

Nanopublication Beyond the Sciences

Patrick Golden, Ryan Shaw

The information expressed in humanistic datasets is inextricably tied to a wider discursive environment that is irreducible to complete formal representation. Humanities scholars must wrestle with this fact when they attempt to publish or consume structured data. The practice of “nanopublication”, which originated in the e-science domain, offers a way to maintain the connection between formal representations of humanistic data and its discursive basis. In this paper we describe nanopublication, its potential applicability to the humanities, and our experience curating humanities nanopublications in the PeriodO period gazetteer.

Nanopublication Beyond the Sciences

Patrick Golden¹ and Ryan Shaw¹

¹School of Information and Library Science, University of North Carolina at Chapel Hill

ABSTRACT

The information expressed in humanistic datasets is inextricably tied to a wider discursive environment that is irreducible to complete formal representation. Humanities scholars must wrestle with this fact when they attempt to publish or consume structured data. The practice of “nanopublication”, which originated in the e-science domain, offers a way to maintain the connection between formal representations of humanistic data and its discursive basis. In this paper we describe nanopublication, its potential applicability to the humanities, and our experience curating humanities nanopublications in the PeriodO period gazetteer.

Keywords: nanopublication, periodization, scholarly communication, time, Linked Data, JSON-LD

1 INTRODUCTION

2 Humanists seeking to integrate their work with digital tools face a common dilemma: How can one publish
3 structured data while keeping a connection to their discursive basis? The kind of information produced in
4 humanistic disciplines, such as biographical details, political and temporal boundaries, and relationships
5 between people, places, and events are inextricably tied to discursive arguments made by human scholars.
6 Converting all the information expressed in scholarly discourse into algorithmically-processable chunks
7 of formal, structured data has proved to be extraordinarily difficult, if not impossible.

8 Rather than attempting to exhaustively formally represent humanistic information, however, a scholar
9 can promote small pieces of information within a work using the practice of *nanopublication* (Mons and
10 Velterop, 2009). Nanopublication represents the provenance of structured assertions as a first-class citizen,
11 critically connected to the production of data. We believe that this emphasis on connecting assertions
12 with authors is well-suited for the needs of humanistic disciplines. By adopting the nanopublication
13 approach, creators of datasets in the humanities can focus on publishing small units of practically useful
14 curated assertions while keeping a persistent pointer to the basis of those claims—the discourse of scholarly
15 publishing itself—rather than its isolated representation in formal logic.

16 We offer an example of this approach in our description of the PeriodO period gazetteer, which collects
17 definitions of time periods made by archaeologists and other historical scholars. In constructing the
18 gazetteer, we sought to make period definitions parsable and comparable by computers while also retaining
19 the broader scholarly context in which they were conceived. We found that a nanopublication-centric
20 approach enabled this practice.

21 In this paper, we describe the concept of nanopublication, its origin in the hard sciences, and its
22 applicability to the humanities. We then describe PeriodO, a historical time period gazetteer we cre-
23 ated using the nanopublication approach. We discuss our experience mapping nonscientific data into
24 nanopublications and offer advice to other humanities-oriented projects attempting to do the same.

25 NANOPUBLICATIONS

26 Nanopublication is an approach to publishing research in which individual research findings are modeled
27 as structured data in such a way that they retain information about their provenance. This is in contrast to
28 both traditional narrative publishing, where research findings are not typically published in a structured,
29 computer readable format, and “data dumps” of research findings which are typically published without
30 any embedded information about their origin or production. The nanopublication approach is motivated by
31 a desire to publish structured data without losing the wider research context and the benefits of traditional
32 scholarly communication (Groth et al., 2010).

33 **Motivation**

34 Nanopublication emerged from a context of data-intensive sciences like genomics and bioinformatics,
35 where recent advances in computational measurement techniques have vastly lowered the barrier to
36 collecting genetic sequencing data. As a result, millions of papers have been published with findings
37 based on these new methods. However, the reported results are almost always published in the form of
38 traditional narrative scholarly publications (Mons et al., 2011). While narrative results can be read and
39 understood by humans, they are not so easily digested by computers. In fields where computationally has
40 been the key to the ability to ask new and broader questions, it should surely be the case that research
41 results are published in such a way that they are able to be easily parsed, collected, and compared by
42 computer programs and the researchers who use them.

43 On the occasions when research data are released and shared, they are often distributed on their own,
44 stripped of their necessary context within a broad research environment (the identity of the researchers,
45 where and how this research was conducted, etc.). In this case, publishing practice has swung too far to
46 the opposite extreme. In the service of creating and sharing discrete datasets, the published results have
47 been stripped of their provenance and their position within the wider scholarly endeavor that culminated
48 in their publication. This contextual information is crucial for researchers to determine the trustworthiness
49 of the dataset and learn about the broader project of research from which they resulted.

50 **Definition**

51 Nanopublication offers a supplementary form of publishing alongside traditional narrative publications. A
52 nanopublication consists of three parts, all representable by RDF graphs:

- 53 1. An assertion (a small, unambiguous unit of information)
- 54 2. The provenance of that assertion (who made that assertion, where, when, etc.)
- 55 3. The provenance of the nanopublication itself (who formed or extracted the assertion, when, and by
56 what method) (Groth et al., 2013)

57 By representing their research in nanopublications alongside their narrative reports, researchers can
58 publish their data in such a way that they remain within their human context while also being easily
59 digested by computer programs.

60 Authors are encouraged to include the smallest possible unambiguous pieces of information as the
61 assertions at the center of a nanopublication. This enables statements of the same fact to be connected
62 with different sources of provenance, thereby potentially augmenting the ability of consumers to judge
63 the quality of that assertion. Groth et al. (2010) call the collection of nanopublications all referring to
64 the same assertion “S-evidence”, and cite the potential benefits of the ability to automatically connect
65 findings across research publications.

66 **Uses**

67 Several European repositories of bioinformatic data have begun to publish their contents as nanopub-
68 lications, including the Biosemantics Group, neXtProt, and DisGeNET¹²³. These publications can be
69 aggregated and connected in larger systems, such as the decentralized reputation system described by
70 Kuhn (2015).

71 **NANOPUBLICATION IN THE HUMANITIES**

72 While the bioinformatics research community has enthusiastically adopted nanopublication, other disci-
73 plines have been slow to follow. Gradmann (2014) suggested that specialized and stable terminologies,
74 as well as sufficient funding to organize these terminologies in formal ontologies, may be prerequisites
75 for the successful deployment of nanopublication. Thus while he expects other scientific, technical,
76 and medical disciplines to eventually embrace nanopublication, he is less sure that nanopublication will
77 work for the humanities. Historians, for example, use relatively little specialized terminology and pride
78 themselves on their ability to use “ordinary language” to represent the past. Even when humanist scholars

¹<http://www.biosemantics.org>

²<http://nextprot.org/>

³<http://www.disgenet.org/web/DisGeNET/v2.1>

79 use specialized theoretical language, their use of this language is often unstable, ambiguous, and highly
80 contested. Perhaps, then, a publishing technique that seeks to eliminate such ambiguity is ill-suited for
81 these fields.

82 A related obstacle to the adoption of nanopublication beyond the hard sciences has to do with
83 differences in the role played by “facts”. Researchers trained in the hard sciences understand their work to
84 be cumulative: scientists “stand on the shoulders of giants” and build upon the work of earlier researchers.
85 While scientists can in principle go back and recreate the experiments of their predecessors, in practice they
86 do this only when the results of those experiments have not been sufficiently established as facts. Efficient
87 cumulative research requires that, most of the time, they simply trust that the facts they inherit work as
88 advertised. Something like this process seems to be assumed by many proponents of nanopublications.
89 For example, Mons and Velterop (2009) claim that a major goal of nanopublication is to “elevate” factual
90 observations made by scientists into standardized packages that can be accumulated in databases, at least
91 until they are proved wrong. These standardized packages can then be automatically or semi-automatically
92 analyzed to produce new factual observations (or hypotheses about potential observations), and the cycle
93 continues.

94 Yet as Mink (1966) observed, not all forms of research and scholarship are aimed at producing
95 “detachable conclusions” that can serve as the basis for a cumulative process of knowledge production.
96 Anticipating Gradmann, Mink argued that

97 Detachable conclusions are possible in science because—and only because—of its theoretical
98 structure. The division of labor in research requires that concepts have a uniformity of
99 meaning, and the methodological problem of definition therefore becomes central. (Mink,
100 1966, 39)

101 He contrasted science to the study of history, which, lacking both explicit methodology and uniform
102 consensus on the meanings of its concepts, does not produce “detachable conclusions”. But this does not
103 mean that historical scholarship fails to produce knowledge, only that it is a separate and autonomous mode
104 of understanding. The goal of most historical scholarship is not to establish conclusions by constructing an
105 explanatory chain of inferences from evidence. Rather the goal is to render what Mink called a “synoptic
106 judgment”, an interpretive act in which the scholar comes to “see together” the disparate observable
107 elements of some phenomena as a synthetic whole. The historian who judges the advent of the printing
108 to have constituted a “communications revolution” (Eisenstein 1979) has not made an inference from
109 the available evidence but has constructed a particular interpretation of that evidence. To communicate
110 her synoptic judgment to others, she cannot simply state her conclusions unambiguously and rely on her
111 audience’s theoretical understanding to make them meaningful; instead she must arrange and exhibit the
112 evidence to help them “see together” what she saw.

113 So is nanopublication a poor fit for fields of knowledge production that do not follow the model
114 of cumulative science? We believe the answer is no. First of all, even Mink did not argue that there
115 were no facts in history, only that the significant conclusions drawn by historians do not typically take
116 the form of factual statements. There are plenty of equivalents in history and the humanities to the
117 databases of curated factual statements that exists in the sciences: prosopographical databases (Bradley
118 and Short, 2005), digital historical gazetteers (Elliott and Gillies, 2011), not to mention the catalogs and
119 indexes of bibliographical data that make humanist scholarship possible (Buckland, 2006). Some of
120 these facts may be vague or uncertain, but as Kuhn et al. (2013) observe, even knowledge that cannot be
121 completely formally represented, including vague or uncertain scientific findings, can benefit from the
122 nanopublication approach. We agree but would go further to say that nanopublication is useful even for
123 information that is neither testable nor falsifiable, exemplified by Mink’s synoptic judgments. We have
124 demonstrated the utility of nanopublications for describing synoptic judgments of historical periodization
125 in a project called PeriodO, which we describe below.

126 PERIODO

127 Motivation

128 In their work, archaeologists and historians frequently refer to time periods, such as “Classical Iberian Pe-
129 riod” or the “Progressive Era.” These time periods are shorthand representations of commonly referenced
130 segments of time and space. While time periods might have commonly understood definitions, they are

131 scattered throughout myriad publications and are often treated as shared, assumed knowledge. This leads
 132 to difficulty and repeated effort when scholars want to visualize their data in space and over time, which
 133 requires mapping these discursive period labels to discrete spatiotemporal ranges (Rabinowitz, 2014).

134 For the PeriodO project, we compiled thousands of definitions of time periods from published sources
 135 within the fields of archaeology, history, and art history. We mapped these time periods to a consistent,
 136 standardized data format and published them as linked open data so that future scholars would be able
 137 to cite these contextualized definitions instead of creating their own ad-hoc period assertions. Users are
 138 able to propose additional period definitions or change existing ones through the PeriodO interface. All
 139 proposed and accepted changes are stored, and each period definition has a history of patch submissions
 140 and approvals.

141 **Data Model**

142 PeriodO models a scholarly assertion about the name and spatiotemporal extent of a period as a period
 143 definition. The basis of a period definition consists of text taken from the original source indicating the
 144 name of the period, its temporal range, and the geographic region to which it applies. Multiple period
 145 definitions from the same source are grouped into a period collection. For example, the article “Domestic
 146 Architecture and Social Differences in North-Eastern Iberia during the Iron Age (c.525–200 BC)” includes
 147 the following sentence:

148 For the Catalan area, the complete system with the four above-mentioned categories is not
 149 as clearly documented before the fourth century as it is during the Classical Iberian Period
 150 (400–200 BC), although differences in the size of the sites, as well as the specialization of
 151 the functions of some settlements, can be already detected during the Early Iberian Period
 152 (525–400 BC). (Belarte, 2008)

153 This sentence contains two assertions defining period extents, so it is modeled in PeriodO as two
 154 period definitions. The first definition has the label “Classical Iberian Period” and its start and end points
 155 are labeled as “400 BC” and “200 BC” respectively. The second definition has the label “Early Iberian
 156 Period” and its start and end points are labeled as “525 BC” and “400 BC” respectively. The spatial extent
 157 of both definitions is labeled as “Catalan area”. Note that all of these labels are taken verbatim from the
 158 source text and should never change.

159 Because they come from the same source, these two period definitions are grouped into a period
 160 collection. The bibliographic metadata for the source article is associated with this period collection.
 161 (In the event that a source defines only a single period, then the period collection will be a singleton.)
 162 Note that belonging to the same period collection does not imply that period definitions compose a
 163 periodization. A periodization is a single coherent, continuous division of historical time, each part of
 164 which is labeled with a period term. A period collection, on the other hand, is simply a set of period
 165 definitions that share the same source. When the period definitions in a period collection do compose
 166 a periodization, this can be indicated through the addition of additional statements relating the period
 167 definitions to one another.

168 Because source languages, dating systems, and naming of geographical regions can vary widely, labels
 169 taken verbatim from source documents are insufficient for indexing and visualization period definitions
 170 in a uniform way. Thus the rest of the PeriodO data model consists of properties added by PeriodO
 171 curators to normalize the semantic content of these textual labels. First, all periods originally defined in
 172 a language other than English are given an alternate English-language label. When a period definition
 173 was originally defined in English, the alternate label may make minor changes for consistency.
 174 For example, the Belarte’s aforementioned definition of the “Classical Iberian Period” period is given
 175 an alternate label of “Classical Iberian”, removing the word “Period” for brevity and consistency with
 176 other definitions. Next, the specification of temporal start and end points is standardized by adding ISO
 177 8601 lexical representations of proleptic Gregorian calendar years⁴: -0399 for “400 BC” and -0199 for
 178 “200 BC”. Finally, descriptions of spatial extent are normalized by adding references to “spatial things”,
 179 typically modern nation-states. In this case both definitions are linked to the spatial thing identified
 180 by <http://dbpedia.org/resource/Spain>. The complete PeriodO representation in Turtle of
 181 Belarte’s collection of period definitions is given in Figure 1.

⁴Proleptic refers to dates represented in some calendar system that refer to a time prior to that calendar’s creation. The Gregorian calendar was adopted in 1582, but most of our dates fall in years prior to that one.

Figure 1. Turtle representation of a PeriodO period collection.

```
182 @prefix skos: <http://www.w3.org/2004/02/skos/core#>.
183 @prefix dcterms: <http://purl.org/dc/terms/>.
184 @prefix foaf: <http://xmlns.com/foaf/0.1/>.
185 @prefix time: <http://www.w3.org/2006/time#>.
186 @prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
187 @prefix bibo: <http://purl.org/ontology/bibo/>.
188 @prefix periodo: <http://n2t.net/ark:/99152/p0v#>.
189
190 <http://dbpedia.org/resource/Spain>
191   skos:prefLabel "Spain".
192
193 <http://dx.doi.org/10.1111/j.1468-0092.2008.00303.x>
194   dcterms:creator <http://id.crossref.org/contributor/maria-carme-belarte-2
195     mkpvn5eyc7oh>;
196   dcterms:issued <"2008"^^xsd:gYear>;
197   dcterms:title "DOMESTIC ARCHITECTURE AND SOCIAL DIFFERENCES IN NORTH-
198     EASTERN IBERIA DURING THE IRON AGE (c.525-200 BC)".
199
200 <http://id.crossref.org/contributor/maria-carme-belarte-2mkpvn5eyc7oh>
201   foaf:name "MARIA CARME BELARTE".
202
203 <http://n2t.net/ark:/99152/p06xc6m>
204   a skos:ConceptScheme;
205   dcterms:source [
206     dcterms:isPartOf <http://dx.doi.org/10.1111/j.1468-0092.2008.00303.x>;
207     bibo:locator "page 177"
208   ].
209
210 <http://n2t.net/ark:/99152/p06xc6mq829>
211   a skos:Concept;
212   periodo:spatialCoverageDescription "Catalan area";
213   dcterms:language "eng-latn";
214   dcterms:spatial <http://dbpedia.org/resource/Spain>;
215   skos:altLabel "Early Iberian Period"@eng-latn, "Early Iberian"@eng-latn;
216   skos:inScheme <http://n2t.net/ark:/99152/p06xc6m>;
217   skos:prefLabel "Early Iberian Period";
218   time:intervalFinishedBy [
219     skos:prefLabel "400 BC";
220     time:hasDateTimeDescription [
221       time:year <"-0399"^^xsd:gYear>
222     ]
223   ];
224   time:intervalStartedBy [
225     skos:prefLabel "525 BC";
226     time:hasDateTimeDescription [
227       time:year <"-0524"^^xsd:gYear>
228     ]
229   ].
230
231 <http://n2t.net/ark:/99152/p06xc6mvjx2>
232   a skos:Concept;
233   periodo:spatialCoverageDescription "Catalan area";
234   dcterms:language "eng-latn";
235   dcterms:spatial <http://dbpedia.org/resource/Spain>;
236   skos:altLabel "Classical Iberian Period"@eng-latn, "Classical Iberian"
237     @eng-latn;
238   skos:inScheme <http://n2t.net/ark:/99152/p06xc6m>;
239   skos:note "Equivalent to Iberian III (450-350 B.C.) and IV (350-200 B.C.)
240     - cf. M. Diaz-Andreu & S. Keay, 1997. The Archaeology of Iberia;
```

```

241         Dominguez in C. Sanchez & G.R. Tsetskhladze, 2001. Greek Pottery from
242         the Iberian Peninsula.";
243     skos:prefLabel "Classical Iberian Period";
244     time:intervalFinishedBy [
245         skos:prefLabel "200 BC";
246         time:hasDateTimeDescription [
247             time:year <"-0199"^^xsd:gYear>
248         ]
249     ];
250     time:intervalStartedBy [
251         skos:prefLabel "400 BC";
252         time:hasDateTimeDescription [
253             time:year <"-0399"^^xsd:gYear>
254         ]
255     ].

```

256 INTERPRETATION AS LINKED DATA

257 We have taken pains to make it easy to work with the PeriodO dataset. In particular, we have tried to make
258 the PeriodO dataset easily usable by developers who do not use an RDF-based tool stack. The PeriodO
259 dataset is published as JSON, which is easily parsed using standard libraries in most programming
260 environments including, of course, web browsers. But while JSON provides an easy and convenient way
261 to work with the PeriodO dataset by itself, we expect that many users will want to combine the PeriodO
262 dataset with the growing amount of scholarly Linked Data being published. Thus we take advantage of
263 the recent W3C Recommendation of JSON-LD (Sporny et al., 2014) to also make the PeriodO dataset
264 available as Linked Data. By providing a JSON-LD context for the PeriodO dataset, we make it usable
265 within an RDF-based stack.

266 RDF Vocabularies

267 The JSON-LD context maps relationships between PeriodO entities to terms from RDF vocabularies. Of
268 these, the most important are SKOS (Hobbs and Pan, 2006). The human-readable labels for a PeriodO
269 definition are mapped to the SKOS `prefLabel` and `altLabel` properties, implying that a PeriodO
270 period definition can be interpreted as a SKOS Concept. The relationship between a period definition
271 and the period collection to which it belongs is mapped to the SKOS `inScheme` property, implying that
272 a period collection is a SKOS `ConceptScheme`. The relationship between a period collection and its
273 source is mapped to the DCMI `source` term, and the various properties in the bibliographic description
274 of the source are also mapped to the appropriate DCMI terms. Finally, the relation between a period
275 definition and its geographical extent is mapped to the DCMI `spatial` term.

276 The relationships between a period definition and the start and end of its temporal extent are respec-
277 tively mapped to the OWL-Time `intervalStartedBy` and `intervalFinishedBy` properties.
278 This implies that a period definition, in addition to being a SKOS Concept, is also an OWL-Time
279 `ProperInterval` (an interval of time having non-zero duration). Importantly, this also implies that
280 the start and end of a period definition's temporal extent are themselves `ProperIntervals`, not points
281 or instants. This is important because the beginnings and endings of historical periods can never be
282 precisely determined. In the example of the Classical Iberian Period given above, both the beginning and
283 the end of the period are interpreted as intervals with a duration of one year. Interpreting period starts and
284 ends as `ProperIntervals` also allows us to make a distinction between the intervals themselves and
285 their descriptions. The intervals themselves are not precisely specifiable, but we can create pragmatic
286 OWL-Time `DateTimeDescriptions` of them for the purposes of comparison and visualization.

287 The start and end of a period definition's temporal extent are themselves intervals with their own starts
288 and ends, so temporal extent can be associated with a maximum of four values. This is interoperable
289 with other proposed representations of fuzzy, imprecise, or uncertain temporal extents, such as the four
290 `start`, `stop`, `earliest`, `latest` keys proposed for GeoJSON-LD (Meeks and Grossner, 2013). In
291 the current PeriodO data set these four properties only have (ISO 8601) year values, because none of our
292 sources specified endpoints at a more granular level than year. However, we expect to have finer-grained
293 values as we add periodizations of more recent history. At that point we will need to decide upon a unit of
294 representation that makes it simple to compare intervals defined at different levels of granularity. Adding

295 complexity to time interval expressions will be possible without changing our underlying data model
296 because of the flexibility of our current approach.

297 The *start*, *latest start*, *earliest end*, *end* approach enables us to represent the most common patterns for
298 defining periods found in our sources. For example a period defined as starting “3000 B.C. (+/- 150 years)”
299 and ending “about 2330 B.C.” can be represented with three values: -3149 , -2849 , and -2329 . Some
300 proposals for representing fuzzy, imprecise, or uncertain intervals, such as Topotime (Kauppinen et al.,
301 2010) propose a method for setting such curves in order to maximize precision and recall with respect
302 to temporal relevance judgments made by experts. We have chosen not to support these more complex
303 representations at this time because we are focused primarily on representing periods as defined in textual
304 sources. Natural language is already a compact and easily indexable way to represent imprecision or
305 uncertainty. Rather than imposing an arbitrary mapping from natural language to parameterized curves,
306 we prefer to maintain the original natural language terms used. However if scholars begin defining periods
307 with parameterized curves (which is certainly possible) then we will revisit this decision.

308 **Modeling provenance**

309 To model the provenance of period assertions, we utilized the Provenance Ontology [cite]. We record
310 each patch to the dataset as a `prov:Activity`. This Activity has `prov:startedAtTime` and
311 `prov:endedAtTime` values representing timestamps when the patch was sent and accepted, re-
312 spectivevely. The activity also has two `prov:used` statements: one which refers to the specific ver-
313 sion of the entire dataset to which the patch was applied (for example, `http://n2t.net/ark:`
314 `/99152/p0d?version=1`), and one referring to the patch itself as a `prov:Entity`. The patch
315 Entity contains a URL to the JSON-Patch file which resulted in the change Activity. Finally, the Activity
316 has `prov:generated` statements for each of the periods collections and period assertions (implied
317 to be of the type `prov:Entity`) that were affected by the given patch. Each of these affected entities
318 has a `prov:specializationOf` statement which refers to the permanent identifier for the period
319 assertion or collection (at no particular version). If they are revisions of an existing entity, they also have
320 `prov:wasRevisionOf` statements that refer to the version that they were descended from.

321 We defined a changelog at `http://n2t.net/ark:/99152/p0h#changelog` that represents
322 he sequential list of `prov:Activity` entities that created the current version of the dataset as an
323 ordered RDF list. In this way, one can reconstruct the origin of each change to the dataset as a whole, or
324 to individual period assertions.

325 **Minting Long-term URLs**

326 In addition to mapping relationships to well-known vocabularies, interpreting PeriodO as Linked Data
327 requires a way to assign URLs to period collections and definitions. As shown in Figure 1, period
328 definitions and period collections in the dataset are given short identifiers: `p06xc6mvjx2` identifies the
329 definition of the Classical Iberian Period, and `p06xc6m` identifies the collection to which it belongs. But
330 these identifiers are only useful within the context of the PeriodO dataset; they are not guaranteed to be
331 unique in a global context and, unless one already has the PeriodO data, one cannot resolve them to obtain
332 representations of the entities they identify. URLs, on the other hand, are globally unique and can be
333 resolved using HTTP to obtain representations; this is the core concept behind Linked Data. So, we need
334 a way to turn the short PeriodO identifiers into URLs.

335 To turn PeriodO identifiers into URLs we rely on the ARK identifier scheme (Starr et al., 2012)
336 provided by the California Digital Library (CDL). First, we include in the JSON-LD context a `@base`
337 value specifying the base URI (`http://n2t.net/ark:/99152/`) to use when interpreting the
338 PeriodO dataset as Linked Data. This allows the short PeriodO identifiers to be interpreted as URLs; for
339 example `p06xc6mvjx2` is interpreted as a relative reference to the URL `http://n2t.net/ark:`
340 `/99152/p06xc6mvjx2`. The host of this URL (`n2t.net`) is the registered name of the CDL’s
341 Name-to-Thing resolver, which is similar to other name resolution services for persistent URLs such as
342 PURL. We have registered with the EZID service a single ARK identifier (`ark:/99152/p0`) with the
343 URL of the HTTP server currently hosting the canonical PeriodO dataset. Thus any request to a URL
344 starting with `http://n2t.net/ark:/99152/p0` will be redirected to that server. An HTTP GET
345 to `http://n2t.net/ark:/99152/p0d.jsonld` will return the entire dataset, while GETting
346 (for example) `http://n2t.net/ark:/99152/p06xc6mvjx2.jsonld` will return a JSON-LD
347 representation of Belarte’s definition of the Classical Iberian Period.

348 PERIOD ASSERTIONS AS NANOPUBLICATIONS

349 We created the PeriodO dataset based on the same core concerns of nanopublication authors: to extract,
350 curate, and publish small, computable concepts from their broader sources while still preserving their
351 provenance. A nanopublication is made up of an assertion, the provenance of that assertion, and the
352 provenance of the nanopublication itself. In PeriodO, these elements come in the following pieces of
353 information:

- 354 • **Assertion:** The definition of a period
- 355 • **Provenance:** The source this period was derived from. This may be a citation of a printed work or
356 a URL for a resource hosted on the web.
- 357 • **Provenance of nanopublication:** The history of the period definition within the PeriodO system,
358 including the date it was added or changed, the identity of the person who submitted or changed it,
359 and the identity of the person who approved additions or changes.

360 Figure 1 above contains two assertions with the same provenance. Each of these assertions would be
361 represented by individual nanopublications. The nanopublication for the Early Iberian Period is shown
362 in Figure 2. While the nanopublication concepts readily map to the nanopublication scheme, we faced
363 several challenges during our creation of the dataset due to its interpretive nature.

Figure 2. Nanopublication of the Early Iberian Period

```

364 @base <http://n2t.net/ark:/99152/> .
365 @prefix : <p06xc6mq829/nanopubl#> .
366 @prefix bibo: <http://purl.org/ontology/bibo/> .
367 @prefix dcterms: <http://purl.org/dc/terms/> .
368 @prefix foaf: <http://xmlns.com/foaf/0.1/> .
369 @prefix skos: <http://www.w3.org/2004/02/skos/core#> .
370 @prefix time: <http://www.w3.org/2006/time#> .
371 @prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
372 @prefix periodo: <p0v#> .
373 @prefix prov: <http://www.w3.org/ns/prov#> .
374 @prefix np: <http://www.nanopub.org/nschema#> .
375
376 :head {
377   <p06xc6mq829/nanopubl> a np:Nanopublication ;
378     np:hasAssertion :assertion ;
379     np:hasProvenance :provenance ;
380     np:hasPublicationInfo :pubinfo .
381 }
382
383 :assertion {
384   <p06xc6mq829>
385     a skos:Concept;
386     skos:inScheme <p06xc6m>;
387     skos:prefLabel "Early Iberian Period";
388     periodo:spatialCoverageDescription "Catalan area";
389     dcterms:language "eng-latn";
390     dcterms:spatial <http://dbpedia.org/resource/Spain>;
391     skos:altLabel "Early Iberian Period"@eng-latn, "Early Iberian"@eng-
392       latn;
393     time:intervalFinishedBy [
394       skos:prefLabel "400 BC";
395       time:hasDateTimeDescription [
396         time:year "-0399"^^xsd:gYear
397       ]
398     ];
399     time:intervalStartedBy [
400       skos:prefLabel "525 BC";

```

```

401         time:hasDateTimeDescription [
402             time:year "-0524"^^xsd:gYear
403         ]
404     ].
405 }
406
407 :provenance {
408     :assertion dcterms:source [
409         dcterms:isPartOf <http://dx.doi.org/10.1111/j.1468-0092.2008.00303.x>;
410         bibo:locator "page 177"
411     ].
412
413     <http://dx.doi.org/10.1111/j.1468-0092.2008.00303.x>
414         dcterms:creator <http://id.crossref.org/contributor/maria-carme-
415             belarte-2mkpvn5eyc7oh>;
416         dcterms:issued "2008"^^xsd:gYear;
417         dcterms:title "DOMESTIC ARCHITECTURE AND SOCIAL DIFFERENCES IN NORTH-
418             EASTERN IBERIA DURING THE IRON AGE (c.525-200 BC)".
419
420         <http://id.crossref.org/contributor/maria-carme-belarte-2mkpvn5eyc7oh>
421             foaf:name "MARIA CARME BELARTE".
422     }
423
424     :pubinfo {
425         <p06xc6mq829/nanopub1> prov:wasGeneratedBy <p0h#change-1> ;
426         prov:generatedAtTime "2015-07-29T21:49:31"^^xsd:dateTime ;
427         prov:wasAttributedTo <http://orcid.org/0000-0002-3617-9378> .
428     }

```

429 **The Unfalsifiable Nature of Time Period Definitions**

430 Unlike data such as measurements of genomic expression or statements of biological causality, much
431 of the information produced in humanist disciplines is not testable or falsifiable. The PeriodO dataset
432 is no different in this regard. Compare the assertion that “malaria is transmitted by mosquitoes” to the
433 one that “there is a period called the Late Bronze Age in Northern Europe, and it lasted from about 1100
434 B.C. to 500 B.C.” Malaria and mosquitoes are two well-defined entities that exist within strict taxonomies
435 reflected the physical world. “Mosquito” and “malaria” are terms that point to positions within these
436 taxonomies. Conversely, the “Late Bronze Age” is a purely discursive construct. Whereas a relationship
437 between the class of insects we call mosquitoes and cases of the illness we call malaria existed prior to its
438 observation by humans, there was no discrete entity called the “Late Bronze Age” before it was coined
439 by those studying that time and place. Consequently, one cannot disprove the idea that there was a time
440 period called the Late Bronze Age from around 1100 B.C. to 500 B.C.; one can only argue that another
441 definition has more credence based on non-experimental, discursive arguments.

442 Kuhn et al. (2013) are concerned that requiring formal representation for all scientific data published
443 as nanopublications “seems to be unrealistic in many cases and might restrict the range of practical
444 application considerably.” We have found the same to be true with our dataset, and argue that the form
445 and scope of nanopublication assertions should ultimately be determined by the practical needs of the
446 researchers who use them. If nanopublications are to expand beyond computational scientific fields,
447 the nature and scope of assertions will vary between applications based on the practical concerns of
448 researchers. For computational biologists, the forms of individual assertions reflect the need to connect,
449 consolidate, and assess trillions of measurements scattered throughout a rapidly growing body of research
450 findings. The goal is to create a global, connected knowledge graph that can be used as a tool for scientists
451 to guide new discoveries and verify experimental results. For a domain like the definition of time periods,
452 the extraction and publication of pieces of information is practically beneficial even if the resulting
453 assertions are not provable, unambiguous or chainable.

454 There is no reason why the assertions at the center of nanopublications must be atomic, unambiguous,
455 and falsifiable. These requirements only matter within certain contexts, such as the connective application
456 required by the practical needs of computational scientists. We must recognize that even discursive data

457 that cannot be combined in such chains of signification can be usefully processed by computer programs.

458 In the PeriodO context we are not concerned with making an exhaustive taxonomy of “correct” periods
 459 or facilitating the “discovery” of new periods (a non sequitur—there are no periods that exist in the world
 460 that are awaiting discovery by some inquiring historian or archaeologist). Rather, we are interested in
 461 enabling the study and citation of how and by whom time has been segmented into different periods. Our
 462 approach to modeling assertions has been guided by this concern.

463 In some sense, the nanopublication focus on provenance is even more important for non-scientific
 464 datasets, since the assertions made therein are so critically dependent on their wider discursive context.
 465 Because subjectivity is inextricable from these sorts of unfalsifiable relationships, it is important to
 466 preserve their provenance and original context in order to judge their quality, trustworthiness, and
 467 usefulness.

468 **The Critical and Unavoidable Role of Curation**

469 Another divergence of the PeriodO dataset from traditional nanopublications is the unavoidable curatorial
 470 work that was necessary to extract practically useful assertions from textual period definitions. In all of
 471 the applications of nanopublications we found, the published assertions typically appeared in the form
 472 of measurements or well-defined relationships between discrete entities. These are types of data which
 473 humans or computers can easily and reliably extract from research findings. Our dataset required explicit
 474 curatorial decisions: a time period exists within a certain spatiotemporal context, and there is no sure way
 475 to discretely, accurately, and unambiguously model such boundaries. While a human might be able to have
 476 a nuanced understanding of temporary and ever-shifting political boundaries or the uncertain and partially
 477 arbitrary precision suggested by “around the beginning of the 12th century BC”, we cannot assume the
 478 same of computers. Therefore, in order for our dataset to be readily algorithmically comparable, we had
 479 to map discursive concepts to discrete values. Our curatorial decisions in this regard reflect a compromise
 480 between uniformity, potential semantic expressiveness, and practical usefulness.

481 As humanist scholars publish their own nanopublications (or linked data in general), they will also go
 482 through a curatorial process due to the interpretive, unstandardized nature of humanistic datasets discussed
 483 above. There is a temptation in this process to imagine perfect structured descriptions that could express
 484 all possible nuances of all possible assertions. However, chasing that goal can lead to overcomplexity and,
 485 in the end, be practically useless. In describing period assertions as linked data, we adopted a schema that
 486 was only as semantically complicated as was *a*) expressed in our collected data and *b*) necessitated by the
 487 practical needs of our intended users. Humanities nanopublication creators should focus on polishing
 488 the usefully comparable parts of their data and not get bogged down in the futile task of perfect formal
 489 representation.

490 In our case, as we started to collect data, we considered the basic characteristics of a dataset that
 491 would be necessary to accomplish automated retrieval and comparison tasks that we believed were most
 492 important. These tasks included:

- 493 • Finding all periods within a certain geographic area. (“What time periods have scholars used in
 494 Northern Europe?”)
- 495 • Finding all periods within a certain span of time. (“What time periods have been used to describe
 496 years between 100 AD to 500 AD?”)
- 497 • Finding how the definition of periods have differed across time/authors, or finding contested period
 498 definitions. (“How have different authors defined the Early Bronze Age?”)
- 499 • Finding periods defined for different languages. (“What time periods been defined in Russian?”)

500 Based on these decisions, we needed to impose some consistent amount of specificity upon the temporal
 501 and spatial coverage of period definitions.

502 Our initial model for temporal mapping was to express the termini of periods as Julian Days represented
 503 in scientific notation. Julian Days are a standard form of time measurement commonly used by astronomers
 504 to represent dates in the far historical past. Julian Days work by counting the number of continuous
 505 days that have passed since January 1, 4713 BC in the Proleptic Julian calendar. Conceptually, this is
 506 a similar measurement to the common Unix time standard, which counts the number of milliseconds
 507 that have passed since midnight GMT on January 1, 1970. The idea is that by counting forward using

508 well-defined units since an accepted epoch, one can get away from the inconsistencies and periodic lapses
509 that characterize different calendrical systems. Representing Julian Days using scientific notation allowed
510 us to express variable levels of uncertainty. See examples of this notation system in Table 1.

Table 1. Example Scientific Notation of Julian Days

Scientific Notation	Julian Day (JDN)	Proleptic Gregorian
1.3E6	Between JDN 1,250,000 and JDN 1,350,000	1150 BC \pm 150 years
1.30E6	Between JDN 1,295,000 and JDN 1,305,000	1150 BC \pm 15 years
1.300E6	Between JDN 1,299,500 and JDN 1,300,500	1150 BC \pm 1.5 years

511 However, in practice, we found this scheme to be overly complex. The necessary imposition of a level
512 of specificity, while theoretically useful in certain cases, was often not appropriate. In almost every single
513 case that we observed, authors did not explicitly state a precise level of uncertainty for their temporal
514 expressions. By adding precise uncertainty ourselves, we would, in effect, have been putting words in
515 authors' mouths. Further, Julian Days are not widely used outside of very specific disciplines, meaning
516 that consumers of our data would have to convert to a more familiar time system before being able to
517 understand or use our data.

518 Instead of the Julian Day model, we settled on the four-part ISO date schema, described above. This
519 model is less expressive for complicated forms of uncertainty, but it is less complex and more easily
520 understood by both our target audience and typical software programs. It was also easy to convert to, since
521 almost all of the periods assertions we observed were drawn from sources based on Western calendars. If
522 our pool of collected data contained periods that had more complex time expressions or were based on
523 varying calendrical systems, we might have used a different, more expressive schema.

524 To encourage a standardized mapping for all period definitions, we build a simple grammar and parser
525 for date expressions that covered the vast majority of our sample data. The parser takes in a string like
526 "c. mid-12th century" and outputs a JSON string consistent with our data model. This parser also gives
527 a naïve interpretation to descriptions like "mid-fifth century", assigning them to the third of the epoch
528 described according to the conventional segmentation of "early" "mid" and "late." "Mid-fifth century"
529 would, then, be parsed as the range of years 401 to 434. Similarly, we created an autocomplete interface
530 to modern political entities to allow users to enter spatial coverage. These techniques result in a practical
531 approximation of spatiotemporal coverage rather than a complete, unambiguous representation. The
532 interface we created to edit period definitions is shown in Figure 3.

533 FUTURE WORK

534 After the initial step of gathering period definitions, we hope to gather information on their citation and
535 use. This would include both studying the historical use of attributed period definitions as well as tracking
536 the citation of PeriodO period identifiers going forward.

Editing period: Early Iberian Period

Label *

eng ▾
latn ▾

Alternate labels

eng ▾
latn ▾

+
-

Locator

URL

Same as (not editable)

Linked data for this period

Spatial coverage

Spatial coverage description

Spatial Coverage Extent

- ×
Spain

Temporal coverage

Parse dates automatically

Start*

Label

Year

Toggle earliest/latest

Stop*

Label

Year

Toggle earliest/latest

Notes

Note

Notes derived from the source

Editorial note

Notes about the import process

Save
Delete
Cancel

Figure 3. Period editing form.

537 CONCLUSION

538 Ultimately, nanopublication is a way to balance the needs of computers for uniformity in data modeling
 539 with the needs of humans to fully understand and judge information based on context. As scholars of all
 540 disciplines continue to integrate computational methods into their work, the need for this balance grows.
 541 This is as true in the humanities and social sciences as it is in the natural sciences. However, different
 542 disciplines have different practical concerns, and their use of nanopublications should reflect this fact.
 543 Implementors of nanopublication systems (and linked data-producing systems as a whole) should worry
 544 about fitting data into precise, minutely-defined models only insofar as it is practically useful for their
 545 intended users to do so.

546 Nanopublication is an important trend which accounts for the creation of “data” within a wider

547 scholarly context. In this way, it echoes old ideas about hypertext which respect the importance of
548 provenance, authorship, and attribution (Nelson, 1999). We hope our work shows that this approach is
549 relevant and feasible even to fields outside of experimental, observable sciences.

550 REFERENCES

- 551 Belarte, M. C. (2008). Domestic architecture and social differences in north-eastern Iberia during the Iron
552 Age (c.525-200 BC). *Oxford Journal of Archaeology*, 27:175–199.
- 553 Bradley, J. and Short, H. (2005). Texts into Databases: The Evolving Field of New-Style Prosopography.
554 *Literary and Linguistic Computing*, 20(Suppl. no. 1):3–24.
- 555 Buckland, M. (2006). Description and Search: Metadata as Infrastructure. *Brazilian Journal of Information
556 Science*, 0(0).
- 557 Elliott, T. and Gillies, S. (2011). Pleiades: an un-GIS for Ancient Geography. In *Digital Humanities
558 Conference 2011*, Stanford.
- 559 Gradmann, S. (2014). From containers to content to context. *Journal of Documentation*, 70(2):241–260.
- 560 Groth, P., Gibson, A., and Velterop, J. (2010). The anatomy of a nanopublication. *Information Services &
561 Use*, 30(1-2):51–56.
- 562 Groth, P., Schultes, E., Thompson, M., Tatum, Z., and Dumontier, M. (2013). Nanopublication Guidelines.
563 Technical report, Concept Web Alliance.
- 564 Hobbs, J. R. and Pan, F. (2006). Time Ontology in OWL. report, W3C.
- 565 Kauppinen, T., Mantegari, G., Paakkari, P., Kuittinen, H., Hyvönen, E., and Bandini, S. (2010).
566 Determining relevance of imprecise temporal intervals for cultural heritage information retrieval.
567 *International Journal of Human-Computer Studies*, 68(9):549–560.
- 568 Kuhn, T. (2015). Science bots: A model for the future of scientific computation? In *WWW 2015
569 Companion Proceedings*. ACM.
- 570 Kuhn, T., Barbano, P. E., Nagy, M. L., and Krauthammer, M. (2013). Broadening the Scope of Nanopub-
571 lications. In *The Semantic Web: Semantics and Big Data*, volume 7882, pages 487–501. Springer.
- 572 Meeks, E. and Grossner, K. (2013). Topotime v0.1 specification.
- 573 Mink, L. O. (1966). The Autonomy of Historical Understanding. *History and Theory*, 5(1):24–47.
- 574 Mons, B., van Haagen, H., Chichester, C., Hoen, P.-B. t., den Dunnen, J. T., van Ommen, G., van Mulligen,
575 E., Singh, B., Hoof, R., Roos, M., Hammond, J., Kiesel, B., Giardine, B., Velterop, J., Groth, P., and
576 Schultes, E. (2011). The value of data. *Nature Genetics*, 43(4):281–3.
- 577 Mons, B. and Velterop, J. (2009). Nano-Publication in the e-Science Era. In Clark, T., Luciano, J. S.,
578 Marshall, M. S., Prud'hommeaux, E., and Stephens, S., editors, *Proceedings of the Workshop on
579 Semantic Web Applications in Scientific Discourse (SWASD 2009)*, Washington, D.C. CEUR Workshop
580 Proceedings.
- 581 Nelson, T. H. (1999). Xanalogical structure, needed now more than ever: parallel documents, deep links
582 to content, deep versioning, and deep re-use. *ACM Computing Surveys*, 31(4es):33–es.
- 583 Rabinowitz, A. (2014). It's about time: historical periodization and Linked Ancient World Data. *ISAW
584 Papers*, 7(22).
- 585 Sporny, M., Longley, D., Kellogg, G., Lanthaler, M., and Lindström, N. (2014). JSON-LD 1.0: A
586 JSON-based Serialization for Linked Data. Technical report, W3C.
- 587 Starr, J., Willett, P., Federer, L., Horning, C., and Bergstrom, M. (2012). A Collaborative Framework
588 for Data Management Services: The Experience of the University of California. *Journal of eScience
589 Librarianship*, 1(2):109–114.