The below document outlines a datasheet for the datasets and other supplemental material created during the Archive of Tomorrow project, namely the JSON metadata files and crawl logs. This datasheet is based on the work done by [Emily Maemura](https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/asi.24048), who looked at adapting [datasheets from the machine learning community](https://arxiv.org/abs/1803.09010) into a humanities setting. This is version 1.0 of the project’s datasheet, circulated in April 2023.

# **Motivation**

**For what purpose was the dataset created?**

[The UK Web Archive](https://www.webarchive.org.uk/en/ukwa/) (UKWA) collects websites and pages published within the UK domain under non-print legal deposit legislation. [The Archive of Tomorrow](https://www.nls.uk/about-us/working-with-others/archive-of-tomorrow/) (AoT) project contributes to and operates under this structure, but focuses on collecting material around health topics. A more detailed explanation of the material that can be found in that collection can be found in the [Scoping Document](https://www.webarchive.org.uk/wayback/archive/20230311105013/https://drive.google.com/file/d/1sSNk9kTEds8Md9D_QN3oeS9aH2bcu3PV/) (this document is most easily accessed by following the link to ‘visit the current, live site’).

This datasheet focuses on the datasets provided as output in this project, namely the JSON metadata file and the crawl logs.

**Who created the dataset?**

Most targets within this collection were collected by the AoT team. However, due to the structure of the UKWA it is possible for the wider web archiving community in the UK who have access to the backend of the UKWA, called ACT, to edit and add targets to this collection, even after the project ends.

It has also been possible for the public to [nominate targets](https://www.webarchive.org.uk/en/ukwa/info/nominate) for the collection.

**Who funded the creation of the dataset?**

The AoT project was funded by Wellcome. The digital preservation and maintenance of the collection will be managed by the partner institutions of the project, namely the British Library (BL) and The National Library of Scotland (NLS).

**Any other comments?**

For more information on the AoT project please consult the Final Report, which will be published in May 2023 and made available through [nls.uk](http://nls.uk).

# **Composition**

**What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)?**

A target comprises of seed URLs which are intended to collect an online work, at title level. This may be a page, e.g. <https://twitter.com/WellbeinginYork/>, or a whole website: <https://www.actiononsugar.org/>. In some cases, a relevant section of a website may be the target: <https://www.imperial.ac.uk/infectious-disease/>. Datasets represent descriptive and technical metadata describing these targets, or track the crawl process through which targets are captured.

**How many instances are there in total (of each type, if appropriate)?**

3400 (as of February 2022). This could still increase, as people can still tag into the collection and the National Library of Scotland intends to return to the collection for a short maintenance review each year.

**Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**

The UK Web Archive aims to capture the whole web domain in the UK. The AoT collection is a subset within that, bounded by its subject matter, and by the collecting period, February 2022 - April2023. In terms of access the collection was made under the [The Legal Deposit Libraries (Non-Print Works) Regulations 2013](https://www.legislation.gov.uk/uksi/2013/777/contents/made); this empowers Legal Deposit Libraries to collect and preserve material, but as it is owned by respective copyright owner, access must be licenced by them. The collection is also expressed in its metadata and catalogue records. Accessible datasets include collection metadata expressed in [JSON](https://www.json.org/json-en.html) format, as well as UKWA crawl logs (covering the period between February - September 2022), which will be hosted at data.nls.uk.

**What data does each instance consist of?**

Archivists create the target, which in turn defines the web capture(s) to be made. The term is used synonymously with all captures made under the target’s rules. The archivist decides what the “work” is that they want to describe as a target, such as a whole website or single page. The frequency is based on the decision of the web archivists (daily, weekly, monthly, six-monthly, quarterly, annually, etc.). These could be captured outside of the active collecting period.

The metadata which explains the target, i.e. description “This is the Twitter profile of the Chief Medical Officer of Scotland” (a one page target), the frequency: “weekly”, and the start and end dates (e.g. 1st March 2021 - ongoing) is expressed in an export in JSON format.

Original crawl metadata is expressed as a text file under headings set out in [Heritrix 3 documentation](https://heritrix.readthedocs.io/en/latest/operating.html?highlight=log#log-files).

**Is there a label or target associated with each instance?**

Yes, but the target stores both the description of the instances, and the instruction to the crawler to create copies at the required frequency. This means that over time the description attaches to several copies of a potentially evolving work.

**Is any information missing from individual instances?**

Web crawlers cannot pick up everything - video embeds, scripted data visualisations, and databases are relevant examples which will not be well captured in this collection. In some cases, no information is captured at all, or only a client error response ([40x](https://developer.mozilla.org/en-US/docs/Web/HTTP/Status#client_error_responses)), at which point Heritrix stops its attempt to capture.

**Are relationships between individual instances made explicit (for example, users’ movie ratings, social network links)?**

No. In about 1000 records, a modified Wikidata ID has been added to create a link to information of an originating organisation, the publisher, in order to create agent information for a catalogue record. This is because the “originating organisation” field is free-text, and entries are therefore not controlled and can’t be used for this purpose. In a few cases, this Wikidata ID will include social media information for these organisations. For about 40 records, where the author is also the author of a book, a VIAF ID has been added for the same reason. VIAF IDs do not link to movie ratings or social network links directly, but do link to other linked data, which may include this information.

**Are there recommended data splits (for example, training, development/validation, testing)?**

No.

**Are there any errors, sources of noise, or redundancies in the dataset?**

Yes. There is a great deal of duplication in the collection, as follow up crawls may be identical where sites are not regularly updated.

The JSON file repeats metadata.

**Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)?**

WARC file can only be taken out one at a time, it is integrated with all the other data.

JSON export file and other metadata is self-contained. Crawl logs are also self contained.

**Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals’ non-public communications)?**

Everything that is collected is within the public domain. But there is information within the collection that could be considered confidential. An internal report, James Waddell’s ‘Defamation in Web Archives’, explores this topic in more detail.

**Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**

Yes. The collection includes a variety of public opinion on the topic of health, sometimes explicit information about health conditions or risks to health.

**If the dataset does not relate to people, you may skip the remaining questions in this section.**

**Does the dataset identify any subpopulations (for example, by age, gender)?**

Yes. There is a lot of collection done around certain groups or demographics. Examples of this are the subcollections on Trans Health, and Parenting and mental health. More information on this can be found in the [Scoping Document](https://www.webarchive.org.uk/wayback/archive/20230311105013/https://drive.google.com/file/d/1sSNk9kTEds8Md9D_QN3oeS9aH2bcu3PV/).

**Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset?**

Yes. Directly from VIAF author identifiers, and indirectly from names in titles and from URLs using the name of the person who owns the website. This includes social media handles identifying the user by their name. Individuals may be mentioned in the description, when explaining what the work is, and who made it.

**Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?**

Yes.

**Any other comments**

No.

# **Collection process**

**How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/ derived from other data (for example, part-of-speech tags, model-based guesses for age or language)?**

It was directly observable. The capture data is produced by external internet users and reproduced within the UKWA. Derivative data reflects elements of the capture data and includes general descriptive cataloging generated by web archivists (for example describing a target’s general topic area).

**What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)?**

Heritrix crawling for capturing the websites, all the metadata associated with a target was manually added to ACT, the tool used to create and update targets for websites and web pages.

Guidance on the use of ACT is available in the [W3ACT github](https://github.com/ukwa/w3act/wiki/W3ACT-User-Guide).

**If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?**

N/A, but it can be noted that this dataset feeds into a wider collection of websites within the UKWA.

A collection is created within the UKWA to increase visibility of a particular topic or theme, addressing platform-specific difficulties in discovering individual resources. The Talking about Health collection exists as a subset of the larger UKWA as well as a resource intended to make available a large variety of websites around a subject. A variety of factors define the scope of what is included - curatorial perspectives, period of collecting, serendipitous discovery - and description is ordinarily based on a limited instance/capture of a website.

**Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?**

The majority of this data reflects a web archive collection built by Web Archivists contracted to work on the project. Smaller contributions to the collection were made by ACT users outside of the project - for example Web Archivists and those in related positions at other institutions - and by researchers working on the project in less formal roles (members of the public nominating websites for inclusion, research fellows). This work was not compensated.

**Over what timeframe was the data collected?**

Extensively over a year, but targets can still be added to the collection.

**Were any ethical review processes conducted (for example, by an institutional review board)?**

Ongoing oversight was provided through reviews administered by the host institution, with a Steering Group; Advisory Board; and Collecting Framework group all meeting regularly to review progress and address any ethical issues. Consultation with research communities and stakeholders was conducted at periodic project workshops..

During the initial project proposal consultation was conducted to ensure that the project would meet appropriate standards and follow relevant safeguards governing ethical research, with some consultees continuing to advise from Steering Group/Advisory Board positions, and an Equality Impact Assessment was produced.

Throughout the project, ongoing reviews were conducted by Web Archivists tasked with making decisions around which websites to collect, assessing publishing concerns related to proposed targets as they came up. The scope of the collection grew throughout the project, often encompassing new topical areas, and where the ethical implications of collecting particular health-related material were considered sensitive these concerns were communicated to the Steering Group and Advisory Board.

**Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?**

Everything which the web archive collects can be regarded as “other sources”. Websites published in the UK are crawled and collected under [The Legal Deposit Libraries (Non-Print Works) Regulations 2013](http://www.legislation.gov.uk/uksi/2013/777/contents/made). This legislation permits legal deposit libraries to collect content published on the open web that can be identified as being published in the UK, with [exceptions](https://www.legislation.gov.uk/uksi/2013/777/regulation/13/made). Excepted material includes anything behind a login, e.g. intranets, or websites which contain mainly video or audio content, or personal data. Prior consent is not required from publishers for copying and preserving UK web material under the legal deposit regulations. It is the web archivist’s responsibility to assess each target for risk, under the rules. As most targets are set and left to run and then only spot checked for quality assurance purposes, it is possible that some personal data may be reproduced on archived web pages. The UK Web Archive operates with this low risk, maintaining a [Notice and Takedown policy](https://www.webarchive.org.uk/en/ukwa/info/notice_takedown/) in case material needs to be taken down under e.g. UK GDPR law.

**Were the individuals in question notified about the data collection?**

No, only if an open access licence was requested. Around 58% of the targets collected were associated with a licence request to individuals making them aware of collection of their website or page. A further 33% received an email making them aware of the specific scope and funding of the project.

**Did the individuals in question consent to the collection and use of their data?**

Content is collected under Non-Print Legal Deposit legislation. Though some individuals were made aware of collection and asked to give permission to grant wider access to archived copies of their site, permission was usually not asked for archiving websites and webpages for legal deposit.

**If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?**

All content is subject to the UKWA’s [Notice and Takedown policy](https://www.webarchive.org.uk/en/ukwa/info/notice_takedown/) and objections may lead to content being removed or disabled from public access. Takedown action is at the discretion of the UK Web Archive. It is likely that if a consenting individual later wished to revoke or alter the specific permissions a resource was made available under, this would be possible through liaison with the UKWA, although not guaranteed.

**Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted?**

Yes, a data protection impact assessment was conducted prior to the project start.

**Any other comments?**

No.

# **Preprocessing/cleaning/labelling**

**Was any preprocessing/cleaning/labelling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**

No preprocessing has taken place on the JSON target metadata dataset; the dataset is an unmodified export of the data produced and maintained in ACT (where web targets are subject to basic quality assurance work - name standardisation, ensuring compliance with baseline descriptive standards, etc). During the development of the API process used to generate these exports, some work was done to prevent the inclusion of blank, ‘null’ objects appearing in the JSON file for each collection; no additional exclusions were made. Additional contextual information about the collection has been added to the spreadsheet representation of the collection, 4028.xlsx, primarily to provide a list of collection titles generally indicated by numerical URIs in the metadata.

The large collection of crawl metadata logs also made available were cleaned to remove IP addresses throughout.

**Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)?**

This dataset represents the full metadata output the project has access to; while it might be possible to extract additional or more granular metadata on these resources from the systems involved, this was not possible during the project. Unredacted crawl metadata could be recovered in a situation where this was deemed valuable.

**Is the software that was used to preprocess/clean/label the data available?**

No.

**Any other comments?**

No.

# **Uses**

**Has the dataset been used for any tasks already?**

Research fellows affiliated with Cambridge University Libraries and Cambridge Digital Humanities have begun experimenting with the dataset in different contexts: one example include Andrea Kocsis's analysis of the availability of trustworthy health information online, detailed [here](https://andreakocsis.notion.site/AoT-Topic-Modelling-results-a1d4163df3a64293bf43584cf29447df). [A series of draft workbooks](https://github.com/LeontienT/ArchiveofTomorrowMetadataNotebooks) researchers might use or further develop to provide ways of working with the dataset are also available; the development of these or comparable workbooks should support the use of the dataset.

**Is there a repository that links to any or all papers or systems that use the dataset?**

If researchers use the material, they may publish the datasets alongside their findings.

**What (other) tasks could the dataset be used for?**

The dataset reflects a project with a subject focus on health information online, and as such could be a valuable source for research on general and specific health concerns as well as web archives more generally. Collection and subject indicators can be used to narrow the dataset to expose certain topical areas, and in this context the dataset is expressive of the range of topics captured in the UKWA more generally.

**Is there anything about the composition of the dataset or the way it was collected and preprocessed/ cleaned/labeled that might impact future uses?**

It is difficult to anticipate the changes that might affect future users, but among the factors that could impact researchers are:

* Changes to, or transition away from, the ACT platform used to capture and catalogue the resources described by the dataset. The introduction of a new platform might, for example, require redirection if links or URIs in the current dataset are followed.
* Changes to, or modernisation of, Non-Print Legal Deposit law or its interpretation, which might, for example, alter whether the dataset’s references to a website’s ‘open access’ status are current.
* Individual takedown requests or changes to the licensing terms imposed on individual targets. These requests might prompt NLS to provide more recent versions of the dataset reflecting appropriate inclusions/exclusions/changes.

**Are there tasks for which the dataset should not be used?**

To use this dataset there should be an understanding of how web archives are created, therefore it may be difficult to do extensive research in change over time, as most of these targets are from 2022-2023.

The dataset’s intrinsic focus on health information, which at times will inevitably be sensitive and personal, means that any tasks focusing on the contents of the websites described should be assessed in accordance with relevant guidelines on ethical research.

**Any other comments?**

No.

# **Distribution**

**Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created?**

There are limited circumstances in which the dataset will be actively distributed - one example would be the recent effort to encourage project Fellows to experiment with the data. It is expected the project will generally be available to interested parties through appropriate institutional data repositories detailed below.

**How will the dataset be distributed (for example, tarball on website, API, GitHub)?**

Datasets will be hosted at data.nls.uk.

**When will the dataset be distributed?**

It is expected a static copy of the data will be available via the channels above by May 2023.

**Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?**

<http://data.webarchive.org.uk/opendata/>

<https://creativecommons.org/licenses/by/4.0/>

**Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**

No; the most likely source of restrictions or alterations to the data would be individual changes mandated by relevant sections of GDPR, for example the redaction of object metadata surfacing Special Category Data.

**Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**

The contents of the collection is variously subject to Non-Print Legal Deposit regulations; General Data Protection Legislation; individual terms of licence agreed with contributors; and general copyright law.

**Any other comments?**

No.

# **Maintenance**

**Who will be supporting/hosting/maintaining the dataset?**

UKWA will be looking after the material, with the digital preservation and technical aspects being handled by the British Library. After March 2023, it will no longer be developed as it was during the AoT project. NLS will commit three days of work per year; this may be working directly with the collection, or interacting with designated contacts to encourage addition of new targets via the shared spreadsheet. Updates to the collection would then be relatively low in cost to partners, and require minimal work to retag new selections at NLS.

**How can the owner/curator/ manager of the dataset be contacted (for example, email address)?**

Contact may be made via the project Discourse page, which will be maintained in future by UKWA partners. Otherwise the Web Archivist at NLS would be a long term contact.

**Is there an erratum?**

At the project level, Archive of Tomorrow maintained an issues log. At the level of the UK Web Archive service, issues and fixes are logged on its Github account: <https://github.com/ukwa>. Data sets will be published at: data.nls.uk; known issues in the set would be included with descriptive metadata for the dataset, alongside version information, e.g. <https://data.nls.uk/data/metadata-collections/national-bibliography-of-scotland/> or http://data.webarchive.org.uk/opendata/ukwa.ds.2/.

**Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)?**

Amendments may be made where it is necessary to update or remove instances from the dataset.

New versions may be published alongside the old, as the collection is updated. Labeling errors would likely be preserved in older versions, for reference; however, it may be considered part of a GDPR query and replaced, in that case.

**If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?**

N/A.

**Will older versions of the dataset continue to be supported/hosted/ maintained?**

It is unlikely that superseded versions of the dataset would be made available.

**If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**

Targets can be added to the collection through the ACT tool with an account, or via the Archive of Tomorrow shared sheet, with agreement and depending on staff resource at the British Library.

**Any other comments?**

No.