



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

Paul WU Horng-Jyh, PhD  
Nanyang Technological University  
hjwu@[ntu.edu.sg](mailto: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



Web Archive  
Singapore

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

The image shows two overlapping Mozilla Firefox browser windows. The top window, titled 'Nicoll Highway Investigations', is a web archive page from 2004. The bottom window, titled 'Ministry of Manpower', is the official MOM website from 2006. The 2006 website features a modern layout with a search bar, navigation tabs, and a sidebar with a 'Progress Package' advertisement. The 2004 window shows a similar but less developed interface.

**Ministry of Manpower - 2006 Website Content:**

- Home > New OSH Framework and Investigations on Nicoll Highway Collapse**
- New OSH Framework**
- Investigations on Nicoll Highway Collapse**
- Workplace Safety and Health Bill**
- Speeches**
  - 20 July 2005 - Keynote Address by Dr Ng Eng Hen, Minister for Manpower and Second Minister for Defence, at the Presentation Ceremony for the Annual Safety Performance Awards 2005, 20 July 2005, 7.30pm, Suntec Singapore International Convention & Exhibition Centre
  - 28 April 2005 - Keynote address by Dr Ng Eng Hen, Minister for Manpower and Second Minister for Education at the Inaugural National Occupational Safety and Health Week on 28 April 2005, 1000 hours at Novotel Clarke Quay Singapore
  - 10 March 2005 - Committee of Supply: Responses by Minister for Manpower, Dr Ng Eng Hen to Members of Parliament on Workplace Safety and Health
  - 10 March 2005 - Ministerial Statement by Minister for Manpower, Dr Ng Eng Hen - A New Occupational Safety and Health Framework
- Press Releases**
  - 28 April 2005 - Inaugural National Occupational Safety & Health

**Ministry of Manpower - 2004 Website Content:**

- Home > Nicoll Highway Investigations**
- Information For**
  - Employers
  - Employees
  - Job Seekers
  - HR Practitioners
- Information On**
  - Work Permit
  - S Pass
  - Employment Pass
  - Occupational Safety & Health
- Resources**
  - e-Services
  - Forms
  - Procedures/Guidelines
  - Legislation
  - Publications
  - Statistics
  - Events
- Press Releases**
  - 13 September 2004 Government Committee of Inquiry into (C824 Of The Circle Line Highway on 20 April 2004
  - 01 September 2004 Committee of Inquiry into the Cause of the Collapse of the Nicoll Highway on 20 April 2004
  - 29 July 2004 Pre-Inquiry into the Cause of the Collapse of the Nicoll Highway on 20 April 2004

MOM circa 2004

MOM circa 2006



# Web Sites Then – 2008, and now 2010

Singapore Government  
Integrity • Service • Excellence  
Contact Us | Feedback | Sitemap

MINISTRY OF MANPOWER  
Welcome to the **NEW & IMPROVED** MOM website

Search powered by Google

Within MOM Website

- Employer
- Employee
- Work Pass
- Workplace Relations and Standards
- Workplace Safety and Health

FAQs (Frequently Asked Questions)

## Workplace Safety and Health

<p>» Planning for a Safe and Healthy Workplace</p> <ul style="list-style-type: none"> <li>» The Occupational Safety and Health (OSH) Framework</li> <li>» Registration for Workplaces (Licences)</li> <li>» Registration for Pressure Vessels and Lifting Equipment</li> </ul> <p>more...</p>	<p>» Maintaining a Safe and Healthy Workplace</p> <ul style="list-style-type: none"> <li>» Safety and Health Management System</li> <li>» OSH Programmes</li> <li>» Incidents - Reporting, Investigations and Work Injury Compensation</li> </ul> <p>more...</p>	<p>» Building Capabilities</p> <ul style="list-style-type: none"> <li>» Developing Competencies through Training</li> <li>» Best Practices</li> <li>» Managing Workplace Hazards</li> </ul>
<p>» Events, Awards and Incentives</p> <ul style="list-style-type: none"> <li>» Events</li> <li>» Awards</li> <li>» Incentives</li> </ul> <p>more...</p>	<p>» Accredited Service Providers</p> <ul style="list-style-type: none"> <li>» List of Accredited Professional Services</li> <li>» Workplace Hazard Service Providers</li> <li>» Application to Provide Equipment &amp; Services</li> </ul> <p>more...</p>	<p>» Reports and Statistics</p> <ul style="list-style-type: none"> <li>» Annual Reports &amp; Health Reports</li> <li>» Workplace Safety Health Statistics</li> </ul>

MOM circa 2008 from Web Archive

Singapore Government  
Integrity • Service • Excellence  
Contact Us | Feedback | Sitemap

MINISTRY OF MANPOWER  
Virtual Centre for your Manpower Needs

Search powered by Google

Within MOM Website

Home | Services & Forms | Statistics | Publications | Legislation | Press | About Us | Careers at MOM | Help

Employer | Employee | Work Pass | Workplace Relations and Standards | Workplace Safety and Health

FAQs (Frequently Asked Questions)

## Workplace Safety and Health

<p>» Planning for a Safe and Healthy Workplace</p> <ul style="list-style-type: none"> <li>» The Occupational Safety and Health (OSH) Framework</li> <li>» Notification &amp; Registration for Factories</li> <li>» Registration for Pressure Vessels and Lifting Equipment</li> </ul> <p>more...</p>	<p>» Maintaining a Safe and Healthy Workplace</p> <ul style="list-style-type: none"> <li>» Safety and Health Management System</li> <li>» OSH Programmes</li> <li>» Incidents - Reporting, Investigations and Work Injury Compensation</li> </ul> <p>more...</p>	<p>» Building Capabilities</p> <ul style="list-style-type: none"> <li>» Developing Competencies through Training</li> <li>» Best Practices</li> <li>» Managing Workplace Hazards</li> </ul>
<p>» Events, Awards and Incentives</p> <ul style="list-style-type: none"> <li>» Events</li> <li>» Awards</li> <li>» Incentives</li> </ul> <p>more...</p>	<p>» Accredited Service Providers</p> <ul style="list-style-type: none"> <li>» List of Accredited Professional Services</li> <li>» Accredited Training Providers</li> <li>» Application to Provide Equipment &amp; Services</li> </ul> <p>more...</p>	<p>» Reports and Statistics</p> <ul style="list-style-type: none"> <li>» Annual Reports &amp; Work Health Reports</li> <li>» Workplace Safety &amp; Health Statistics</li> </ul>

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

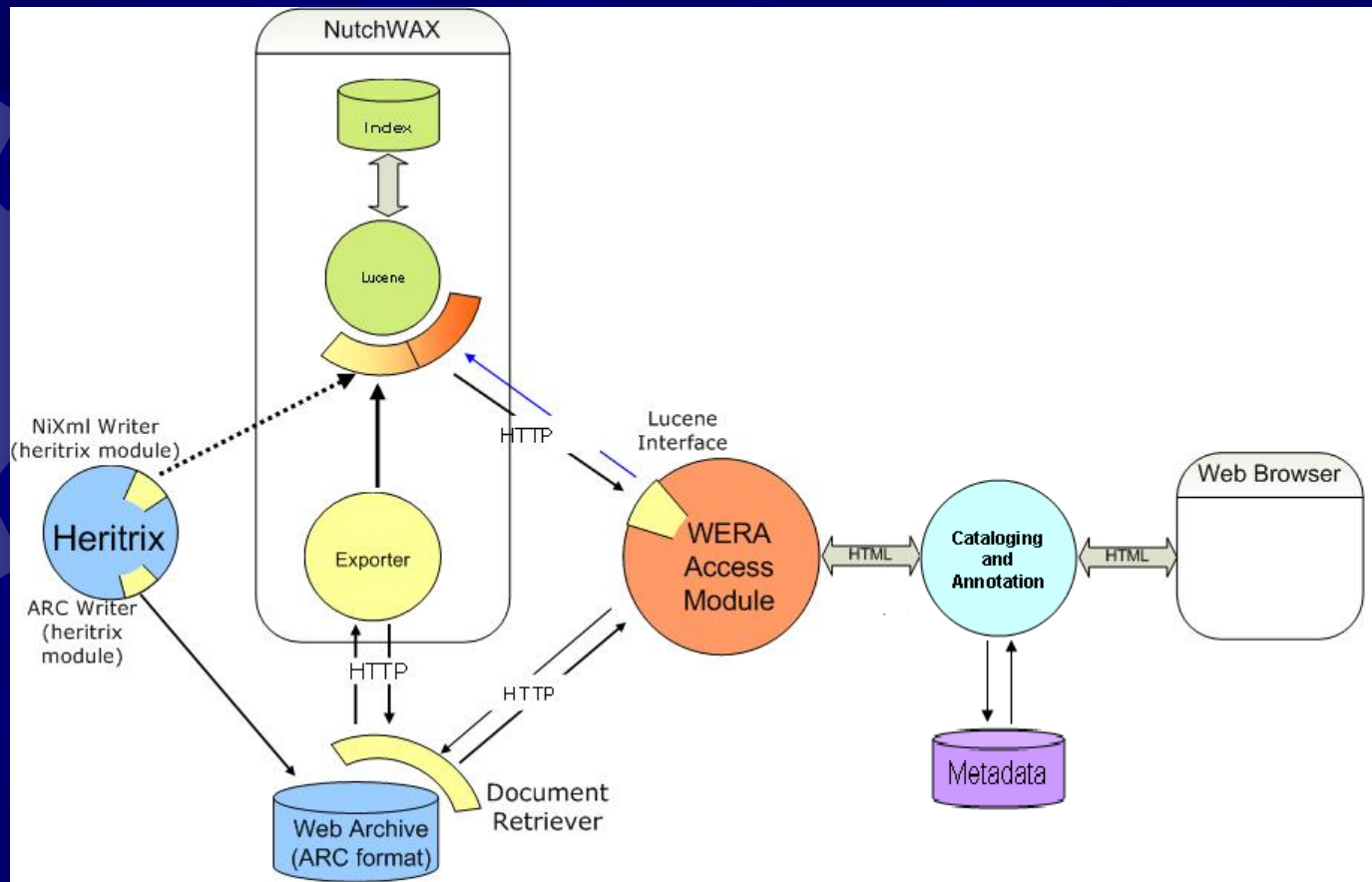
# Accessing Live Web Sites v.s. Web Archives

- 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 communication ensure the activities of MOM divisions are reported timely
  - Minister speech
  - Committee of Inquiry Reports
  - Press releases
  - Hearing of reports
  - FAQ regarding how public and participate
    - The process
    - The public hearing announcements
    - Other questions the public may have



# MOM Organization Chart - An Ontology

## MOM



Corporate Communication

OSH Division

# "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 Daiqi	63171610	hawazi@mom.gov.sg

Senior Management & Their Personal Assistants		
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 screenshot shows a Mozilla Firefox browser window displaying a web page titled "Annotation - Mozilla Firefox". The address bar shows the URL "http://localhost:8080/annotation/viewall.php?i=3". The page content is divided into several sections:

- PORTFOLIO**: A table of subjects and departments. The text "Occupational Safety and Health" is highlighted in yellow and enclosed in a red box. A red arrow points from this box to the text "Annotation as evidence" written in red.
- Subjects**: A list of subjects including Manpower Planning and Policy, Labour-Management Relations, National Wage Guidelines, Occupational Safety and Health, Employment Services, Labour Research and Statistics, etc.
- Departments/Divisions**: A list of divisions including MANPOWER PLANNING AND POLICY DIVISION, LABOUR RELATIONS AND WORKPLACES DIVISION, FOREIGN MANPOWER MANAGEMENT DIVISION, INTERNAL AUDIT UNIT, etc.
- Statutory Boards**: A list of boards including WORK PASS DIVISION, OCCUPATIONAL SAFETY & HEALTH DIVISION, INCOME SECURITY POLICY DEPARTMENT, etc.
- Right Sidebar**: A tree view of the MOM Ontology. The text "Occupational Safety" is highlighted in yellow. Other highlighted terms include "FOREIGN MANPOWER".

At the bottom of the browser window, there is a "NOTICE" bar and a "Done" status bar.

The Context – A highlighted text in a Web Page

The MOM Ontology

# Annotating OSH Records with Metadata

- Records of
  - Minister Speech,
  - Press Release,
  - FAQcan be annotated with evidence and metadata

Nicoll Highway Investigations - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://web.archive.org/web/20041012085523/www.mom.gov.sg/MOM/CDA/0,1858,7494-----,00.html

Customize Links Free Hotmail my del.icio.us My Yahoo! Windows Marketplace Windows Media Windows Yahoo! Bookmarks Yahoo! Mail Yahoo! News Yahoo!

MINISTRY OF MANPOWER

Lifelong Employability Great Workplaces

05 July 2006, Wednesday

About MOM Our Departments Press Room Careers at MOM

Search MOM

Home > Nicoll Highway Investigations

Committee of Inquiry's (COI) Interim Report

- Interim Report of the Committee of Inquiry into the incident at the MRT Circle Line Worksite (C824 Of The Circle Line Project) that led to the collapse of Nicoll Highway on 20 April 2004

Speech

- Statement by Dr Ng Eng Hen, Acting Minister for Manpower, On Safety at Workplaces at Parliament on 19 May 2004
- Opening Statement by Manpower Minister Ng Eng Hen - Government response to interim report of the Committee of Inquiry

Press Releases

- 13 September 2004 Government Response to Interim Report of the Committee of Inquiry into the incident at the MRT Circle Line Worksite (C824 Of The Circle Line Project) that led to the collapse of Nicoll Highway on 20 April 2004
- 01 September 2004 Committee of Inquiry into the Incident at the MRT Circle Line Worksite that led to the Collapse of the Nicoll Highway on 20 April 2004

Frequently Asked Questions

- What is MOM's role in the Nicoll Highway Collapse?
- Who is in this Committee of Enquiry?
- What will this Committee of Inquiry do?
- Has the date and venue of the Inquiry been fixed?
- Is the Inquiry bearing open to the public?

Speech File of OSH

Press Release File of OSH

FAQ File of OSH

MOM Polls

Will you use the WoM Fund to implement flexible work

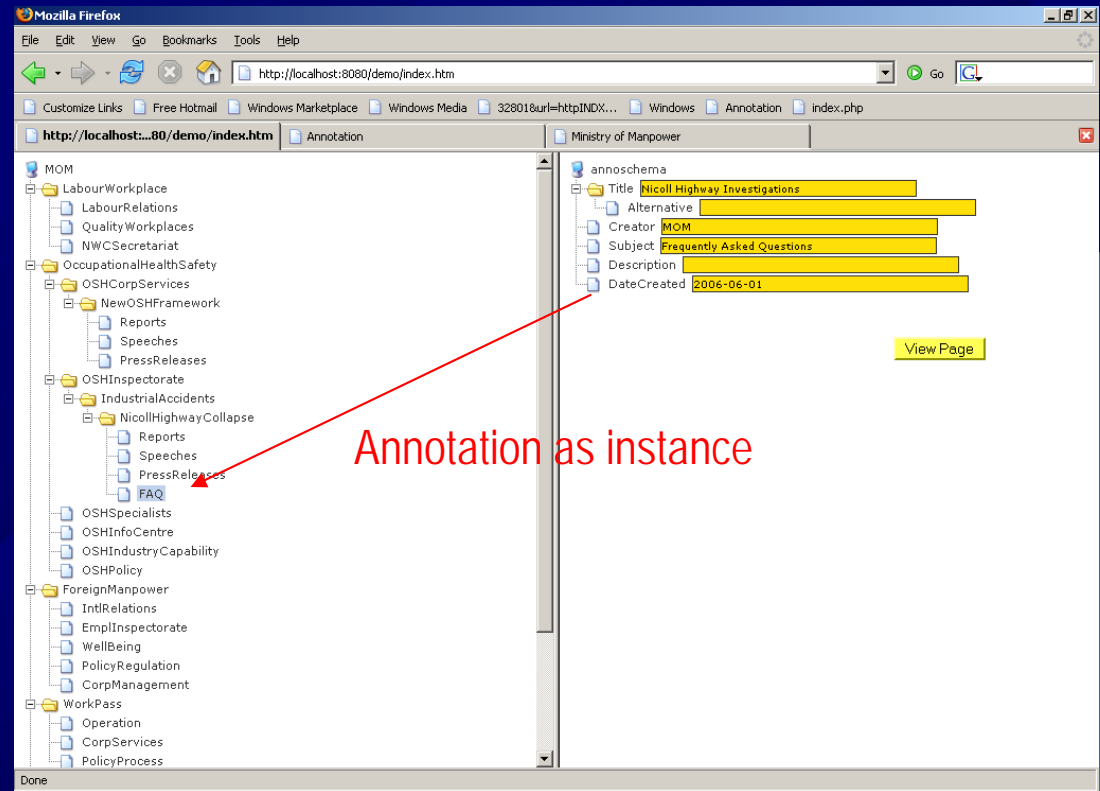
# Content Comparison: Before and After

- 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](mailto:hjwu@ntu.edu.sg)