Annotating Web Archives — structure, provenance, and context through archival cataloguing

Paul WU Horng-Jyh, PhD
Nanyang Technological University
hjwu@ntu.edu.sg

Agenda

- Background: Motivation & Objectives
  - Web Archives of Singapore (WAS)
- Understanding Web archives
  - An example of Singapore's eGovernment
  - Archival Cataloguing (Arizona Model)
- WAWI Web Annotation System
  - Review of Previous Work
    - Cataloging, Annotation & Semantic Web
- Concluding Remarks

Web Archives of Singapore (WAS) – A joint project between NLB and NTU

- Objective:
  - Capturing web as a record of Singapore's national identity
  - Pilot targeting 1,200 Singapore (.sg) websites
- Theme-based: Singapore Heritage
  - Websites produced by government ministries, departments and their agencies as well as academic, educational, commercial, community and social organizations are to be included
  - Websites with content reflecting Singapore's culture, history, politics, economics and social conditions
  - Topical websites documenting current events eg. Asian tsunami and Singapore election 2006
  - Websites of conference proceedings, newsletters and online publications that have ceased print coverage
  - Excluded: Blogs, CAMS, Discussion lists, chat rooms, bulletin boards, news groups, Product listings, Illegal and inappropriate websites, compilations of information from other sources
- URL: http://was.nlb.gov.sg/

R&D Motivation

- Initiate first Singapore national web archives
  - Preserve snapshots of Singapore's Web heritage
  - Preserve evidence for cataloging decision based on web context – web annotation
  - Facilitate collaborative and community based cataloging process
- Increase access to web archives content – beyond index and full text search
  - Organizing web archives for research and analysis

NLB
National Library Board
Singapore

Singapore Government
Integrity • Service • Excellence

Web Archive Singapore

Web Archive Singapore

NLB's Web Archive is a rich collection of some 1,000 Singapore-related online contents that showcase various facets of Singapore life. Ranging from subjects such as government administration and education to popular topics like arts and recreation, these websites were carefully selected to be part of the nation's documentary heritage.

Visit the Archive and explore Singapore's heritage in the making.

Note:
This service is best used with Internet Explorer version 6 and above.
Users are encouraged to clear the PC cache and access the Web Archive through NLB proxy.
Users should reset to the original proxy settings before exiting from the Web Archive.
Opened windows may not work properly.

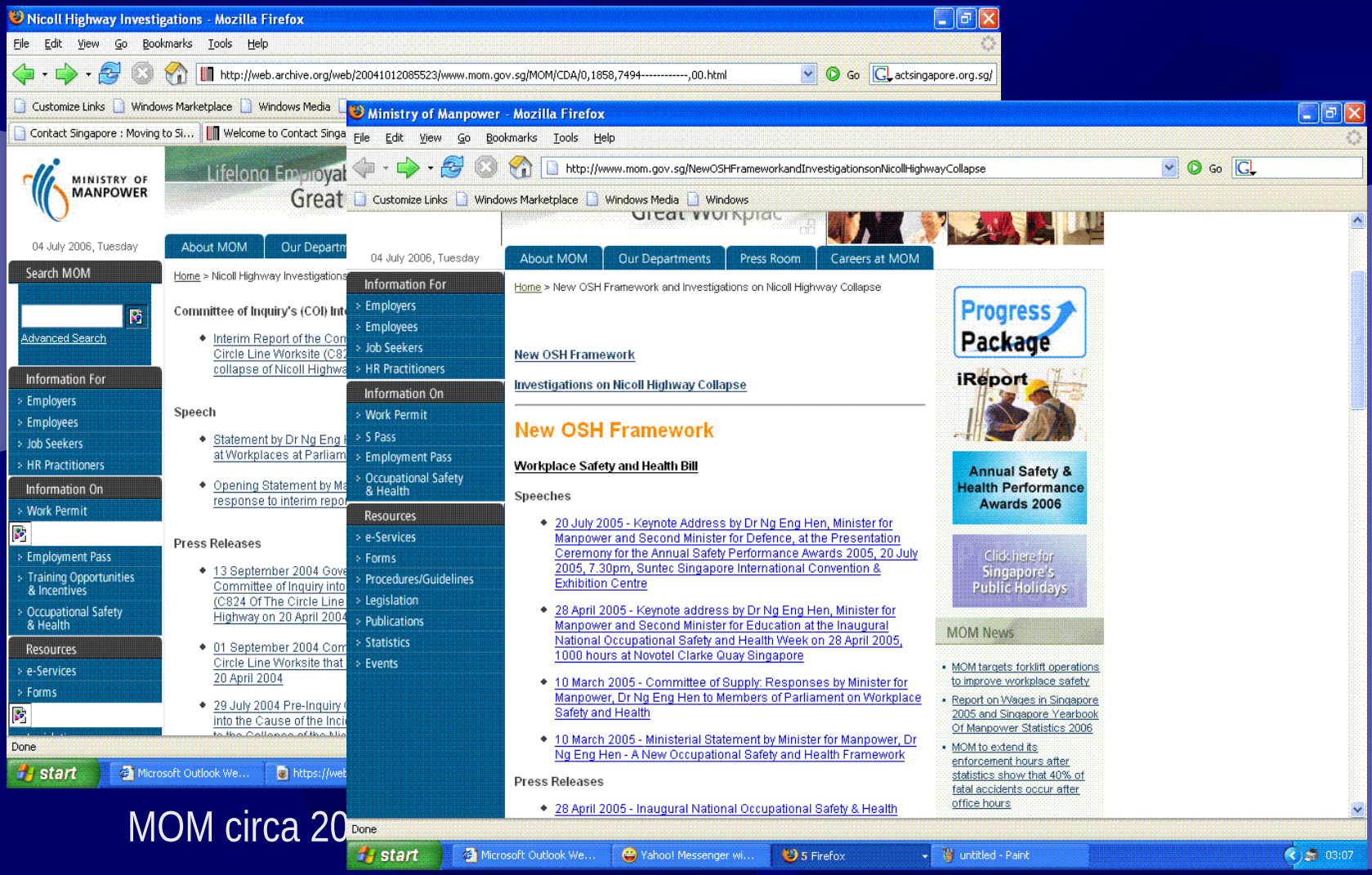
You can either visit the Web Archive as per **normal**, or through NLB **proxy**. Accessing through NLB proxy will ensure that the contents displayed are from the Web Archive.

Access through NLB Proxy

An Example: Ministry of Manpower (MOM) Web Archives

- Nicoll Highway Collapse Incident
  - A high profile work site accident in Singapore:
    - Occurred in 2004 involving major
      - government agencies and
      - multinational corporations
    - Further delayed already delayed engineering work
- Help trigger legislating of a new Law:
  - Workplace Safety and Health Bill

Web Sites Then – 2004, and 2006



MOM circa 20

MOM circa 2006

Web Sites Then – 2008, and now 2010

MOM circa 2008 from Web Archives

MOM circa 2010 from Current Website

Content Comparison: Before and After

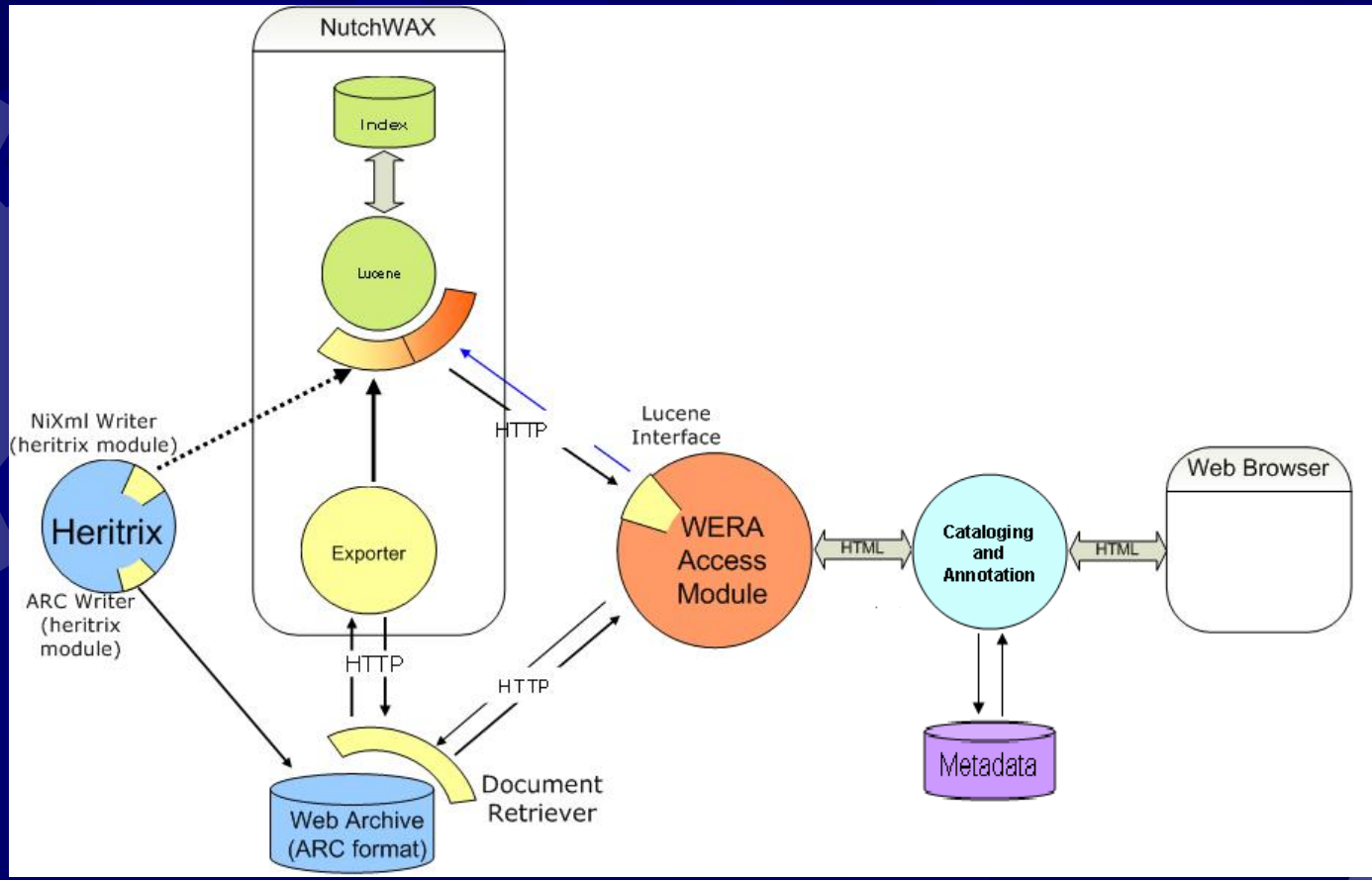
- Content stabilized between 2008 and 2010
- Content that is in common between 2004 and 2006:
  - Minister's Speeches
  - Press Release
- Content that was available 2004, but not 2006:
  - Committee of Inquiries (COI) reports
  - FAQ
- Content that is 2006, but not 2004:
  - Workplace Safety and Health Bills

Accessing Live Web Sites v.s. Web Archives

- A Web site is a single version
  - Reflecting the current concern of the content provider
- A Web archives contains many versions of the same web site
  - Reflecting the evolution of a Web site
  - Bringing to bear what was in the past with the current
    - Better understanding the present based on the past

- Need information architecture beyond current index & full text search
- Semantic Web with Ontology-aware annotation

WAWI Annotation and Cataloging - Integrated with IIPC platform



IIPC stands for International Internet Preservation Consortium
WAWI stands for Web Annotation for Web Intelligence

Key Features in WAWI

- Context-aware annotation
- Ontology-aware annotation

- Implemented through the following:
  - Relate web content to semantic content in the metadata
    - “Recordlizing” Web content based on “semantic content”
    - Provide flexible and precise annotation of the evidence
      - Render agreement, disagreement and different granularities of evidence
    - Detecting change for
  - Relate metadata to ontology
    - Capture Structural, Provenancial Relations

Understanding MOM

- The mission of Ministry of Manpower (MOM) is to achieve a globally competitive workforce and great workplace, for a cohesive society and a secure economic future for all Singaporeans.
- Occupational Safety and Health (OSH), a division of MOM, promotes OSH at the national level. It works with employers, employees and all other stakeholders
  - to identify, assess, and manage workplace safety and health risks so as to eliminate death, injury and ill-health.
- OSH Inspectorate, a department within the OSH Division focusing on the reduction of safety and health hazards
- It does so by providing advice and guidance through
  - inspections of workplaces,
  - investigating accidents and
  - enforcing the relevant laws.

Understanding MOM (cont'd)

- The corporate communicatoin ensure the activities of MOM divisions are reported timely
  - Minister speech
  - Communittee of Inquiry Reports
  - Press releases
  - Hearing of reports
  - FAQ regarding how public and participate
    - The process
    - The public hearing announcments
    - Other quesitons the public may have

MOM Organization Chart - An Ontology

"Annotatable" from Government Directory Website

Political Appointees & Their Personal Assistants

Post	Name	DID	Email
Minister	GAN Kim Yong	63171601	gan_kim_yong@mom.gov.sg
Minister of State	LEE Yi Shyan	63171603	lee_yi_shyan@mom.gov.sg
Senior Parliamentary Secretary	HAWAZI Daipi	63171610	hawazi@mom.gov.sg

Senior Management & Their Personal Assi

Post	Name	DID
Permanent Secretary	LOH Khum Yean	63171688
Deputy Secretary	Aubeck KAM Tse Tsuen	63171626
Quality Service Manager	Mrs Roslyn TEN, PPA(G)	1800-5386930(Toll-Free)
Press Secretary to Minister	Ms Farah ABDUL RAHIM	63171637
Personal Assistant to Minister	Ms Lilian NG	63171601
Personal Assistant to Minister of State	Ms Stella WOO	63171603
Personal Assistant to Senior Parliamentary Secretary	Ms Patricia ONG	63171610
Personal Assistant to Permanent Secretary	Ms Dorothy CHUA	63171688
Personal Assistant to Deputy Secretary	Ms Maxine KWAN	63171626

Portfolio

Subjects

Manpower Planning and Policy	International Manpower Programme	Manpower Development Programme
Labour-Management Relations	Regulation of Trade Unions	Employment Terms and Conditions
National Wage Guidelines	Administration of the Central Provident Fund	Occupational Safety and Health
Work Injury Compensation	Employment Services	Regulation of Employment Agencies
Work Permits for Foreign Workers	Labour Research and Statistics.	

Departments/Divisions **Statutory Boards**

MANPOWER PLANNING AND POLICY DIVISION	WORK PASS DIVISION	INTERNATIONAL MANPOWER DIVISION
LABOUR RELATIONS AND WORKPLACES DIVISION	OCCUPATIONAL SAFETY & HEALTH DIVISION	CORPORATE SERVICES GROUP
FOREIGN MANPOWER MANAGEMENT DIVISION	INCOME SECURITY POLICY DEPARTMENT	LEGAL SERVICES DEPARTMENT
INTERNAL AUDIT UNIT	WORKPLACE POLICY AND STRATEGY DIVISION	

OSH Division

URL: http://www.sgdi.gov.sg/

Existing Annotation Models

- Context-less
  - Additional information is added without reference of the web content
  - E.g. WebArchivist.org's drill search
- Context-aware
  - Information is added with web content - in a manner of like the literary warrant in library science
  - E.g. Annotea in Semantic Web community

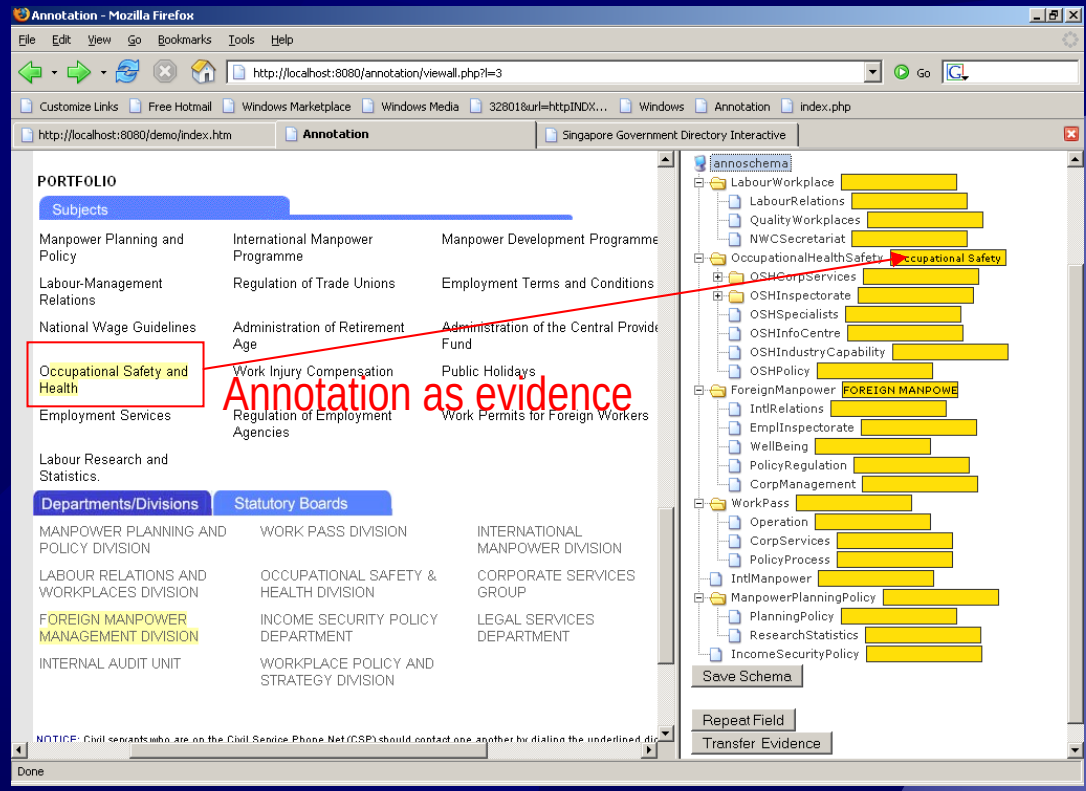
WAWI Annotator
- Limitation of Annotea

- Precision is based on X-Path, “text node,” rather than “text position,” is the smallest reference unit
- Annotation is a single marker, boundary/extent is not indicated

- WAWI Annotator
  - Page coordinate as reference (per character unit)
  - Annotation Graph (AG) representation of annotation

Annotating (and Monitoring) MOM Ontology

- The highlighted text is captured as evidence of the ontology
- So, if there is a change on the left panel, an alert will be sent saying the ontology may need to be modified



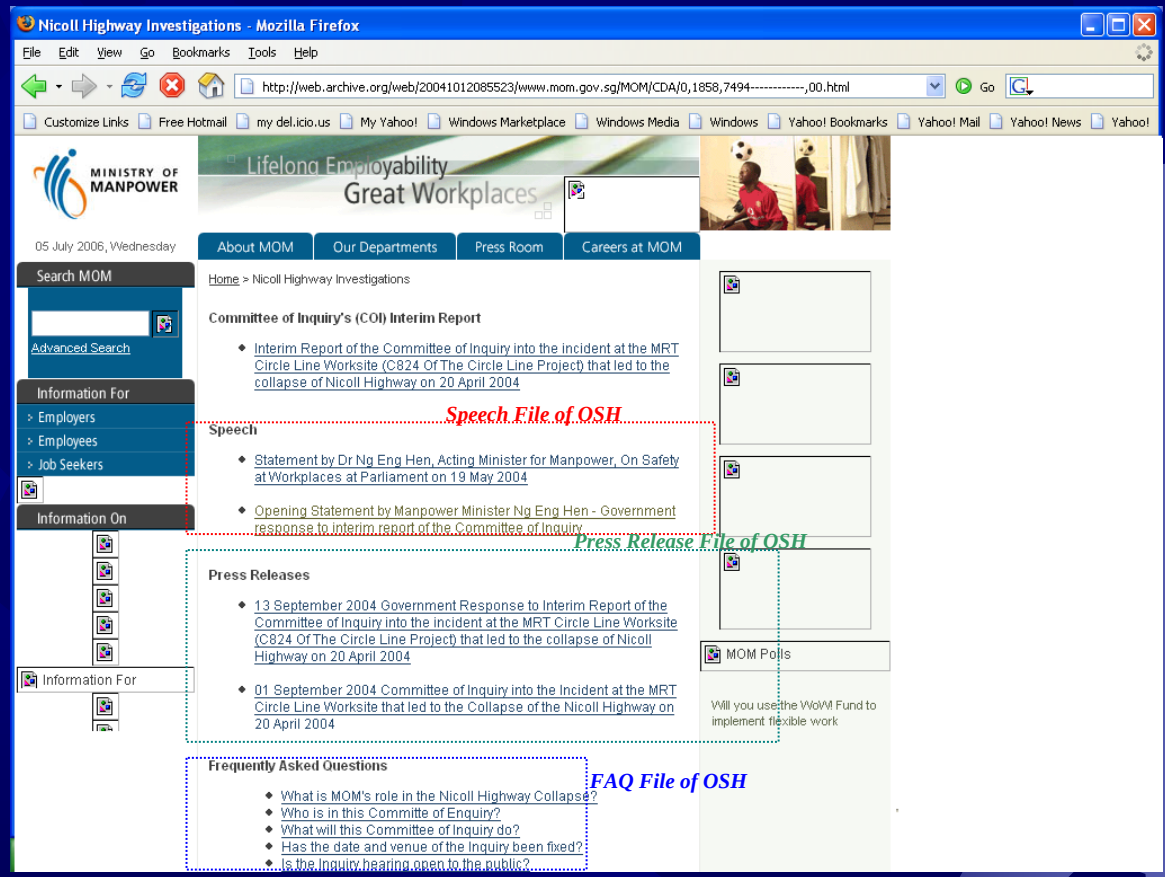
The Context – A highlighted text in a Web Page

The MOM Ontology

Annotating OSH Records with Metadata

- Records of
  - Minister Speech,
  - Press Release,
  - FAQ

  can be annotated with evidence and metadata

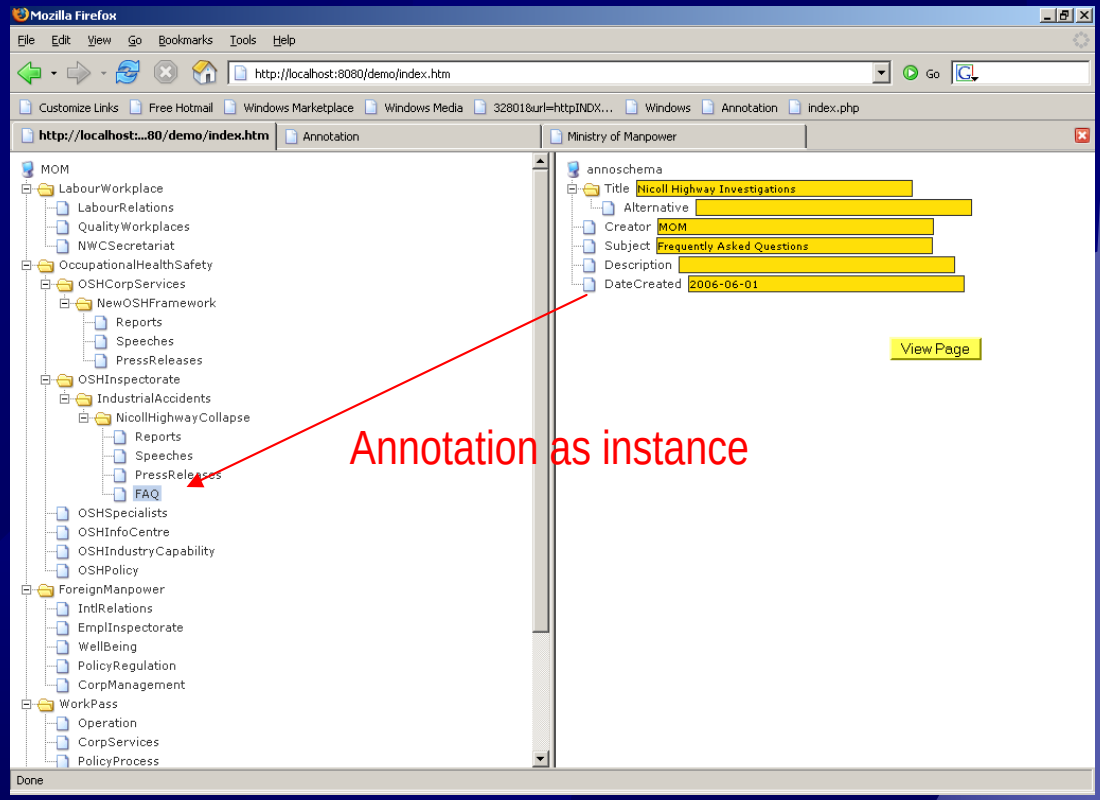


Content Comparison: Before and After

- Content stabilized between 2008 and 2010
- Content that is in common between 2004 and 2006:
  - Minister's Speeches
  - Press Release
- Content that was available 2004, but not 2006:
  - Committee of Inquiries (COI) reports
  - FAQ
- Content that is 2006, but not 2004:
  - Workplace Safety and Health Bills

Annotating Relational Metadata

- The OSH/FAQ metadata can be annotated, as an instance, of the FAQ in the MOM ontology metadata
- The relational metadata is represented using ref field and metadata id.



FAQ of the MOM ontology metadata on the left panel

OSH/FAQ metadata on the right panel

Key Features in WAWI

- Context-aware annotation
- Ontology-aware annotation

- Implemented through the following:
  - Relate semantic content in the metadata to the web content
    - Render agreement, disagreement and different granularities of evidence
    - Provide flexible and precise annotation of the evidence
  - Relate ontology to metadata in a relational metadata

Concluding Remarks

- Completed Web Archives of Singapore (WAS), jointly with National Library Board
- WAWI platform under development
  - Context-aware annotation [Allow Recordlization and Change Detection]
  - Ontology-aware annotation [Allow Structural and Provenancial Organization]
- Embarking in Internet Research on
  - eLearning Recordskeeping System
  - Policy making and communication in eGovernment
  - eSocial Science in general

Thank You

contact: hjwu@ntu.edu.sg