World-Historical Gazetteer Project Plan
February 2018

Our plans for the World-Historical Gazetteer are taking shape. Here is a broad picture of what we plan to build over the next 2 1/2 years, followed by some details.

OVERVIEW

The World-Historical Gazetteer project (WHG) will assemble a data store of place names and annotations, develop graphical and programmatic interfaces for using and contributing to it, and foster a community of interest committed to sustaining the project beyond the three-year term of our NEH grant. The WHG is a linked open data project, with software maintained in a public repository. Among the guiding principles of our data and software development efforts are usefulness and usability.

The spatial scope of WHG is global. Its temporal scope, though not strictly bounded, will focus on the centuries after 1500 CE. Naturally, the data store cannot be comprehensive in any sense after three years, however the system is being designed to accommodate contributions indefinitely. We are aiming to have significant depth at launch for a few region/period combinations, including the trans-regional Atlantic World, terrestrial and maritime Eurasian connections, and Early Modern Europe.

ARCHITECTURE

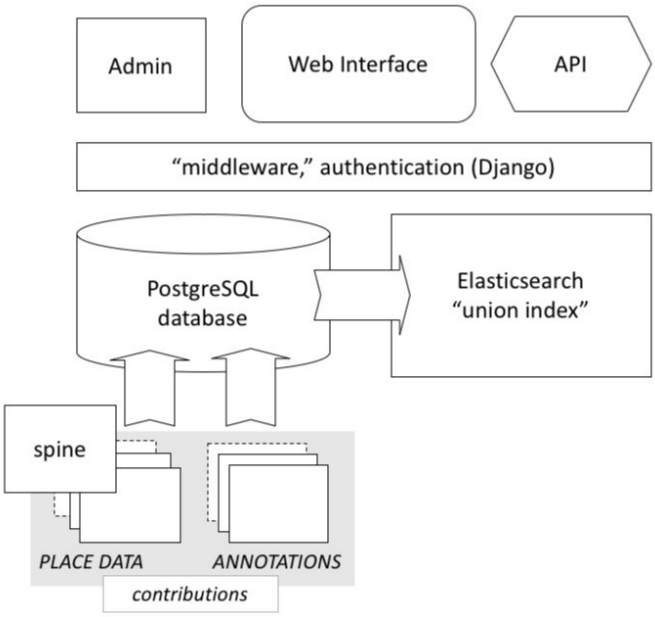

Fig. 1 - Schematic Architecture of WHG System

The structure of the WHG data store will, by design, closely resemble that of the Pelagios project's Peripleo, joining it as the second place-centered hub in an emerging Linked Pasts Network. As illustrated in Figure 1, WHG will be a) an aggregator of

gazetteer data from multiple sources, and b) a repository of contributed annotations linking published records of historical "items" (e.g. events, works, datasets, people) to place identifiers in the gazetteer. In time, contributions of any size meeting scholarly criteria will be facilitated. Our core gazetteer "spine" and initial additional sources are largely set for now, as described in the "Data" section below.

We are coordinating with Pelagios to ensure that the interconnection formats used by Peripleo and WHG for both gazetteer and annotation contributions will be mutually intelligible, if not identical.

THEMES

While WHG and Peripleo technical implementations will be similar in many respects, their content will differ in scope and theme. Peripleo has focused to date on the ancient Mediterranean, and has begun to expand somewhat in space and time. As the hyphen in our name indicates, WHG will have a distinctively world-historical perspective. It is global, not temporally bounded, and will foreground connections between places in several ways.

Movement. Historical routes and route systems will be a featured place type, and we will actively solicit annotations of data about journeys (pilgrimages, exploration, etc.) having waypoint places within our "union index." Place records will thereby include journeys and routes for which they were an attested waypoint.

Setting. Inhabited places have an environmental context; the physical geographic features in our spine gazetteer (rivers, mountains, biomes) will allow us compute these spatial relations and include them explicitly in place records.

Containment. We will actively gather and solicit time-indexed regionalization data to expose relationships between places falling within named regions of all kinds.

Fig. 2 - Regions in D-Place (partial). One of many distinctive world regionalizations.

We will also experimentally link place names with Library of Congress subject headings and name authorities, enabling automated bibliographic listings.

DATA

The WHG data store will comprise a "spine" gazetteer of approximately 20,000 places and two indexes: 1) a union index of gazetteer data from contributors and public sources, and 2) an index of annotated "items" associated with places. The spine gazetteer will naturally be among the initial datasets indexed.

Master copies of data for the WHG spine, contributed gazetteer records, and annotation data will be held in a PostgreSQL relational database, from where they can be readily manipulated and exported in a variety of formats.

To the extent possible, place records in the spine will be matched to identifiers from the Getty Thesaurus of Geographic Names (TGN), DBpedia, GeoNames, Wikidata, and Library of Congress (LoC) name and subject authorities. This linking will provide many additional name variants, and the potential for experimental text mining and analysis. For example, article text from Wikipedia can provide references to people and events related to a given place.

Once published, the WHG spine will remain essentially fixed, but the system is being designed to accommodate both big and small contributions to the union index on an ongoing basis. Several initial contributors are listed below, and discussions are underway with several large projects currently aggregating data for particular region/period combinations.

Spine

Initial sources for the spine include:

- Atlas of World History (Black, 1999; ~10,000 cultural and natural features)
- D-Place (~ 2000 "societies," related languages; regions)
- Getty TGN (several thousand places sourced to historical maps, texts, and indexes)
- Natural Earth (modern rivers, lakes, mountain ranges, coastlines, country boundaries)
- World Resources Institute (254 major watersheds)
- World Wildlife Federation (83 global ecoregions)
- NOAA (ocean currents)
- DBpedia (for name variants and abstracts)

Additional indexed data

Sources for non-spine data to be indexed so far include:

- Getty TGN (a larger subset, probably 200k-300k names)
- TGAZ (temporal gazetteer of >71k names; offshoot of Harvard's China Historical GIS)

- "The Alcedo Gazetteer" (Diccionario geográfico-histórico de las Indias Occidentales ó América, 1786; ~10k places)
- Voyages: Trans-Atlantic Slave Trade Database (~800 places)
- Old World Trade Routes (OWTRAD; ~4k settlements, ~5k routes)
- Linked Places (named historical routes)
- Seshat Global History Databank (400+ historical polities, "natural geographic areas")
- Selden Map of China (ports and other settlements; routes; data partner: Robert Batchelor)
- Atlas of Maritime Buddhism (ports; data partners: Lewis Lancaster and Jeanette Zernecke)

MODELS & FORMATS

The WHG conceptual model of place corresponds with that of the Pleiades project, as "a geographical and historical context for Names and Locations." Places are the settings for earthly phenomena, and place names—closely tied to human experience—are one kind of answer to "where" questions. Places typically have multiple names in multiple languages, which, along with their physical extent (location, size and shape) can change over time. It is useful in gazetteers to assign one or more types to places (e.g. settlement, port, river), which can also change over time. We are adopting a subset of the hierarchical "placetypes" found in the Getty's Art and Architecture Thesaurus (AAT). Locations are modeled as positions on the earth surface, represented as geographic coordinates, and time-indexed when possible. For many historical places, locations are limited to estimated centroids. Point, line and area geometry in this context are all normally imprecise. Historical route features may be more or less "geographically embedded"; that is, waypoints are always tied to places and/or locations but the actual paths between them may be unknown and therefore indicated only by straight lines.

Our conceptual model of place will be represented formally in multiple formats within the system and in data exports. The default serialization offered will be GeoJSON-LDT, a work-in-progress extension of GeoJSON-LD that allows "when" elements in several locations within a record. CSV and RDF will also be made available. The following is a shorthand "schematic" view of a Place record in the WHG spine (subject to modification in the near future; see the GeoJSON-LDT repository for details of "when"):

```
FeatureCollection {
   @context: "http://geojson.org/geojson-ld/geojson-context.jsonld",
   @context: "http://whgazetteer.org/models/whgazetteer.jsonld",
   features: [{
      type: Feature,
      geometry: {
        type: GeometryCollection,
        geometries: [
            {  type: [Point|Line|Polygon],
               coordinates: [[x, y], ...],
               loc_attestations: [
                  {source, contributors: [], when: {}, certainty}],
            }
       ],
      },
      when: {#computed from the "when"s of name and loc attestations},
      properties: {
         name_attestations: [
          {name, language, when: {}, isPreferred, source, certainty}],
         relations: [
            {relation, placeUri, when: {}, source, certainty}],
         snippets: [
          {uri, language, description, source, certainty}],
         links: [
          { type: [exactMatch|closeMatch|seeAlso|seeFurther], uri,
            [who, when, note]}],
         modifications: [
            {timestamp, who, note}],
         note: ""
      }}
  ]
}
```

INDEXES

Data for places and annotations will be ingested to a PostgreSQL database, then exported to a JSON format for ingest into high-speed Elasticsearch indexes which the graphical web interface and API will draw more or less directly from.

Union index of places

The JSON format for the place index will aggregate all records for a given place in an "is_conflation_of" element. So, if the spine has a record for Abydos (Egypt) and two contributors include a corresponding Abydos record, the index record will hold three shortened records, each potentially having distinct source(s), geometries, temporal expressions, alternate spellings, etc., and both GUI and API will represent that union of records. This corresponds to the model developed and implemented by Rainer Simon for Peripleo, along these lines:

```
{
   id: "whg:1234",
   representative_title: "",
   representative_geometry: [[x,y]],
```

```
   suggest: ['variant1, variant2, ...'], #for autocomplete
   temporal_union: [min_yr, max_yr],
   is_conflation_of: [
      {
         id: "",
         dataset: "",
         uri: "",
         title: "",
         names: [{lang: "", name: ""}, ...],
         exact_matches: [{uri: "", name: ""}, ...],
         close_matches: [{uri: "", name: ""}, ...],
         place_types: [],
         when: {{start}, {end}},
         descriptions: [{descrip: "", lang: ""}, ...]
      }
   ]

}
```

Annotations index

The format for contributed annotations is a work in progress.

API

All indexed data (places and annotations) will be made accessible via an Application Programming Interface. The API will return sets of records filtered in various ways, in multiple formats, including GeoJSON-LDT (readable by all GeoJSON-compatible software), CSV, and RDF. Complete data dump files will also be made available.

WEB APPLICATION (GUI)

We have compiled a list of 75 "user stories" representing desired functionality for a web interface as conceived and expressed by our project team and Advisory Committee, growing out of discussions at a Sep 2017 gathering in Pittsburgh. Those functional requirements will be met by the API and web app features in the following categories:

- **Modal "splash" entry page** highlighting pilot applications
- **Initial heat map** w/filters + zoomTo() for sources, regions
- **About section** [motivation; methods; timeline; criteria; participants; contributors; documentation (models & vocabularies, platform); contact]
- **Contribute section** [how to (contributions, data dev); suggested additions; recently added]
- **Resources page** [gazetteers; bibliography]
- **Blog**
- **Search** [w/autocomplete, against all indexed records]
- **Advanced Search** [filters: place type, timespans, datasets; generate an API query]
- **Disambiguation page**
- **Place record landing page**
  [names; description(s); maplet w/all attested locs + full screen option + print map

button; temporal visualization; links to contributors' records; connections (routes, geo features); annotated items (events, works, datasets); bibliography; Share button; Propose Annotatation button]

- **User registration** (user, editor, contributor, admin)
- **MyPlaces** feature [save records to named shareable, mappable sets; Download button