Goal of WP8

• Test LiWA technology in a real case scenario
  – content-wise: a concentrate of problems in a manageable collection
  – context-wise: an active archiving institution setting
Corpus

• Targetting social web, web 2.0
  – These types of websites are very often rich in content (interactive maps, photos, videos, etc.)
  – content can change at a fast rate (comments, discussions, polls, etc.).
Corpus

• Initially we envisaged natural disasters and their web response.
  – will depend on what comes up
• National Library of Czech republic will create the thematic collection (Year 3)
Partners

• NLP and MZK have a long experience in archiving the Web (since 2000)
  – They are currently undergoing a restructuring of their workflow interface and are willing to integrate LiWA modules
  – They do both selective and holistic archiving (Czech domain)
  – They are members of IIPC and apply mainstream approaches and tools

• Hanzo has a long experience in archiving social web (hanzoweb.com)
Towards a professional UI framework

• What we found is:
  – Cross-project need for interfaces, on underlying process (crawl, content assessment, optimization requiring user feedback)
    • WP2 adversarial gathering management module
    • WP3 Spam assessment interface
    • WP5 Semantic browser
  – Several partners are working on this currently (EA, NLP, MZK, HZ)

• Mutualize and expand effort in this domain
Mock-up

- Goal: integrate LiWA in an archiving institution real workflow (NLP WA Admin)
  - define appropriate steps where LiWA modules can be used
  - describe interaction with professionals
  - show potential benefit for professionals in the testing institution
The initial page of NLP workflow management tool WA Admin

Quick search
Open tasks for curator
Step # 1 - Curator nominates a website:

- Curator enters basic “metadata” (title, publisher, URL) into the system
Step # 1 - Curator nominates a website (cont.):

- Curator can add more “metadata” or edit a website entry later at any stage during the workflow
Step # 2 - Nominated sites are appraised:

- WA Admin lists all newly nominated sites awaiting appraisal by curators
- Curators appraise each of them
Step # 3 - Contacting publishers:

- WA Admin lists all approved websites
- Publishers are sent requests for permission to harvest
Step # 4 – Obtaining a permission:

- WA Admin lists all websites under negotiation with publishers
- Curator can add permission to a website or mark it as “rejected” or “no response”
Step # 5 – Cataloguing:

- WA Admin displays list of all websites that need to be catalogued
Step # 6 – Selective Harvesting:

Websites are passed to the crawl engineer for scheduled harvesting. He pulls a list of websites to be harvested from WA Admin and runs the crawls using Heritrix (via standard web-based UI).
Step # 7 – Quality assurance:

- After a harvest, all websites are checked for potential problems
- WA Admin displays a list of all websites that need QA
Step # 7 – Quality assurance:

- Wayback displays a list of all harvested instances of the website
Step # 7 – Quality assurance:

- Curator navigates within the website in a browser-like fashion
- He can switch between different harvested instances using a timeline at the top of the screen
Step # 7 – Quality assurance

Checking previous instances with liwa temporal difference analyser. Breach in continuity can be due to the lack of completeness but also to radical change in the site technology, which are also good indicators of potential crawling problems. This helps focus QA for repetitive crawls, for instance limiting QA time in case no significant change is identified here.

click on graph to directly access version of site
Step # 7 – Quality assurance

Finding temporal incoherencies with LiWA coherence graph. This graph indicates where in the site some temporal incoherence patterns can be found, and allows the QA operator to click on nodes to see the problematic sections of the site.
Step # 7 – Quality assurance (cont.) – WP2 advanced link extraction:

- For sections which weren’t harvested due to Javascript, etc. Curator clicks a button that runs the virtual browser that generates a list of links on the page; missing links are added to the seed list for a re-crawl or next harvest.

Control screen

List of links discovered

Challenging site to archive
Step # 7 – Quality assurance:

- A new ticket is open for every QA issue in Trac
- It is then passed onto the crawl engineer who looks into it and tries to resolve the issue (see next screen)
**Ticket #186 (new defect)**

![Trac interface](http://icestina.upol.cz/)

**Reported by:** gruber  
**Assigned to:** brokes

**Priority:** minor  
**Milestone:**

**Component:** QA  
**Version:**

**Severity:** severe  
**Keywords:**

**Cc:** krach

**Description:**

Odkazy z levého menu směřují pouze na živý web...

**Attachments**

- [Attach File]

**Change History**

- 10/15/2008 09:37:55 PM changed by brokes  
  - **cc set to krach**
  - component changed from design to QA

Luději jsem cvičnou skřízen a jsem schopen web sledit v plném rozsahu. Odkazy v levém menu (které jsou tvořena flashem) nefungují ve vyjádření, ten není schopen modifikovat urážky flash součtu a ty odkazy mohou tam napsat návod.  

Lukasi, asi existuje rešení?

- 12/07/2008 11:21:48 PM changed by trach  
  - Nevím o nen, pokud vím, tak flash skřížit neumíme a ani ho neumíme přehradit.

**AddChange #186**

**Comment** (you may use **WikiFormatting** here):

- [WikiFormatting]
Step # 7 – Quality assurance (cont.) – WP3 spam detection:

- QA is the best suited part of the workflow for the learning stage of the spam detection engine – whenever a curator finds spam or other irrelevant content during QA, he can use a widget to mark it as such and document the relevant feature set for enhancing detection accuracy.
Step # 8 – Semantic search:

- Users are able to make semantic search that expands query in time
- Analysis of terminology evolution over time