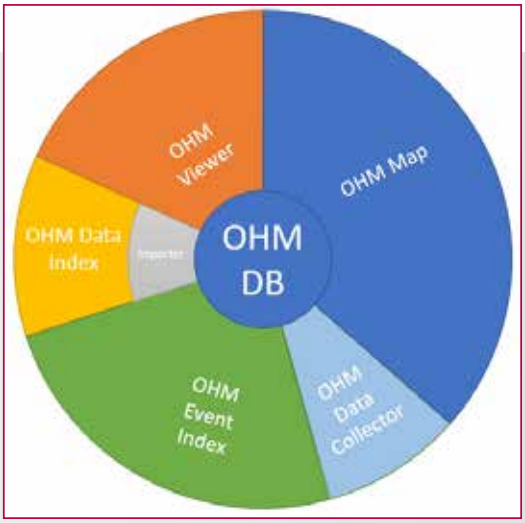
Open History Map – Status of the Project

by Marco Montanari, Lucia Marsicano, Raffaele Trojanis, Silvia Bernardoni, Lorenzo Gigli

Open History Map, an open map of the past that was already presented as a concept a few years ago, is now in its first year of functioning infrastructure and collects around 150GB of data from around 90 sources. The platform is open in all of its aspects and enables research groups to create new importers for their own open datasets.



Open History Map aims to be a platform that collects and displays data about our past with modern tools and considering several user groups as reference. This means that, unlike classic "collaborative" approaches, it does not rely on a users single contribution but more on blocks of contributions that are bound to research and to possible publications. This also means that despite the original approach wanted to be very ontology-centric [Montanari et al. 2015] and [Bernardoni et al. 2017], this maximalist requirement had to be changed to a more realistic and bottom-up approach, where several research groups use different ontologies and different representations for similar aspects, as each research group can give a very specific insight into one specific point of view on the same phenomenon in history or, on the contrary, similar points of view on very distant phenomena over time.
This led us to a less strict integration methodology, where the ontology is still there, on our side, to enable the interpretation of the data in different ways and to guide users in the interoperability aspects of data cleaning. On the other hand, this lead us to a completely different approach to the infrastructure and to the architecture of the platform. Specifically, it meant Open History Map was no longer just "the map" and the data connected to it, but also the classification system, the importers, the methodology, the tools around the map itself. Choosing to open the doors to difference made the platform more robust and more solid. This also opened up for different elements we had not, at the beginning, considered. The "map" usually contains only "things that are there", i.e. buildings, trees, objects that shape the world. What about all those elements that do shape the world in a non-direct way? Part of the context we aspired to create about things happening in our past with the original project is defined by the events that happened, because the "material" world and the "ephemeral" world are incredibly interwoven and interconnected.
All these elements brought us to rework some of the concepts behind the platform itself. Open History Map is now an ecosystem of tools and points of view on the various aspects of data. This locates the project at a crossroads of several branches of knowledge and study, that can be summed up in the broad concept of Digital/Spatial Humanities. Nonetheless the platform is built with the idea not to force providers to bring data in a specific form, but to have every provider keeping her ontologies and formats in order to facilitate the work on the leaves of the ecosystem despite giving more work to the data integrator [Zundert, 2012].

Specifically, the ecosystem is now comprised of the following components:

- Open History Map Data Index, to classify data sources and papers;
- Open History Map Data Importer, to define transformations and extractors for the datasets;
- Open History Map Viewer, to display photos, paintings and reconstructions of moments or views of the past located correctly in time and space;
- Open History Map Event Index, to collect references to anthropic and natural events in time and their connections and relationships;
- Open History Map, the core map of the project.

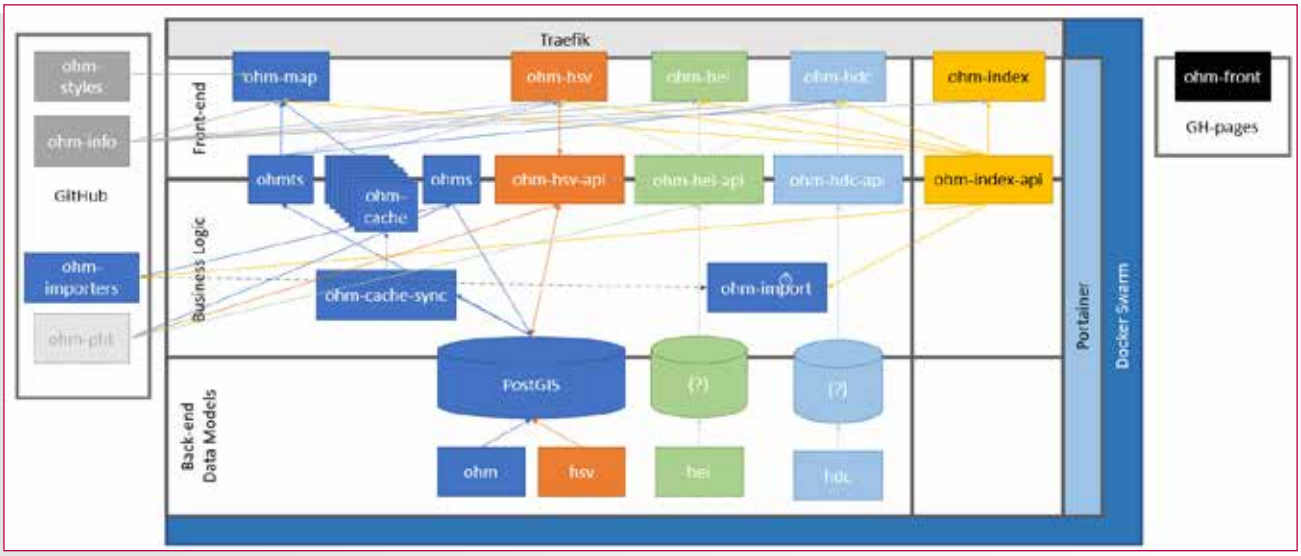
In addition to these already existing parts, the following components are either planned or built but being rewritten to be integrated into the broader platform:

- Open History Map Data Collector, to contain statistical data about several phenomena about the various time periods in the various places collected;
- Open History Map Public History Toolkit, a tool to help researchers and groups to collect data about historical events and periods in a structured manner;
- GeoContext [Marsicano 2018], a tool to create simple visualizations of self contained research data; this will be integrated into the OHM Public History Toolkit in order to make the datasets created explorable and navigable as single platforms.

All of these components are different points of view and different ways to read the data contained in the core data storage. Technically it is not just one database but it is a constellation of databases interconnected via high level APIs enabling the maximization of the data throughput for the end-user.
Every single one of these components has a well defined structural core ontology, on which we can rely to do the majority of the work, and a weaker "content" ontology part, where elements can be mapped or assigned in various ways and with varying degrees of quality. The core part is the one that the components use to define the main activities for their own function and the general purpose APIs that interconnect the infrastructure, the rest gives the possibility to expand and adapt the specific dataset or data point into a more complex and advanced structure without requiring the infrastructure to change. This flexibility is radically important in the first phases of the project, as the ontology can emerge from the data collected and can be fixed over time.
The core map has the largest amount of data, being a collection of polygons, lines and points representing the structure of the past in its various moments described by historians, documents and digitized maps. Technically, the information stored is described in a way that is very similar to the openstreetmap ontology, except for the fact that for every dataset imported the source is an identifier that references one specific dataset described in the Data Index.
The other major data collectors are the Event Index and the Viewer. For each the geometries are simpler, being points, and the API transforms these points into more complex structures connected either by the same subject, by context or by other references. This gives the possibility to visualize connections between items and create lines and convex containers representing the area of a specific phenomenon (i.e. a war, a cultural presence) or the track of a specific object over time (ship, person, army unit).
The smallest data collection is the one related to the Data Index. It contains only metadata about the data sources that are imported into the system. This component is one of the core elements of the whole system because it is on this that the definition of the system relies, as the infrastructure does not in itself have OHM-owned data. The importer relies on this component to define the import methodologies for specific datasets.

THE DATA INDEX

The original Open History Map infrastructure did not define a repository for data sources, while the data quality had to be considered as a secondary element in the definition of the map itself. The infrastructure already had a defined hierarchy of data quality definitions, but this definition was more addressed towards the quality of the single specific geographic information, than the source itself as a whole. The change of paradigm in the import of the data required a radical change in the quality assessment itself, because obviously reliable primary sources are way more useful than unreliable ones, and these are themselves more useful than reliable hearsay sources, for example.
For this reason, the definition of source reliability and source quality depend on several important factors and are the result of a series of simplifications based on the most common models for the evaluation of data quality. In addition to a general model for Digital Humanities, a broader information and data quality paradigm was analyzed, looking also into metrics for the world wide web, as many elements could be applied in the Data Index, as it is more generic than a specific collector.
This considered, [Knight, s.d.] defines 20 dimensions for Information and Data Quality. Some of these are very web-oriented (i.e. Consistency, Security, Timeliness and others) and as such not relevant in our specific case. Other dimensions are, on the other hand, very topic specific (i.e. Concise, Completeness, Relevancy) and this depends on what the research we are collecting is about: a study of a very



specific topic relevant to a small number of cities in the ancient roman age will be way more concise than a broad study on a very common phenomenon in modern ages.
In [Akoka et al. 2021] a 7-level hierarchy is defined for the categorization of imprecise temporal assertions in an application to the prosopographical database definition area. Based on these two approaches, with the latest iteration of the Open History Map platform, the Data Index was introduced, creating the structure to collect, classify and evaluate the data quality of the various sources that the platform collects and imports. The importance of an external metadata information collector is incredibly important in order to be able to define methodologies and criteria to uniform the import modes for the single datasets or dataset groups.
The dimensions defined for this collector are divided into three classes, Objective, subjective and process and these into six dimensions. The subdivision is the following:
Space coverage is done using Geonames Identifiers and as such creating a dynamically populated tree of areas covering the various researches. For the period a beginning time and end time of the period covered are the elements classified.
The topics that are being collected as of now are the following:

- **Agriculture** - These datasets and researches regard the area of agriculture, such as for example identification of agricultural land use or of crops cultivated in the past in specific areas.
- **Climate** - These datasets and research cover the changes in climate in the past.
- **Economy** - These datasets and researches regard the changes in economy and the possible normalizations of the past with modern economic conditions.
- **Entertainment** - These datasets and researches regard the world of entertainment, from circus to theatre to cinema.
- **Industry** - These datasets and research cover the changes in the industrial complex of the past.
- **Infrastructures** - These datasets and researches regard the transportation infrastructure and its evolution. It is further subdivided in air, water and land transport.
- **Politics** - These datasets and researches regard the changes in borders.
- **Religion** - These datasets and researches cover the distribution and activities related to religion
- **Urban** - These datasets and researches regard urban evolution over time.
- **Ephemeral** - These datasets and researches regard movements of people, ships and anything mobile
- **War** - These datasets and researches regard detailed aspects of war.
- **Geography** - These datasets and researches regard the changes in nature, both totally natural and anthropic

Class	*Dimension*	*Meaning*
Objective	Spatial Coverage	What area or areas does the research cover?
	Period	What time frame or time frames does the research cover?
	Topic	What topic or topics does the research cover?
	Subtopic	What kind of information is collected about the topic?
Process	Reliability	How reliable is the research? What methodologies have been used?
Subjective	Quality	How precise is the information collected?

All of these topics have specific sub-classifiers defining the kinds of elements collected. Specifically, these sub-classifiers are:

- **Location** - These researches and datasets give us the punctual localization of analyzed items.
- **Structure** - This subtopic regards research that resulted in the creation of open datasets of the planimetries of buildings, cities, areas.
- **Model** - This subtopic regards activities that resulted in the creation or the collection of 3D models for specific buildings or items of the main topic.
- **Events** - These datasets and researches give us timelines (as datasets or not) about the specific topic in the specific analyzed period.
- **Usage** - These datasets and researches give us context and modes for the fruition.
- **Indexes** - These datasets and researches analyze specific indicators, such as for example salary over time or GDP normalized or population.
- **General** - These datasets and researches covering the main topic in general or without a specific viewpoint.

These four levels of classification enable the positioning of the specific dataset in a multidimensional grid. This enables the creation of the main visualization of the Data Index.

The quality and reliability evaluation are given as two values ranging from 1 to 6. The reliability score is divided as follows:

- 6 - Academic peer-reviewed research / Excavation Report;
- 5 - In-period source material;
- 4 - Review papers;
- 3 - Non peer-reviewed research;
- 2 - Local public history activities, eventually supervised, but not guaranteed in any way;
- 1 - Hearsay and oral tradition.

While the data quality is divided as follows:

- 6 - Precise dataset with well defined and documented tools, with a validatable high level accuracy
- 5 - Lower fine granularity of data; data is available but less precise.
- 4 - Verifiably incomplete data, some information is missing for accurate identification
- 3 - Uncertain data, specification of uncertainty in a major part of the data.
- 2 - Uncertain data with no references to other datasets for cross-validation.
- 1 - Low quality dataset, with obvious discrepancies and/or errors

The reliability descriptor represents the intrinsic quality of the methodology applied in the publication of the data. On the other hand the data quality depends on the precision of the published data.
The current infrastructure is currently based on the storage of items in a public Zotero collection where all the collected researches are described as structured tags.
The following is the description of the CLIWOC dataset:
The datasets can not be described from the front-end and for this reason these are currently defined by a name and, in addition to the previously mentioned ontology a descriptor for the single dataset or for the whole collection of datasets associated with the project.

A CLOUD FIRST ARCHITECTURE FOR THE DH

One of the key principles of cloud-first infrastructure design is to rely on pre-existing services already available within the cloud infrastructure we are working with. This is in principle optimal for high burn-rate startups with huge funding but it might not be an ideal solution for low budget digital humanities projects. For this reason we designed an infrastructure that could take advantage of an abstraction of this basic principle and of other general principles of software architecture in order to create an efficient architecture that complies with the needs and requirements that define the Open History Map platform in all of its aspects [Montanari 2021] but can also offer some of the high-level tools that cloud infrastructures typically lack for use in Digital Humanities.
One of the main factors taken into account during the architectural planning process was the importance of delaying and deferring decisions as much as possible both for us, using the architecture as well as for the architecture expandability itself [Martin 2017]. This principle is true when designing an application, an API and a complex architecture. For this reason most of the API orchestration is delegated to the various interfaces that use the data and compose the element in the way that fits best. For example, the map front-end uses the Data Index API in order to display information about a given source, using its id as symbolic reference (which is in itself a symbolic reference to the identifier defined in Zotero). This means a partial delegation of the knowledge of the inner workings of one specific part of the infrastructure to other elements that might not have to know, in principle, anything of it, but on the other hand, being part of the same ecosystem means that many elements can be cross referenced between various interfaces.

Space coverage	*ohm:area*	geonames:6295630
Time coverage	*ohm:from_time*	1750
	ohm:to_time	1855
Topic classification	*ohm:topic*	ephemeral
	ohm:topic:topic	location
Data quality	*ohm:source_quality*	6
Data reliability	*ohm:source_reliability*	5

The infrastructure is divided into several macro-areas, each of whom covers one specific aspect of the platform. Wherever possible the vertical architecture of the macro-area has been structured with the same pattern:

- Database (postgres/mongodb/redis/filesystem/influxdb)
- Writing API, that also controls the database initialization infrastructure (a python/flask microservice)
- Reading API (a python/flask microservice) with eventually Tile Server
- Front end (an angular application)

This very basic template may or may not have all of its components, as for example, a service might not need to write to a backend, while another service might only be using interaction-less writing operations, and as such might not need a direct reading API. Obviously, if the reading operations are simple enough, they are integrated into the writing component and vice-versa.
The less cross-depending macro-area is the already explained Data Index, being simply a visualization of the data collected in the Zotero collection representing the sources used for populating all other areas. The infrastructure is totally stateless, it has no persistent database and downloads the current collection of sources once the docker image starts, creating just an on-the-fly database. The api only exposes the endpoints used by the interface and the APIs for other macro-areas to get source-specific data to display on their interfaces. All areas beyond this rely on its presence and on its being up-to-date.
The main user of the Data Index API is the Data Importer, an interface-less system that does the heavy lifting of transforming and importing data into the various databases coordinating the use of the various APIs. This service is stateless as well, not having a persistent database, yet it does connect to the local Docker socket, as it uses the Docker-in-docker methodology to spawn the machines that do the real ETL operations as well as the test database and test-APIs in case of import testing. The ETL code is downloaded from a repository and for each source identifier or source dataset identifier taken from the Data Index API the importer builds the specific docker image and launches it, writing, if already tested, directly on the various production APIs.

The other areas are all full stack infrastructures, as all rely on one or more databases, APIs and specific front-ends. Starting from the map, the data is stored in a PostGIS database configured automatically via the API module and already set up to be distributed across a cloud infrastructure via partitioning. The partitioning configuration is done both on layers (that are bound to topics) and year the dataset starts being valid. Setting up a variable grain to the time-dependent partitions enables a major optimization of the resources for the infrastructure, giving the possibility to move the storage of data-heavy periods (wars, moments of major changes) into separate databases. The API relies, in the writing part, on the possibility of using a redis-based buffer to have a workload manager deciding when to import specific items. This is very important because many polygons are very detailed, and the possibility to do an indirect upload of data gives the client better feedback on operations even in very complicated cases. The tile server is completely separated from the rest of the API and uses a stored procedure defined in the db-initialization part of the API. This enables an enormous optimization of the activities, as it relies on the database-native generation of MVT tiles for the specific requested tile.
Beyond the main map, all other tools are always cartographic as well, being the Event Index and the Historical Street View. Both contain mainly points, in contrast with the multiple types of geometries stored by the main map. The data, in the two cases, is bound to mapping events in time and (not always) in space and mapping documentation of the form of the world in history, respectively.
The Event Index collects data from various sources enabling the location of specific events in time and space, but also tracking particular subjects in their activities in time. For example it is possible to visualize the course of a specific ship over time, such as, in this case, the course of the Endeavour over its travels to New Zealand.
In addition to movements, the layer also contains data about events caused by anthropic causes of change (wars, battles, murders, births and deaths) as well as by natural causes (quakes, volcanic eruptions, various forms of disasters). This layer is a pure collection of time-space coordinates with general information about the event and a reference to the external source. This block contains a

modified tileserver, slightly different from the main map one, optimized for point management.
Finally, the Historical Street View macro-area is, like the Event Index, simply a point storage for not just space and time coordinates, but also the reference to external sources for documents, being photos, paintings of views, videos. These multimedia sources are not stored or cached in the system and are visualized directly from other providers. This is very important in order to guarantee the maximum independence from local storage.
In conclusion, the defined architecture guarantees a great amount of flexibility in case of additions to the system as well as simplicity, based on the templatization of the single structures. The transformation of the various parts into microservices gives the whole system additional reliability and resilience, enabling possible changes to the implementation and horizontal growth without stopping the infrastructure.

References

Bernardoni, Silvia, M. Montanari, & R. Trojanis. «Open History Map». Archeologia e Calcolatori n. XXVIII.2 - 2017. Edizioni All'Insegna del Giglio, 20172-01-01. https://doi.org/10.19282/AC.28.2.2017.44.
Montanari M., Raffaele Trojanis, Silvia Bernardoni, & Luca Tepedino. «Open history map - a new approach to open access for archaeology and cultural heritage». In 2015 Digital Heritage, 2:479–80, 2015. https://doi.org/10.1109/DigitalHeritage.2015.7419557.
Zundert, J. «If You Build It, Will We Come? Large Scale Digital Infrastructures as a Dead End for Digital Humanities». Historical Social Research / Historische Sozialforschung 37, n. 3 (141) (2012): 165–86.
Akoka J., Isabelle Comyn-Wattiau, Stéphane Lamassé, & Cédric Du Mouza. 'Conceptual Modeling of Prosopographic Databases Integrating Quality Dimensions'. Journal of Data Mining & Digital Humanities Special Issue on Data Science... (7 May 2021): 5078. https://doi.org/10.46298/jdmdh.5078.
Franke M., Ralph Barczok, Steffen Koch, & Dorothea Weltecke. 'Confidence as First-Class Attribute in Digital Humanities Data'. In Workshop on Visualization for the Digital Humanities (VIS4DH), 2019.
Knight, S.. 'Developing a Framework for Assessing Information Quality on the World Wide Web', n.d., 14.
Naumann F., & Claudia R. 'Assessment Methods for Information Quality Criteria', 27 October 2000.
Therón S., Roberto, Alejandro Benito Santos, Rodrigo Santamaría Vicente, & Antonio Losada Gómez. 'Towards an Uncertainty-Aware Visualization in the Digital Humanities'. Informatics 6, no. 3 (September 2019): 31. https://doi.org/10.3390/informatics6030031.
Martin, Robert C. *Clean Architecture : A Craftsman's Guide to Software Structure and Design*. 1° edition. London, England: Addison-Wesley, 2017.

Abstract

Open History Map, an open map of the past that was already presented as a concept a few years ago, is now in its first year of functioning infrastructure and collects around 150GB of data from around 90 sources. The platform is open in all of its aspects and enables research groups to create new importers for their own open datasets. In addition to that, OHM enables the visualization of "ephemeral" datasets, i.e. representation of vicinity for historical characters and vehicles, battles and events. The present work will analyze the status of the project and the contributions it is doing to the general DH and PH sector, specifically on source quality management and general cloud first architectures.
OHM is based on the collection of open datasets available online. The geographic precision as well as the informational quality varies a lot between sources, research teams, projects. These factors higlight the need of a tool to manage the data quality, which we called OHM Open Data Index, (https://index.openhistorymap.org) where we collect all sources we find and all datasets we import in order to analyze and display the general quality and/or lack of data.
The complexity of the infrastructure behind a project such as Open History Map required an original and cloud-first approach, enabling the optimization of every single aspect of the development as well as the deployment and the usage of the system. For this reason a cloud-first approach was used, trying to harness all the features of the most common FLOS software platforms in order to maximize the quality of the final product.

Keywords

Digital Humanities; Digital Archaeology; GIS; GLAM; Data quality; Software Architecture

Author

Marco Montanari, Lucia Marsicano, Raffaele Trojanis,
Silvia Bernardoni, Lorenzo Gigli
University of Bologna