Building Hybrid Online/Offline Scholarly Information Resources

Ryan Shaw & Patrick Golden

The PeriodO project

http://perio.do
• Pros and cons of common publishing strategies
• A “new” strategy: hybrid online/offline publishing
• Case study: the PeriodO period gazetteer
• Discussion
Scholarly information resources

- Research datasets
- Scholarly collections
- Knowledge organization systems
Research datasets

Scholarly collections

Stanford Digital Forma Urbis Romae Project

ID AND LOCATION
Stanford # 10g(*)
AG1980 # 10g
PM1960 # 10 g
Slab # VIII-3
Adjoins 10aa 10f 10h 483ab

CONDITION
Located true
Incised true
Surviving true
Subfragments 1
Plaster Parts 0
Back Surface smooth
Thickness not yet available
Slab Edges 2
Clamp Holes 0
Tassello no

TECHNICAL INFO
Scanner gantry

Photograph (Mosaic) (570 KB)
Full resolution photo | Note about photographs

IDENTIFICATION
Large section of the Subura neighborhood (Subura) on the Oppian Hill, including the Clivus Suburanus (clivus Suburanus) and a small bath (balneum)

INSCRIPTION
None

3D Model Full model | PLY (159 MB)
Download the viewer | Note about 3D models

Knowledge organization systems

Mincius (river)

The Mincio river is a tributary of the Po river.

Canonical URI for this page:
http://pleiades.stoa.org/places/383708

Locations:
- location of Mincius fl. (modern)

Names:
- *Mincius* (30 BC - AD 640)

Place type:
- river, drainage

Makes a connection with:
- Padus/Eridanus (river)

Has a connection with:
- *Mantua* — by M. Pearce — last modified Nov 14, 2015 04:48 PM
Mantua (modern Mantova) is a city of Etruscan origin that received Roman citizenship in 49 B.C.

http://pleiades.stoa.org/places/383708
Publication strategies

• Bulk download

• Database-driven website
Constant availability

Distributed computing is a pain.

• Wifi can be flaky
• Local networks go down
• ISPs go down
• Hosts go down
• Domain names expire
Performance

“An interesting observation about network-based applications is that the best application performance is obtained by not using the network.”

Roy Fielding

Prerequisites for imports via MWDumper

Before using mwdumper, your page, text, and revision tables must be empty. To empty them, do this (note that this will wipe out an existing wiki):

```
In SQL: DELETE FROM page; DELETE FROM text; DELETE FROM revision;
In maintenance directory: php rebuildall.php
```

Import dump files with MWDumper

Sample command line for a direct database import:

```
java -jar mwdumper.jar --format=sql:1.5 pages_full.xml.bz2 |
mysql -u <username> -p <databasename>
```

or

```
cd mwdumper/src
javac org mediawiki dumper Dumper java
cd ..
java -classpath ./src org mediawiki dumper Dumper --format=sql:1.5 pages_full.xml.bz2 |
mysql -u <username> -p <databasename>
```
Versioning problems

• How to know which version of a dataset you are using?

• How to know when a new version of a dataset is available?

• How to know what has changed?
Provenance issues

• Few datasets publish fine-grained attribution data

• Presumably this data is available somewhere, but it may not be easily accessible
Database-driven website

LAMP Architecture
- Linux - OS
- Apache - Web
- MySQL - DB
- PHP - Script

Your online exhibit is one click away

Let Omeka.net host your collections, research, exhibits, and digital projects.

Sign Up!

Collaborate in class with Omeka

Have you thought about using Omeka.net with your students this fall? With one-click options to install plugins, sites are quick to set-up and easy to manage. Your students can sign up for their own Omeka.net accounts and build their own sites as well as collaborate on sites which you own. Small groups can work...
Simpler identity and version control
Expensive

Greater initial investment
Creating a database-driven website, even using a CMS, takes more time than just uploading a data dump

Greater ongoing costs
Hosting fees, keeping servers patched and secure
Preservation headaches

“While page collectors provide a good service, especially for sites that do not allow access to the database that drive them, they fail on several fronts.”

Paul Jones

http://ibiblio.org/pjones/blog/the-web-going-dark-preserving-and-serving-aging-websites/
Database-driven website

Pros:

• User-friendly

• Simpler identity and version control

Cons:

• Greater initial investment

• Greater ongoing costs

• Preservation headaches
Bulk download

Pros:

- Constant availability
- Speed

Cons:

- Not easily usable for non-programmers
- Versioning and provenance issues
A “new” strategy

• Constant availability and speed of bulk downloads
• User-friendliness of a database-driven website
• Centralized identity control only where needed
Hybrid online/offline resources

• Work with a local copy of the entire resource

• Bundle the raw data with an offline client for browsing, sorting, filtering, search, & visualization

• Only use the server to manage changes
Goals

• Document definitions of historical period names, in a machine-friendly way

• Provide stable identifiers for period definitions so data curators can reference them unambiguously

• Provide tools to find and compare period definitions and to understand patterns of periodization
Initial sources

- ARIADNE network
- British Museum
- German Archaeological Institute
- Archaeology Data Service (UK)
- English Heritage
- Hypermedia Research Unit at the University of South Wales
- Pleiades / Pelagios
- Fasti Online
- ArcheolInf project
- University of Oxford’s CLAROS project
- UCLA Encyclopedia of Egyptology
- Levantine Ceramics Project
Data model

To be included in the gazetteer, a definition must

• give the period a name

• impose some temporal bounds on the period

• have some implicit or explicit association with a geographical region, and

• have been formally or informally published in some citable source.
Periods are concepts

&lt;p06v8w44jbs&gt;
  a skos:Concept ;
  skos:prefLabel "Неоліт" ;
  skos:altLabel "Neolithic"@eng-latin, "Неоліт"@ukr-cyrl .

• skos:prefLabel is used for the name of the period exactly as given in the original source.

• skos:altLabel is used for language-specific names of the period, assigned by PeriodO curators.
Periods are concepts

• **skos:note** is used for notes about the period definition that appeared in the original source.

• **skos:editorialNote** is used for administrative or editorial notes added by the PeriodO curators.
Temporal extent

- `skos:prefLabel` is used to textually describe the interval exactly as given in the original source.

- `time:hasDateTimeDescription` is used to describe the interval in a more structured fashion.

```xml
<node id="node1" label="Transitional (Philia Culture)"
      starts="2600/2500 BCE"
      ends="2300 BCE"
      description="earliest: -2599, latest: -2499"
      />
```
Spatial extent

- `skos:prefLabel "Urartian Eastern Anatolia (900–600 BC)"`;
- `periodo:spatialCoverageDescription "Eastern Anatolia"`;
- `dcterms:spatial <http://dbpedia.org/resource/Armenia>,`;
  `|<http://dbpedia.org/resource/Turkey> .`

- `periodo:spatialCoverageDescription` is used to textually describe the spatial extent exactly as given in the original source.

- `dcterms:spatial` is used to link a period definition to descriptions of locations in gazetteers.
Source

<p0rqpwq4prr>
  skos:prefLabel "Deuxième époque" ;
  skos:inScheme <p0rqpwq> .
</p0rqpwq>

<p0rqpwq>
a skos:ConceptScheme ;
dcterms:source [
  dcterms:isPartOf <http://www.worldcat.org/oclc/60543214> ;
  bibo:locator "page 137"
].
</p0rqpwq>

<http://www.worldcat.org/oclc/60543214>
  dcterms:creator <http://viaf.org/viaf/22159335> ;
  dcterms:issued "2004"^^xsd:gYear ;
  dcterms:title "Le tophet de Salammbô à Carthage : essai de reconstitution" .
</http://viaf.org/viaf/22159335>
  foaf:name "Bénichou, Hélène." .

```

"@context": { "_JSON-LD context elided" },
"id": "p0fh3zc", 
"type": "PeriodCollection"
"source": { 
"title": "The Celts in Iberia: An Overview",
"creators": [ 
{ "name": "Alberto J. Lorrio" },
{ "name": "Gonzalo Ruiz Zapatero" } ],
"yearPublished": 2005,
},
"definitions": { 
"p0fh3zcqs6h": { 
"id": "p0fh3zcqs6h",
"type": "PeriodDefinition",
"label": "Early Iron Age",
"originalLabel": { "en-latin": "Early Iron Age" },
"spatialCoverage": [ 
{ "id": "http://dbpedia.org/resource/Galicia_(Spain)",
"label": "Galicia" } 
],
"start": { 
"in": { "year": "-0799" },
"label": "800 B.C." 
},
"stop": { 
"in": { "year": "-0399" },
"label": "400 B.C." 
}
}
```
[{
    "path": "/periodCollections/p0fh3zc/definitions/p0fh3zcqs6h/stop",
    "op": "add",
    "value": {
        "in": {
            "year": "-0449"
        },
        "label": "450 B.C."
    }
},
{
    "path": "/periodCollections/p0fh3zc/definitions/p0fh3zcqs6h/editorialNote",
    "op": "add",
    "value": "The end of this period was incorrect due to a data entry error."
}]

The PeriodO client

http://n2t.net/ark:/99152/p0
Challenges

• IndexedDB support on the client

• Client-side RDF libraries still maturing

• Cross-origin resource sharing (CORS) support on Linked Data servers

• HTTPS issues
Workflow

1. User downloads latest canonical dataset.
2. User proposes changes to the canonical dataset by POSTing patches to the server.
3. User authorized as curator reviews and accepts proposed changes.
4. Users may request data from, and propose changes to, multiple servers.
5. Patches can also be exchanged between users via email or any other communication channel.
Identity management

- California Digital Library's EZID service for managing ARK identifiers
- ORCID for managing contributor identities
ARK Identifiers

http://n2t.net/ark:/99152/p0rqpwq4prr

name mapping authority hostport

name assigning authority number

name
Reducing risk of failure

1. Dataset

2. Dataset with version control

3. Collaboratively-maintained dataset with version control

4. User-friendly browsing / sorting / filtering

5. User-friendly visualization

6. User-friendly editing
Related and future developments
Linked Data Fragments

Linked Data Fragments client
Enter or choose a SPARQL query below and see how your browser solves it using only triple pattern fragments.

Choose datasources:

Type a SPARQL query:
SELECT ?dessert ?fruit
WHERE {
      dbpedia-owl:ingredient ?fruit.
}

...or pick an example query:
Desserts made with plants

Stop execution

Query results:
?dessert: http://dbpedia.org/resource/Apple_cake
?fruit: http://dbpedia.org/resource/Apple

http://linkeddatafragments.org/concept/
Dat

InterPlanetary File System (IPFS)

IPFS Alpha | Why We Must Distribute The Web

https://ipfs.io

https://youtu.be/skMTdSEaCtA?t=714
“An architecture for the Web must... be designed with the context of communicating large-grain data objects across high-latency networks and multiple trust boundaries.”

Roy Fielding

<table>
<thead>
<tr>
<th>Label</th>
<th>Earliest start</th>
<th>Latest stop</th>
</tr>
</thead>
<tbody>
<tr>
<td>2nd Millenium BCE</td>
<td>-2000</td>
<td>-1000</td>
</tr>
<tr>
<td>2nd Millenium BC Egypt (2000-1000 BC)</td>
<td>-2000</td>
<td>-1000</td>
</tr>
<tr>
<td>2nd Millenium BC Levant (2000-1000 BC)</td>
<td>-2000</td>
<td>-1000</td>
</tr>
<tr>
<td>3rd millennium BC</td>
<td>-3000</td>
<td>-2000</td>
</tr>
<tr>
<td>4th millennium BCE</td>
<td>-4000</td>
<td>-3000</td>
</tr>
<tr>
<td>13th Century AD Eastern Mediterranean (AD 1200-1300)</td>
<td>1200</td>
<td></td>
</tr>
<tr>
<td>16th Century</td>
<td>1500</td>
<td></td>
</tr>
<tr>
<td>17th Century</td>
<td>1600</td>
<td></td>
</tr>
<tr>
<td>17th Century (1600 - 1699)</td>
<td>1600</td>
<td></td>
</tr>
<tr>
<td>17th Century: 1st half (1600 - 1649)</td>
<td>1600</td>
<td></td>
</tr>
<tr>
<td>17th Century: 1st quarter (1600 - 1624)</td>
<td>1600</td>
<td></td>
</tr>
<tr>
<td>17th Century: 2nd half (1650 - 1699)</td>
<td>1650</td>
<td></td>
</tr>
<tr>
<td>17th Century: 2nd quarter (1625 - 1649)</td>
<td>1625</td>
<td></td>
</tr>
<tr>
<td>17th Century: 3rd/4th quarter (1625 - 1649)</td>
<td>1625</td>
<td></td>
</tr>
<tr>
<td>17th Century: 3rd quarter (1625 - 1649)</td>
<td>1625</td>
<td></td>
</tr>
<tr>
<td>17th Century: 4th quarter (1675 - 1699)</td>
<td>1675</td>
<td></td>
</tr>
</tbody>
</table>

Thanks

http://perio.do
https://github.com/periodo
@perio_do
ryanshaw@unc.edu
ptgolden@email.unc.edu