A Taxonomic Approach to the Organization of Penn State Web Space

Michael Pelikan, Jin Ma, Margaret Smith, and James Leous

The Pennsylvania State University
Forming the Posse

- Formed by Dr. Russell Vaught, Associate Vice Provost for Information Technology as a result of his Web report to the Provost

- His assignment:
  - make finding specific pages easier
  - simplify Web content management tasks
  - come up with something durable

- Vaught held a series of one–on–one meetings in 2003, pulled together small group in 2004
The Team

• James Leous, Manager, Research Programmer, Information Technology Services
• Jin Ma, Metadata Librarian
• Sam Haldeman, Program Coordinator, Information Technology Services
• Richard Pearce, Director, Business and Finance, Auxiliary and Business Services
• Margaret Smith, Manager of Publications, Information Technology Services
• Michael Pelikan, Technology Initiatives Librarian
• Russell Vaught, Associate Vice Provost for Information Technology
The Penn State Web

Extent of Penn State Web

- 24 campus locations outside of University Park
- 12 colleges
- Graduate School
- Medical School
- Law School
- more than 1,000,000 public, non–personal web pages
- “first stabs” at Content Management popping up like mushrooms!
Initial Findings

Group’s initial findings

- near–total lack of control on:
  - page titles
  - link labels
  - standardized or authoritative form for units, departments, etc.
Initial Findings, cont’d

Group’s initial findings (cont’d)

- increasing adoption of CMS systems across Penn State
  - dozens? hundreds?
- ever–evolving search engines with differing treatment of tags
- Need to be CMS and search engine agnostic!
Initial Targets

The Group’s initial targets (2004):

- Rules to govern authorized form for name for any entity in the University
Initial Targets

The Group’s initial targets (2004):

- Rules to govern authorized form for name for any entity in the University
- a taxonomic rendering of those terms, where useful
Initial Targets

The Group’s initial targets (2004):

- Rules to govern authorized form for name for any entity in the University
- A taxonomic rendering of those terms, where useful
- A plan to derive benefit from such terms, even in the present environment
Initial Targets

The Group’s initial targets (2004):

- Rules to govern authorized form for name for any entity in the University
- A taxonomic rendering of those terms, where useful
- A plan to derive benefit from such terms, even in the present environment
- A plan anticipating their broader application in foreseeable Web management systems
Initial Targets

The Group’s initial targets (2004):

- Rules to govern authorized form for name for any entity in the University
- A taxonomic rendering of those terms, where useful
- A plan to derive benefit from such terms, even in the present environment
- A plan anticipating their broader application in foreseeable Web management systems
- A set of practices to ease their maintenance over time
Tools c. 2004

Potential implements to wield, at hand in 2004:

- Google Appliance
- LDAP
- Zope–based CMS effort
- concept of CMS as both a means and an opportunity to implement
Google Fly in the Ointment

In the first year of the “Tags” committee work, Penn State purchased a Google Appliance. This meant:

- The PageRank\(^{TM}\) algorithm determines quality of search
- Keywords/META tags were much less important

Overall, our constituents were much “happier” with the quality of the searches, but we had a harder time with “fixing” pages. We suggested a “Sponsored Links” on the Google Appliance.
The Fix is On

Without a “Sponsored Links” option the process of making the search find the “proper” academic department is very manual. Could the University Directory (LDAP) help us?

LDAP is populated by our CACTUS Accounts database. The LDAP table comes from several authorities:

- Student Data from Enrollment Management
- Employee Data from Office of Human Resources
- Medical School/Center information from MSHMC
- Departmental Information from Call Center
Combined Search

Can we create a combined search page which incorporates Google result set with a return from LDAP?

- Such a page would satisfy our need for getting the “right” answer if LDAP departmental entries were properly populated.
- The key is to have the proper synonyms for searches
- Can EduOrg LDAP schema help us?
Controlling Terminology

Predictable content retrieval
  - based upon data in elements or fields
  - common or mappable elements necessary but not sufficient by themselves!

Need for consistently applied terms
  - controlled vocabulary
  - authority control, if feasible
Formal Names, Legal Names

- Departments, Colleges, Locations, Units, etc.
- The So and So Institute of the Donor–Name College of Subject Area of the Pennsylvania State University
- Non–intuitive unit names:
  - General Stores (office supplies)
  - Fleet Services (rental cars)
Ownership

Who “owns” the authorized terms?

- As of now, list is maintained by the Office of Publications
- The Call Center provides a means to update some information (Consulting and Support Services, a unit of Information Technology Services)
Formal Terms
Distillation of complicated, formal terms

- must be done by hand
- must use consistent practices
- must result in unique terms
  - or must represent context for non–unique terms
  - meaning thesaural or taxonomic representation is indispensable
- need for mechanism by which common names or “nicknames” resolve to authorized term
Tools c. 2005

Implements to wield, at hand in 2005:

- The “ear” and interest of Google (??)
- Day Communiqué, licensed by Penn State Libraries, but for University–wide use
- Library local LDAP implementation
- Engagement of Old Main as a University–level strategic project
Google

The “ear” of Google

- Sponsored Links?
- Database connectors (including LDAP?)?
- Will they let us at the controls??
Google

The “ear” of Google

- Sponsored Links?
- Database connectors (including LDAP?)?
- Will they let us at the controls??
  - “Naaaaaaah!”
Day Communiqué

- “Content Bus”
  - LDAP connector
  - Workflow modules
  - Workflow elements are Content

- “Glossary”
  - pre–populated check boxes
  - capture locations, Organizational Units, etc.
  - representation of hierarchical relation
    - `<dcterms:isPartOf>ISRARTOF</dcterms:isPartOf>`
    - `<dcterms:hasPart>HASPRT</dcterms:hasPart>`

- metadata is Content
Old School

Good old fashioned Library work

- round up, wrangle and corral Unit, Department, College, Location names, etc.
- populate LDAP OUs with standardized, authorized form (OUs as DNs == main entry)
- populate LDAP “nickname” elements with access points resolving to authorized form
- similar authorized labels needed for page “function” (resource, transaction, etc.)
Progress

The Taxonomic Tags Group is making substantial progress!

- How can this be?
  - small, interdisciplinary group
  - a blank sheet of paper
  - fortunate convergence of circumstances and technology
  - sheer, innocent audacity
Questions and Discussion