NatSCA News

Title: Website Conservation: A Case Study on the NatSCA Website

Author(s): Edward Baker

Source: Baker, E. (2009). Website Conservation: A Case Study on the NatSCA Website. *NatSCA News*, *Issue 18*, 53 - 54.

URL: http://www.natsca.org/article/136

Website Conservation: A Case Study on the NatSCA Website

Edward Baker

NatSCA website Editor
The Natural History Museum, Cromwell Road, London, SW7 5BD
Email: Edward.baker@nhm.ac.uk

Abstract
The NatSCA website has been upgraded to the Scratchpad system, allowing for easier and quicker maintenance and the introduction of several new features. Although this article describes only the NatSCA website, many of the principles and methods used are applicable to many different website projects (atomisation, search, metadata, link rot, reusing data, redirects, best practice).

Need for change
On taking over the NatSCA website I immediately noticed some problems and omissions.

- Lack of search capability
- No data about the contents of publications
- Possibility for large amount of 'link rot'
- Data not atomised (hard to make 'deep links')
- Sustainability (the website was hard to maintain (‘conserve’) and develop

The Scratchpad system
The Scratchpad project was chosen as the software (and hosting) for the new website. The Scratchpad project (http://scratchpads.eu) is a set of tools for making biodiversity data available. Although NatSCA is not solely about biodiversity, the tools developed by the project have proved to be very useful. The new NatSCA site uses the bibliography and taxonomy management tools of the Scratchpad to manage publications (the publication, volume, issue and article tree is managed using the taxonomy tools). The Scratchpad also allows for the definition of custom content types (with custom fields) to enable the easy management of standard pages, job adverts, events, committee member biographies, etc.

Scratchpads are built on the Drupal content management (http://drupal.org) system using a MySQL database (both open source projects). The use of a robust relational database and content management system allows for the site to be content rich and easy to use, while simultaneously increasing functionality and reducing the time needed to maintain it. The software and hosting are provided for free. Using open source software means that there is a community of hundreds of developers that are continually providing prompt solutions to problems and developing new functionality.

Identity
The domain name natsca.org was previously used by NatSCA for its website. Due to the details for the domain being lost when it became time to renew ownership, the domain was purchased by another organisation. This meant that the site was only accessible at www.nhm.ac.uk/hosted_sites/NatSCA - hardly an easy address to remember. On moving the content to the Scratchpad server the site temporarily took the address natsca.myspecies.info (sub-domains of myspecies.info being the standard domains used by Scratchpads). The committee then decided to purchase the domain natsca.info. I registered the domain on behalf of the committee and the details needed to renew the domain have been lodged with several members of the committee to prevent this domain being lost.

Search Capability
The new site is fully searchable using standard keyword searches (like you would search on google.com). New content is regularly indexed automatically to speed up the searching process.

Publication data
Article level meta-data has been entered for all NatSCA publications. This allows for searching by author, volume, issue and title. The usefulness of this can be seen by searching for a term such as 'A*nthrenus*'. This system could be much improved by the inclusion of keywords and abstracts - although this would be a time

intensive operation. The Scratchpad bibliography module is used to store this data in a way that can be easily indexed by search engines.

An interface for browsing articles by publication, volume and issue has been implemented and can be accessed via the 'Publications' link at http://natsca.info. It is also possible to download article meta-data in formats suitable for reference manages such as JabRef and Endnote.

Meta-data is provided for all articles, PDFs are available for each issue of a publication. NatSCA News is released on a rolling system (there is a delay between publication and when PDFs are available for download).

Link rot
Link rot is the process by which links between websites break over time. This is caused by a number of factors; the address of a webpage changing, a webpage being deleted, a website moving to a new domain name.

Moving the site from the nhm.ac.uk domain to natsca.info could have caused a large number of links across the internet to become broken (the situation was further complicated by an intermediate web address of http://natsca.myspecies.info). This was prevented by arranging with the NHM IT team to redirect all NatSCA traffic from the old domain to the new one. In this way a visitor to www.nhm.ac.uk/hosted_sites/NatSCA/[path to page] would be seamlessly transferred to natsca.info/[path to page]: by following an 'old' link you will still end up on the correct page.

In the old site there was no real archiving of old content. Whenever possible, content that has become outdated (e.g. past events, information requests, etc.) should stay online. There is always a possibility that someone has linked to that information, and that people may follow that link to the NatSCA website. The new site allows these pages to kept online even though they are no longer available via the standard site menus. When appropriate a notice will be displayed stating that the content is out of date, and links provided to relevant current content.

Atomising data
The use of defined content types and fields in the new site ensures that the data is stored in a well structured format. This allows data to searched, sorted and retrieved more easily. Previously all event information was stored as a plain document, with multiple events on the same page. Now each event is stored separately, with separate fields for date, location, etc. To display a list of upcoming events a query is run against the database to extract all content of the event type that start in the future. The individual pages for each event remain after the event as an archive - preventing link rot.

Having separate pages for individual entities also allows for deep linking - you can link to the precise piece of information you need rather than, for example, a list of events, publications or people.

Reusing data
When you view a page on the new website what you see is generated by running a set of queries on the underlying database. By varying the queries run and the way the result is formatted, it is possible to reuse the same data in a number of different ways. In this way it is possible to easily generate RSS feeds from any content type (for example publications or jobs). The site also has the ability to automatically post new content to the NatSCA JISC mailing list if desired. These functions are not yet completely implemented - if you have a particular need for a particular service please e-mail me to prioritise (edward.baker@nhm.ac.uk).

Everything in one place
A search of the web for the Insect Collections Managers Group, (whose web content is hosted by NatSCA) revealed an out of date (circa 1997/8) site as being the number one result in Google. This old site was hosted by the Natural History Museum and after consultation with Mike Sadka we determined that the site was not being maintained and that the content should be transferred to the new site. Once the content was transferred Mike created a redirect from the old content to the new so that http://www.nhm.ac.uk/hosted_sites/uksf/icmg.htm redirects to http://natsca.info/ICMG.

The new NatSCA website is still work in progress. Any comments and feedback from NatSCA members are welcome.