BUILDING A BIODIVERSITY CONTENT MANAGEMENT SYSTEM FOR SCIENCE, EDUCATION, AND OUTREACH

CS Parr[1]*, R Espinosa[2], T Dewey[3], G Hammond[3], and P Myers [3,4]

[1]*Human-Computer Interaction Lab, UMIACS, Univ. of Maryland, College Park, MD 20742*
Email: csparr@umd.edu
[2]*Information Technology Central Services, University of Michigan, Ann Arbor, MI 48105*
Email: roger@umich.edu
[3]*Museum of Zoology, Univ. of Michigan, Ann Arbor, MI 48109*
Email: {lqb, gstarrh}@umich.edu
[4]*Department of Ecology and Evolutionary Biology, Univ. of Michigan, Ann Arbor, MI 48109*
Email: pmyers@umich.edu

ABSTRACT

We describe the system architecture and data template design for the Animal Diversity Web, an online natural history resource serving three audiences: 1) the scientific community, 2) educators and learners, and 3) the general public. Our architecture supports highly scalable, flexible resource building by combining relational and object-oriented databases. Content resources are managed separately from identifiers that relate and display them. Websites targeting different audiences from the same database handle large volumes of traffic. Content contribution and legacy data are robust to changes in data models. XML and OWL versions of our data template set the stage for making ADW data accessible to other systems.

Keywords: Database design, scalability, education, ontologies, biodiversity, interoperability

1 INTRODUCTION

Recent years have seen an explosion of digitally available information about biological diversity (Bisby, 2000). At this stage in the field of biodiversity informatics, there are multiple, often redundant databases, and work has begun in earnest to establish standards to allow them to be federated so that information retrieval across sources can be efficient. At the same time, access to natural history data about organisms is important to three distinct audiences with different needs: 1) the scientific community, especially those seeking coded data for large scale ecological or organismal analyses, 2) educators and learners in formal education settings, and 3) the general public. Our challenge has been to design a system that efficiently accommodates the data needs of these audiences. Below we describe our project, the Animal Diversity Web, and detail the implementation of a system architecture and data template design that supports highly scalable resource building and flexible delivery. The details of our architecture and data template design may serve as models to other biologists designing knowledge bases. In addition, though our system was not designed explicitly for interoperability, we believe that the technology is now available to make the contents of our database accessible to other computer systems.

1.1 Animal Diversity Web

The Animal Diversity Web (ADW) is an online resource providing information on extant taxa in the kingdom Animalia from all over the world. Content includes media, text, keywords, quantitative fields describing basic natural history and conservation status, a glossary, and a taxonomic database used for validating and organizing content. A large part of the content is provided by university undergraduates who submit reports on species as part of their course requirements. This content is edited by their instructors, and then edited again by a team of biologists at the University of Michigan. Experts at the University of Michigan and elsewhere provide content at higher taxonomic levels. The ADW project currently maintains

two parallel websites – the ADW, aimed at adults and intended primarily for undergraduate education and outreach, and the BioKIDS Critter Catalog, aimed at 10 to 12 year olds involved in an inquiry-learning biodiversity curriculum.

1.2 System requirements

A truly scalable, flexible biodiversity information system meets four main requirements: 1) it supports large numbers of authors and editors, 2) it allows managers to modify or add new data models (e.g. add, split or lump keywords, add new conservation lists or a physiology section, etc.) while preserving the integrity of legacy data, 3) it allows managers to deliver content to audiences with differing levels of subject expertise, or to otherwise change presentation at will, and 4) it supports sophisticated querying for inquiry learning or data harvesting for scientific studies.

1.3 Related work

Many web sites are designed to deliver natural history information about organisms. FishBase (Froese & Pauly, 2004) and AmphibiaWeb (AmphibiaWeb, 2004), for example, provide in-depth information on particular subsets of related taxa. OBIS (Ocean Biographic Information System) (OBIS, 2004) and FWIE (Fish and Wildlife Information Exchange) (Conservation Management Institute, 2001) Master Species File systems have a broader taxonomic scope. These systems, maintained in large relational databases, are created by and aimed primarily at experts. They often rely on an extensive controlled vocabulary of technical terms that is relatively static. The Tree of Life website includes full text descriptions more accessible to broad audiences. Its focus is on conveying information on evolutionary relationships among organisms and the characteristics supporting those hypotheses of relationships. The distributed nature of this system (taxonomically-related pages are maintained by experts on their local systems, then federated) offers high scalability.

A content management system similar to ADWs has been developed at University of Washington (Cherry et al., 2003). Its goal is to provide a flexible learning platform supporting multiple authors. This system, zBento, is designed to accommodate multiple domains, but not multiple audiences. In addition, zBento is not explicitly designed to maintain long term data using evolving data models. SenseLab (Marenco et al., 2003), uses an evolvable system designed to provide web access to an expert-oriented neuroscience database that is part of the Human Brain Project. Their semantic tagging approach is similar to ours.

2 IMPLEMENTATION

2.1 System implementation

The ADW approach can best be summarized as an application of the "loose coupling" philosophy (Weinberger, 2002) to content management. Content objects, or nodes, e.g. a photograph, sound, or other rich media file, or a paragraph of text or keyword pertaining to an organism, are managed together in a single object-oriented database. They are coupled, or related, by three kinds of identifiers. Semantic identifiers are concepts defined in an ontology/thesaurus and used to tag a node. From the user's perspective, this occurs via the process of filling out a data template. Taxon identifications tag the node with its biological taxonomic source -- a species or a higher level biological name. A route identifier, such as which audience education level or geographic region should see the node, specifies which website the node should appear in.

The "looseness" of the coupling refers to the fact that nodes are, in effect, managed separately from the identifiers used to relate and display them. Contributors and editors can manipulate the nodes and staff can modify data templates, taxonomic sources, and site display stylesheets. In practice, each tag on a node is merely an id number that points to the definition of the identifier.

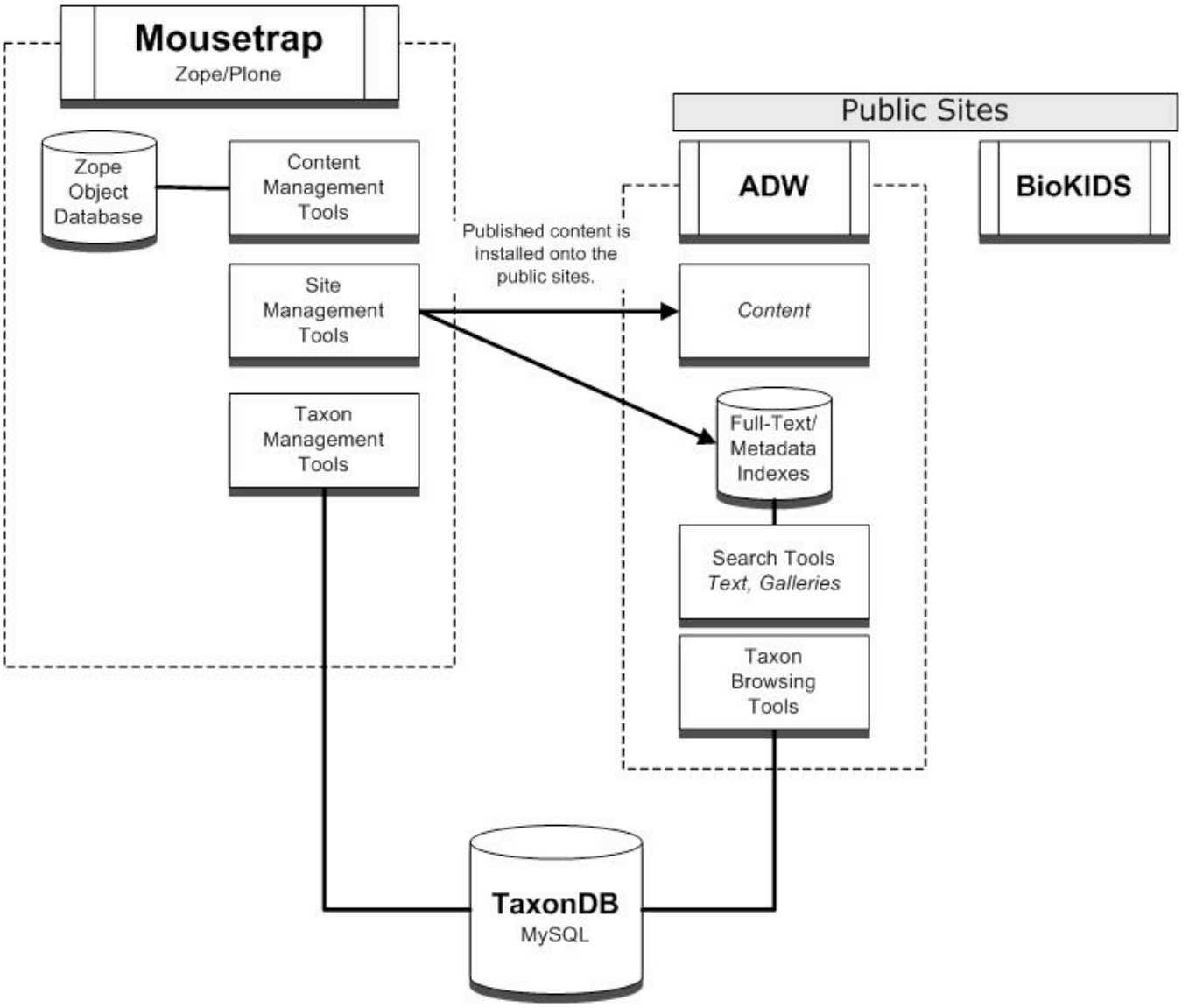

Figure 1. ADW architecture. Mousetrap is our online development environment, providing tools to allow contributors and editors to manipulate content. TaxonDB is a relational database providing both a taxonomic authority for content developers in Mousetrap, and a means of browsing the public sites taxonomically. The public sites are the content-rich pages and searching and browsing tools available to the general public, each customized to different audiences. As an example, the ADW site is expanded to show its subparts.

Our system architecture is shown in Figure 1. Mousetrap, available online to registered contributors and editors, is a customization of the Plone content management system. Its content management tools provides services to manage contributor information and access, file uploading and image processing, content metadata, and routing of nodes to particular websites. Nodes are managed as Zope objects. Mousetrap provides tools for customizing our public sites, such as style sheets. It also includes tools for managing the taxonomic database. TaxonDB is a MySQL relational database of biological names and their hierarchical or parent-child relationships. TaxonDB was built by integrating a number of publicly available datasets (Parr et al., 2004). It serves both as an authority for taxonomic identification and as a source of page organization in the published sites. Support in TaxonDB for multiple hierarchies provides flexibility in how we present the tree of life. The public sites, each built to serve a particular audience, are the third major part of the system. Figure 1 shows our current ADW

and BioKIDS sites, but any number of targeted sites are possible. The sites share some content, but also house content and tools specific to their intended audiences. Full-text and metadata searches for public sites are serviced by Swish-E indexes (http://swish-e.org).

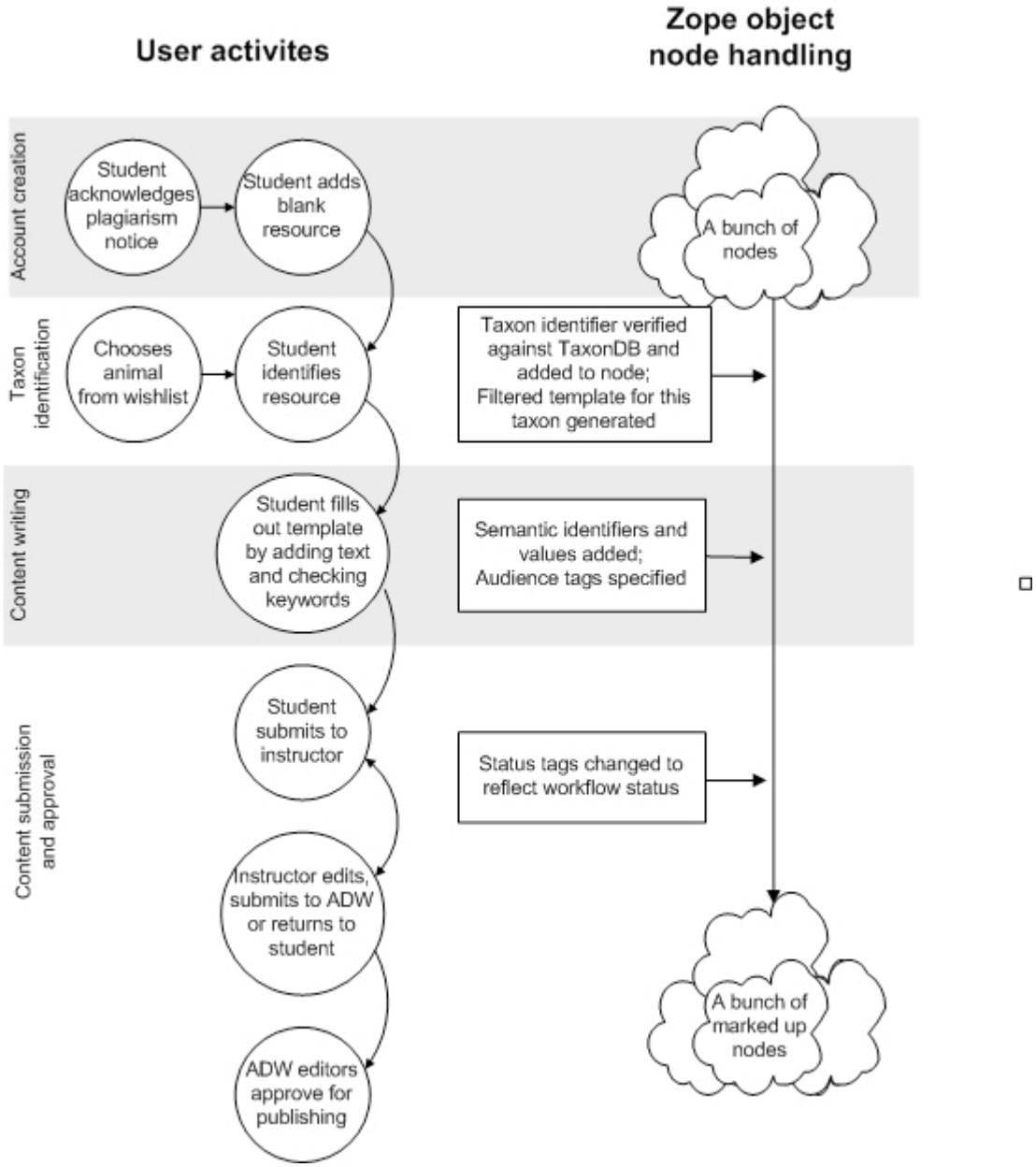

Figure 2. Content creation workflow in ADW. In this example of a taxon account, workflow from the user perspective is shown on the left. Workflow from the content management system perspective is on the right.

Figure 2 shows the workflow that supports the creation and manipulation of nodes by contributors and editors. Each node or collection of nodes (a media file, taxon account, etc.) must be identified with a valid taxonomic identifier from TaxonDB. Mousetrap prevents redundant taxon accounts and enforces that Latin name spellings be consistent with our authority. If the node is to be a taxon account, an appropriate data template is generated which is customized for the taxonomic group (see *taxon filters*, below). By filling in a blank data template including checking keyword boxes, contributors are actually creating a collection of text nodes and adding semantic identifiers to them. Editors may indicate the audience of a particular paragraph is college-level only and add a parallel paragraph appropriate for other audiences. Appropriate status tags are added as the nodes pass through the workflow from contributor to instructor to ADW editors, ensuring that content is

available to appropriate people for modification. Other kinds of nodes (sound clips, photographs) may have entirely different templates but pass through similar workflow.

After being entered and edited on Mousetrap, content is published via the following process. The simple plain-text markup language our contributors use (based on reStructuredText) is rendered into HTML. Search indices are updated to allow searching, and external and internal links are created and tested. The account is transformed to semantically mark up the content, resolving pointers to identifiers, and routed to the presentation stylesheet appropriate for each site. The system also uses TaxonDB to organize the resources for display in a Linnean hierarchy. Dynamic pages (e.g. image galleries and "feature" pages) are first generated when a user asks for a page, through a servlet reading the indices. The page is then written to the filesystem so the next request is static. Thus the public sites write themselves as they are used.

Because legacy data templates remain archived, legacy data remains semantically related. For example, in the current version of our data template, we might attach the semantic identifier "hermaphrodite" to nodes identified to the taxon "woolly slug." Later, we might decide to require contributors to specify "simultaneous" or "sequential" hermaphrodism, so we alter our data template. The legacy data entered under the previous template remains semantically tagged, so queries can still find these nodes and we can continue to display them, if we desire, in appropriate places on a web page. In addition, the new keywords are available to editors of legacy content. We may decide that this information is not appropriate for display to younger audiences, and so remove routing identifiers so as not to show it to them, or to show them a simpler synonym. We may re-identify all the nodes from the "woolly slug" to a more recent name simply by managing the taxonomic database.

The integration of semantic markup and nodes occurs at the lowest level; most of the system works with this combined XML so this approach could therefore be achieved in any environment with good XML support.

2.2 Data template implementation

Taxon accounts form the core data objects in the ADW natural history database. Information in the current taxon account template is organized into as many as 18 sections describing important aspects of animal biology. Section topics include distribution, physical description, reproductive biology, lifespan, behavior, food habits, predators, ecosystem roles, economic importance to humans, and conservation status. Template section choice was driven primarily by the goal of organizing the incredible breadth of natural history patterns in the animal kingdom into manageable, related pieces that could be consistently recorded across a wide range of animal taxa. The organization of the template in this way facilitates two activities. First, it allows the use of the ADW by both scientific researchers and educators as a source of data on animal behavior, ecology, and evolution. Second, it supports reliable addition of new content to the ADW by student contributors who are not technical experts and often lack access to some kinds of sources. For example, although we could add sections to the ADW data template covering population genetics, physiology, etc., those kinds of information are often only available for a limited set of organisms, may be available only in primary literature, and often require advanced training to understand and summarize. The dynamic features of the template and its legacy consistency make it possible for the template to be continually modified for new purposes.

The most important part of each section of the template is a block of searchable text, written by the account author. This text contains all the information presented in the section. Each section also has a list of controlled vocabulary keywords, unique to the section and may include data fields, where authors enter numerical data (e.g. mass, basal metabolic rate) or small items of text that address particular points (e.g. names of known predators, breeding season). The use of controlled vocabulary keywords avoids problems of synonymy and varying parts of speech (e.g. "hibernates" vs. "hibernation"), thus improving the accuracy of data searches. Hierarchical keywords are employed as appropriate, for instance, a taxon coded as eating mollusks (molluscivore) is automatically tagged as a carnivore as well.

This template structure facilitates accurate data searches by allowing users to search in specific natural history fields, for particular natural history descriptors (keywords), for data ranges (e.g. birds with wingspan 25 to 50 cm), and for combinations of these. In addition, the template structure acts as a guide to contributors, ensuring that a broad suite of natural history data is considered in writing about an animal taxon.

Contributors provide standard-format reference entries to document all information used in creating accounts. These references are managed separately and directly linked to the relevant taxon account section. Contributors select from a list of reference types (journal article, book, web resource, etc.) and are then supplied with a reference template with the fields and format appropriate for that reference type. Once all references are entered contributors then select relevant references from a list appearing within each template section. This process facilitates uniformity and consistency of both reference format and citation style within text sections. Online references are available as hotlinks.