Disambiguation: the Key to Information Architecture?

Chris Lüer
Ball State University
clueer@bsu.edu

August 5, 2006

Outline

1. What is disambiguation?
2. Traditional disambiguation techniques
3. Disambiguation in Wikipedia
4. A genealogy wiki
5. Information architecture, or is the whole Web a wiki?

What is disambiguation?

- A relation
  - word -> article
- Example
  - **mercury (planet)**
  - **mercury (mythology)**
  - **mercury (element)**
- Wiki idea
  - every page has a unique name
  - pages can be referred to even if they don't exist
- Wikimedia mechanisms
  - disambiguation pages
    - **mercury**
    - **Michael Jackson (disambiguation)**
      - through "For other uses, see Michael Jackson (disambiguation)."
  - header template: "For the foo, see bar."

What is good disambiguation?

- Article name must be unique
- It must be clear which article is meant
  - **mercury (science)** ??!
- Article name must be guessable
  - so people can find the article
  - so people can link to it
  - ideally, use of disambiguation page not required
- Redirects may help
- A system is needed
  - naming conventions in WP
- Importance of good disambiguation
  - avoiding bad links
  - avoid having to read disambiguation pages

Traditional disambiguation

- Monarchs
  - **Elizabeth II of the United Kingdom**
  - **Luis XIV of France**
  - sometimes historical, sometimes invented by historians and lexicographers
  - POV issue: which predecessors are counted?
  - POV issue: claimants
- Animal species
  - **Homo sapiens**
  - **Felis silvestris**
  - Established by Linné (18th century), standardized and interpreted by International Commission on Zoological Nomenclature (ICZN)
- Köchel directory
- Chemical nomenclature

Disambiguation in Wikipedia

- Established naming conventions work well
  - rulers, species
  - stretched beyond limits
    - minor rulers, pretenders
- By class
  - **mercury (element)**
  - **Thomas (Apostle)**
- By theme
  - **Mercury (mythology)**
  - **reservoir (water)**
- Specialized scheme for cities / towns
  - **Cambridge, Massachusetts**
  - but: **Frankfurt (Oder)**
- Specialized scheme for people
  - **John Taylor (1781-1864)**

Disambiguation in Wikipedia

- Guessability
  - historical rulers, towns in USA: user can guess location
  - why **Mercury (mythology)**, not **Mercury (god)**?
  - why **mercury (element)**, not **mercury (chemistry)**?
- Three competing schemes for people
  - John Taylor (poet) (1580-1654), English poet
  - John Taylor (1704-1766), English classical scholar
  - John G. Taylor, British neural-network researcher
  - which is better?
    - cause of fame = easier
    - lifetime = more likely to be unique
    - middle name = only if well-known

A genealogy wiki

- Disambiguation is important in an encyclopedia, but much more so in other applications
  - encyclopedia: wide range of well-known topics
- Genealogy
  - study of genetic relations between people
  - large numbers of dead, non-famous people
- Disambiguation issues
  - people
    - 1000s of John Taylor
  - places
    - towns that were given up
  - calendar...

Disambiguation of persons

- Genealogy = all people who were ever documented
- Encyclopedia = all famous people
- Market research and other business applications = all living adults in a certain group
- Possible schemes
  - Name
  - SSN: recent USA only, privacy concerns
  - profession, location, nationality: changeable, imprecise
  - birth and death year: often only one known
- Best generic solution
  - **Michael Jackson, 8/29/1958, Gary, Indiana**
  - if birth date not available, death date
  - both birth and death date unnecessary
  - effectiveness depends on disambiguation and precision of birth place

The whole Web a wiki?

- Disambiguation: an easy yet efficient way to identify stuff
- Can it be extended?
- Identifying Web pages
  - by URL = unique, cumbersome
  - by search engine keywords = imprecise, easy
  - tagging (metadata) = more precise, but not unique
- Disambiguation as the best of both worlds?
  - **Chris' homepage**
  - **Harvard University homepage**
  - **Wikipedia page on Michael Jackson**
  - **Foo's Nature paper on Bar**

Between URLs and the Semantic Web

- Naming should be a contract between reader and publisher
  - not a decision solely to be made by publisher
- The big contribution of wikis:
  - well-practiced trade-offs
- Unlike: URLs
  - based on domain names
  - first-come first-serve approach
- Unlike: keywords or tags
  - no effort at uniqueness
  - no way to enforce correctness
- Wiki-like technologies suitable for many networks
  - specialized peer-to-peer networks
  - internal Webs
  - wiki-like disambiguation does not require wiki editing

Summary & Questions

- Disambiguation is core feature of wikis
  - useful, but issues exist
  - guessability
  - uniqueness
- Naming conventions required
  - social process
- Can disambiguation be used elsewhere?
  - URLs -> keyword search -> tagging -> wiki-like page names with disambiguation
  - promising approach for other applications
  - combines precision with ease of use