The Web Integrity Project at the Sunlight Foundation compared the text on webpages in the Department of Health and Human Services' website, HHS.gov, before January 20, 2017, against their September 2019 versions. We analyzed the text in the body of 1,889 webpages, limiting our analysis to pages that both existed before President Trump's inauguration, and continued to exist at the same URL in September 2019. Our analysis included terms relating to a wide variety of issues, including sex education, contraception, and science. This analysis reveals:

- The use of the term “faith-based” increased by 35%, while the use of the terms “science-based” and “evidence-based” declined slightly.

The term “faith-based” (in both its hyphenated and non-hyphenated forms) appeared 64 times in the pre-inauguration versions of the HHS webpages included in this analysis. Today, “faith-based” appears 98 times in those pages.

The increase resulted largely from HHS's new strategic plan (compare the FY 2014 - 2018 plan to the FY 2018 - 2022 plan). For example, the “Strategic Goal 2” page was re-written between February 26, 2018, and March 5, 2018 (according to our web monitoring software; see the Internet Archive’s Wayback Machine captures for February 2, 2018 and August 28, 2018) to replace “Strategic Goal 2: Advance Scientific Knowledge and Innovation” with “Strategic Goal 2: Protect the Health of Americans Where They Live, Learn, Work, and Play.” The new version of the page used the term “faith-based” 16 times, whereas the old version did not use the term at all.

- The terms “success sequence” and “sexual risk avoidance” made their debut.

The term “success sequence” was added to the “Strategic Goal 3” page in February 2018, and a second instance was added in March 2019 (see the Internet Archive’s Wayback Machine captures from February 2, 2018 and May 3, 2019). The term describes a purportedly ideal
sequence of life events — completing schooling, obtaining a full time job and marrying, and only then having children — that proponents argue can help prevent poverty. The term “sexual risk avoidance” was added to the “Program Priorities” page, a page that outlines the priorities of the Office of Population Affairs (OPA) for the current year’s Title X program, when the FY 2018 priorities were replaced with the FY 2019 priorities (see the Internet Archive’s Wayback Machine captures for November 16, 2017 and November 17, 2017).

● The use of the term “abstinence” increased, although from a small base.

The number of times “abstinence” appeared on the 1,889 webpages studied increased from nine instances to twelve. The term, in the context of “abstinence counselling,” was added to OPA’s “Program Priorities” page when the 2018 priorities were replaced with the 2019 priorities (see the Internet Archive’s Wayback Machine captures for November 16, 2018 and November 17, 2018). The term, in the context of “adolescent abstinence counseling,” was also added to the “Statute and Regulations” page between May 30, 2017 and June 30, 2017 in the description of Section 1001 of Title X.

Our methodology, detailed below, excludes newly-created pages from the analysis. As a consequence, the website’s topic-specific “Abstinence” page was not included. (Based on an analysis of menu changes in the “Pregnancy Prevention” section of the website, this page was created between August 23, 2017 and August 24, 2017.) While this page did not contribute to our overall analysis of the use of the word “abstinence” on the HHS site, we note that the first Internet Archive’s Wayback Machine capture of the page, from June 13, 2018, refers to “abstinence” 16 times and “abstinent” five times on the body of the page.

● The use of the term “marriage” and “married” increased.

The use of the terms “marriage” and “married” increased 25% and 21%, respectively. Much of this increase stems from changes made to the “Strategic Goal 3” page between February 2, 2018 and May 3, 2019, in which 10 instances of “marriage” and 3 instances of “married” were added to the page. The usage of the terms “traditional” and “traditional marriage” did not change.

● There was a large decline in the use of the term “gender” and a small decline in the use of “sex” (in any context).

The usage of the term “gender” declined by 25% since inauguration, with 68 fewer references appearing on the HHS webpages studied. The use of the term “sex” (in any context) also declined slightly (with 11 fewer references) and the one instance of the term “biological sex” that existed before inauguration was removed.

Most of the decrease in the use of the three terms occurred in the “Civil Rights for Individuals & Advocates” section of HHS.gov, maintained by the Office of Civil Rights. This section of the
HHS site details the nondiscrimination provisions contained in Section 1557 of the Affordable Care Act.

The Web Integrity Project previously reported on these changes, which foreshadowed by almost two years the release of a new rule on Section 1557 that reduced the scope of sex discrimination protections in healthcare. In the six webpages in the “Civil Rights for Individuals & Advocates” section that relate to sex descrimination and the Affordable Care Act, there are now 63 fewer references to “gender,” 25 fewer references to “sex,” and 1 less reference to “biological sex” (Table 1).

Table 1: Reduced References to “gender,” “sex,” and “biological sex” on the “Civil Rights for Individuals & Advocates” section of HHS.gov.

<table>
<thead>
<tr>
<th>Page Title</th>
<th>gender</th>
<th>sex</th>
<th>biological sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Section 1557 of the Patient Protection and Affordable Care Act</td>
<td>2</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Section 1557: Frequently Asked Questions</td>
<td>28</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>Section 1557: Coverage of Health Insurance</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Section 1557: Protecting Individuals Against Sex Discrimination</td>
<td>5</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>OCR Enforcement under Section 1557 of the Affordable Care Act Sex</td>
<td>16</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Sex Discrimination</td>
<td>9</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>63</td>
<td>25</td>
<td>1</td>
</tr>
</tbody>
</table>

- Increased use of “Sexually Transmitted Disease” and reduced use of “Sexually Transmitted Infection.”

The usage of the terms “Sexually Transmitted Disease(s)” and “STD(s)” increased 20%, from 74 uses to 89 uses, while the use of the terms “Sexually Transmitted Infection(s)” and “STI(s)” decreased 20% from 44 to 35 uses.

- The use of the terms “abortion” or “abortions” increased slightly.

There were three more references to “abortion” in the HHS webpages examined in September 2019 compared to before January 20, 2017. Two of those additional references expressed prohibitions on the use of funds for abortion services (on the Program Priorities page, compare November 16, 2018 and November 17, 2018, and the “Statute and Regulations” page, compare May 30, 2017 and June 30, 2017).¹ The third additional reference comes from the Office for Civil

¹ Note: two references to “abortion” were added to the “Statute and Regulations” page between May 30, 2017 and June 30, 2017. Later, one reference to abortion on that page, which existed during the Obama administration was removed (compare May 31, 2018 and June 2, 2018).
Rights’ “OCR News Releases” page, which currently includes a link to a news release titled “OCR Issues Notice of Violation to the University of Vermont Medical Center After It Unlawfully Forced a Nurse to Assist in Abortion.”

- The use of the terms “teen(age) pregnancy” and “teen birth(s)” increased.

The number of references to “teen” or “teenage pregnancy” increased 12%, largely as a consequence of two changes. First, the addition of a section titled “Grants Awards to Replicate Teenage Pregnancy Prevention Programs” to the “Recent Grant Awards” page between July 11, 2019 and July 12, 2019, which includes four references to “teen(age) pregnancy.” Second, the replacement of content at https://www.hhs.gov/ash/oah/ when the Office of Adolescent Health (OAH) was folded into the Office of Population Affairs.2 The former OAH homepage was replaced with OPA content, which included four more references to “teen pregnancy” in links to statistics and information on grant programs.

References to “teen birth(s)” debuted on the HHS webpages studied, being added three times to the “Strategic Goal 3” page (compare February 2, 2018 and May 3, 2019) and once on the “Office of Adolescent Health News” page3 when the Office of Adolescent Health part of the HHS website was overhauled to the standard HHS.gov format (between January 27, 2017 and April 27, 2017). In the overhaul, two references to “teen pregnancy” were also added to the page.

- There was a decrease in the use of the term “contraceptive(s),” but an increased use of the terms “birth control,” “pregnancy prevention” and “condom.”

There was a 17% decline in the number of references to “contraceptive(s)” in the pages studied. On the “Program Priorities” page there are now fewer references to “contraceptive(s),” with the FY2017 version using the term “contraceptive(s)” five times; the 2018 version making no reference to “contraceptive(s);” and the 2019 version using the term “contraception” once. Most of the decline in the use of the term “contraceptive(s)” was, however, due to the removal of technical notes about SAS datasets and variable labels in the “Performance Measures” section of the website.4 The 30% increase in the use of the term “birth control” and the 24% increase in the use of the term “condom” were consequences of changes made to the “Pregnancy Prevention” section of

---

2 The webpage at the URL https://www.hhs.gov/ash/oah/ was titled “Office of Adolescent Health” and is now titled “Office of Population Affairs.”

3 After the OAH folded into OPA, the webpage was re-titled “News.”

4 For example, between October 10, 2017 and October 11, 2017, this note containing six uses of the term “contraceptive” was removed from the “Postpartum Most or Moderately Effective Contraceptive Methods” page:

There are two different SAS programs for the postpartum contraceptive measures; one is for 3-day postpartum contraceptive use and the other is for 60-day postpartum contraceptive use. Each program uses the variable MEM_MethodMost for the Contraceptive Care: Postpartum Most & Moderately Effective Contraceptive Methods measure and the variable LARCMeasure for the Contraceptive Care: Postpartum Access to LARC measure.
the website. For example, on the “Female Sterilization” page, three additional items were added to a list of contraceptive options, each with “birth control” in their name and two references to “condom” were added (in the context of the relative effectiveness of female sterilization and condoms, and the need to use a condom to “reduce the risk of STDs”) between November 16, 2017 and November 17, 2017.

Methodology

The Sunlight Web Integrity Project has developed a content analysis methodology that enables the exploration of broad language trends over time for an agency’s entire website or, for thematic analyses, on selected pages across agency websites. This method could be used, for example, to compare the frequency with which bilateral relationships are mentioned relative to multilateral treaties on the State Department’s website from one presidential administration to another, or to explore how terms relating to crime and punishment are used on a set of webpages relating to criminal justice, before and after the passage of a criminal justice reform bill.

In this analysis, we worked with Annalisa Merelli at Quartz to develop a list of terms relating to sex education and family planning that we expected might have been added or removed from content on HHS.gov since January 20, 2017. In all, we identified about 70 terms (excluding permutations and plurals):

- abortion(s)
- abortion counseling
- abstinence
- abstinence only
- abstinence plus
- biological sex
- birth control
- Chlamydia
- comprehensive sex(uality) education
- condom
- contraception/ contraceptive(s)
- contraceptive pill
- counseling
- crisis pregnancy center(s)
- day after pill
- emergency contraception
- evidence-based
- faith-based
- fertility-awareness
- gay
- gender
- Gonorrhea
- HIV/AIDS, AIDS, HIV
- human papillomavirus, HPV
- intercourse
- intrauterine device, IUD
- Levonorgestrel
- male contraception/ contraceptive
- marriage
- married
- natural family planning
- not married
- pill
- plan b
- pregnancies/pregnancy/pregnant
- pregnancy center/s
- pregnancy prevention
- reproductive health
- safe sex, safer sex
- same sex marriage
Once we have identified a domain or issue we wish to examine, our methodology involves four key steps. First, we first generate a set of webpages, identified by their URL, to compare at least two periods of time. Second, we generate a list of terms that includes terms we wish to assess (the “test terms”) as well as several control terms (terms that are unrelated to the test terms). Third, we scrape the text from the body of those pages (to exclude commonly repeated parts of the webpage like headers, side menus, and footers) for the two date ranges using snapshots captured by the Internet Archive’s Wayback Machine, counting terms from the list as we go. Fourth, we compare the term counts from the two date ranges and present data about relative and absolute changes.

The methodology is spelled out in detail below.

**Generating a set of URLs**

In order to maximize the extent to which we have a set of comparable pages within the two time periods, we generate a list of URLs for pages that both (1) were already live during the first time period in the analysis (the “pre-” date range) and (2) still existed during the second time period (the “post-” date range). Depending on the scope of the analysis, the list of URLs may be generated by a search of the usa.gov search engine for topic relevant pages, from the web crawling activities of our web monitoring software, or by using an application developed by Gayatri Balasubramanian that crawls the Internet Archive’s Wayback Machine. For all URLs in the list, these rules apply:

1. Only URLs that were live in both time ranges of the analysis, and which were captured by the Internet Archive’s Wayback Machine in both time ranges, are included.
2. Only URLs from federal government (.gov) domains are considered. That is, web pages belonging to state or local governments are systematically excluded.
3. Non-HTML formats (pdf, doc, docx, rtf) are excluded from the analysis as the program developed for this analysis does not yet support them.
4. Non-English websites were excluded from the analysis.

This methodology excludes some pages that may be of interest to the analyst. For example, in this analysis of the HHS website, the “Emergency Contraception” page was excluded because it had moved from one URL to another, and so criterion 1 was not fulfilled. The page included references to terms of interest including “emergency contraception,” “Plan B,” and “Levonorgestrel.” Exclusions of this type represent an unavoidable limitation on any analysis arising from the data.

**Text content analysis**

To analyze the text from the set of URLs, WIP developed a program in Python, which adapts EDGI's CTRL-F scripts to extend its functionality to support sentiment analysis and content extraction of URLs across different domains. The program returns the most recent “viable” Wayback Machine snapshots within the two provided timeframes (pre- and post-). Then, the program stores a “snapshot object” with attributes for the pre- and post- states of the URL. One of these attributes is the “visible text” of the web page (ie. the text that a user can see while navigating the web page) in the main body of the webpage. In order to extract the visible text, we exclude footers, navigation menus, headers, and footers using variations of their HTML tags.

The program counts the number of instances in which each subject term appears in the visible text of the page, and stores information that allows for the posterior analysis of the text as well as other statistics. Each term is counted exactly as it appears. The count of the term “pizza” does not include any instances of the term “pizzas,” and vice versa. However, the count of the term “pizza” counts all instances of the terms “pepperoni pizza,” and “meat lovers’ pizza.”

The program only outputs counts for URLs that had both pre- and post- attributes. URLs for which either the pre- or post- Wayback Machine snapshot was unavailable, redirected, or returned an error are excluded from the analysis.

Our approach has limitations. We are only able to compare web pages stored by the Wayback Machine for both the pre- and post- date ranges at the same URL. This may mean our methodology understates the extent of changes on an agency website, because the largest changes in content are likely to result from the creation of new pages at new URLs, and the removal of pages so that an old URL returns a 404 “page not found” error. Both of these types of changes (removed and added pages) are systematically excluded from our analysis. Additionally, our analysis cannot guarantee a perfect comparison of the pre- and post- set of captures, because sometimes when content is relocated from an existing URL (“A”) to a new one (“B”), URL A remains live though devoid of content or contains new unrelated content. Nonetheless, this methodology produces highly comparable sets of URLs, and we believe we get close to a true comparison of content that existed in one point in time versus an earlier time.
Additionally, for multi-agency analyses, our program may deal with agency-to-agency variability less than perfectly. HTML standards vary significantly even within departments, which makes it more difficult to accurately extract only the “visible text” from URLs on different domains in one function.
The Web Integrity Project is a project of the Sunlight Foundation.

Our mission is to monitor changes to government websites, holding our government accountable by revealing shifts in public information and access to web resources, as well as changes in stated policies and priorities. We work with journalists to make our findings public, and we produce policy analyses to evaluate and recommend changes to web governance practices and help ensure access to valuable web resources.

We would like to thank our team of volunteers and our partners at the Internet Archive’s Wayback Machine.

Contact us at webintegrity@sunlightfoundation.com

Follow us on Twitter @SunWebIntegrity