



Web archiving programme

Request asked a series of questions about our web archiving programme

Information provided:

We do very much aim to preserve the original archival order of the websites we collect - as you say, we believe that this is the best way to ensure that they can be viewed in their original context. All our websites are catalogued under the ZWEB lettercode. Within this are 7 series, corresponding to the 7 thematic clusters we use for selection (e.g. ZWEB 1 is Defence and Foreign Policy). Each website is catalogued as a subseries within this. Every snapshot of that website is then catalogued as a piece within that subseries (e.g. ZWEB 1/14 is the snapshot of the Ministry of Defence website from 3 November 2003. We do therefore arrange the websites by their provenance, although the series level is an artificial distinction, and we do tend to separate the websites from other records produced by those departments. It may be useful to read our selection policy for websites, which is available at <http://www.nationalarchives.gov.uk/recordsmanagement/selection/pdf/osp27.pdf>.

...The majority of our sites are collected through a contract with the Internet Archive, and collected at either weekly or six-monthly frequencies. In both cases, all sites on the list are collected in a single crawl (e.g. all sites on the weekly list are collected in one weekly crawl). This does create a single ARC file for each crawl. We may also collect specific sites using other methods, in which case we will simply crawl that site. In fact, there is no problem with having sites split between multiple ARC files, and this is quite normal. The Wayback Machine software we use retrieves the files from the relevant ARC file using an index which is maintained for the whole collection.

... We do catalogue down to the level of the individual snapshot. However, if a website had several subdomains, these would be treated as a single snapshot and given one catalogue reference.

... 'Cabinet Office (4 sites) means that we collect four different websites, all owned by the Cabinet Office. They are collected and catalogued as separate sites, but would be collected in a single crawl

... Pages means website snapshots, so '2003: 6' pages means that we have six different snapshots of the same website from 2003. This part of the archive uses the Internet Archive's Wayback Machine software, so it is difficult for us to change the terminology.

... We apply our standard catalogue metadata, which is based on ISAD(G) and EAD. We currently collect very little metadata at the level of the individual snapshot, although we are investigating the collection of more detailed metadata.

...TNA is using a version of PANDAS for some of its web archiving, through the UK Web Archiving Consortium (www.webarchive.org.uk). We have changed the subject headings to suit the needs of the UK, but the majority of government websites are catalogued under the heading of "Government & Politics", with sub-categories such as "Central Government" and "Local Government". We were very constrained in the categories we could create for two reasons: firstly, the PANDAS software places limits on the number of levels of subcategory, and on the number of terms within each category, and secondly, as a consortium which involves both libraries and archives, we had to find a compromise which would meet

everyone's needs. Ideally, we would have created sub-categories to reflect the thematic clusters I mentioned in 1 above, but this was not possible.

Date of disclosure: 19 December 2005