Finding Culture in Data Preparation: Interdisciplinarity and the Phone Book

Wrisley, David Joseph
NYU Abu Dhabi, United Arab Emirates
djw12@nyu.edu

Ampofo, Prince
NYU Abu Dhabi, United Arab Emirates
pla252@nyu.edu

Mirza, Wajahat
NYU Abu Dhabi, United Arab Emirates
mwm356@nyu.edu

Table of contents

1. Introduction

Our paper emerges from the imaging and data preparation phases of the Abu Dhabi Calling! Project, studying late twentieth-century telephone directories from Abu Dhabi, the capital city of the United Arab Emirates (UAE). The printed directories (1970-2005), housed in the Frauke Heard-Bey and David Heard Collection of the Archives and Special Collections Division of NYU Abu Dhabi Library, contain hundreds of thousands of names of individuals and businesses. These data provide a rich opportunity to study the spatial and demographic features of the capital, creating a kind of urban "social map" [*kharita ijtima'iyya*] (Al-Douaihi 2014). The data document a period of rapid development, during which the population increased twentyfold, from 62000 in 1970 to nearly 1.5M in 2021. The Abu Dhabi telephone directory bears an organizational structure imported from a Western context, but which is not a perfect container for the complexity of information about this non-Western city. Seen from this perspective, our paper argues that the cultural complexities of sources must be reflected not only in project design, but also in stepwise project progress (digitization, cleaning, modeling and analysis), especially as digital humanities methods are increasingly adapted to non-Western settings and sources.

Researchers have used city directories as historical documents in Western cities (König 2004; Khayrallah Center 2015; de Maupeou / Saint-Raymond, 2013; Milliken 2018; Wolf et al. 2020), extracting location-based information, indicators of profession as well as religious, sectarian or racial identification. To our knowledge, ours is the first project working with telephone directories from an Arab world setting. We have confronted numerous challenges, however, in reproducing such research in our context. First, the limited digitization facilities in our institution, combined with the restrictions of pandemic, have led us to create our own on-demand digitization workflow. Second, our phone directories contain no street addresses; instead for subscribers we have post office box numbers and phone numbers. Third, the Abu Dhabi telephone directories contain a diversity of naming practices (in the order, length and structure of names) which complicate downstream analysis. This means that even when data are extracted, cleaned and structured, other ways of imputing location and identity—key steps of spatial research—must be developed.

2. Complexities of digitization and OCR

As of December 2020, the imaging of the telephone directories has been completed. It was carried out during the pandemic in a safe corner of Archives and Special Collections with a Nikon D7000 camera and Sofortbild, producing images of 12-15 MB with a resolution of 4928x3264 per page and an estimated dpi of 400. Spine tightness led to unavoidable truncation and image warp for some pages.

A typical page in the directories contains four left- and right- justified columns as well as a subscriber name, post office box and a telephone number. Whereas initial attempts at OCR suggested that the segmentation of the multi-column page would be easier in Abbyy FineReader, we have opted for Tesseract to keep our project as close to open tools as possible and to combine it with other automated pre- and post-processing steps (Wolf et al. 2020). Additional points of interest in the OCR of these documents include block advertisements—quite common in the early 1980s—the size and position of which vary as the telephone directories become more voluminous. Furthermore, commercial listings are characterized by special fonts, outlining and color, which exacerbate issues in the OCR steps of greying and image thresholding. We have opted for a workflow of semi-automated dewarping for the pages, combined with automated cropping into columns and zooming, a process which for now excludes the block advertisements in favor of the subscriber lists.

Extraction of the subscriber data from the HOCR layer with accuracy has proved to be a time- consuming part of our OCR pipeline, requiring significant trial and error. We have instituted a post-processing and data cleaning step using OpenRefine, including dedoubling and removing artefacts of the OCR process using merging, clustering and regular expressions. We estimate there to be data loss from the imaging process of approximately 10% and 15-20% from the HOCR extraction. Notable in this process is how indentation across multiple lines can lead to error in the case of households or businesses with multiple phone numbers for the same subscriber and also with Muslim names which do not follow a first-, middle- and last-name convention, but instead use a chain of names.

3. Disambiguating and spatializing in a superdiverse context

Our OCR output has a simple data structure, separating the initial "name string" from the post office box and phone number as shown in Figure 1. We have opted for this structural simplicity since naming practices in the directories are not uniform across the various religious and ethnic communities. We anticipate a significant amount of community consultation in the subsequent analysis phase in order to understand the complexity of these name strings and to categorize them.

Mohd Kidrowski	Bxx 34I	445****
Fatima Al Rahwan (Mrs)	Box 2T3	772****
Saleh Ahmed Al Sheikh Abdul Razaq Al Jabari	Bx 222	6632**
Golden Sh33p Shawarma	Rax //8	43****
Sheikh Shaheer Al Musali	Box 1343	432****

Figure 1: An example of fake OCR-generated data in tabular format with the initial "name string" (left)

In the first column of Figure 1 multiple layers of cultural information are both revealed and masked. Some subscribers are listed first name first, followed by a family name or chain of names, whereas others are listed by surname first, reflecting the diversity of naming practices for peoples of Arab and Asian origin living in Abu Dhabi. The directories also contain a number of identity markers, the spatial and identity-based dimensions of which are of interest to us, including parenthetical mentions of gender (Mrs, Ms, Dr Mrs), as well as honorifics for the engineering, medical, legal and military professions (Eng, Dr, Atty, Lieut).

Common names such as Muhammad are systematically abbreviated (Mohd), masking different cultural norms of transliterating this very common male given name (e.g. Mohamed, Mohammed). In other cases, ways of transliterating family or first names from other writing systems into Latin script can suggest the origin of the subscriber. We have experimented with, but not finalized, an annotation system which might help us to crowd-tag the different kinds of names present in the phone directories and to experiment with semi-supervised classification. Indeed, input from many sides will be required to mitigate bias in such classification models and careful attention must be paid to certain pages of the directories where bias has already been introduced by OCR.

Second, the lack of street addresses in the UAE means that a fundamentally different approach to spatialization will need to be adopted for this project. In fact, many Arab countries still today do not have postcode and house-specific addressing systems on which postal services rely in the rest of the world (UPU 2017). Other Arab world telephone directories provide a more straightforward mode for spatial analysis. Take, for example, Alexandria and Tunis where there is a street address and number, or Riyadh where a landmark style, numberless addressing practice is adopted, such as "Tarek Ben Zayed Street, near Khaldiya School." Initial attempts to reconstruct a general map of telephone exchanges in Abu Dhabi (the digits of the phone number indicating the neighborhood) using data extracted from the Overpass API have produced promising results with clusters of similar exchanges as shown in Figure 2. We hope this will allow a rough spatialization of subscribers in the city.

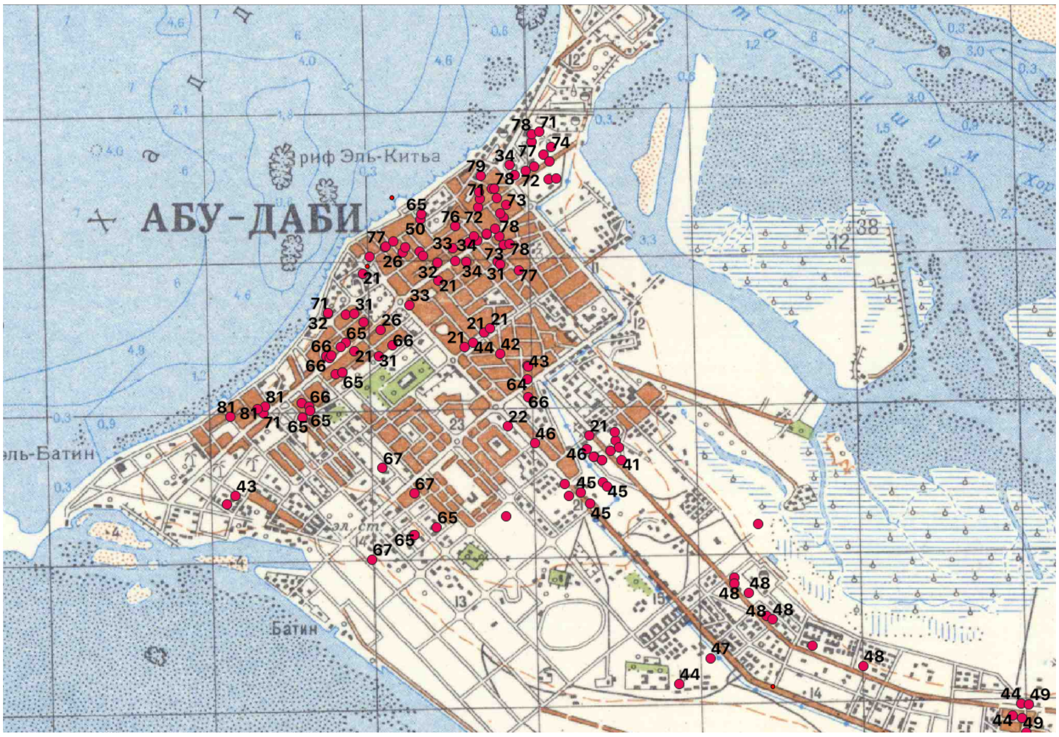

Figure 2: Reconstructing the exchange map for Abu Dhabi. Point data of corner groceries (baqala-s) from OpenStreetMap overlaid on a georeferenced 1:100,000 scale topographic map produced by the Soviet military from the early days of the city in 1978

We also expect that the order of the post office boxes, along with the classification of subscriber name by national or religious origin, will point to residential and commercial networks in the city. Initial analysis of portions of the subscriber data has revealed two points of interest that invite further inquiry. Multiple subscribers using the same phone and post office box numbers seem to indicate patterns in cohabitation or service sharing of postal and telecommunications infrastructure in the face of large demographic changes. Also, the rapid growth of subscribers of certain letters of the alphabet in the 1990s hint at an influx of people of specific religious and national origins.

4. Conclusion

Abu Dhabi Calling! is currently in the post-imaging, data extraction and cleaning phase. We argue that the more mundane phases of digital projects—the data cleaning and preparation phases—are rich moments to stay attuned to culture. Our paper looks at many of the socio-technical challenges of digitizing and studying a global genre—the telephone directory—from a local context. Beyond the obvious importance to local contemporary history and engaging Gulf urban studies with digital, spatial humanities methodologies, the importance of the project, we believe, lies in the expansion of methodologies for studying of phone directories outside the Western context and in scenarios where kinds of data relied upon by spatial humanities research are missing.

Appendix A

Bibliography

1. **Al-Douaihi, Hamoud** (2014): "Names and Addresses Disclosed, Without any Difficulty!" [in Arabic], in: *Al-Riyadh* < https://www.alriyadh.com/959334> [15.03.2021].
2. **Khayrallah Center for Lebanese Diaspora Studies** (2015): "The Syrian Business Directory" < https://lebanesestudies.omeka.chass.ncsu.edu/collections/show/41> [15.03.2021].
3. **König, Mareike** (2004): “Georg Kibler, Möbelbauer, rue de Charonne 39: Adreßbuch der Deutschen in Paris für das Jahr 1854”, in: *Francia* 30, 3: 145-158.
4. **de Maupeou, Félicie / Léa Saint-Raymond** (2013): "Les 'marchands de tableaux' dans le Bottin du commerce: une approche globale du marché de l’art à Paris entre 1815 et 1955", in: *Artl@s Bulletin* 2, 2: Article 7.
5. **Milliken, Genevieve** (2019): “Directory of Churches and Religious Organizations in New Orleans, 1941”

<https://github.com/GenevieveMilliken/WPA_Directory_of_Churches_New_Orleans_1941> [15.03.2021].

6. **NYU Spatial Data Repository** (1978): *Soviet Topographic Map of UAE (1:100k)*, Sheet 7-40-121.
7. **UPU - Universal Postal Union** (2017): *Postal Networks: Actors in the Social and Economic Development of the Arab Region*. *Regional Development Plan 2017-2020*. < https://www.upu.int/UPU/media/upu/files/postalSolutions/developmentCooperation/rdpArabRegion20172020En.pdf> [15.03.2021].
8. **Wolf, Nicholas / Chioh, Wesley / Balogh, Stephen / Spaan, Bert** (2020): "New York City Directories Extracted Persons Entries, 1850-1890," New York University Faculty Digital Archive, <http://hdl.handle.net/2451/61521> [15.03.2021].