"With the 5ame name and adrvocation of S.Juan there is another one, in the sámeprovince"- towards a digital edition of the historical-geographical dictionary of the Indies by Antonio de Alcedo

Stangl, Werner

werner.stangl@gmail.com
EHESS-CRH et CNRS-CREDA, Paris, France, projet TopUrbi

Brando, Carmen

carmen.brando@ehess.fr
EHESS-CRH et CNRS-CREDA, Paris, France, projet TopUrbi

Zúñiga, Jean-Paul

jean-paul.zuniga@ehess.fr
EHESS-CRH et CNRS-CREDA, Paris, France, projet TopUrbi

Haedo, Anahi

anahi.haedo@ehess.fr
EHESS-CRH et CNRS-CREDA, Paris, France, projet TopUrbi

Sometimes, the path “towards an edition” is truly a winding and interrupted one. Some years ago, the first author of the proposal wanted to exploit the rich information of the late 18th-century, 5-volume Diccionario histórico-geográfico de las Indias occidentales by Antonio de Alcedo, in order to add relevant, contemporary descriptions for the territories and settlements contained in his historical database HGIS de las Indias - but only got partial results due to lacking OCR quality and the problems of named entity recognition in the dictionary. Later, he joined the "LatAm Gazetteer" effort initiated by Gimena del Río Riande and Ben Brumfield aimed at using the dictionary (or its later English edition) as the backbone for a Latin American domain within World Historical Gazetteer (Karl Grossner). This effort, too - due to a lack in substantial funding and the complexities of crowdsourcing -, did not evolve into a concise resource. Idiosyncrasies of 18th-century lexicography and knowledge representation, particularities in Alcedo's structuring of the work, and layers of translation are further complications which make working with the resource a significant challenge.

The "TopUrbi" project led by Jean-Paul Zúñiga at the EHESS, Paris, now offers an environment for a more concentrated take on this monumental historical oeuvre, overcoming the multiple historiographic, programmatic, semantic challenges for modelling, correcting, annotating, and reconsiling/geoparsing the resulting text, digital edition and historical analysis which had so far obstructed major advances. The proposed short paper will guide through our process/workflows:

First, we will show the **TEI-model** of the dictionary, which must cope with the fact that the TEI module for dictionaries was designed for language dictionaries which contain semantic units/concepts, while a gazetteer essentially contains entities.

Second, we will sketch our efforts for **correcting OCR errors**, which of course concentrate on proper names of people and places as well as numbers. Our methods for this include three phases: Bulk replacements, algorithmic detection of likely errors, and close reading organized as a collaborative effort of project participants and students (Omeka-S, Scripto).

Third, we will outline our methods for **extracting structured information** bits from the more than 19000 entries, such as the type of place, its geographic situation, and full name (which might be different from the entry’s lemma). This task is particularly challenging: one particular issue is the existence of multiple entries under a single lemma which share a homonymic core part in their names (Alcedo has 116 toponyms under JUAN), another issue is the frequent use of relative language when referring to previous text elements or entries (“another one, in the same province”), a third one the considerable variety in the phrasing and order that such elements are presented by Alcedo.

Fourth, we will present our procedures and categories of **annotation** for named entities within the entries using natural language processing and machine learning tools.

Fifth, we will detail our strategies for **matching geographic entities** with historical gazetteers (HGIS de las Indias, Atlas digital da América lusa, WHGazetteer). A major (and often rather disregarded) concern in geoparsing are false matches, an issue that we will underpin. We will then sketch our quite conservative approach and the introduced checks to minimize false matches or to flag “grey areas” that require special attention or disambiguation, and to create tiers for the results of the algorithmic matching process. Serious challenges for this process are the only partially compatible spatial concepts between Alcedo and the gazetteers, the diverse provenance and temporal distance of information in Alcedo, which result in conflicting concepts even within the dictionary itself. Due to the scale and complexity, we limit this (for now) to settlements in the “Spanish sphere” and only to the entities of the entries themselves, not simple mentions within the text.

Sixth, we will show our approach for **manually reviewing elements** which cannot be matched. For these cases, we use the information provided by Alcedo and extracted by us as structured data to create dummy coordinates as a “starting hypothesis” for manual review. We will use the historical gazetteers, modern gazetteers/mapping services (OSM, GoogleMaps, Geonames, Getty-TGN...), georeferenced historical maps, historical sources, and modern research to understand the nature of the unmatched entries, to match them manually where possible and to locate and categorize the remaining ones.

The outcome of this complex effort will ultimately be a full digital edition of the Alcedo dictionary in TEI and a specific Alcedo gazetteer with structured and linked data.

Bibliography

Alcedo, Antonio de (1786-1789): *Diccionario Geográfico-Histórico De Las Indias Occidentales Ó América : Es Á Saber: De Los Reynos Del Perú, Nueva España, Tierra-Firme, Chile, y Nuevo Reyno de Granada. Con La descripcion de sus Provincias, Naciones, Ciudades, Villas, Pueblos, Rios, Montes, Costas, Puertos, Islas, Arzobispados, Obispados, Audiencias, Vireynatos, Go-*

biernos, Corregimientos, y Fortalezas, frutos y producciones; con expresion de sus Descubridores, Conquistadores y Fundadores: Conventos y Religiones: ereccion de sus Catedrales y Obispos que ha habido en ellas: Y Noticia de los sucesos mas notables de varios lugares : incendios, terremotos, sitios, é invasiones que han experimentado: y hombres ilustres que han producido. Escrito por el Coronel Don Antonio de Alcedo, Capitan de Reales Guardias Españolas. 5 vols, Madrid: Various Printers.