuart.dunn@kcl.ac.uk

References

JRS: *Journal of Roman Studies*

RIB: *Roman inscriptions of Britain*

[37] Dunn and Hedges, 'Engaging the crowd' (n. 5 above) 3.

Anderson, S., T. Blanke, and S. Dunn, 'Methodological commons – arts and humanities e-Science fundamentals', *Philosophical Transactions of the Royal Society A* 28 368.1925 (2010) 3779-96.

Bücheler, F., *Carmina Latina Epigraphica* (Leipzig 1897-1927).

Buonocore, M., '*Carmina Latina Epigraphica* regionis IV Augustae. Avvio ad un censimento', *Giornale di Filologia Italiana* vol. 49 (1997) 21-50.

Carletti-Colafrancesco, P., M. Maddaro, and M. L. Ricci, *Concordanze dei Carmina Latina Epigraphica* (Bari 1986).

Copeland, T., 'Presenting archaeology to the public: constructing insights on-site', in *Public Archaeology*, ed. N. Merriman (London 2004) 132-44.

Crane, G., D. A. Smith, and C. Wulfman, 'Building a hypertextual digital library in the humanities: a case study on London', *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, (Roanoke VA 2001) 426-34. Available online at: <http://portal.acm.org/ft_gateway.cfm?id=379756&type=pdf&CFID=13792823&CFTOKEN=80903370> [accessed: 15th March 2011].

Cugusi, P., *Carmina Latina Epigraphica Provinciae Sardiniae* (Bologna 2003).

Cugusi, P., *Carmina Latina Epigraphica Pannonica* (Bologna 2007).

Cugusi, P., *Carmina Latina Epigraphica Moesica* (Bologna 2008).

Cugusi, P., *Carmina Latina Epigraphica Thraciae* (Bologna 2008).

Dunn, S., 'Space as an artefact: a view of neogeography from the digital humanities', in *Digital research in the study of classical antiquity*, ed. G. Bodard and S. Mahony (Farnham 2010) 53-69.

Dunn, S., and M. Hedges, 'Engaging the crowd with humanities research', report for the Arts and Humanities Research Council *Connected Communities* programme (online at <http://crowds.cerch.kcl.ac.uk>).

Ell, P. S., 'GIS, e-Science and the humanities grid', in *The Spatial Humanities*, ed. D. J. Bodenhamer, J. Corrigan, and T. M. Harris (Bloomington IN 2010) 143-66.

Fernández-Martinez, C., *Carmina Latina Epigraphica de la* Bética Romana (Sevilla 2007).

Jessop, M., 'The inhibition of geographical information in digital humanities scholarship', *Literary and Linguistic Computing* 23.1 (2008) 39-50.

Kittur, A., E. Chi, B. A. Pendleton, B. Suh, and T. Mytkowicz, 'Power of the few vs wisdom of the crowd : Wikipedia and the rise of the bourgeoisie', *Algorithmica* 1.2. (2007) 1-9.

Levy, P., *Cyberculture* (Paris 1997).

Lock, G., *Using computers in archaeology: towards virtual pasts* (London 2003).

Rettberg, S., 'All together now: collective knowledge, collective narratives, and architectures of participation', in *Digital Arts and Culture* (Copenhagen 2005) 1-2. Available online at: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.5802&rep=rep1&type=pdf> [accessed: 15th March 2011].

Scharl, A., and K. Tochtermann, *The geospatial web. How geobrowsers, social software and the Web 2.0 are shaping the network society* (London 2007).

Shirky, C., 'The political power of social media: technology, the public sphere and political change', *Foreign Affairs* 90.1 (2011) 28-41.

Siorpaes, K., and M. Hepp, 'myOntology: the marriage of ontology engineering and collective intelligence', in *Proceedings of Bridging the Gap between Semantic Web and Web*, (2007).

Talbert, R. J. A., ed., *Barrington atlas of the Greek and Roman world.* (2000).

Tarte, S., 'Digitizing the act of papyrological interpretation: negotiating spurious exactitude and genuine uncertainty', *Literary and Linguistic Computing* 26.2 (2011).

Tupman, C., 'Contextual epigraphy and XML: digital publication and its application to the study of inscribed funerary monuments', in *Digital Research in the Study of Classical Antiquity*, ed. G. Bodard and S. Mahony (Farnham 2010) 73-86.

INDEX

The cover image is of a torso of Pothos (Roman 1st century BC – 1st century AD) in the Museu Calouste Gulbenkian, Lisbon, Portugal.

Photo © Simon Mahony.
All rights reserved.

This edited volume collects together peer-reviewed papers that initially emanated from presentations at Digital Classicist seminars and conference panels.

This wide-ranging volume showcases exemplary applications of digital scholarship to the ancient world and critically examines the many challenges and opportunities afforded by such research. The chapters included here demonstrate innovative approaches that drive forward the research interests of both humanists and technologists while showing that rigorous scholarship is as central to digital research as it is to mainstream classical studies.

As with the earlier Digital Classicist publications our aim is not to give a broad overview of the field of digital classics; rather, we present here a snapshot of some of the varied research of our members in order to engage with and contribute to the development of scholarship both in the fields of classical antiquity and Digital Humanities more broadly.

BICS SUPPLEMENT 122
ISBN 978-1-905670-80-2
xvi + 162 pp, colour images, index

To find out more about our books, and our journal, the Bulletin of the Institute of Classical Studies, and to order books online, please visit our website.

You can also order books by emailing us at the
Publications Department,
Institute of Classical Studies,
School of Advanced Study,
University of London,
Senate House, Malet Street,
London WC1E 7HU, UK

web https://ics.sas.ac.uk/publications
email sas.publications@sas.ac.uk