Additional Datasets for PP 5531

1. **Civilian harm in Ukraine.** Researchers at Bellingcat and contributors to its Global Authentication Project have assembled a map and dataset of 1,000+ incidents "that have resulted in potential civilian impact or harm since Russia began its invasion of Ukraine." They include incidents "where rockets or missiles struck civilian areas," "where attacks have resulted in the destruction of civilian infrastructure," and/or where visual evidence depicts civilian injuries or "immobile civilian bodies." The information, collected from public sources and vetted by Bellingcat, includes each incident's date, location, description, sources, type of area affected, and type of weapon system (if known). [h/t Philip Bump].

2. **Government contracting data, cataloged.** The nonprofit Open Contracting Partnership has launched a registry of government procurement datasets that use its Open Contracting Data Standard (featured in DIP 2020.02.26). The registry contains 100+ entries so far, across 50+ countries — from Argentina's national roads authority and the city of Buenos Aires to Zambia's Public Procurement Authority. You can filter the listings by dataset recency, update frequency, and the data types included (parties, awards, documents, amendments, et cetera). [h/t Georg Neumann]

3. **Congressional votes and ideology.** The Voteview project "allows users to view every congressional roll call vote in American history," and places those votes in the context of ideology estimates along a liberal-to-conservative spectrum. The core estimates come from DW-NOMINATE, a method developed by the project's directors emeritus, Keith T. Poole and Howard Rosenthal. Voteview's bulk data includes ideology estimates for every member of the House and Senate since 1789, every vote taken in either chamber, and every member's position on those votes. [h/t Philip Bump]

4. **EPA-regulated facilities.** The US Environmental Protection Agency's Facility Registry System "provides Internet access to a single source of comprehensive information about facilities, sites or places subject to environmental regulations or of environmental interest." It includes each entity's name, type, location, industry, regulatory programs, and more. That information, which spans millions of facilities, is "subjected to rigorous verification and data management quality assurance procedures." The records also provide facilities' ID numbers from other EPA systems, such the agency's Risk Management Program database featured in last week's edition. [h/t Michael Allen]

5. **Foreign direct investment.** Among its many activities, the Bureau of Economic Analysis collects data on US-owned businesses' direct investments abroad and those by foreign-owned businesses in the US. The agency provides statistics based on that data, aggregated by industry and country. The metrics, available as spreadsheets and interactive tables, include cumulative investments, overall transaction flows, foreign affiliates' employee counts, and more. **As seen in**: "Why are US companies investing more abroad?" (USAFacts).

6. **India's high-court judges.** Know Your High Court Judges, a multi-institution collaboration, has compiled biographical and professional data on all 1,700+ judges appointed to India's high courts (each state and union territory's top appellate court) between late 1993 and mid-2021. The details, based on a range of official sources, include each judge's name, gender, date and place of birth, dates of appointment and retirement, courts served, prior experience, and more. [h/t Apoorv Anand via csv,conf,v7]

7. **Women's well-being.** Camille Belmin et al.'s LivWell dataset presents "a set of key indicators on women's socio-economic status, health and well-being, access to basic services and demographic outcomes" in 447 regions of 52 countries from 1990 to 2019. The indicators include, for example, rates of home ownership, educational attainment, and domestic violence; they're based primarily on data from the Demographic and Health Surveys Program, a USAID-funded initiative that, since 1984, "has provided technical assistance to more than 400 surveys in over 90 countries, advancing global understanding of health and population trends in developing countries." **Read more**: An introductory Twitter thread from Belmin.

8. **Money in UK politics.** The Westminster Accounts, a recent collaboration between Sky News and Tortoise Media, examines the flow of money through UK politics. It does so by combining data from three key sources: the Register of Members' Financial Interests, Electoral Commission records of donations to parties, and the Register of All-Party Parliamentary Groups. You can search and explore the results through the collaboration's interactive database. They haven't published a downloadable version yet, but Simon Willison has extracted one. **Read more**: The project's methodology.

9. **Government open-source policies.** Researchers at the Center for Strategic and International Studies have compiled a dataset of 660+ government policies regarding open-source software proposed at the national

level since 1999. The [dataset](#) indicates each proposed policy's country, year, title, issuing authority, stated objective, status, and more. The countries with [the largest number of entries](#): Argentina, the US, and South Korea. [h/t [Kevin Xu](#)]

10. **Drone registrations.** Through a FOIA request to the Federal Aviation Administration, [Ben Welsh](#) has [obtained data](#) on all "small unmanned aircraft" (popularly known as drones) actively [registered](#) with the agency through 2021 — nearly 900,000 registrations in all. The spreadsheets, provided separately for [commercial](#) and [recreational](#) registrations, list each drone's type and model, [Remote ID](#) capability, and registration date, plus the registrant's city, state, and postcode. **Related**: Welsh, a journalist and [newly-licensed](#) commercial drone pilot, has [filed many data FOIA requests](#), including [one](#) that pried loose the [FAA's drone accident database](#) (released, alas, as a PDF). [h/t [MuckRock](#)]

11. **Indonesia earthquake intensities.** [Gempa Nusantara](#), a database compiled by [Martin et al.](#), uses historical documents to catalog 7,300+ "macroseismic effects" of 1,200 earthquakes near Indonesia during four-century span, from 1546 to 1950. It provides summaries of the local reports and categorizes the effects according to the [European Macroseismic Scale](#), which focuses on the intensity of ground-shaking and [potential impacts on buildings and terrain](#).

12. **Semiconductor logistics.** The [Emerging Technology Observatory](#), a [new initiative](#) based at Georgetown University, aims to provide a "public platform for high-quality, actionable data resources on the global emerging technology landscape." Its [Advanced Semiconductor Supply Chain Dataset](#), which you can [download](#) and [explore online](#), contains "manually compiled, high-level information about the tools, materials, processes, countries, and firms involved in the production of advanced logic chips." [h/t [Zach Arnold](#)]

13. **Big emitters.** [Climate TRACE](#), a [nonprofit coalition](#) launched in 2020, uses satellite imagery, sector-specific datasets, and [other sources](#) to estimate greenhouse gas emissions in detail. Their most recent inventory, [released last week](#), highlights 70,000+ individual sites that "represent the top known sources of emissions in the power sector, oil and gas production and refining, shipping, aviation, mining, waste, agriculture, road transportation, and the production of steel, cement, and aluminum." You can [download the data](#), [explore sector- and country-level estimates](#), and [browse a map of the sites](#). **Read more**: [Coverage in the New York Times](#). [h/t [Ian Johnson](#)]