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

Journal:	*Historical Methods: A Journal of Quantitative and Interdisciplinary History*
Manuscript ID	VHIM-2017-0034.R2
Manuscript Type:	Original Article
Keywords:	data harvesting, RDF triplestore, database ontology, database metamodels, data curation

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

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 & Knowledge and Data Engineering Group, School of Computer Science and Statistics, Trinity College Dublin, Ireland (rob.brennan@scss.tcd.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, School of Humanities, De Havilland Campus, University of Hertfordshire, Hatfield, AL10 9EU, UK and Institute of Cognitive and Evolutionary Anthropology, Oxford University, Oxford OX4 1QH, UK (pieter.francois@stb.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)

[†]Corresponding author

ABSTRACT

New advances in computer science address problems historical scientists face in gathering and evaluating the now vast data sources available through the Internet. As an example we introduce Dacura, a dataset curation platform designed to assist historical researchers in harvesting, evaluating, and curating high-quality information sets from the Internet and other sources. Dacura uses semantic knowledge graph technology to represent data as complex, inter-related knowledge allowing rapid search and retrieval of highly specific data without the need of a lookup table. Dacura automates the generation of tools to help non-experts curate high quality knowledge bases over time and to integrate data from multiple sources into its curated knowledge model. Together these features allow rapid harvesting and automated evaluation of Internet resources. We provide an example of Dacura in practice as the software employed to populate and manage the Seshat databank.

KEYWORDS

data harvesting; RDF triplestore; database ontology; database metamodels; data curation

Current developments in computer science provide new ways of harvesting, storing, and retrieving data from the Internet that have the potential to transform how literature reviews and data harvesting are undertaken in the historical sciences. Dacura is a data curation platform that uniquely reflects three of these developments: (1) a data-model based on a knowledge graph (as opposed to the standard column and row data structure), (2) the use of Web Ontology Language (OWL) for data definition, and (3) an automated process based on sematic reasoning for weeding out the thousands of on-line and database hits not directly related to a problem of interest and/or of dubious accuracy. Dacura was built in tandem with the Seshat databank, which is being constructed to coordinate quantitative historical and archaeological data to statistically test models of historical dynamics (Turchin et al. 2015). We introduce both Dacura and Seshat here as concrete examples of how the advances in computer science noted above might be employed by historical researchers.

We begin with the basic problem the Dacura data curation platform is intended to address: the overabundance of unevaluated information available to researchers. As an example, consider a researcher who wants to gather quantitative data to answer a specific question, let's say whether population increased on island of Hawaii prior to the emergence of states ca. 1500 CE, as would be predicted by a simple population-pressure model (e.g. Diamond 1997, 284-292). If she were to simply type "ancient Hawaii population" into Google, she would obtain nearly 250,000 results (some discussing modern demographics) with no easy way of knowing which of the many thousands of results on ancient

Hawaii would provide the information she needs, nor which of them would provide reliable information (the Wikipedia page on "Ancient Hawaiian Population", for example, provides only high estimates and apparently from only one source; the inability to clearly identify the source of the data is itself a serious problem). If this researcher were to use Google Scholar instead, the results would be fewer (around 165,000), 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 60,000 articles to churn through.

The example above illustrates a central problem in contemporary historical research: the Internet and open-access publishing provide researchers abundant information on virtually any topic of interest, but there is no quality assurance, 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 tool that allows a researcher to build a dataset containing high-quality structured information which provides specific pieces of information needed to answer such questions. 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 a search for Hawaii population estimates, as our hypothetical researcher might perform, not only recovers all information on Hawaii population estimates but does not also retrieve unrelated information on other demographics or other locations. In short, 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 quite a few extant search tools that provide this functionality: rapid retrieval of specific, quality information across domains. Considering only archaeological search tools (the first author is an archaeologist), a good example is eHRAF (Human Relations Area Files; hraf.yale.edu) which provides 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 archives. 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 of high-quality data easy, and that means the data provided by eHRAF grows slowly and eHRAF must charge a user fee.

An alternative model of a search tool providing rapid retrieval of specific, high-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 free retrieval of information from tDAR (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 data repository 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, but 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 while providing an assurance of data quality for those data contributed directly to Open Context, though not to linked data, which we see as a serious limitation.

We present here what we argue is a more comprehensive approach to the problem of retrieving specific, high-quality information across domains than the three outlined above (and there are many other excellent programs and data repositories we could have cited)—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) which we introduce below. By being developed and implemented in tandem with a data-heavy research initiative, Dacura has benefitted from the ongoing identification of problems and shortcomings that gathering and managing large and complex historical data entail, and thus serves as a good example of a resource that would be of use to academic researchers.

Knowledge graph technology is increasingly being used by companies such as Google (N.d.) and Facebook (N.d.) to manage and structure the vast and diverse sources of information they curate. Traditional SQL storage solutions based on tables and rows are insufficiently expressive to capture the structure or semantics of the information that they manage and the complexity of the relationships between things. However, to date, knowledge graph technology remains very technical and difficult to use except by very large and technically savvy corporations. Dacura's design goal is to help make knowledge graph technology available to researchers in such a way that it does not require tremendous expertise to use. Dacura achieves this by automating much of the software required to harvest and curate data from the semantic model, providing users with easy to use tools and interfaces which do not require them to understand the underlying technology.

We do not intend this article to be simply an advertisement for Dacura, but rather we use Dacura 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 the historical sciences.

DACURA

Dacura is a data curation platform designed to assist historical researchers in creating and curating high-quality datasets in the form of rich semantic knowledge graphs [Endnote 1]. The basic idea is simple—the researcher starts by defining the structure of the data that they would like to collect. The system uses this information to support the user in discovering, harvesting, filtering, correcting, refining and analyzing information from the Internet in order to compile the highest quality information possible. The details provided by the researcher include basic information such as the definition of the fundamental entities of interest (e.g. Hawaii), 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. numeric), and relationships with other entities both within and beyond the dataset itself (e.g. Hawaii is in Polynesia).

The process of defining the structure of the desired data is one of the strengths of this approach. All historical scientists know that data must be property questioned before they can speak. Because Dacura requires the precise shape of the desired data to be specified before performing a search, Dacura encourages researchers to think carefully about the nature of the desired data and the means by which they will be questioned before starting to collect those data. Such preparation before data collection saves time and effort in what is often the most difficult task in historical research—identifying useful information sources. Dacura's simple and user friendly interface (discussed below) makes the process of data definition easy, and, because Dacura provides a flexible search structure, this process can be iterative, changing as data are interrogated 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 town should not be greater than the population of the region that it is in). In contrast with the unstructured natural language strings that drive most 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 "Hawaii is Polynesia", "Hawaii has Island", or "Polynesia has Island" (www.w3.org/TR/rdf11-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, Polynesia, Hawaii, and Island are all linked so that no indexed search is required to identify Hawaii as a Polynesian Island.

OWL ontologies are used in Dacura to enable semantic reasoning in quality control and data harvesting; that is, if there are conflicts between triples Dacura identifies and marks them as conflicts for further evaluation (see dacura.scss.tcd.ie/ontologies/dacura-130317.ttl). 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. A video example of Dacura's operation is provided at https://youtu.be/AEb1wF3jAgk. A key aspect of Dacura is that harvested data, including those harvested through Linked Data, are evaluated for quality through both automated and human evaluators as part of the system's workflow. Thus Dacura not only harvests data quickly and easily, it also evaluates it for quality.

The Dacura workflow breaks the process of data creation and curation down into 4 stages, as illustrated in Figure 1. 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 automated agents to search the Internet, to manual specification of information sources by researchers. 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 data specification.

[Figure 1 about here]

The second stage in the Dacura data creation and curation process is knowledge extraction. This involves extracting the precise information from harvested sources into the structure required by the researcher's data 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 ensure 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. One of Dacura's strengths is its focus on data quality, providing both automated and manual tools to ensure that harvesting captures accurate and complete information that conforms to the researcher's data 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 Hawaii population estimates), 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.

In the case of Seshat experts are solicited by the Seshat management team based on their publication record in the area(s) they are asked to review. All have Ph.D.s in their field and the vast majority have academic appointments. Experts are periodically evaluated to identify individuals whose input appears inconsistent with harvested data on a regular basis. One might be concerned that developing a pool of experts by solicitation might not be effective, but the quality of those currently volunteering can be seen in the list 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. Most projects employing systems like Dacura will not likely be as large in scope as Seshat, and only one expert may be required to evaluate the data—perhaps only the researcher herself.

Finally, Dacura supports a variety of output tools to make data available to third parties in a range of formats. Dacura publishes its curated data as Linked Data so other users and platforms (such as Open Context, discussed above) can readily access it. Dacura also 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 data in unforeseen ways. For human users, Dacura can produce graphs, charts, maps and other visualizations to provide users with easy-to-understand insights into the data.

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 2 shows the meta-model being used to implement Seshat: Global Historical Databank (Turchin et al. 2015, Francois et al., 2016; see also dacura.scss.tcd.ie/ontologies/seshat-130317.ttl). Seshat (seshatdatabank.info) was designed 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, including the possible role of religion, ritual, warfare, agriculture and other clusters of variables in the rise of social complexity. To date Seshat has been used to identify a single dimension of complexity that explains about three-quarters of the variation in human social organization (Turchin et al. 2018) and has also been used to demonstrate that the hypothesized socio-political transformations of the Axial Age took place at varied times across Eurasia, encompassing more than a millennium (Mullins et al. 2018). Exploring such large-scale questions with appropriate statistical techniques requires data that are both valid and reliable; that is, data that define what the researcher thinks they define and that are measured in the same manner across cases.

Dacura, also under development when Seshat was being planned, seemed an excellent platform for Seshat. The two groups of researchers decided to develop Dacura and Seshat in tandem, so that both would inform the other. Because computer scientists and historical scientists do not regularly work together, the cooperation between the Dacura and Seshat research teams proved extremely fruitful, allowing the Dacura team to understand the needs of historical scientists, and the Seshat team to understand both the possibilities and limitations of triplestore data harvesting and curation.

[Figure 2 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 3 shows entity subclasses that have been created for the current entity classes listed above.

[Figure 3 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 2, 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.

Populating Seshat: The Dacura Workflow Model

As an example of how an historical scientist might employ Dacura to populate a dataset, Figure 4 illustrates how data 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 automated analysis 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 automated searching of the Internet for Cultural Domain information as classified through the OCM thesaurus. At this point a researcher can also have Dacura search for Cultural Domain information through both Internet and print sources. Dacura compares information on a specific Cultural Domain, identified in Figure 4 as Archaeological Data, with values in DBpedia to determine if linked values should be included from other sources, and then evaluated by an automated analyst for consistency. Inconsistent data (data with sematic or value conflicts) are output 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. Regardless, all the harvested values are included in Seshat and marked as either canonical or non-canonical so that other researchers can return to those values and revise or augment their interpretation. Canonical values are also exported to DBpedia to assist other researchers and future searches.

[Figure 4 about here]

Researchers may also input their own data, including images, media, and shapefiles. Incorporating new data allows other researchers to access it through Linked Data, making it widely available. Perhaps more importantly, data incorporated into Seshat or other Dacura-generated dataset can be continuously improved as Dacura allows researchers to comment on and re-evaluate previously incorporated data. We suggest this allows Dacura to not only create and curate large datasets, but also to ensure the continuous improvement of data quality.

Using Seshat: Outputs from Dacura

Our researcher interested in Hawaii population estimates would be able to quickly identify accurate and fully referenced estimates through Seshat (or her own Dacura-generated dataset). She would open Seshat through Dacura (a mock-up is provided at http://dacura.scss.tcd.ie/seshat/) [Endnote 2], select the Natural-Geographic Area for Hawaii, select the Polity sub-class of the Human Population inhabiting Hawaii for the time period of interest, and then select the population variable. A video illustrating this procedure is available at https://youtu.be/tZNcQNq2Mp0. The data on population she would obtain in this case would be taken from the Seshat data repository created with Dacura through the data harvesting and verification process described above. But our researcher could also create a new ontology using Dacura to conduct her own unique search, as discussed above and illustrated in the video mentioned earlier at https://youtu.be/AEb1wF3jAgk.

Our researcher would have a wide range of possible outputs from her search, whether in Seshat or with data she employed Dacura to harvest from other sources. 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 visualizations. Important for historical researchers, SPARQL works with GeoSPARQL to allow data integration into geographic information system using well-understood OGC query standards (GML, WKT, etc.). Raw textual, media, or numeric data produced through Dacura can be browsed, searched, and selected, allowing our researcher the ability to access the sections of texts, images, media, or datasets she finds most useful. Dacura also allows harvested or input material (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 a spreadsheet or statistical package and our researcher could then run any analysis she required to answer her question. Figure 5 shows a simple line graph of Hawaii population estimates derived through Dacura and Seshat with data output to a csv file and graphed using an Excel spreadsheet.

The answer to the researcher's initial question would be that population indeed increased prior to the first state system on the island of Hawaii, as expected under the population-pressure model. This is not a particularly impressive result in itself (and would in fact be quite simplistic—see Field, Ladefoged, and Kirch, 2011), but consider that our researcher would have been able to compile these data in a matter of minutes, be confident of their quality, and have access to all the metadata attached to them.

[Figure 5 about here]

CONCLUSIONS

The Internet provides historical scientists abundant information, but often the information is too abundant, and usually lacks quality control. Dacura was designed to address these problems. 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 (dacura.scss.tcd.ie/ontologies/dacura-130317.ttl) allows researchers to readily identify and retrieve information directly relevant to their 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 a unique model for new computer-based methods of historical and archaeological data handling.

In this way, Dacura represents a set of important new tools for the historical sciences. As Kintigh et al. (2015, 3) have recently pointed out "[historians and] archaeologists are increasingly challenged as they acquire, manage, and analyze large volumes of disparate data." Dacura illustrates several responses to this problem. Specifically, Dacura incorporates (1) a semantic knowledge graph technology based on an RDF triplestore, (2) the use of Web

Ontology Language (OWL) for data definition, (3) semantic reasoning as the foundation of an automated process for data evaluation, and (4) output following Linked Data standards.

OWL data definition and knowledge graph technology allow the enormous volume of data available to historical researchers to be quickly and easily reduced to the most important information on a given question, and then output in a variety of useful formats. Semantic reasoning provides mechanisms for rapid data evaluation and ongoing curation. And Linked Data standards allow other researchers to readily access data that has already been harvested and evaluated.

In addition, the recent advances in computer science available in platforms like Dacura offer a way to provide useful and accurate historical data to scholars who are not historical scientists. It has long been a frustration among historical researchers that data which can provide 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 and history. But it is also not surprising, as historical data can be hard to access and hard to understand by non-professionals (Kintigh et al. 2015, 2). By providing an automated means of harvesting, evaluating, and exporting historical data that has been evaluated for accuracy, platforms like Dacura offer both a means and a model for economists, political scientists, ecologists, geographers, and others to access and explore the rich and valuable record of human history.

ACKNOWLEDGEMENTS

This work was supported by a John Templeton Foundation grant to the Evolution Institute, entitled "Axial-Age Religions and the Z-Curve of Human Egalitarianism," a Tricoastal Foundation grant to the Evolution Institute, entitled "The Deep Roots of the Modern World: The Cultural Evolution of Economic Growth and Political Stability," an ESRC Large Grant to the University of Oxford, entitled "Ritual, Community, and Conflict" (REF RES-060-25-0085), a grant from the European Union Horizon 2020 research and innovation program (grant agreement No 644055 [ALIGNED, www.aligned-project.eu]), and an European Research Council Advanced Grant to the University of Oxford, entitled “Ritual Modes: Divergent modes of ritual, social cohesion, prosociality, and conflict.”

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. We gratefully acknowledge the contributions of our team of research assistants, post-doctoral researchers, consultants, and experts. Additionally, we have received invaluable assistance from our collaborators. Please see the Seshat website (www.seshatdatabank.info) for a comprehensive list of private donors, partners, experts, and consultants and their respective areas of expertise. Finally, we want to thank the anonymous reviewers whose insightful comments allowed us to substantially improve this paper.

ENDNOTES

1. The Dacura software was developed as part of the European ALIGNED Horizon 2020 project. All of the software, as well as other useful tools for managing semantic datasets and knowledge graphs, is available with an open source license through the project's web site at http://aligned-project.eu/open-source-tools/. However, this still requires users to configure and install their own knowledge-graph server, which remains a complex undertaking. It is our goal to also make the system available to researchers through the web as service in such a way that no technical knowledge is required to use it. We anticipate releasing a pilot version of this service in the middle of 2018. In the meantime, for any researchers who are particularly interested in seeing and using the web service, we are running an ongoing series of trials with Seshat and other collaborators and are open to new research collaborations. For updates check the Dacura website at http://dacura.cs.tcd.ie/ or by email at dacura@scss.tcd.ie.

2. The version of Dacura shown on the site is a mock-up that lacks full functionality and is intended solely to give readers a sense of what the Dacura interface and output might look like when installed on their institutional computer. This version searches only within a portion of the Seshat databank and only returns simple html output. As discussed in the article, the fully-functional version of Dacura does much more than the mock-up provided here.

REFERENCES CITED

Diamond, J. 1997. *Guns, germs, and steel*. New York: W.W. Norton.

Facebook, N.d. Introducing graph. https://www.facebook.com/graphsearcher/. Accessed 1/3/2018.

Field, J. S., T. Ladefoged, and P. Kirch. 2011. Household expansion linked to agricultural intensification during emergence of Hawaiian archaic states. *PNAS*, 108(18), 7327-7332.

François, P., J. Manning, H. Whitehouse, R. Brennan, et al. 2016. A macroscope for global history: Seshat Global History Databank. *Digital Humanities Quarterly* 10(4).

Google. N.d.. The knowledge graph. https://www.google.com/intl/bn/insidesearch/features/search/knowledge.html. Accessed 1/3/2018.

Mullins, D.A., D. Hoyer, C. Collins, T. Currie, et al. 2018. A systematic assessment of "Axial Age" proposals using global comparative historical evidence. *American Sociological Review* (in press).

Kansa, E.C. 2010. Open Context in context: Cyberinfrastructure and distributed approaches to publish and preserve archaeological data. *SAA Archaeological Record* 10(5):12-16.

Kintigh, K.W., J. Altschul, A. Kinzig, W.F. Limp, et al.. 2015. Cultural dynamics, deep time, and data. *Advances in Archaeological Practice* 3:1-15

Murdock, G.P. 1983. *Outline of world cultures, 6th edition*. New Haven, CT.: Human Relations Area Files.

Peregrine, P.N. and M. Ember (editors). 2001. *Encyclopedia of prehistory, 9 vols.* New York: Kluver Academic / Plenum Publishers.

Turchin, P., R. Brennan, T. Currie, K. Feeney, et al. 2015. Seshat: The Global History Databank. *Cliodynamics* 6: 77-107

Turchin, P, T. Currie, H. Whitehouse, P. Francois, et al. 2018. Quantitative historical analysis uncovers a single dimension of complexity that structures global variation in human social organization. *PNAS* 115(2): E144-E151. https://doi.org/10.1073/pnas.1708800115.

Watts, J. 2011. Building tDAR: Review, reduction, and ingest of two reports series. *Reports in Digital Archaeology* 1:1-15.

Whitehouse, H., P. François, and P. 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: The four stages of the Dacura data curation process.

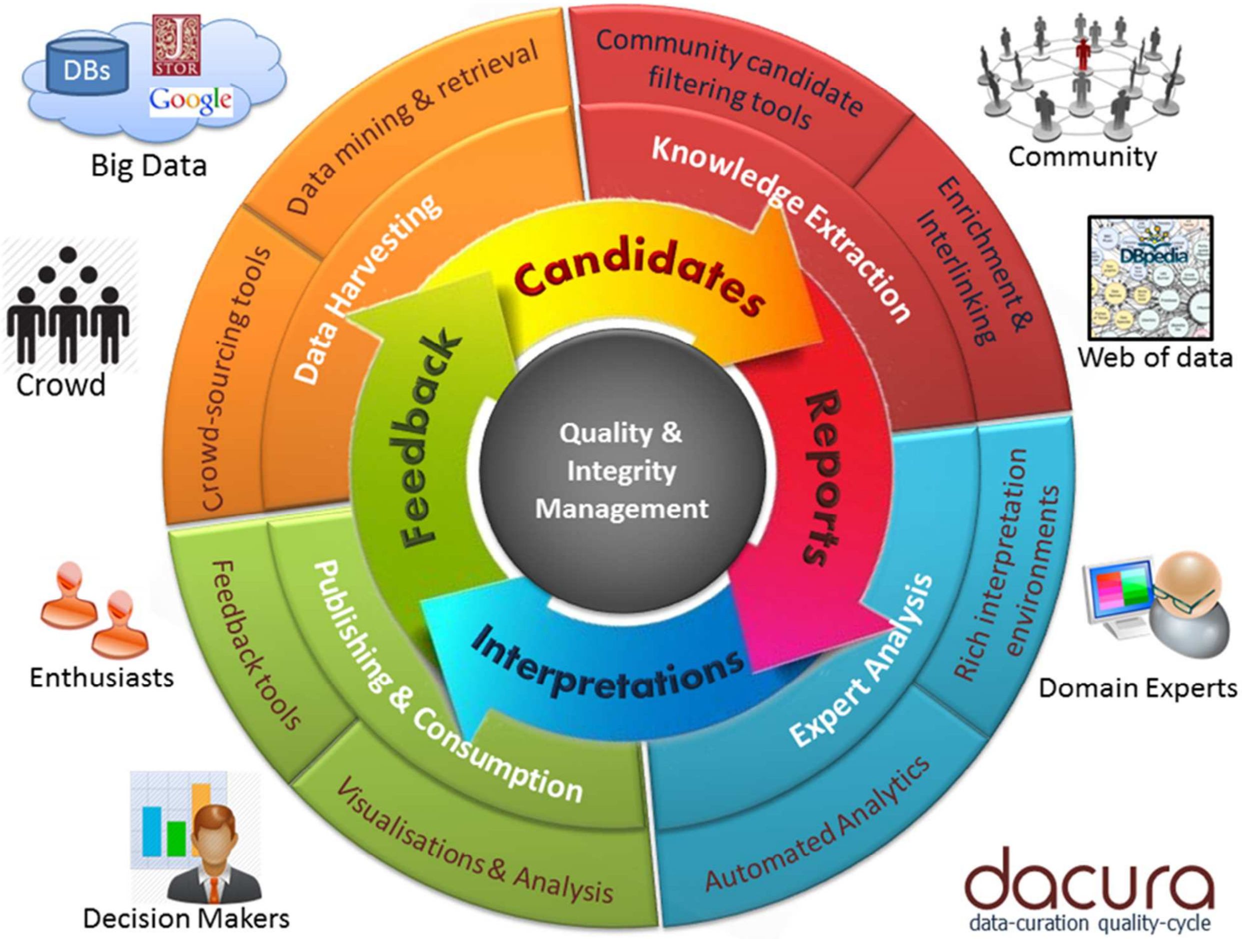


Figure 2: Metamodel for Seshat: The Global History Databank.

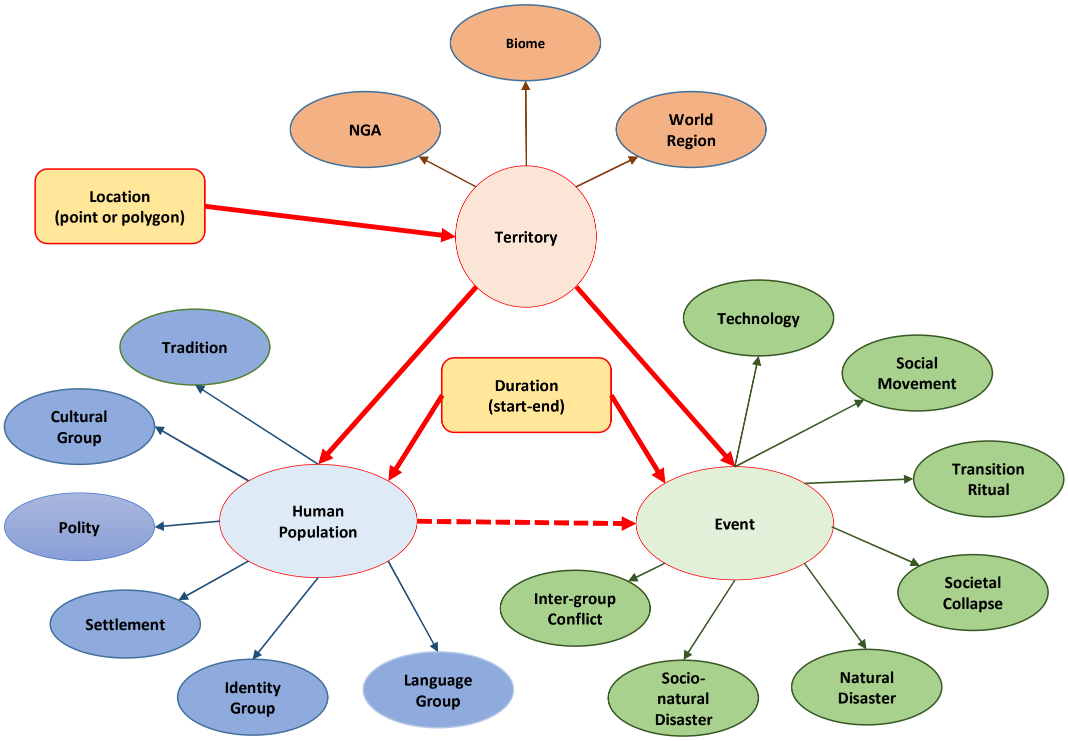


Figure 3: Detail of the Human Population entity, showing current entity classes and subclasses.

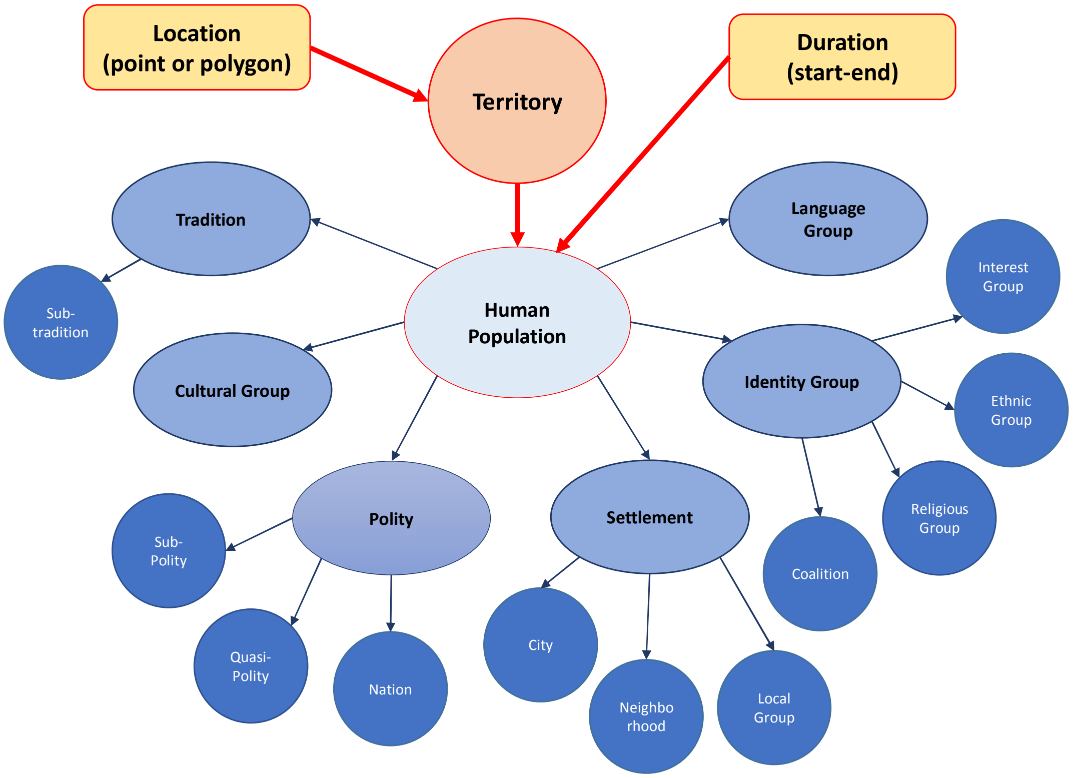


Figure 4: Workflow for the incorporation of numerical data for the Archaeological Tradition entity class into Seshat through Dacura. Human characters marked "A" represent automated data evaluators; the character marked "E" represents a human scholar or "expert" evaluator.

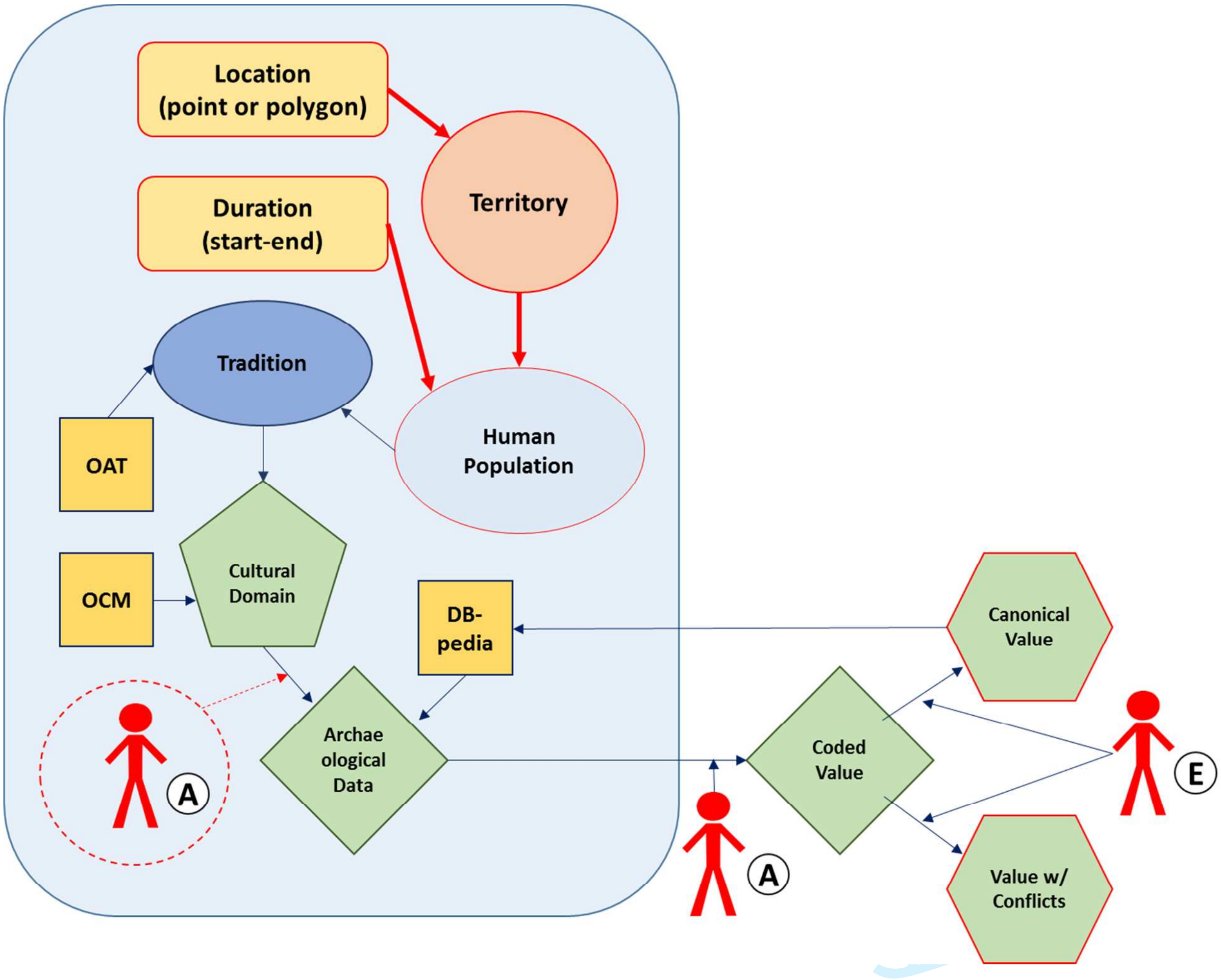
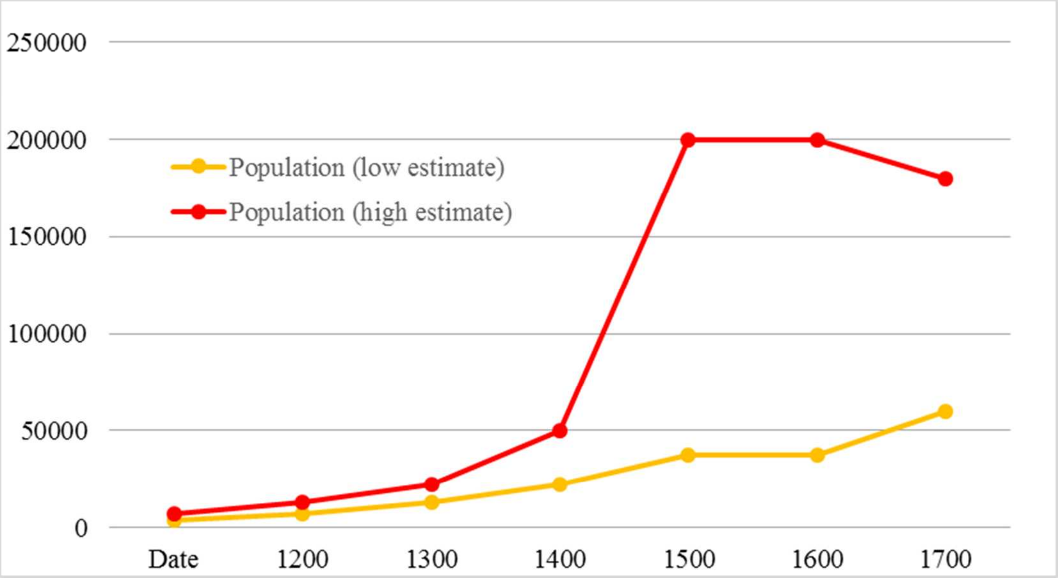


Figure 5: Population dynamics on Big Island Hawaii from 1200 to 1700 CE.


Population (low estimate)
Population (high estimate)
Date