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.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1284v1 | CC-BY 4.0 Open Access | rec: 6 Aug 2015, publ: 6 Aug 2015

Nanopublication Beyond the Sciences

Patrick Golden[1] **and Ryan Shaw**[1]

[1]**School 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

INTRODUCTION

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

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

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

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

NANOPUBLICATIONS

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

Motivation

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

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

Definition

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

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

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

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

Uses

Several European repositories of bioinformatic data have begun to publish their contents as nanopublications, including the Biosemantics Group, neXtProt, and DisGeNET[1][2][3]. These publications can be aggregated and connected in larger systems, such as the decentralized reputation system described by Kuhn (2015).

NANOPUBLICATION IN THE HUMANITIES

While the bioinformatics research community has enthusiastically adopted nanopublication, other disciplines have been slow to follow. Gradmann (2014) suggested that specialized and stable terminologies, as well as sufficient funding to organize these terminologies in formal ontologies, may be prerequisites for the successful deployment of nanopublication. Thus while he expects other scientific, technical, and medical disciplines to eventually embrace nanopublication, he is less sure that nanopublication will work for the humanities. Historians, for example, use relatively little specialized terminology and pride 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

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1284v1 | CC-BY 4.0 Open Access | rec: 6 Aug 2015, publ: 6 Aug 2015

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

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

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

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

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

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

PERIODO

Motivation

In their work, archaeologists and historians frequently refer to time periods, such as "Classical Iberian Period" or the "Progressive Era." These time periods are shorthand representations of commonly referenced segments of time and space. While time periods might have commonly understood definitions, they are scattered throughout myriad publications and are often treated as shared, assumed knowledge. This leads to difficulty and repeated effort when scholars want to visualize their data in space and over time, which requires mapping these discursive period labels to discrete spatiotemporal ranges (Rabinowitz, 2014).

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

Data Model

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

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

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

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

Because source languages, dating systems, and naming of geographical regions can vary widely, labels taken verbatim from source documents are insufficient for indexing and visualization period definitions in a uniform way. Thus the rest of the PeriodO data model consists of properties added by PeriodO curators to normalize the semantic content of these textual labels. First, all periods originally defined in a language other than English are given an alternate English-language label. When a period definition was originally defined in English, the alternate label may make make minor changes for consistency. For example, the Belarte's aforementioned definition of the "Classical Iberian Period" period is given an alternate label of "Classical Iberian", removing the word "Period" for brevity and consistency with other definitions. Next, the specification of temporal start and end points is standardized by adding ISO 8601 lexical representations of proleptic Gregorian calendar years[4]: `-0399` for "400 BC" and `-0199` for "200 BC". Finally, descriptions of spatial extent are normalized by adding references to "spatial things", typically modern nation-states. In this case both definitions are linked to the spatial thing identified by `http://dbpedia.org/resource/Spain`. The complete PeriodO representation in Turtle of Belarte's collection of period definitions is given in Figure 1.

[4]Proleptic 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.

```
@prefix skos: <http://www.w3.org/2004/02/skos/core#>.
@prefix dcterms: <http://purl.org/dc/terms/>.
@prefix foaf: <http://xmlns.com/foaf/0.1/>.
@prefix time: <http://www.w3.org/2006/time#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix bibo: <http://purl.org/ontology/bibo/>.
@prefix periodo: <http://n2t.net/ark:/99152/p0v#>.

<http://dbpedia.org/resource/Spain>
   skos:prefLabel "Spain".

<http://dx.doi.org/10.1111/j.1468-0092.2008.00303.x>
   dcterms:creator <http://id.crossref.org/contributor/maria-carme-belarte-2
       mkpvn5eyc7oh>;
   dcterms:issued <"2008"^^xsd:gYear>;
   dcterms:title "DOMESTIC ARCHITECTURE AND SOCIAL DIFFERENCES IN NORTH-
       EASTERN IBERIA DURING THE IRON AGE (c.525-200 BC)".

<http://id.crossref.org/contributor/maria-carme-belarte-2mkpvn5eyc7oh>
   foaf:name "MARIA CARME BELARTE".

<http://n2t.net/ark:/99152/p06xc6m>
   a skos:ConceptScheme;
   dcterms:source [
     dcterms:isPartOf <http://dx.doi.org/10.1111/j.1468-0092.2008.00303.x>;
     bibo:locator "page 177"
   ].

<http://n2t.net/ark:/99152/p06xc6mq829>
   a skos:Concept;
   periodo:spatialCoverageDescription "Catalan area";
   dcterms:language "eng-latn";
   dcterms:spatial <http://dbpedia.org/resource/Spain>;
   skos:altLabel "Early Iberian Period"@eng-latn, "Early Iberian"@eng-latn;
   skos:inScheme <http://n2t.net/ark:/99152/p06xc6m>;
   skos:prefLabel "Early Iberian Period";
   time:intervalFinishedBy [
     skos:prefLabel "400 BC";
     time:hasDateTimeDescription [
       time:year <"-0399"^^xsd:gYear>
     ]
   ];
   time:intervalStartedBy [
     skos:prefLabel "525 BC";
     time:hasDateTimeDescription [
       time:year <"-0524"^^xsd:gYear>
     ]
   ].

<http://n2t.net/ark:/99152/p06xc6mvjx2>
   a skos:Concept;
   periodo:spatialCoverageDescription "Catalan area";
   dcterms:language "eng-latn";
   dcterms:spatial <http://dbpedia.org/resource/Spain>;
   skos:altLabel "Classical Iberian Period"@eng-latn, "Classical Iberian"
       @eng-latn;
   skos:inScheme <http://n2t.net/ark:/99152/p06xc6m>;
   skos:note "Equivalent to Iberian III (450-350 B.C.) and IV (350-200 B.C.)
        - cf. M. Diaz-Andreu & S. Keay, 1997. The Archaeology of Iberia;
```

```
      Dominguez in C. Sanchez & G.R. Tsetskhladze, 2001. Greek Pottery from
      the Iberian Peninsula.";
  skos:prefLabel "Classical Iberian Period";
  time:intervalFinishedBy [
   skos:prefLabel "200 BC";
   time:hasDateTimeDescription [
    time:year <"-0199"^^xsd:gYear>
   ]
  ];
  time:intervalStartedBy [
   skos:prefLabel "400 BC";
   time:hasDateTimeDescription [
    time:year <"-0399"^^xsd:gYear>
   ]
  ].
```

INTERPRETATION AS LINKED DATA

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

RDF Vocabularies

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

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

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

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1284v1 | CC-BY 4.0 Open Access | rec: 6 Aug 2015, publ: 6 Aug 2015

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

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

Modeling provenance

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

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

Minting Long-term URLs

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

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

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1284v1 | CC-BY 4.0 Open Access | rec: 6 Aug 2015, publ: 6 Aug 2015

PERIOD ASSERTIONS AS NANOPUBLICATIONS

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

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

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

Figure 2. Nanopublication of the Early Iberian Period

```
@base <http://n2t.net/ark:/99152/> .
@prefix : <p06xc6mq829/nanopub1#> .
@prefix bibo: <http://purl.org/ontology/bibo/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix time: <http://www.w3.org/2006/time#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix periodo: <p0v#> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix np: <http://www.nanopub.org/nschema#> .

:head {
   <p06xc6mq829/nanopub1> a np:Nanopublication ;
      np:hasAssertion :assertion ;
      np:hasProvenance :provenance ;
      np:hasPublicationInfo :pubinfo .
}

:assertion {
   <p06xc6mq829>
      a skos:Concept;
      skos:inScheme <p06xc6m>;
      skos:prefLabel "Early Iberian Period";
      periodo:spatialCoverageDescription "Catalan area";
      dcterms:language "eng-latn";
      dcterms:spatial <http://dbpedia.org/resource/Spain>;
      skos:altLabel "Early Iberian Period"@eng-latn, "Early Iberian"@eng-
          latn;
      time:intervalFinishedBy [
        skos:prefLabel "400 BC";
        time:hasDateTimeDescription [
          time:year "-0399"^^xsd:gYear
        ]
      ];
      time:intervalStartedBy [
        skos:prefLabel "525 BC";
```

```
        time:hasDateTimeDescription [
          time:year "-0524"^^xsd:gYear
        ]
      ].
}

:provenance {
   :assertion dcterms:source [
    dcterms:isPartOf <http://dx.doi.org/10.1111/j.1468-0092.2008.00303.x>;
    bibo:locator "page 177"
   ].

   <http://dx.doi.org/10.1111/j.1468-0092.2008.00303.x>
      dcterms:creator <http://id.crossref.org/contributor/maria-carme-
          belarte-2mkpvn5eyc7oh>;
      dcterms:issued "2008"^^xsd:gYear;
      dcterms:title "DOMESTIC ARCHITECTURE AND SOCIAL DIFFERENCES IN NORTH-
          EASTERN IBERIA DURING THE IRON AGE (c.525-200 BC)".

   <http://id.crossref.org/contributor/maria-carme-belarte-2mkpvn5eyc7oh>
      foaf:name "MARIA CARME BELARTE".
}

:pubinfo {
   <p06xc6mq829/nanopub1> prov:wasGeneratedBy <p0h#change-1> ;
      prov:generatedAtTime "2015-07-29T21:49:31"^^xsd:dateTime ;
      prov:wasAttributedTo <http://orcid.org/0000-0002-3617-9378> .
}
```

The Unfalsifiable Nature of Time Period Definitions

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

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

There is no reason why the assertions at the center of nanopublications must be atomic, unambiguous, and falsifiable. These requirements only matter within certain contexts, such as the connective application required by the practical needs of computational scientists. We must recognize that even discursive data that cannot be combined in such chains of signification can be usefully processed by computer programs.

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

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

The Critical and Unavoidable Role of Curation

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

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

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

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

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

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

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1284v1 | CC-BY 4.0 Open Access | rec: 6 Aug 2015, publ: 6 Aug 2015

well-defined units since an accepted epoch, one can get away from the inconsistencies and periodic lapses that characterize different calendrical systems. Representing Julian Days using scientific notation allowed 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 ± 150 years
1.30E6	Between JDN 1,295,000 and JDN 1,305,000	1150 BC ± 15 years
1.300E6	Between JDN 1,299,500 and JDN 1,300,500	1150 BC ± 1.5 years

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

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

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

FUTURE WORK

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

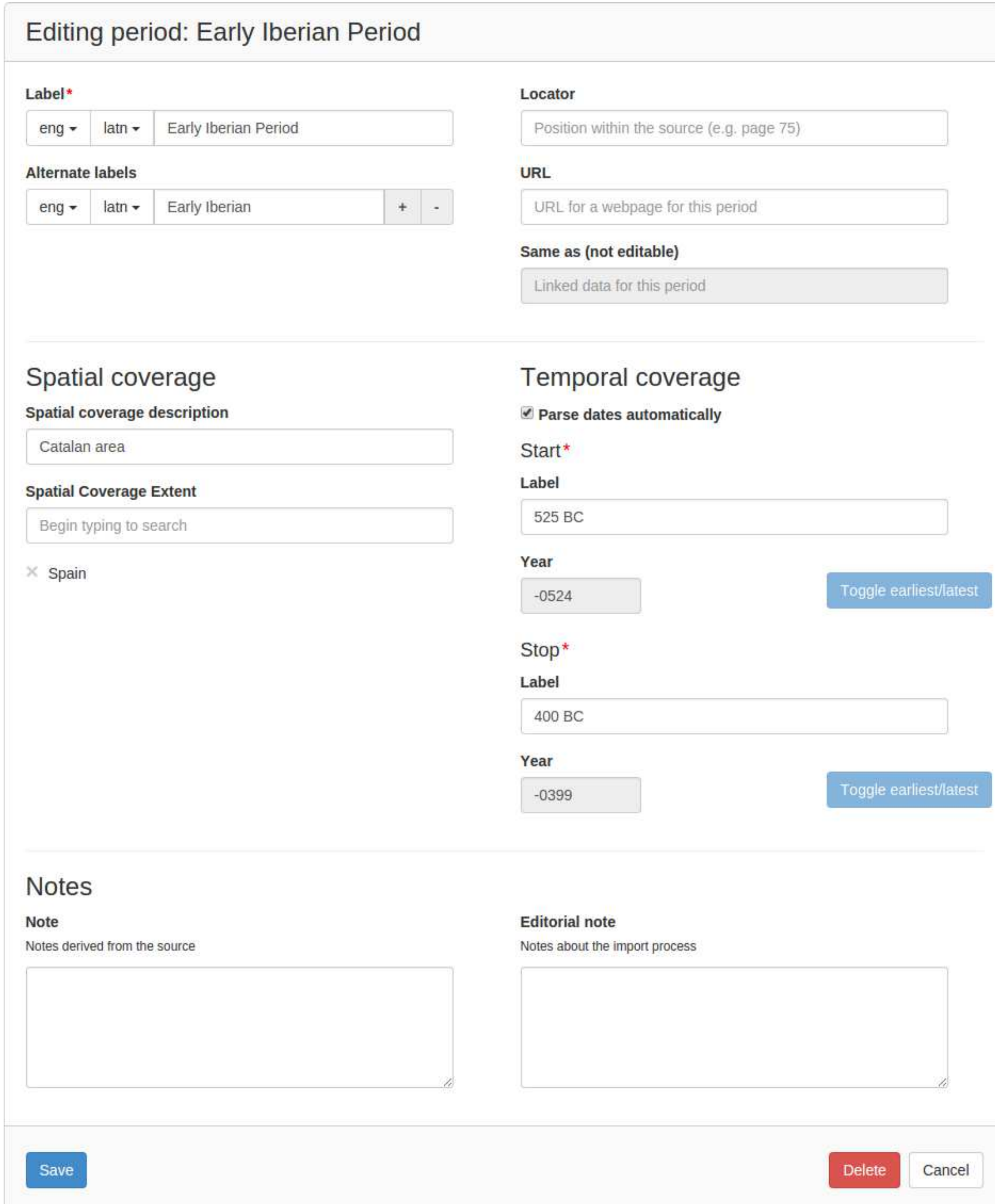

Figure 3. Period editing form.

CONCLUSION

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

Nanopublication is an important trend which accounts for the creation of "data" within a wider

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1284v1 | CC-BY 4.0 Open Access | rec: 6 Aug 2015, publ: 6 Aug 2015

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

REFERENCES

Belarte, M. C. (2008). Domestic architecture and social differences in north-eastern Iberia during the Iron Age (c.525-200 BC). *Oxford Journal of Archaeology*, 27:175–199.

Bradley, J. and Short, H. (2005). Texts into Databases: The Evolving Field of New-Style Prosopography. *Literary and Linguistic Computing*, 20(Suppl. no. 1):3–24.

Buckland, M. (2006). Description and Search: Metadata as Infrastructure. *Brazilian Journal of Information Science*, 0(0).

Elliott, T. and Gillies, S. (2011). Pleiades: an un-GIS for Ancient Geography. In *Digital Humanities Conference 2011*, Stanford.

Gradmann, S. (2014). From containers to content to context. *Journal of Documentation*, 70(2):241–260.

Groth, P., Gibson, A., and Velterop, J. (2010). The anatomy of a nanopublication. *Information Services & Use*, 30(1-2):51–56.

Groth, P., Schultes, E., Thompson, M., Tatum, Z., and Dumontier, M. (2013). Nanopublication Guidelines. Technical report, Concept Web Alliance.

Hobbs, J. R. and Pan, F. (2006). Time Ontology in OWL. report, W3C.

Kauppinen, T., Mantegari, G., Paakkarinen, P., Kuittinen, H., Hyvönen, E., and Bandini, S. (2010). Determining relevance of imprecise temporal intervals for cultural heritage information retrieval. *International Journal of Human-Computer Studies*, 68(9):549–560.

Kuhn, T. (2015). Science bots: A model for the future of scientific computation? In *WWW 2015 Companion Proceedings*. ACM.

Kuhn, T., Barbano, P. E., Nagy, M. L., and Krauthammer, M. (2013). Broadening the Scope of Nanopublications. In *The Semantic Web: Semantics and Big Data*, volume 7882, pages 487–501. Springer.

Meeks, E. and Grossner, K. (2013). Topotime v0.1 specification.

Mink, L. O. (1966). The Autonomy of Historical Understanding. *History and Theory*, 5(1):24–47.

Mons, B., van Haagen, H., Chichester, C., Hoen, P.-B. t., den Dunnen, J. T., van Ommen, G., van Mulligen, E., Singh, B., Hooft, R., Roos, M., Hammond, J., Kiesel, B., Giardine, B., Velterop, J., Groth, P., and Schultes, E. (2011). The value of data. *Nature Genetics*, 43(4):281–3.

Mons, B. and Velterop, J. (2009). Nano-Publication in the e-Science Era. In Clark, T., Luciano, J. S., Marshall, M. S., Prud'hommeaux, E., and Stephens, S., editors, *Proceedings of the Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009)*, Washington, D.C. CEUR Workshop Proceedings.

Nelson, T. H. (1999). Xanalogical structure, needed now more than ever: parallel documents, deep links to content, deep versioning, and deep re-use. *ACM Computing Surveys*, 31(4es):33–es.

Rabinowitz, A. (2014). It's about time: historical periodization and Linked Ancient World Data. *ISAW Papers*, 7(22).

Sporny, M., Longley, D., Kellogg, G., Lanthaler, M., and Lindström, N. (2014). JSON-LD 1.0: A JSON-based Serialization for Linked Data. Technical report, W3C.

Starr, J., Willett, P., Federer, L., Horning, C., and Bergstrom, M. (2012). A Collaborative Framework for Data Management Services: The Experience of the University of California. *Journal of eScience Librarianship*, 1(2):109–114.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1284v1 | CC-BY 4.0 Open Access | rec: 6 Aug 2015, publ: 6 Aug 2015