IFLA Section
News Media

IFLA International News Media Conference 2024

Kitchen Gadgets: Newspaper Recipe Data and AI

Melissa Jerome
Latin American & Caribbean Collection, University of Florida, Gainesville, Florida, USA.
E-mail address: mmespino@ufl.edu

Sarah Tew
US Caribbean & Florida Digital Newspaper Project, University of Florida, Gainesville, Florida, USA
E-mail address: sarahetew@ufl.edu

Abstract:

Recetas de las Américas *[https://recetas.domains.uflib.ufl.edu/] is a bilingual web project where users can view, browse, filter, and print recipes published between 1954 and 1960 in the newspaper,* Diario las Américas *(Miami, FL, USA). Launched in October 2022, the project is currently undergoing an expansion to include more than 300 recipes, all from* Diario las Américas *and which are available through Chronicling America, a newspaper database managed by the Library of Congress in Washington, D.C., USA.*

Recetas *is a valuable project that calls attention to communities and contributors who have been historically underrepresented by traditional historical narratives, namely the Latin American immigrant community in Florida, particularly the Latina community, and women editors who were excluded from the newsroom and relegated to so-called "soft news" sections including "Del Hogar" where these recipes were published.*

The lightweight and mostly free and open-source technical infrastructures for data manipulation and website generation in the original version of Recetas *follow minimal computing principles and have the potential to help facilitate multilingual digital publishing by underrepresented and under-resourced communities on historically marginalized topics. In scaling up the project, however, we are forced to confront AI's potential to perpetuate existing inequalities and create new avenues for exclusion and elision from the historical record.*

In this paper, we discuss experimenting with AI tools to directly restructure, clean, and translate recipe text as well as to write Python code to perform these actions using traditional natural language

processing resources. We will also present the findings and outputs from explorations of AI tools to generate PDF recipe cards and images of the recipes.

Keywords: Historical Newspapers, Culinary Heritage, Multilingual Publishing, Natural Language Processing, AI Image Generation.

Introduction

Recetas de las Américas is a bilingual digital web project showcasing recipes from *Diario de las Américas*, a historic Florida newspaper which was digitized through the US Caribbean & Florida Digital Newspaper Project (USCFDNP) at the University of Florida George A. Smathers Libraries. Within *Diario de las Américas*, the first Hispanic newspaper in the USCFDNP archive, the "Receta del día" (Recipe of the Day) feature emerged, inspiring our team to create a digital project celebrating culinary heritage. *Recetas* was developed based on 53 recipes that were manually harvested and translated. Through meticulous data cleaning and AI-driven innovation, we hope to expand the project and enrich user experience. With the imminent addition of over 350 recipes, our vision extends beyond preservation, aiming to foster culinary exploration and scholarly engagement. *Recetas de las Américas* epitomizes the transformative potential of digital humanities, connecting cultures across time and space.

About the Project

Recetas de las Américas was designed to showcase recipes originally published in the first decade of Diario de las Américas' publication. Issues of the daily newspaper from 1953-1963 were digitized through the National Digital Newspaper Program (NDNP). Established in 2003 as a collaboration between the Library of Congress and the National Endowment for the Humanities, the NDNP aims to provide enduring, free access to historical newspapers published across the United States and its territories via a national database called *Chronicling America*. For more than twenty years, the NDNP has expanded significantly, encompassing institutional partners from all 50 states and two territories. These partners have contributed over 21 million newspaper pages, featuring content in more than 10 languages to *Chronicling America.*

Our project, the US Caribbean & Florida Digital Newspaper Project (USCFDNP), based at the University of Florida (UF), has been an active participant in the NDNP since 2013. Over the past decade, we have contributed more than 500,000 pages of historical newspaper content from 1836 to 1963. Through partnerships with the University of Puerto Rico-Rio Piedras and the University of the Virgin Islands, we have successfully digitized historic newspapers from Puerto Rico and the Virgin Islands, alongside those from Florida. These digitized newspapers offer a diverse array of voices and perspectives, featuring languages beyond English such as Spanish, Italian, and Danish. The content within these resources not only illuminates past events but also provides context, shaping our understanding of the present and future. While our primary aim is to digitize and make available historical newspaper content, we recognize the importance of raising public

awareness about these invaluable resources. We continually seek innovative ways to engage researchers and the public in utilizing newspapers, hence the creation of *Recetas*.

As we delved into the selected newspapers during the microfilm digitization process, the concept of *Recetas* began to take shape. Our attention was drawn particularly to the newspaper *Diario de las Américas*, where a recurring feature caught our eye: the "Receta del día" or "Recipe of the Day", printed in nearly every issue. These recipes adhered to a consistent article structure in the "Del Hogar" ("Of the Home") section making them easily recognizable amidst the newspaper's diverse content. Given that *Diario de las Américas* was the first Hispanic newspaper to be digitized for our project, this culinary section held special significance, resonating with our Co-Principal Investigator, Melissa Jerome, who is of Hispanic descent. The recipes offered a comforting sense of cultural familiarity, further highlighting their value within the digitized archives. Recognizing the potential significance of these culinary snapshots from the past, we were determined to promote their use and appreciation. This led us to conceive the idea of developing a dedicated website to showcase these recipes as standalone entities. By doing so, we aimed to elevate these culinary treasures from mere historical artifacts to accessible resources that could be appreciated and utilized by a wider audience.

Building *Recetas*

The existing *Recetas* website is created from an Excel spreadsheet using Oxygen XML Editor and Hugo static site generator. It allows users to read recipe text, view images of the original newspaper recipe, search and browse recipes by title, and filter recipes by tags for course, ingredients, dietary restrictions in English and Spanish. In the current website, all 53 recipes were manually harvested, cleaned and restructured by Melissa Jerome and Sarah Tew. The recipes were translated into English using Microsoft Excel's built-in translation feature then edited by the team. For more information of how the site was generated, see the Receta's GitHub Repository[1] and a tutorial created by Sarah Tew [English tutorial [2] and Spanish tutorial[3]]. In September 2024 we will relaunch the *Recetas* website to include over 350 more recipes from *Diario de las Américas*. Our immediate goal is to create a dataset and website containing every digitized recipe from *Diario* published between 1953-1963. Our long-term goal is to ingest recipes from all newspapers digitized through the NDNP grant which were published in Florida, Puerto Rico, and the US Virgin Islands.

As we began planning the September expansion, it became clear that we needed more efficient methods for harvesting, cleaning, restructuring and translating newspaper recipe content and generating new assets such as the printable PDF recipe cards in English and Spanish. This paper will cover our work thus far—cleaning and restructuring textual data and PDF generation.

[1] Sarah Tew, Public Recetas Github (2024) https://github.com/SarahTew/public-recetas

[2] Sarah Tew, "From Spreadsheet to Mutlilingual Website Using Oxygen XML Editor & Hugo Static Site Generator." https://sites.google.com/ufl.edu/dlf2023/home.

[3] Sarah Tew, "De una hoja de cálculo a un sitio web multilingüe utilizando el editor XML de Oxygen y el generador de sitios estáticos Hugo." https://sites.google.com/ufl.edu/dlf-2023-es/home

Nature of the data

Prior to the AI experiments, Melissa Jerome and Sarah Tew manually harvested 415 recipes from *Diario de las Américas*. The team entered the URL of each recipe, the title, and the raw OCR recipe text into a CSV file and then automatically populated unique recipe IDs in Microsoft Excel[4].

To work with our existing website pipeline, the data needed to be transformed in several ways. We needed to clean and correct the OCR text, remove non-recipe text (eg. country of orgin for the recipe), separate ingredients from directions, convert recipe ingredients into generic lists, add separators between items, and translate the text from Spanish to English. We also sought to develop an automatic way to create printable PDF recipe cards in both English and Spanish for each recipe.

The recipes in *Diario las Americas* are good candidates for a project like this since they are extremely formulaic and relatively short. The typical format of a “Receta del día" is a single recipe beginning with a list of ingredients immediately followed by a paragraph of directions. A small minority of recipes are written in a completely narrative format, with the ingredients and quantities mixed within the instructions. There are also a small number of recipes that contain multiple recipes— usually a main recipe and a recipe for a sauce or icing —or one main recipe followed by variations of that same recipe, most often for beverages like coffee or smoothies.

Because the quality of the original newspaper, microfilm, and digital files were high, the OCR text was generally good. The most frequent OCR errors were random single characters (eg. * . / ~ -), articles, and fractions. Random characters were most often caused by minor printing imperfections and occurred more frequently at the end of lines. Article errors were usually a case of the OCR encoding an ‘L’ or ‘l’ as ‘I’ or ‘i’. Recipes used either a “/” or “–” to represent a fraction. The OCR text was better for fractions using the hyphen but frequently misinterpreted the fractions with a slash as some variation of “Vi” or “Va”.

Single, extraneous non-alphanumeric characters were simply removed from the string. The most recurring article errors were corrected through simple substitution. The fractions were more difficult as the OCR often did not contain any digits and the correction could not be effectively predicted from the text alone. Instead, the code searches for these instances and returns a list of recipe IDs that require this to be manually corrected through comparison with the digital image.

We found it necessary to clean the text at multiple stages in the cleaning and restructuring process. We did a first-round cleaning to prepare the text for initial restructuring then once non-recipe text had been separated, we did another round to clean up any extraneous characters.

In the future, we will also explore translation as a possible solution to remaining OCR errors. When translating the original 53 recipes automatically in Excel in 2022, we found that the translation into English resolved typos and spelling errors present in either the original printed Spanish newspaper text or the OCR. We believe translation may be a useful cleaning tool for the recipes, especially since the grammatical structure and style is simple. The recipes are mostly

[4] Sarah Tew, Recetas CSV in Recetas Github (2024) https://github.com/SarahTew/public-recetas/blob/main/source-data/recetas.csv

lists of concrete nouns and short imperative commands which translate well without much nuance.

Experimenting with AI - Natural Language Processing

Microsoft Copilot was extremely efficient in cleaning, restructuring, and translating the raw OCR text when given as plain text in the chat box. While it could access and read the recetas.csv file via the GitHub link[5], it could not effectively navigate the data table, so its accuracy plummeted.

When pasting into the chat box, Copilot quickly and accurately separated recipes into ingredients and directions, simplified the ingredients list by removing quantities, and translate into English and back into Spanish from English. It was more effective when set to the “more precise” conversation style. When using the “more creative "and “more balanced” settings, it would often deviate from the target recipe by adding new ingredients and rewording steps. Since Copilot includes citations, it became clear that it was generating new and hybrid recipes by adding information compiled from similar recipes online. This issue was resolved by switching to “more precise” and explicitly instructing Copilot to only use the input text only. Copilot handled all text formats including narrative recipes, lists, and combined recipes with the same high-level accuracy.

The limiting factor for using Copilot directly was the input and output formats. While it was able to access the CSV file in GitHub, it was unable to effectively navigate and iterate over each recipe. It was always inconsistent and frequently incorrect even with simple prompts such as “Tell me the title of the recipe with ID r2”.

Sarah Tew also experimented with Google’s Gemma which integrates directly into Colab, Googles hosted Jupyter Notebook service. The cost of compute credits to utilize Gemma directly in Colab was prohibitive and this method was quickly abandoned. Use of UF's supercomputer for research also costs money which is beyond the USCFDNP budget and therefore not a viable option for this project.

During our AI experiments for this project, we were informed that UF was negotiating an institutional license for Copilot integration into Microsoft 365 Apps for faculty and staff. Once this is available to us, we will see how well the Copilot integration in Microsoft Excel allows us to clean, restructure, and translate the recipe data. We hope that this integration will allow us to apply Copilot’s natural language processing capabilities more effectively within data tables.

To overcome the limitations of currently available tools and work more effectively at scale, Sarah Tew used Copilot to coauthor code to clean and correct OCR text in Spanish, remove non-recipe content such as notes on cuisine type, serving size, and source, split ingredient lists and directions, and generate lists of recipes with OCR fraction errors and narrative-style recipe text

[5] Sarah Tew, Recetas CSV in Recetas Github (2024) https://github.com/SarahTew/public-recetas/blob/main/source-data/recetas.csv

so they can be checked and restructured manually. Please see the Colab notebook for detailed descriptions and the final code.[6]

Ultimately, Sarah found that the most effective use of Copilot for the *Recetas* project now was prompting it to help me write Python code that she copied, edited, and ran in Google Colab to clean and restructure the OCR text. Co-authoring code with Copilot helped her achieve her goals much more quickly and efficiently than if she had continued coding alone. Giving more specific prompts, precisely describing errors and their sources yielded more efficient code and better corrections by Copilot. Sarah Tew already had intermediate Python skills and familiarity with Google Colab from an MA in digital humanities. She did not have much experience with natural language processing, however. Before UF provided access to Copilot, she had already started to write code herself to clean and restructure the data, but the pace was so slow it was not feasible to continue the expansion project considering her other job duties. Since using Copilot, writing code has become much more efficient, freeing up time to do more with *Recetas*. She also feels she have increased her coding skills and computational problem solving through reading and editing AI-generated code and the explanatory comments she prompted Copilot to include. Using Copilot to write and edit Python code for natural language processing tasks would have been much less useful and more frustrating if she did not already have some previous coding experience.

Experimenting with AI – PDF Generation

Prior to launching *Recetas*, Melissa Jerome manually created 4x6inch double-sided recipe cards that resembled traditional recipes cards but were available in PDF format for *Recetas* users to download and print. Recognizing the time-consuming nature of this process, she decided to explore automating it with AI. To kickstart this automation, Melissa tested creating these recipe card PDF files using both ChatGPT and Copilot. She began with ChatGPT, asking if it could generate a 4x6 traditional recipe card with the raw OCR text. At first, ChatGPT responded "Certainly! Please provide the recipe and I'll format it into a traditional 4x6 recipe card for you". However, after supplying the recipe name, ingredients, and directions, ChatGPT produced a hyperlinked text for a file that could not be downloaded or opened.

Generating PDFs with Copilot offered a different experience. Melissa experimented with generating PDFs using a variety of inputs. Initially, using the "more balanced" option, she asked Copilot to produce a PDF for a recipe. To streamline the process and save time from manually extracting recipe text, Melissa provided Copilot with a link to a recipe on *Recetas* and requested it to extract the recipe information from there. However, this resulted in Copilot producing the same webpage it was given. Subsequently, Melissa attempted to use the "more precise" and "more creative" options available in Copilot, hoping for varied results. Unfortunately, in these instances, Copilot consistently responded that it does not "have the capability to create a PDF directly from a webpage".

After reverting to "more balanced", Melissa proceeded to test PDF generation by directly supplying recipe text for the selected recipes, one at a time, instead of directing Copilot to the

[6] Sarah Tew, Recetas Colab notebook (2023) https://colab.research.google.com/drive/1HJ036_cCPkk85XAi6_gKD0-8dIcFAWPt

Recetas website. The testing involved alternating between providing Copilot inputs in English, in Spanish, and in Spanglish. In these initial tests, Copilot generated a new copy of the provided recipe that could be downloaded and saved as a TXT, Word Document, or PDF. It performed well in producing files that were accurate and consistent with the language of the input, however this was not without issues.

Upon initial review, the files created by Copilot appeared to correspond to the provided recipe. However, they included edits to the ingredients list and/or the directions that Copilot introduced. Additionally, Copilot did not consistently provide the same output despite receiving identical inputs. Furthermore, Copilot appended short messages to the end of the recipes (e.g. “Hope you enjoy!”), followed by emojis it selected to accompany the recipe. These could not be removed, despite requesting that these be omitted.

Experimenting with AI - Image Generation

While the recipes themselves lacked accompanying photographs in their original newspaper format, we enhanced the *Recetas* user experience by adding images of the original "Receta del día" column to the website. Additionally, we actively encourage individuals who try out the recipes to submit their own photos for display on the homepage. This feature allows users to see real-life renditions of the dishes and fosters a sense of community engagement. However, because the original recipes lack photos and not all recipes are represented in user submissions, we wanted to explore the potential of AI image generation to create visual representations of the recipes. Given the availability of recipe names and ingredients, some of which are commonly known or easily understandable, we anticipated that AI image generation could effectively develop images that would visually bring these recipes to life.

Melissa conducted tests using three different platforms, all with free versions: Copilot, DeepAI, and Adobe Firefly. Given the bilingual nature of *Recetas*, Melissa provided prompts in both English and Spanish (at different times) across all three platforms to assess their performance across languages and input types. In testing, she experimented with various text inputs for a handful of recipes, ranging from straightforward and commonly known to those that might post challenges for the AI models without additional context. Initially, Melissa solely provided the recipe names. Then, she changed to providing the ingredients list and directions instead of the name. Finally, she shifted to providing the recipe name, along with the ingredients list and directions.

Melissa began image generation testing with Copilot, as the team was already utilizing it for testing NLP and PDF generation tasks. One notable aspect of Copilot's image generation feature is its integration within the same interface used for AI interactions, allowing for seamless transitions between text-based tasks and image production within a single session. Overall, Copilot performed well across all three input methods. When presented with the full recipe, it often generated infographic-style images, while when presented with just a recipe name it generated more realistic photos of dishes. However, these photos occasionally featured visuals of ingredients not included in the recipes. For instance, in a photo for "huevos a la gitana," Copilot

included an array of bell peppers which were not on the recipe ingredient list. Despite this, Copilot demonstrated proficiency in seamlessly switching between English and Spanish, producing fairly accurate images for Spanish text inputs.

Unlike Copilot's image generation, which is embedded within a chatbot, DeepAI and Adobe Firefly provide standalone image generation capabilities. DeepAI offers two free models (standard and HD), options for preference (speed or quality), and more than 100 styles to choose from (although "food" is not among the options). On the other hand, Adobe Firefly offers two content types (art and photo), allows for style customization (e.g., watercolor, 3D, pencil), and permits users to upload a photo for reference. Despite their differing features, DeepAI and Adobe Firefly functioned similarly with the prompts we supplied. Both platforms exhibited responsiveness to Spanish prompts, but they performed better in creating realistic food images when provided with text in English. While DeepAI and Adobe Firefly were generally able to produce realistic images for most of the recipes, they encountered some difficulties with certain prompts, resulting in images of uncooked dishes or images that appeared artificial. However, both platforms often enhanced the photos by adding appealing visuals in the background.

All three platforms demonstrated better performance with English inputs compared to Spanish. For instance, they encountered difficulty understanding the term "Bolas de nieve," resulting in images of snowballs. However, when the word "cookies" was added or when asked to generate images for "snowball cookies," accurate photographs of the recipe were produced. Additionally, the platforms faced challenges with certain Spanish words, particularly adjectives. For example, when prompted to create images for "tortas de naranja," they generated images of orange-colored foods instead of orange cakes, which were produced when provided with the recipe name "Orange Cakes" in English. Furthermore, the platforms struggled with generating images for recipes containing ingredients that are culturally tied to the Latin American community. *Recetas* currently includes a few recipes that call for using animal brains, and the platforms yielded mixed results. Copilot flagged our text input as censored; Adobe Firefly created generic images unrelated to the recipes; while DeepAI came the closest, generating images that somewhat resembled the dishes. Overall, DeepAI and Adobe Firefly outperformed Copilot, providing the most accurate and realistic images for both English and Spanish recipes we provided, albeit with their own limitations.

Conclusion

AI tools provided mixed results for the *Receta* project. Microsoft Copilot had excellent natural language processing skills in English and Spanish and, when given recipe text for one or more recipes, was able to separate the ingredients from the directions, create a simplified list of ingredients, and translate recipes into English and back into Spanish with high fidelity to the original text. It struggled to navigate the CSV, however, therefore limiting its applicability at scale.

Copilot's difficulty working directly with tabular data coupled with the lack of transparency concerning Copilot's methods for natural language processing led to a preference for using AI to generate Python code so we could perform and better understand our own data cleaning and processing. Copilot generated overall high-quality code which, when combined with Sarah Tew's expertise, did achieve our goals for data cleaning and restructuring. We have not yet used AI to write code for automatic translation but will in the near future.

PDF generation with AI was less successful and ultimately did not achieve our desired output. We're exploring new approaches, both alternative programs and moving away from testing URL and OCR inputs to using XML.

Image generation exposed differences in how AI chatbots responded to English and Spanish content. They performed better with English text input, but the results were too inconsistent to use reliably at our scale of hundreds of recipes. The ethical and artistic questions raised by AI image generation is beyond the scope of this paper but are also a concern.

Overall, AI tools did help us "level-up" our skills and create code and images of higher quality in less time than we would have been able to do by ourselves. The quality of the AI outputs and the time saved by using them is directly proportional to the quality of user prompting and background understanding of the desired and actual outputs. AI tools have the potential to save time and expand what users can do, but they are not a substitute for human intelligence.

References

Adobe Firefly. https://www.adobe.com/products/firefly.html. Accessed May 16, 2024.

"AI Services," University of Florida Information Technology, https://it.ufl.edu/ai/

"Chronicling America," Library of Congress. https://chroniclingamerica.loc.gov/

Copilot. https://copilot.microsoft.com/. Accessed May 16, 2024.

DeepAI. https://deepai.org/. Accessed May 16, 2024.

"National Digital Newspaper Program," Library of Congress. https://www.loc.gov/ndnp/about.html

Tew, Sarah, and Jerome, Melissa. "Recetas de Las Américas." 2022. https://recetas.domains.uflib.ufl.edu/.

Tew, Sarah. "De una hoja de cálculo a un sitio web multilingüe utilizando el editor XML de Oxygen y el generador de sitios estáticos Hugo." https://sites.google.com/ufl.edu/dlf-2023-es/home

Tew, Sarah. ”From Spreadsheet to Mutlilingual Website Using Oxygen XML Editor & Hugo Static Site Generator.“ https://sites.google.com/ufl.edu/dlf2023/home

Tew, Sarah. Recetas Colab notebook (2024). https://colab.research.google.com/drive/1HJ036_cCPkk85XAi6_gKD0-8dIcFAWPt?usp=sharing

Tew, Sarah. Public Recetas, (2024), GitHub repository. https://github.com/SarahTew/public-recetas.

“US Caribbean & Florida Digital Newspaper Project.” 2015. https://ufndnp.domains.uflib.ufl.edu/about/