Longitudinal dataset of patents related to the Web

This document describes our longitudinal dataset of US patents related specifically to the Web. The dataset contains 20,493 patents filed between years 1990 through 2013. Table 1 provides the specifications for this dataset and Table 2 describes the full list of attributes in each patent.

The dataset collection, processing and evaluation procedure is discussed more fully in our article:

Priestley, M., Sluckin, T. & Tiropanis, T. (2020). Innovation on the Web: The end of the S-curve? *Internet Histories*. [in press]

Corresponding author:
Maria Priestley (mp5g15@soton.ac.uk) – Web and Internet Science Group, University of Southampton

There is an additional GitHub repository containing all Python scripts that were used to collect and process the original patent data: https://github.com/mpriestley/web_patents
This GitHub repository includes our keyword filtering procedure and other scripts to create a data frame with monthly patenting rates (‘patents_by_month.csv’). This is where the input for our Code Ocean capsule comes from.

Table 1. Specifications

Subject	Economics and Econometrics, Computer Science (General), History
Specific subject area	Corporate innovation in the Web industry
How data were acquired	Most of the data were downloaded from the patentsview.org online platform in August 2018. The PatentsView API and USPTO website were accessed using Python scripts. Some new attributes were constructed by us to enable a broader range of analyses for ongoing work (these are marked with "created attribute" in Table 2.).
Data format	JSON
Parameters for data collection	Our methodology for selecting patents related specifically to the Web was based on a keyword filtering algorithm. The keywords were informed by technical infrastructures, software, data formats and protocols operant in the Web. The keywords are discussed in our paper and all Python scripts for selecting and processing the data can be found in our additional GitHub repository.
Data source	PatentsView online platform, USPTO website
Data accessibility	The data are shared publicly on the Code Ocean platform. Repository name: Longitudinal study of Web-related patents
Related research article	Priestley, M., Sluckin, T. & Tiropanis, T. (2020). Innovation on the Web: The end of the S-curve? *Internet Histories*. [in press]

Table 2. Patent data attributes

Attribute	**Description**
app_date	Date the patent application was filed
app_month	Patent application date to the nearest month - all days are set to 1 for ease of use (created attribute)
assignee_organization	Organization name, if the patent assignee is an organization (names were cleaned and de-duplicated by us)
assignee_country	Assignee’s country as listed on the patent
cpc_combo	Sequence of Cooperative Patent Classification (CPC) codes as one list (created attribute)
cpc_sequence	Order of the CPC classification in the list of classifications
cpc_group_id	CPC group ID
cpc_subgroup_id	CPC subgroup ID
cpc_subsection_id	CPC subsection ID
classes	List of classes in the patent (CPC subsection IDs) (created attribute)
cited_patents	List of all cited patent numbers
description	Full text description of the patent (scraped from the USPTO website)
insample_citations	Citations only to patents that are in the present sample (created attribute)
inventor_id	Unique inventor ID
patent_abstract	Abstract of the patent
patent_number	US Patent number, as assigned by USPTO
patent_title	Title of the patent
standards	List of W3C standards and Web 2.0 items that are mentioned in the patent text (created attribute)