

Title: Wikimedia Public (Research) Resources

Author: Diego Saez-Trumper / Wikimedia Foundation

Abstract:

The Wikimedia Foundation's mission is to disseminate open knowledge effectively and globally. In keeping with this mission, the Wikimedia Foundation support research in areas that benefit the Wikimedia community. We aim to make any work with our support openly available to the public. At the same time that we do a minimalist user data collection, all the material (text and multimedia) available in our projects is public and reusable by everybody. Moreover, all the revisions history and interactions among users are also public, and we offer a set of tools for accessing such data. In this tutorial we are going to give an overview on all the data sources, and a detailed explanation of how to interact with this content, including data and tools such as the Wikipedia Dumps, Quarry (SQL Replicas), Pageviews, PAWS (Jupyter Public Notebooks), Wikimedia Commons (multimedia content) and WikiData.

Speaker: Diego Saez-Trumper / Wikimedia Foundatioj

An outline of the tutorial:

* Introduction to Wikimedia Projects

* Overview of Wikimedia's dataset and tools:

- **Static Dumps:** Full Wikipedia dumps, where to get and how to parse them.
- **MediaWiki Utilities:** The python package to interact with Wikimedia Utilities.
- **Wikimedia API:** The Wikimedia API for accessing data.
- **Pageviews API:** How to check a detailed pageview count for any Wikipedia Page
- **SQL Replicas / Quarry:** The web interface to interact with Wikimedia SQL servers.
- **Clicks:** Explanation of the click dataset (navigation path within Wikipedia).
- **Event Stream:** Explanation of the (live) Event stream dataset.
- **Wikidata:** How to interact with this (semantic) knowledge base.
- **Wikimedia Commons:** A huge source of annotated images and videos.
- **ORES:** Public Machine Learning based quality control systems.
- **PAWS:** Introduction to the public Jupyter Notebooks.

* Hands-on session with PAWS

Length of the tutorial: 3 hours