DACURA: A NEW SOLUTION TO DATA HARVESTING AND KNOWLEDGE EXTRACTION FOR THE HISTORICAL SCIENCES

Peter N. Peregrine, Rob Brennan, Thomas Currie, Kevin Feeney, Pieter François, Peter Turchin, and Harvey Whitehouse

Peter N. Peregrine, Lawrence University, 711 E. Boldt Way, Appleton WI 54911 and Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM, 87501 (peter.n.peregrine@lawrence.edu).

Rob Brennan, ADAPT, School of Computing, Dublin City University, Ireland (rob.brennan@dcu.ie)

Thomas Currie, Department of Biosciences, University of Exeter—Penryn Campus, Cornwall, TR10 9FE, UK (t.currie@exeter.ac.uk)

Kevin Feeney, Knowledge and Data Engineering Group, School of Computer Science and Statistics, Trinity College Dublin, Ireland (kevin.feeney@scss.tcd.ie)

Pieter François, Institute of Cognitive and Evolutionary Anthropology, Oxford University, Oxford OX4 1QH, UK. (pieter.francois@anthro.ox.ac.uk)

Peter Turchin, Department of Ecology and Evolutionary Biology, University of Connecticut, 75 N. Eagleville Road, Storrs, CT 06269-3042 (peter.turchin@uconn.edu)

Harvey Whitehouse, Institute of Cognitive and Evolutionary Anthropology, Oxford University, Oxford OX4 1QH, UK. (harvey.whitehouse@anthro.ac.uk)

Archaeologists are both blessed and cursed by the information now available through the Internet. We are blessed by the pure abundance of articles, images, and data that we can discover with a simple search, but we are also cursed by the difficult process of parsing those discoveries down to those of scholarly quality that relate to our specific interests. As one cure for this curse we introduce Dacura, a dataset curation platform designed to assist researchers from any discipline in harvesting, evaluating, and curating high-quality information sets from the Internet and other sources. We provide an example of Dacura in practice as the software employed to populate and manage the massive Seshat databank of historical and archaeological information.

Los arqueólogos se bendecido y maldecidos por la información ahora disponible a través de Internet. Somos bendecidos por la pura abundancia de artículos, imágenes y datos que podemos descubrir con una simple búsqueda, pero nosotros también estamos maldecidos por el difícil proceso de análisis de los descubrimientos hasta las de calidad académica que se relacionan con nuestros intereses particulares. Como una cura para esta maldición introducimos Dacura, una plataforma de comisariado de conjunto de datos diseñada para ayudar a los investigadores de cualquier disciplina en la recolección, evaluación y comisariado de sistemas de información de alta calidad de Internet y otras fuentes. Le ofrecemos un ejemplo de Dacura en la práctica como el software empleado para rellenar y gestionar el databank de Seshat masivo de información histórica y arqueológica.

If this paper seems like an advertisement for a particular software product we do not intend it to be taken that way. What we primarily want to do is introduce to archaeologists who may not be familiar with current developments in computer science a new way of harvesting, storing, and retrieving data from the internet that we believe has the potential to transform how archaeological literature reviews and data harvesting are done. These developments are reflected the two core features of the Dacura data curation platform—a "graphical" data structure (as opposed to the standard column and row data structure) and an automated process for weeding out the thousands of on-line and database hits not directly related to a problem of interest and/or of dubious accuracy. These developments have been put into practice in the Seshat databank. This is why we focus on Dacura and Seshat here. We are hopeful that other researchers will create different solutions to the problem we address, but Dacura and Seshat are available today and, we think, are potentially important resources for archaeologists.

We begin with the basic problem Dacura is intended to address: the overabundance of unevaluated information available to researchers. As an example, consider a researcher who wants to build a database on a particular topic, such as population estimates for the ancient North American city of Cahokia. If she were to simply type "Cahokia population" into Google, she would obtain over 200,000 results, some discussing the population of the modern Illinois town of Cahokia, and with no easy way of knowing which of the many thousands of results on ancient Cahokiawould provide the information she needs, nor which of them would provide reliable information. If this researcher were to use Google Scholar instead, the results would be fewer (around 6000), and although she could expect somewhat better quality, there would remain the daunting task of identifying papers and books directly relevant to her interests. Even JSTOR, with quality-ensured content, would proffer around 900 articles to churn through.

The example above illustrates a core problem in contemporary research: the Internet and open-access publishing provide researchers abundant information on virtually any topic of interest, but there is no quality assurance for Internet search results, and even where quality can be assumed (as in peer-reviewed open-access publications), the amount of information is often overwhelming. What is needed is a search tool that provides a middle-ground—easy searching, an assurance of quality, and a manageable body of results. Such a search tool requires a carefully designed hierarchical structure (ontology) to allow a scholar to easily dig down through results to those that are directly relevant to his or her research. This search tool also requires detailed indexing across result domains so that "apples" not only recovers all information on "apples" but also information that does not retrieve "oranges" when applied to particular domains. In other words, such a search tool must be able to apply an integrated thesaurus or set of thesauri as part of the basic search routine.

There are a number of extant search tools that provide this functionality: rapid retrieval of specific, quality information across domains. For example, eHRAF (Human Relations Area Files; hraf.yale.edu) maintains two archives of documents (ethnographic and archaeological, respectively) organized using detailed ontologies (the *Outline of World Cultures* and *Outline of Archaeological Traditions*) and employing a rich thesaurus (the *Outline of Cultural Materials*). Individual paragraphs from nearly three-quarters of a million pages of archaeological and ethnographic primary and secondary source documents are indexed in eHRAF and can be easily searched and retrieved at varying levels of detail using hierarchical and Boolean search strategies. The results are specific, of excellent quality and specificity, and manageable in number. However, the range of results is limited to the documents that have been included in the eHRAF database. The reason eHRAF provides such excellent information retrieval is that the

information has been extensively pre-processed to the extent that every document has been individually placed into the ontology and every paragraph in every document individually indexed by Ph.D.-holding anthropologists. In short, a huge amount of work is required to make search and retrieval easy, and that means the data provided by eHRAF grows slowly and eHRAF cannot afford to be open-source.

An alternative model of a search tool providing rapid retrieval of specific, quality information across domains is tDAR (the Digital Archaeological Record; www.tdar.org). Like eHRAF, entire documents (including raw datasets, shapefiles, and the like) are available through tDAR, and are organized within a basic ontology. Unlike eHRAF, these documents are not processed by tDAR staff (although there is review of the processing to ensure it has been done correctly), but rather the individuals who submit documents complete a metadata form which is attached to the document (Watts 2011). This allows the number of documents in tDAR to increase relatively rapidly, and also allows tDAR to remain open source (there are modest fees for contributing documents). However, because contributors provide the ontological and indexing information themselves, the level of detail and accuracy vary, meaning that searches may not retrieve all relevant documents. And, like eHRAF, the available information is limited to the documents within the database.

Open Context (www.opencontext.org) is another excellent open-source data repository for archaeology that is similar to tDAR, but which provides several additional features that expand its range beyond archaeological data. Like tDAR, archaeological data are contributed for a modest fee. Unlike tDAR, Open Source editors work with contributors to create the metadata and clean the data sources for publication on the web, and the data sources themselves are evaluated for their importance; that is, not all data sources are published, only those that peer

reviewers think will be of use to the broader field. Once incorporated into Open Context, data sources are linked to related data sources on the web by Linked Data standards (Kansa 2010). This allows Open Context to expand beyond the archived data, overcoming a limitation of both eHRAF and tDAR.

We present here what we argue is a more comprehensive approach to the problem of retrieving specific, quality information across domains than the three outlined above—a set of data harvesting, evaluation, assembly, and output processes that has been implemented in Dacura (dacura.cs.tcd.ie)and which itself is being employed as the managing software for the Seshat databank (seshatdatabank.info. By being developed and implemented as an integral part of a data-heavy research initiative, Dacura has benefitted from the on-going identification of problems and shortcomings that gathering and managing large and complex data entail, and thus serves as a good example of a resource that would be of use to academic researchers. We do not intend this article to be simply an advertisement for Dacura, but rather we use Dacura and the Seshat databank to illustrate an approach to harvesting, evaluating, and retrieving data from the Internet or any "big data" source that has been made possible by new advances in computer science and that we believe will have profound impact on archaeological research.

DACURA

Dacura is a dataset curation platform designed to assist researchers from any discipline in creating and curating high-quality datasets. The basic idea is simple—the researcher starts by defining the precise structure of the dataset that they would like to collect. The system uses this detailed information to support the user in discovering, harvesting, filtering, correcting, refining

and analyzing information from the web in order to compile the highest quality information possible with which to populate the dataset. The details provided by the researcher includes basic information such the definition of the fundamental entities of interest (e.g. Mississippian archaeological sites), the properties of those entities in which she or he is interested (e.g. population estimates), the datatypes and desired units of each property (e.g. descriptions, counts, kilos) relationships with other entities both within and beyond the dataset itself (e.g. Mississippian incorporates Cahokia; Cahokia isDescribedBy http://wikipedia/Cahokia).

The process of defining the structure of the desired dataset is one of Dacura's strengths. In order for Dacura to search the Internet to harvest data, the desired data must be carefully considered by the researcher, as well as the structure and type of data desired. This process forces the researcher to carefully design his or her research question and to carefully consider what types of analyses are necessary to answer that question. It is important to point out, however, that because Dacura provides such a flexible search structure, data harvesting can be an iterative process, changing as data are evaluated and as questions become more focused.

Dacura encodes the structure of the dataset defined by the researcher as a semantic web ontology according to the Web Ontology Language (OWL) standard of the World Wide Web Consortium (W3C), the main international standards body for the web. OWL is a rich and flexible language which allows a wide variety of constraints and inference rules to be specified on the data to be collected (e.g. the population of a site should not be greater than the population of the region that it is in). In contrast with the unstructured natural language strings that drive search engine results, the highly structured and precisely specified nature of ontological dataset specifications can be exploited by the computer to provide much greater specificity in results.

The richer the structural specification, the easier it is for the system to automate the harvesting of data and the generation of useful tools with which to analyze, improve and curate it over time.

Dacura is based on semantic web technology. At its core is a Resource Description Framework (RDF) triplestore, a specific form of graph database (as opposed to a two-dimensional column and row database used in most spreadsheets) in which data are identified by a subject-predicate-object combination like "Cahokia is Mississippian", "Mississippian has City", or "Cahokia is City" (www.w3.org/TR/rdr11-concepts/). The subject-predicate-object structure can be understood as nodes-edges-properties within a three-dimensional graph which represents and stores data. The graphic structure of an RDF triplestore allows for index-free adjacency, meaning every subject-predicate-object triple directly links to related subject-predicate-object triples so that no index lookups are necessary. In the example above, Mississippian, Cahokia, and City are all linked so that no indexed search is required to identify Cahokia as a Mississippian city.

OWL ontologies are used in Dacura to enable semantic reasoning in quality control and data harvesting. Dacura is designed to produce and consume data in line with the linked open data principles. This makes it easy to import information from existing structured information sources and to enrich curated datasets by interlinking them with publicly available Linked Data sources (e.g. DBpedia or wikidata, the linked data versions of Wikipedia) and datasets curated by Dacura can similarly be easily linked. An overview of Dacura's structure is presented in Figure 1.

[Figure 1 about here]

The Dacura workflow breaks the process of dataset creation and curation down into 4 stages, as illustrated in Figure 2. The first stage is data harvesting: identifying sources of high

quality information with which to populate the dataset. Dacura supports a number of approaches to data harvesting: from identifying relevant data in known public data sources, to deploying agents to search the Internet,, to manual specification of information sources by curators. The goal of the system is to automate, as much as possible, the identification of the sources of information that will be needed in order to populate the dataset. In this stage, the goal is not to find documents about the entities in which one is interested, but to find specific sources of information which can populate the properties and relationships that a researcher has defined in their dataset specification.

[Figure 2 about here]

The second stage in the Dacura dataset creation and curation process is knowledge extraction. This involves extracting the precise information from harvested sources into the structure required by the researcher's dataset specification. Although Natural Language Processing and other artificial intelligence technologies continue to improve all the time, they remain error prone and thus, in order to produce high quality data, some human input is normally required to filter out false positives. Dacura employs tools to support both human users and automated agents in screening, filtering, improving, annotating and interlinking candidate records to produce knowledge reports; that is, authoritative accounts of the relevant knowledge contained in a source, enriched through links into the web of data.

The third stage in the Dacura process is perhaps the most important for ensuring data quality: expert analysis. Dacura focuses strongly on dataset quality, providing both automatic and manual tools to ensure that datasets provide accurate and complete data that conforms to the dataset specifications. Initial data evaluation is performed through automated tools, which use semantic consistency checking and validity testing to reconcile various data points into a

composite account that represent the tools' best estimate of authoritative data which accurately represents reality. These composite accounts are reviewed by experts in the data domain (like our hypothetical researcher interested in Cahokian population), allowing the expert to correct misinterpretations and identify disagreements between the expert and the automated tools. Experts can create their own personal interpretations (for example, by specifying that only particular sources should be trusted) and overlay this on the dataset to produce a custom dataset, representing their view on what the data should be. The experts currently volunteering as data evaluators for the Seshat data bank are listed at http://seshatdatabank.info/seshat-about-us/contributor-database/. The number (77 at the time of this writing) and range of expertise of these volunteers illustrates that it is quite feasible to incorporate expert evaluation into a data harvesting system like Dacura.

Finally, Dacura supports a variety of output tools to make datasets available to third parties in a range of formats. Dacura publishes its curated datasets as Linked Data and provides a SPARQL endpoint, a query language for RDF graphs, which supports sophisticated filtering and retrieval of data. This allows intelligent applications to interact with the datasets in unforeseen ways. These datasets are produced in accordance with the principles of Linked Data which allows them to interact with the wider semantic web. For human users, Dacura can produce graphs, charts, maps and other visualizations to provide users with easy-to-understand insights into the data in a dataset. Data for graphs or other outputs can be browsed, searched, and selected providing users with the ability to access the sections of datasets they find most useful. Dacura also allows datasets or subsets of them to be exported in a wide range of formats for external analysis, including geographic information systems and statistical packages such as SPSS and R.

IMPLEMENTING DACURA: THE SESHAT META-MODEL

As an example of how Dacura works in practice, Figure 3 shows the meta-model being used to implement Seshat: Global Historical Databank (Turchin et al. 2015, 2016). Seshat (seshatdatabank.info) is intended to bring together into one place a comprehensive body of knowledge about human history and prehistory for the purpose of empirically testing hypotheses about cultural evolution (e.g. "How and under what circumstances does prosocial behavior evolve in large societies?" "What roles do religion and ritual activities play in group cohesion and cultural development?" "What is the impact of climatic and the environmental factors in societal advance?"). Testing these hypotheses with appropriate statistical techniques requires data that are both valid and reliable; that is, data that define what the researcher thinks they define ("apples" rather than "oranges") and that are measured in the same manner across cases. Cases must also be chosen carefully to ensure that units of analysis are equivalent across cases and domains Figure 4 presents an overview of Seshat.

[Figures 3 and 4 about here]

There are two fundamental pieces of information upon which Seshat cases are based: a **Location** and a **Duration**. A location is a point or polygon anywhere on the earth's surface, and defines an entity called a **Territory**. Three entity classes of Territory have been defined in Seshat (more may be defined later as Seshat expands):

(1) *Natural-Geographic Areas* (NGA), which are a contiguous area roughly 100 by 100 kilometers encompassing a reasonably homogenous ecological region.

(2) *Biomes*, which encompass a contiguous biotic region or region of similar climatic conditions.

(3) *World Regions*, which may be pre-defined entities such as nations or states, or can be defined by other specific criteria.

A Duration can be a single date or a date range. Adding a Duration to a Territory entity class defines one of two temporally bounded entities: (1) a **Human Population,** which is group of humans in a defined territory during a specified period of time; and (2) an **Event**, defined as an occurrence taking place in a specific territory in a specific period of time.

Seshat provides the ability to create entity classes within Human Populations and Events for specific research questions. Within the Human Population entity, current entity classes are:

(1) *Tradition*, which is defined as a human population "sharing similar subsistence practices, technology, and forms of socio-political organization that are spatially contiguous over a relatively large area and which endure temporally for a relatively long period of time" (Peregrine and Ember, 2001:ix). For this entity class there is a formal sampling universe for selecting cases, the *Outline of Archaeological Traditions* (hereafter OAT) (hraf.yale.edu/online-databases/ehraf-archaeology/outline-of-archaeological-traditions-oat/) and a formal thesaurus for coding data, the *Outline of Cultural Materials* (hereafter OCM) (hraf.yale.edu/online-databases/ehraf-world-cultures/outline-of-cultural-materials/).

(2) *Cultural Group*, which is a human population sharing norms, beliefs, behaviors, values, attitudes, etc. The primary sampling universe for this entity class is the *Outline of World Cultures* (hereafter OWC)(Murdock, 1983) and the thesaurus is the OCM.

(3) *Polity,* which is a human population that is a politically independent unit with a shared system of governance. This is an example of an entity class created for a specific research project. The sample consists of 30 cases selected for characteristics of sociopolitical organization and geographic location (Turchin et al., 2015). The primary thesaurus for this entity class is the OCM.

(4) *Settlement,* is a human population in a physical location and material facilities ranging in size and complexity from a temporary camp to a great metropolis. Because of the great range of settlements that could be coded, there is no defined sampling universe for the entity class. The primary thesaurus is again the OCM.

(5) *Identity Group,* which is a human population with a shared sense of being part of the same group. Like Polity, this entity class was created for a specific set of research projects and the sample is opportunistic (see Whitehouse, Francois, and Turchin, 2015). There is no formal thesaurus, though the OCM is used for some domains.

(6) *Linguistic Group*, which is a human population with a common language. The sampling universe for this entity class is *Ethnologue* (www.ethnologue.com), but there is no formal thesaurus (again, the OCM is being used for some domains).

In addition, subclasses can be added to entity classes to provide for more specific sets of data. Figure 5 shows entity subclasses that have been created for the current entity classes listed above.

[Figure 5 about here]

The Event entity obviously encompasses an almost infinite range of possible entity classes and subclasses. To maintain some order the event class in DBpedia is used

(mappings.dbpedia.org/server/ontology/classes/) as a basic ontology. As shown in Figure 5, the current entity classes for the Event entity include:

(1) *Inter-group Conflict*, such as a war, a battle, a feud, or the like.

(2) *Socio-Natural Disaster*, such as a famine, or epidemic.

(3) *Natural Disaster*, such as a drought, a flood, an infestation, a volcanic eruption, etc.

(4) Societal Collapse

(5) *Transition Ritual*, such as a marriage, a coronation, or an initiation.

(6) *Social Movement*, including physical movements like migration, but also social movements such as revitalization, millenarianism, strikes, etc.

(7) *Technological*, such as inventions, discoveries, innovations, and the like.

The Event entity classes have little data at present, but the classes will be populated (and new entity classes possibly created) as data on the Human Population entity are generated.

Populating Seshat: The Dacura Workflow Model

Figure 6 illustrates how archaeological data information for the Tradition entity class is incorporated into the Seshat databank through Dacura. The area in the blue rectangle can be entirely automated, while the area outside the blue rectangle requires analysts and experts to ensure the Seshat data are valid and reliable. Starting at the top of the blue rectangle, a Human Population entity is defined by a Duration within a Territory. The characteristics of the Human Population entity are then classified through the OAT thesaurus to define a Tradition entity class. Data mining begins by searching the Internet for Cultural Domain information as classified through the OCM thesaurus. At this point a researcher can also search for Cultural Domain information through both Internet and print sources. Information on a specific Cultural Domain,

identified in Figure 6 as Archaeological Data, is compared with values in DBpedia to determine if linked values should be included from other sources, and then output from Dacura and evaluated by an analyst for consistency. The output is next sent to the researcher or an expert on the Cultural Group or Cultural Domain for evaluation. The researcher or expert either decides upon a canonical value for the Cultural Domain or, if there are conflicts that cannot be resolved, a non-canonical value is given. In either case, the value is included in Seshat, and marked as either canonical or non-canonical. Canonical values are also exported to DBpedia to assist other researchers and future searches.

[Figure 6 about here]

Using Seshat: Outputs from Dacura

Our hypothetical researcher wanting to build a dataset about population estimates for Cahokia (which is included in Seshat as polity) would use the OAT to identify a Tradition entity class to search based on its Location and Duration:

OAT Tradition (Entity Class): Mississippian

Location (NGA Territory Class): Central Mississippi River Valley from northern Illinois to the Gulf, and the Ohio River valley south throughout southeastern North America.

Duration: 150-1500 CE

She might then narrow the search by either finding pre-defined entity sub-classes or define her own, such as:

Sub-Tradition (Tradition Entity sub-Class): Cahokia Mississippian

Location (Polity Human Population Entity Class): Greater Cahokia (a 15km diameter region centered on the Cahokia Central Administrative Complex, located in Collinsville, Illinois).

Duration: 150-1500 CE

Or, perhaps even more specifically:

Phase (Tradition Entity sub-Class): Lohmann-Stirling

Location (Settlement Human Population Entity Class): Cahokia Central Administrative Complex in Collinsville, Illinois.

Duration: 1050-1200 CE

Once having identified the Entity Class or sub-Class upon which to seek information, our researcher would then employ the OCM, another existing definition, or a definition he or she creates themselves to identify a Cultural Domain upon which to search for information on population. If using the OCM, that Cultural Domain would be:

OCM 161: Population. This category includes enumerations and estimates (with dates); density (e.g., arithmetical, for arable land); population trends; etc.

A search on this Cultural Domain would then produce Archaeological Data on Population for the researcher-defined Entity Class. This information may have previously been evaluated through the Dacura workflow process, and/or the search request might initiate a new Dacura search that would run through the process before providing output to the researcher.

[Figure 7 about here]

Our researcher would have a wide range of possible outputs from her search. As noted earlier, Dacura publishes datasets as Linked Data and employs SPARQL for output. SPARQL is a query language for RDF graphs which can produce documents and raw data sets but also

graphs, charts, maps and other visualisations. Important for archaeologists, SPARQL works with GeoSPARQL to allow data integration into geographic information system using well-understood OGC query standards (GML, WKT, etc.). Raw textual or numeric data produced through Dacura can be browsed, searched, and selected, allowing our researcher the ability to access the sections of texts or datasets she finds most useful. Dacura also allows texts or datasets (or subsets of them) to be exported in a wide range of formats for external analysis. For example our researcher might want numerical data on population estimates as output for statistical analysis. Dacura would produce a comma-delimited file that could be ported directly into R, and our researcher could then run any analysis they required to answer their question. Figure 8 shows a simple R line graph of Cahokia population estimates derived through Dacura using Seshat. This is not a particularly impressive result in itself, but consider that our researcher would have been able to compile these data in a matter of hours and be confident of their quality.

CONCLUSION

The Internet provides scholars abundant information, but many times the information is too abundant, and usually lacks quality control. Dacura was designed to address this problem. It provides a way to harvest information from the Internet easily, with an assurance of quality, and with a manageable body of results. Dacura's carefully designed ontology allow researchers to immediately identify and retrieve information directly relevant to his or her research. Dacura's integrated thesauri and RDF triplestore structure removes the need for detailed indexing across result domains so that all information on a given subject, even information that might not be obviously related or indexed as related, is retrieved. And Dacura offers a wide range of possible

outputs, from texts to visualizations to spreadsheets. Dacura is not the only data harvesting and curation package available, but because it has been developed hand-in-hand with the Seshat databank, it provides both a unique model for new computer-based methods of archaeological data handling.

In this way, Dacura represents an important new tool for archaeologists. As Kintigh et al. (2015:3) have recently pointed out "archaeologists are increasingly challenged as they acquire, manage, and analyze large volumes of disparate data." Dacura provides one answer to this problem. Dacura allows archaeologist to harvest data from established resources like tDAR and HRAF as well as input their own data and to identify data sources on the Internet that might not be otherwise discovered. Dacura allows the enormous volume of data available to archaeologists to be quickly and easily reduced to the most important information on a given question, and then output in a variety of useful formats. Dacura also provides a flexible data management tool that, perhaps most importantly, provides access for other researchers on the Internet to data that has already harvested and evaluated (such as those in Seshat).

In addition, Dacura represents a way to provide useful and accurate archaeological data to scholars who are not archaeologists. It has long been a frustration among archaeologists that our data, which can provides both a diachronic record of cultural stability and change and empirical examples of practices that have been successful or unsuccessful in human societies, has not been widely used outside of archaeology. But it is also not surprising, as archaeological data can be hard to access and hard to understand by non-archaeologists (Kintigh et al. 2015: 2). By providing a semi-automated means of harvesting, evaluating, and exporting archaeological data, Dacura provides both a means and a model for economists, political scientists, ecologists, geographers, and others to access and explore the rich and valuable record of the human past.

It is important to note, however, that there is a danger in data banks like Seshat and programs like Dacura that populate them. Because data are readily available through established data banks, researchers may choose to use the extant information rather than to create a new dataset structure and perform a new search. This tends to codify existing data, when in fact those data can be erroneous or superseded by new information. One of the benefits of tools like Dacura is that because they can make specifying dataset structure and performing searches both highly specific and relatively simple, researchers will choose to create new dataset specifications rather than using existing ones that might be related to, but not specific to, their research questions. Researchers will thus be encouraged to develop new information for data banks like Seshat. This will prevent extant data from being codified and provide an ever-expanding realm of usable information on the human past for both archaeologists and non-archaeologists to employ.

Acknowledgements

The authors wish to thank the participants of a workshop held at the Santa Fe Institute May 4-6, 2015 during which the needs for harvesting and integrating quality information was discussed and the Seshat meta-model developed. The Seshat project is funded by the John Templeton Foundation. Dacura was developed in the Knowledge and Data Engineering Group at Trinity College Dublin as part of the ALIGNED project funded by European Union Horizon 2020 program under project number 644055 and the ADAPT Centre for Digital Content Technology, SFI Research Centres Programme (Grant 13/RC/2106) co-funded by the European Regional Development Fund.

Data Availability Statement

The Seshat data bank can be accessed at http://seshatdatabank.info/. Information on Dacura can be obtained at http://dacura.cs.tcd.ie/.

REFERENCES CITED

Murdock, George Peter

1983 *Outline of World Cultures, 6th edition*. Human Relations Area Files, New Haven, CT.

Kansa, Eric C.

2010 Open Context in Context: Cyberinfrastructure and Distributed Approaches to Publish and Preserve Archaeological Data. *SAA Archaeological Record* 10(5):12-16.

Kintigh, Keith W., Jeffrey Altschul, Ann Kinzig, W. Fredrick Limp, William K. Michner, Jeremy Sabloff, Edward Hackett, Timothy Kohler, Bertram Ludäscher, and Clifford Lynch

2015 Cultural Dynamics, Deep Time, and Data. *Advances in Archaeological Practice* 3(1):1-15

Peregrine, Peter N. and Melvin Ember (editors)

2001 *Encyclopedia of Prehistory*, 9 vols. Kluver Academic / Plenum Publishers, New York.

Turchin, Peter, Rob Brennan, Thomas Currie, Kevin Feeney, Pieter François, Daniel Hoyer, Joseph Manning, Arkadiusz Marciniak, Daniel Mullins, Alessio Palmisano, Peter Peregrine, Edward A.L. Turner, and Harvey Whitehouse

2015 Seshat: The Global History Databank. *Cliodynamics* 6(1): 77-107

François, Pieter. Joseph Manning, Harvey Whitehouse, Rob Brennan, Thomas Currie, Kevin Feeney and Peter Turchin.

2016 A Macroscope for Global History: Seshat Global History Databank. *Digital Humanities Quarterly* 10(4).

Watts, Joshua

2011 Building tDAR: Review, Reduction, and Ingest of Two Reports Series. *Reports in Digital Archaeology* 1:1-15.

Whitehouse, Harvey, Pieter François, and Peter Turchin

2015 The Role of Ritual in the Evolution of Social Complexity: Five Predictions and a Drum Roll. *Cliodynamics* 6(2):199-216.

Figure 1: An overview of Dacura (https://youtu.be/MQo4FFLAcPE)

Figure 2: The four stages of the Dacura data curation process.

Figure 3: Metamodel for Seshat: The Global History Databank

Figure 4: An introduction to Seshat (https://player.vimeo.com/video/132643397)

Figure 5: Detail of the Human Population entity, showing current entity classes and subclasses.

Figure 6: Workflow for the incorporation of numerical data for the Archaeological Tradition entity class into Seshat through Dacura.

Figure 7: Population dynamics at Cahokia from 150 to 1500 CE.

Figure 2

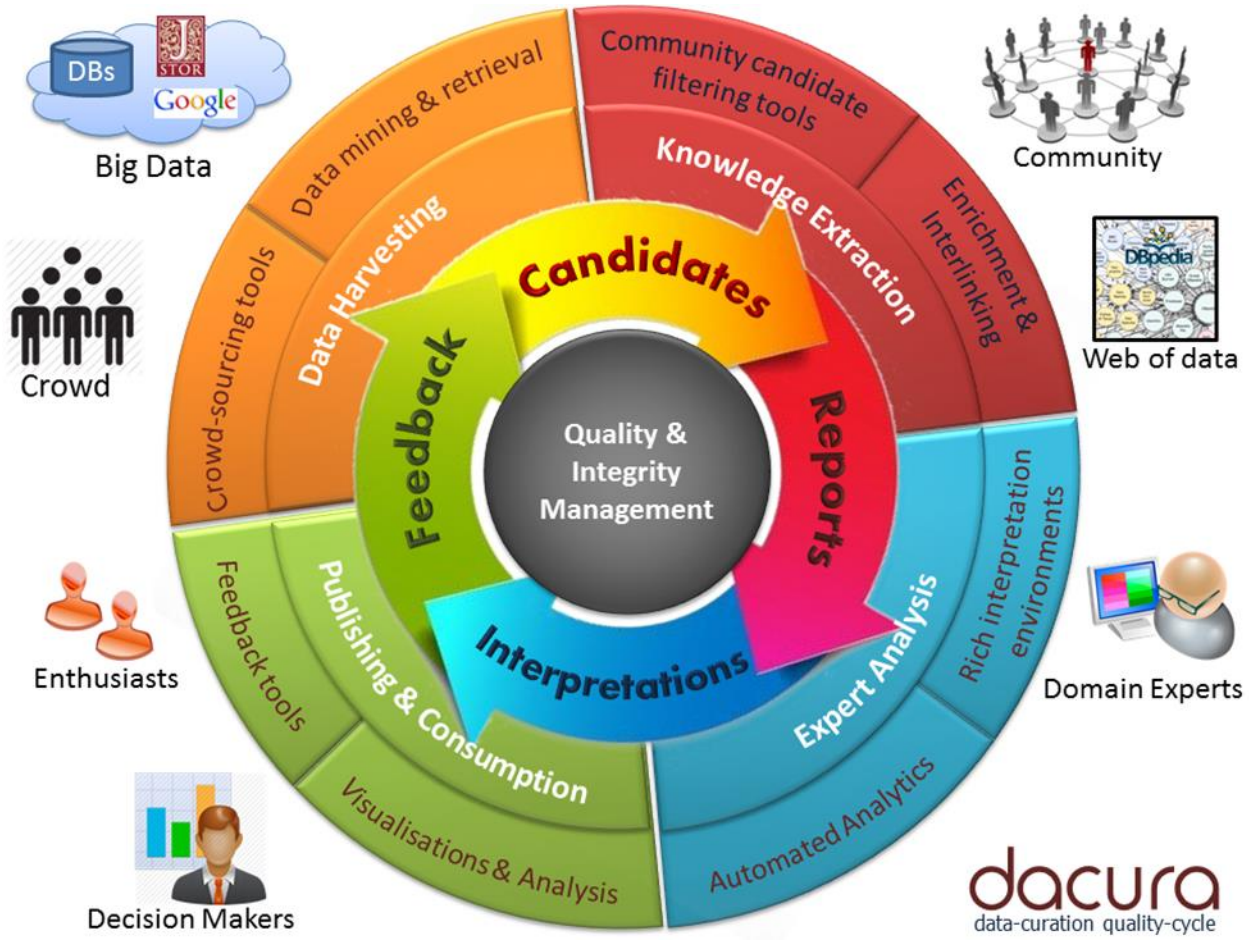

DBs
JSTOR
Google
Big Data
Community
Crowd
DBpedia
Web of data
Enthusiasts
Domain Experts
Decision Makers
Data mining & retrieval
Community candidate filtering tools
Crowd-sourcing tools
Data Harvesting
Knowledge Extraction
Enrichment & Interlinking
Candidates
Feedback
Reports
Quality & Integrity Management
Interpretations
Feedback tools
Publishing & Consumption
Expert Analysis
Rich interpretation environments
Visualisations & Analysis
Automated Analytics
dacura
data-curation quality-cycle

Figure 3

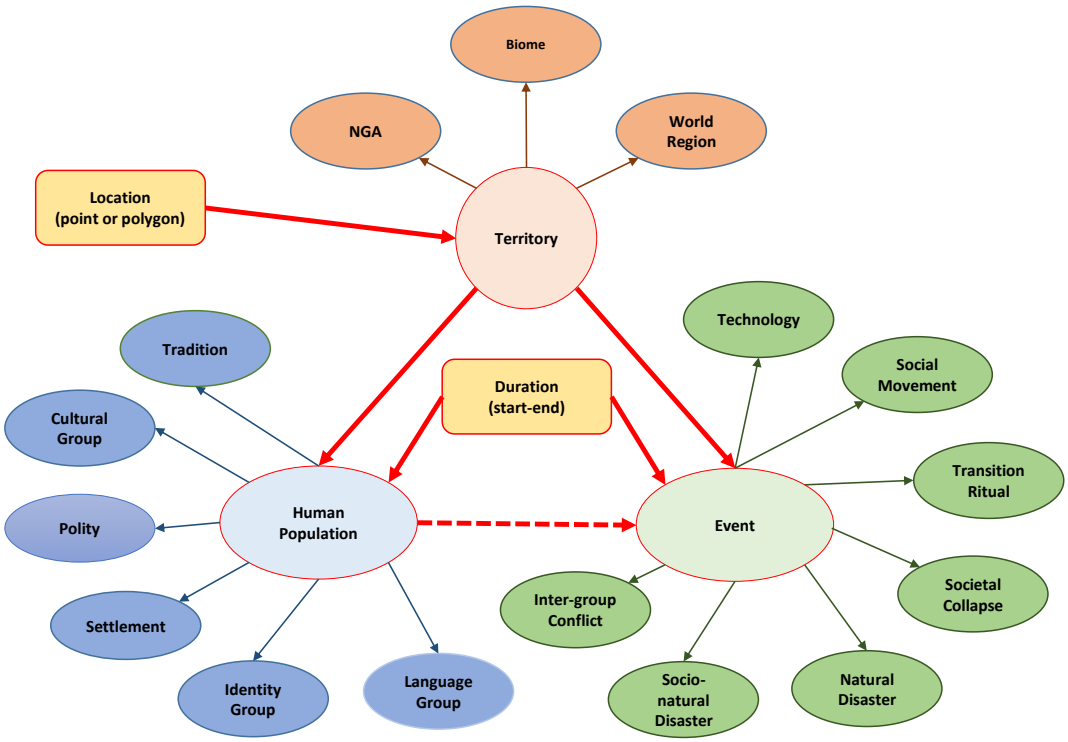

Biome
NGA
World Region
Location (point or polygon)
Territory
Technology
Tradition
Social Movement
Duration (start-end)
Cultural Group
Transition Ritual
Human Population
Event
Polity
Societal Collapse
Inter-group Conflict
Settlement
Identity Group
Language Group
Socio-natural Disaster
Natural Disaster

Figure 5

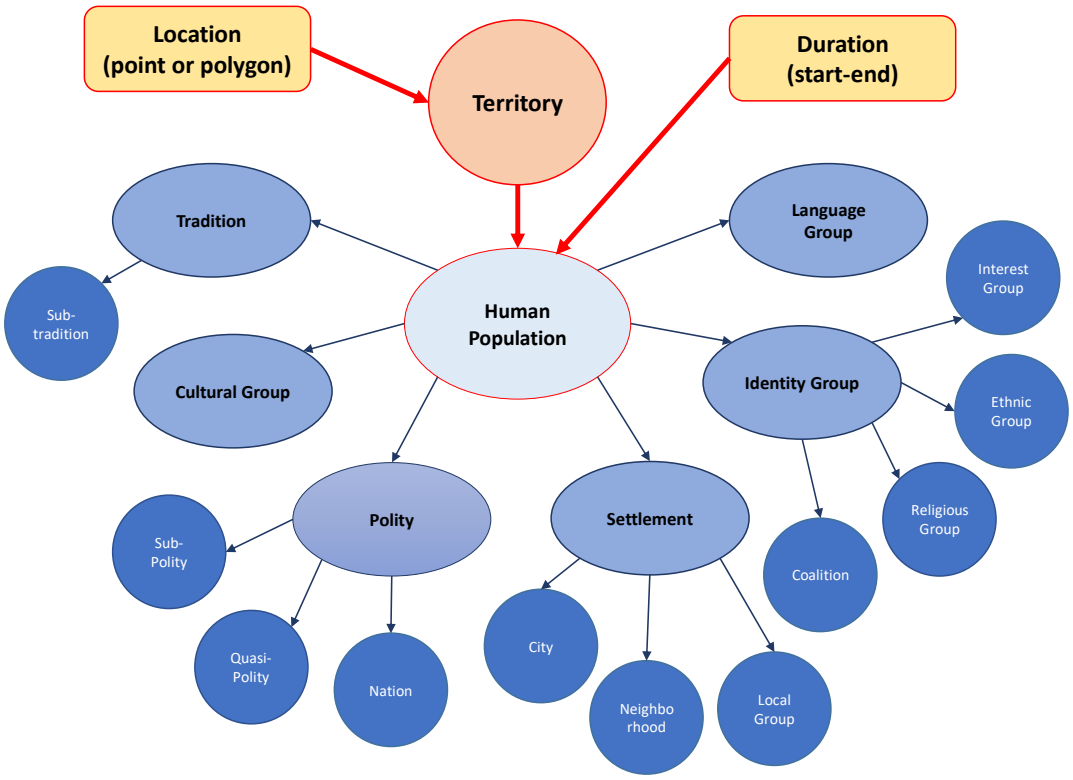

Location
(point or polygon)
Territory
Duration
(start-end)
Tradition
Sub-
tradition
Language
Group
Interest
Group
Human
Population
Cultural Group
Identity Group
Ethnic
Group
Religious
Group
Polity
Sub-
Polity
Quasi-
Polity
Nation
Settlement
City
Neighbo
rhood
Local
Group
Coalition

Figure 6

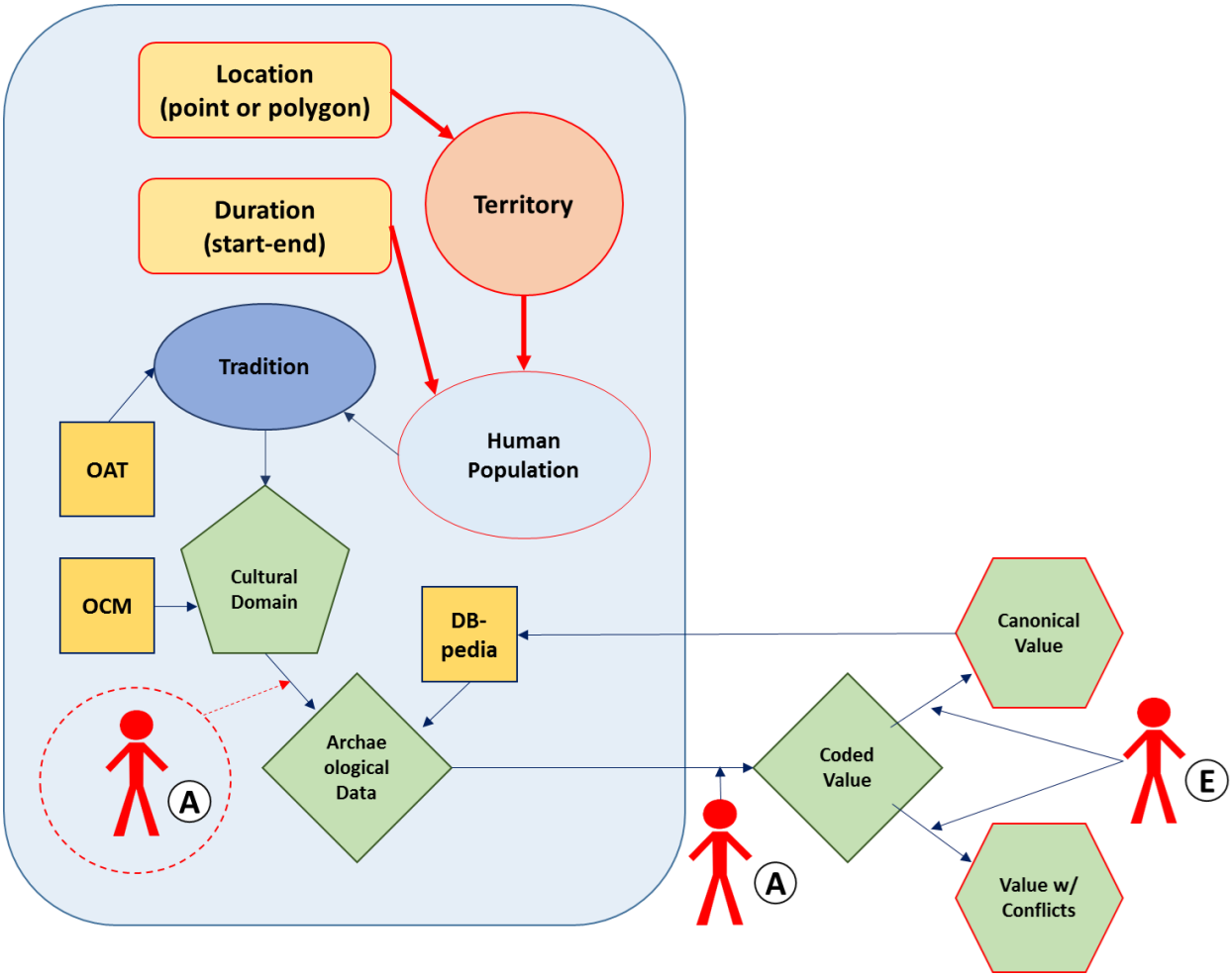
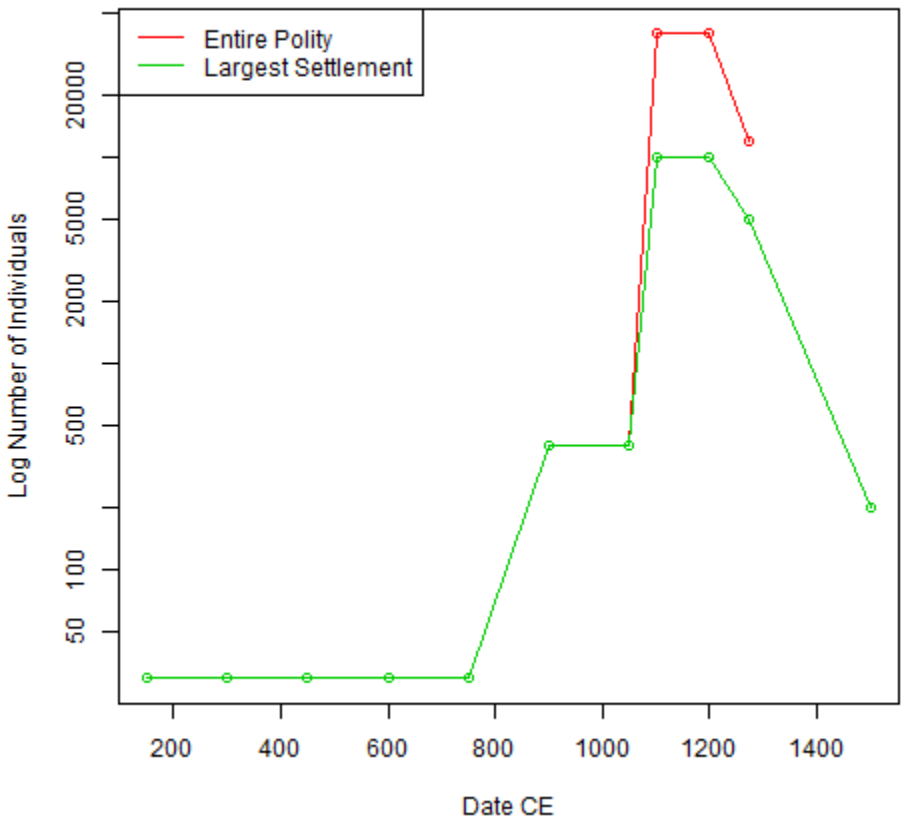

Location
(point or polygon)
Duration
(start-end)
Territory
Tradition
OAT
Human
Population
OCM
Cultural
Domain
DB-
pedia
Archae
ological
Data
A
Canonical
Value
Coded
Value
E
A
Value w/
Conflicts

Figure 8


Entire Polity
Largest Settlement
Log Number of Individuals
50
100
500
2000
5000
20000
200
400
600
800
1000
1200
1400
Date CE