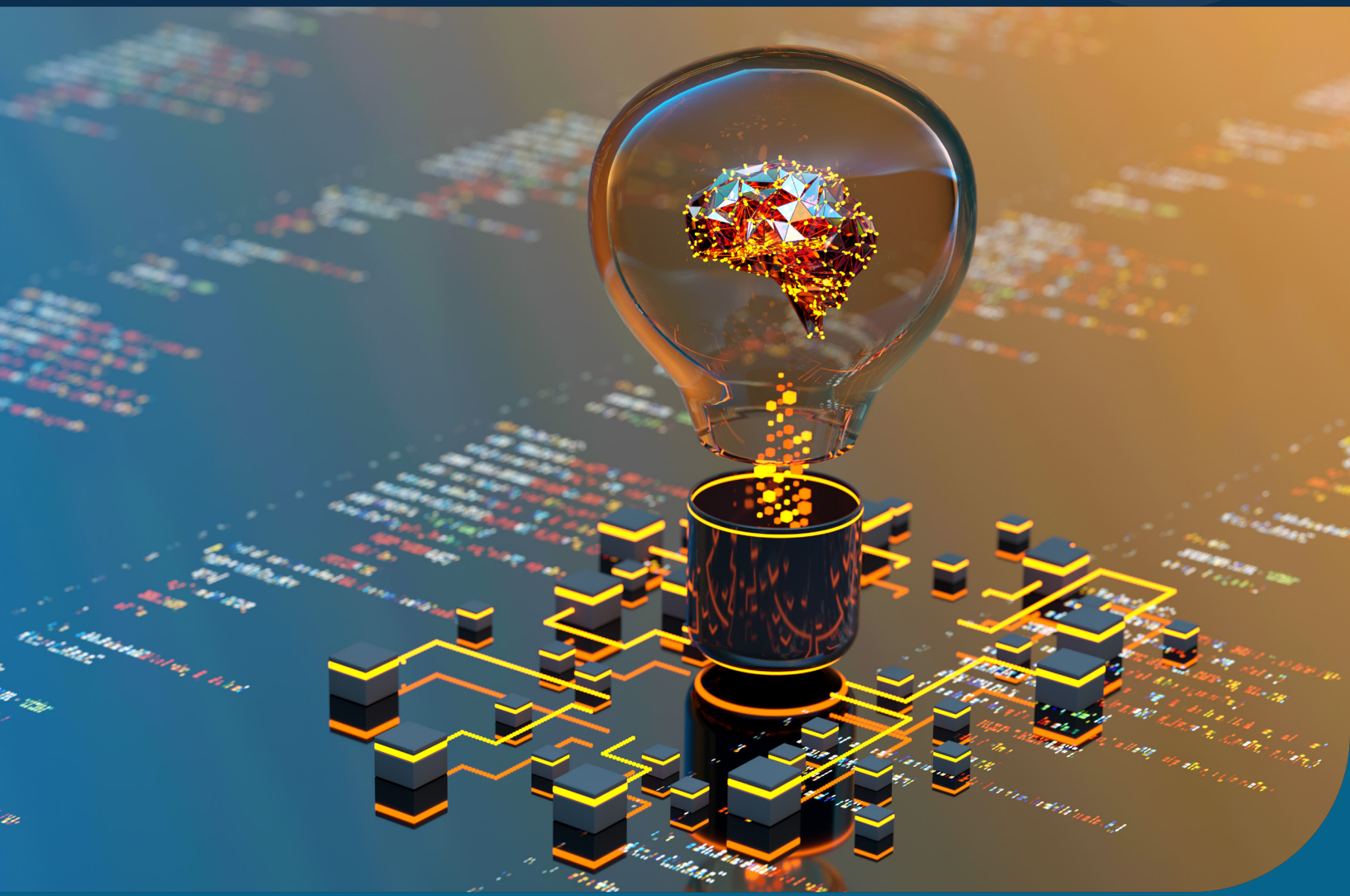
LEVERAGING AI FOR DEMOCRACY

CIVIC INNOVATION ON THE NEW DIGITAL PLAYING FIELD

// BETH KERLEY / CARL MILLER / FERNANDA CAMPAGNUCCI



NED NATIONAL ENDOWMENT FOR DEMOCRACY
SUPPORTING FREEDOM AROUND THE WORLD

FORUM INTERNATIONAL FORUM FOR DEMOCRATIC STUDIES

FROM DATA DESERTS TO AI OCEANS

// **FERNANDA CAMPAGNUCCI**

In an era of rapid technological change, global democratic backsliding, and political polarization, democratic societies face a host of vexing challenges. To build public understanding of these issues and help democratic institutions arrive at well-informed, effective responses, high-quality data is essential. When governments make meaningful, open data on topics of critical popular interest publicly available, it empowers civil society, scholars, and other stakeholders to not only scrutinize the work of authorities, but also to find solutions and co-create policies collaboratively. In short, **open data helps to ensure that democracy delivers**.

Per the principles established by the international community one decade ago, making data open means that it "can be freely used, modified, and shared by anyone for any purpose."[1] While open data alone is not a panacea, its absence is a barrier to addressing our most pressing challenges, from information manipulation to climate change. Without access to accurate data, journalists cannot verify information in a timely manner. Researchers are less likely to uncover new insights that could help prevent or mitigate disasters and epidemics. Policymakers struggle to learn from the experiences of other jurisdictions, and citizens have fewer resources to examine existing inequalities in their communities or monitor their governments' activities.

Over the past decade, Open Knowledge Brasil (OKBR) has dedicated its efforts to advocating for open data from government entities in the country. As in many jurisdictions worldwide, although there has been some progress at the federal level, finding accessible and usable data becomes increasingly challenging in state or local governments. At these lower levels of governance, significant barriers confront those in search of data that can be readily incorporated into third-party analyses—a gap that can hinder democratic decision making on issues of public concern. These barriers have an outsized impact given Brazil's federal political system, in which more than 5,500 municipalities have autonomy to deliver public services and define policies in crucial fields such as the environment, housing, culture, and education.

Although there has been some progress at the federal level, finding accessible and usable data becomes increasingly challenging in state or local governments.

"Data deserts" is an expression that aptly describes the landscape in these cities, where open data for most sectors is lacking: Relevant information is invisible to our eyes, or at least out of our reach. Occasionally, a mirage appears—an open database, but without a proper format that allows for cross-referencing with other data or conducting analyses. To be effectively reusable, data must to be structured. In other words, data should be presented in formats where information is organized into fields with clear relationships and significance (like spreadsheets or other kinds of databases). In Brazil's municipalities, however, while systems used by public agencies are producing a growing volume of data, poor data governance make structured data relevant to topics of public interest scarce and difficult to access. A recent assessment by our organization shows that there is still a long road ahead in seeking to close these gaps. Even São Paulo, the largest metropolis in Latin America, did not clear the minimum bar for data openness.[2]

To work around the limitations of published data, organizations like OKBR have relied on freedom of information requests and bottom-up tactics, such as crowdsourcing, collective data mapping, and building citizen sensing technologies. These and other strategies, however, require us to devote tremendous effort and resources to cleaning and structuring datasets. **Thanks to recent advances in artificial intelligence (AI), we can now approach these challenges differently**.

The rapid evolution of AI tools is changing the game for government transparency work. **About seven years ago, we started to explore possibilities for automated data analysis and anomaly detection using AI to flag suspicious government transactions, irregularities, or potential instances of corruption**. These capabilities would allow civic organizations and government watchdogs to identify priority areas for investigation and monitoring. Still, we needed data sources to fit our statistical models. This requirement restricted our field of action to places where structured data were available, leaving the data deserts behind. Recent advances, particularly in "foundation models" and generative AI, are eroding these constraints, making it increasingly feasible to extract valuable insights from unstructured data.

FROM STRUCTURED DATA TO NATURAL LANGUAGE

OKBR's experiences, and particularly two of our flagship projects—"Serenata de Amor" ("Love Serenade") and "Querido Diário" ("Dear Diary")—offer an illustration of how AI is transforming the open government landscape, and, as a result, opening up new possibilities for tech-enabled accountability work.

Serenata de Amor,[3] launched in 2016, is a pioneering project that uses machine learning, one of the foundations of mid-2010s AI, to monitor and classify the expenses of Brazilian Congressional representatives. By analyzing expense reports and receipts, the AI-powered system identifies potential indicators of dubious transactions, such as excessive spending on meals or travel. Despite its success as a reference in the field of civic AI, there are technical limitations to the scope and replicability of this project.

Serenata was limited to monitoring a specific aspect of government spending—Congress members' expenses—based on a structured dataset. The group of civic hackers who initially launched the project had to extract information from images of receipts published on the congressional website. In the face of public pressure for greater transparency, however, the legislature eventually began providing higher quality data through an application programming interface (also known as an API—a mechanism that enables third-party applications to retrieve information directly from a system and make use of the data it contains with near real-time updates).

This development sparked widespread enthusiasm for our approach, with individuals across the nation expressing interest in replicating it at the level of municipal legislative chambers or city halls. **Though the project's code was openly accessible, it proved impossible to replicate in other environments without access to similarly structured data sources, which are rare**.

Once the National Congress began providing usable data, the technical challenges in Serenata de Amor were relatively straightforward, as the AI application dealt primarily with structured data and the application at scale of simple statistical regression models. For instance, to check if a meal expensed by a member of congress was unusually costly, Serenata reviewed historical spending patterns within the same category. Advancing to more intricate models would require locating additional data sources for cross-referencing, as well as enhancing the technical expertise on our team.

In 2021, inspired in part by an interest in leveraging advances in AI tools for language processing, OKBR shifted its focus to a new project, Querido Diário,[4] which holds greater potential but presents different challenges.

Much of the public information available at the municipal level, even when published, is not as neatly structured as the datasets upon which our Serenata project relied. **Querido Diário sets out to tackle data scarcity challenges by aggregating and analyzing unstructured information sourced from municipal**

gazettes across Brazil. These daily gazettes, also known as "official diaries," serve as repositories where cities publish information—including the text of new laws, summaries of public purchases, and lists of civil servants who are on leave—in the form of text-heavy PDF files.

Brazil's municipal gazettes exemplify the problem of "unstructured information"—here, referring to freely written text, set out in whatever order its authors deem appropriate. The announcement of a new contract signed by a city, for example, can take many forms when it appears in a municipal gazette. The company being hired may be referred to under its whole commercial name, its trademark, its tax registry number, or an abbreviated version of each. Numerical units may be expressed verbally (e.g., "one thousand and three packs of coffee") or in other ways. The formats used in these documents will also vary across cities, or even when the civil servant who usually writes the entry is out of the office. As a result, traditionally it has been difficult for machines to extract meaning from unstructured information automatically, even if a person can read and comprehend it with ease.

Before figuring out how to make computers read that mass of information, we needed to source the data and set it free from its Gutenbergian cage. To do so, we have leveraged a community of dozens of volunteers, who constantly develop web scrapers to extract text from the municipal PDFs and render it accessible, within an open infrastructure, for anyone to access and repurpose. Anyone can look up keywords in a search bar or utilize filters built into the interface to find information within thousands of files, or a bot can be connected to the infrastructure and scan through all the information at once. Since its inception, the project has undergone continual evolution. Presently, it encompasses data from over 410 cities, home to 30 percent of the Brazilian population.

With this mostly unstructured text in an open infrastructure, we now have an ocean of data to navigate and explore with the help of AI. Natural language processing (NLP) models can be used to process and make sense of this data on a scale that would be impossible even for thousands of human volunteers—and large language learning models (LLMs) have the potential to amplify these efforts even further.

Traditional NLP techniques necessitate developers knowing in advance what they want to search for in the text. When given clear instructions of this kind, a traditional NLP model can, for instance, identify contracts related to climate change mitigation within a gazette and list the names of all companies mentioned in the given document. LLMs, powerful simulators of language, can potentially go further. They offer three clear advantages: simpler prompting, greater capacity to analyze relationships among entities (significant objects or pieces of information in the text), and the ability to summarize findings from search results as well as explain how they connect to other contexts. For example, a citizen might ask which contracting company was hired to clean a river and receive a response explaining what each contractor involved was supposed to do, even referring to the history of previous contracts to see if there were costly extensions.

Natural language processing models can be used to process and make sense of this data on a scale that would be impossible even for thousands of human volunteers—large language learning models have the potential to amplify these efforts even further.

NOT THERE YET: NAVIGATING CIVIC AI

Despite the clear potential of AI in civic work, high costs and several other major challenges hinder its widespread adoption and effectiveness.

NLP in Portuguese: Traditional NLP models often struggle to achieve satisfactory accuracy in Portuguese, particularly for domain-specific tasks related to the civic sphere. Training such models requires huge bases of words painstakingly classified by humans (e.g., identifying and defining names, adverbs, and government-related actions). Business incentives have made such lexicons widely available in English, but Portuguese-speaking countries have never had the resources to create them at the necessary scale. Thus, pretrained language models are still underdeveloped and may not perform well on tasks involving Brazilian Portuguese—a common frustration encountered with current NLP systems when working in many languages other than English.

Lack of donor support for critical tasks: To make data available for deploying and fine-tuning AI tools, civic organizations need to classify the information contained in large datasets (such as municipal gazettes) manually. This task is laborious and requires qualified personnel. Moreover, domain experts—for instance, lawyers or specialists in agriculture or education policy—need to review the classifications as well as the outputs of such models. The time of those experts may be more expensive than the technology itself, but these efforts are crucial to ensure that AI systems achieve the needed level of accuracy. Donors often do not understand these needs and are reluctant to fund unglamorous work, like data infrastructure construction and management, that entail investing in process rather than final products.

Infrastructure costs: The infrastructure costs associated with running AI solutions can be prohibitive, especially for smaller civic organizations with limited resources. General-purpose models, such as ChatGPT, are not suited for tasks where accuracy is paramount, since they often return made-up results with no basis in fact. Specifically trained AI models tend to provide more pertinent results. The cost of fine-tuning AI models for specific tasks is decreasing, but cloud services able to support this work usually charge in U.S. dollars, and exchange rate fluctuations may further exacerbate this challenge. Difficulties related to the structure of public-sector data can also play a significant role in infrastructure costs: Government agencies change the formats and sources of the data they publish frequently, which requires civic organizations to re-train and adapt their models.

Hiring qualified personnel: The demand for AI expertise far exceeds the supply, and civic organizations generally find it difficult to compete for talent with better-resourced private sector companies. When dealing with legal data, there is an additional hurdle of recruiting domain experts, including in fields such as data privacy, to review the pertinency of the machine-generated output.

PATHWAYS FORWARD

To address many of these technical and financial barriers to leveraging the full capacity of AI for civic work, OKBR deploys a variety of strategies. These approaches include:

Partnerships with universities: Collaboration with academic institutions provides access to cutting-edge research and expertise in AI and NLP. OKBR has established a program in which professors and researchers from diverse fields can work to tackle specific civic challenges in collaboration with our team and within their regular curriculum, thereby gaining insights from real-world problems.

Open-source code and community collaboration: Making AI algorithms and tools open-source allows for broader collaboration and contributions from the community. To this end, OKBR is committed to sharing its code. We also leverage a Discord channel boasting nearly 1,500 members, predominantly from technical backgrounds, for community technical collaboration.

Providing free and accessible training: Offering free and accessible training programs for developers and AI enthusiasts can empower individuals to contribute to civic AI projects. More than five hundred individuals have taken the “Python for Civic Innovation” course offered by the School of Data (OKBR’s educational program), which teaches the programming language applied to our NLP projects. Some of these students became active volunteers. More broadly, digital rights organizations with educational programs, or those focused on training people in digital and data literacy skills, can enhance the civic AI ecosystem by incorporating contributions to actual civic technology projects into their syllabi.

By leveraging AI techniques, organizations like Open Knowledge Brasil can unlock valuable insights from unstructured data and scale up public oversight of government activities. Other organizations in the region are also beginning to tap into the civic potential of AI: Latin American organizations participating in the EmpatIA initiative,[5] for instance, developed prototypes for various AI-powered applications designed to address public issues such as air pollution and public health. Governments and universities are undertaking additional exploration with generative AI, although the current state of this technology means these projects are most likely still too experimental for release.

Navigating the technical, financial, and ethical challenges of civic AI requires innovative solutions and collaboration. Through partnerships, open-source initiatives, and accessible training programs, **we can harness the full potential of AI for civic technologies that promote transparency, accountability, and democracy**.

From Data Deserts to AI Oceans: Harnessing Artificial Intelligence for Government Transparency

1 For more information, please see Open Knowledge's definition of "Open Data:" https://opendefinition.org/.

2 Launched in June 2024, the "Open Data Index for Cities" is the first comprehensive assessment on open data from capital cities and fourteen public policy areas. Additional information can be found on Open Knowledge's Brasil's webpage about the index: https://indicedadosabertos.ok.org.br. (Original source material in Portuguese.)

3 For more information, please visit Serenata de Amor's webpage: https://serenata.ai/. (Original source material in Portuguese.)

4 For additional information about OKBR's Querido Diário project, please view this webpage: https://queridodiario.ok.org.br/en-US/sobre.

5 For more information about the EmpatIA initiative and its projects, please consult: www.empatia.la/en/proyectos/.