European Web Archive

An Initiative for an FP6 Integrated Project

Andreas Rauber

Dept. of Software Technology
Vienna University of Technology
A-1040 Vienna, Austria

http://www.ifs.tuwien.ac.at/ifs
Overview

- why an EoI? why an Integrated Project?
- 2 views of the project
  - a vertical view (applications)
  - a horizontal view (tasks)
- goals to reach out for
- a few words on the consortium
- next steps
Motivation

- increasing need for and interest in web archiving systems
- several rather isolated projects
- networking, sharing information and experience -> NoE
- development and improvement of new and existing modules -> IP
A Vertical View of EWA - Applications

- archive of the web as digital cultural heritage
- special collections / vertical portals on scientific publications, arts, politics, events,...
- technology evolution and saturation, market predictions
- language evolution, influence of new media
- web sociology: communities on the web
- web design: periods and their characteristics
- and many more...
A Horizontal View of EWA - Tasks (1/3)

(1) acquisition of information:

- source selection:
  manual, open, semi-automatic

- data collection:
  deposit, harvesting, focused crawls, acquisition of dynamic objects, session-filming, ...
A Horizontal View of EWA - Tasks (2/3)

(2) archive organization and preservation:
- scalable storage techniques: 
  *information GRIDS, HDD arrays, tape robots,...*
- maintenance: 
  *meta-data generation, coding, storage, exchange, maintenance,...*
- migration and refresh: 
  *storage media migration*
- access preservation: 
  *system emulation, format conversion*
(3) archive usage, interoperability, mining:

- information retrieval:
  
  *indexing, search engine, audio & image retrieval,...*

- information access:
  
  *new interfaces, interoperability, thematic portals, semantic web, ontologies*

- information mining:
  
  *topic detection and tracking, language evolution, ...*

- technological development:
  
  *technologies employed, market shares, ...*

- society analysis:
  
  *communities on the web - graph, digital divide*
Goals

- archival system: 
  *tools and components*

- best practice reports and standards: 
  *guides, advice, framework, lessons-learned*

- training material: 
  *workshops, tutorials, cooperation infrastructure*

- state-of-the-art research: 
  *information organization, preservation, knowledge mining, visualization,...*

- european web archive: 
  *a distributed, collaborative archive*
Consortium

- 12 national, state and university libraries
- 10 universities and research centers
- 5 companies
- EoI consortium, but NOT project consortium:
- need additional partners, additional expertise,...
Next Steps

- wait for results of EoI exercise -> mid-September
- digicult meeting in Luxembourg in November
- -> project proposal?
- -> ?
questions?