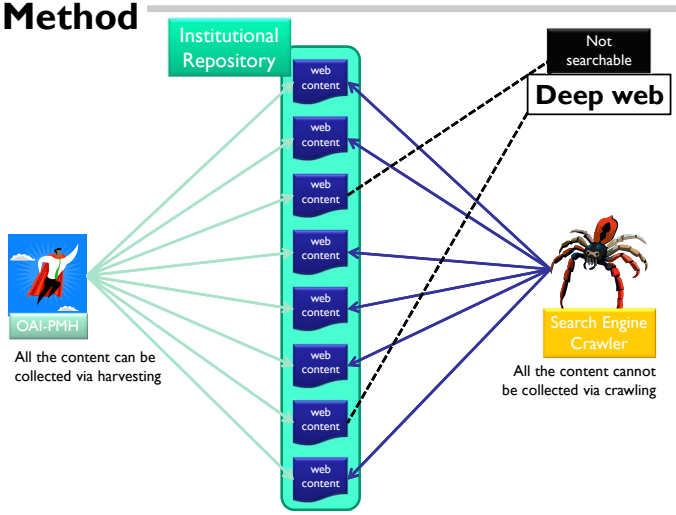
The Deep Web in Institutional Repositories in Japan

Teru Agata (Asia University) Yosuke Miyata (Keio University) Atsushi Ikeuchi (Tsukuba University) Shuichi Ueda (Keio University)

Purpose

- Investigating the current status of the deep web is important for both the general public and researchers.
- We calculate the extent of the deep web based on the content of searchable IRs in Japan, using a more appropriate interval and exhaustive search with three major search engines (Google, Yahoo!, and Bing).

Method



IR

- Harvest metadata from 92 IRs in Japan on April 11, 2009
  ⇒404,431 URLs

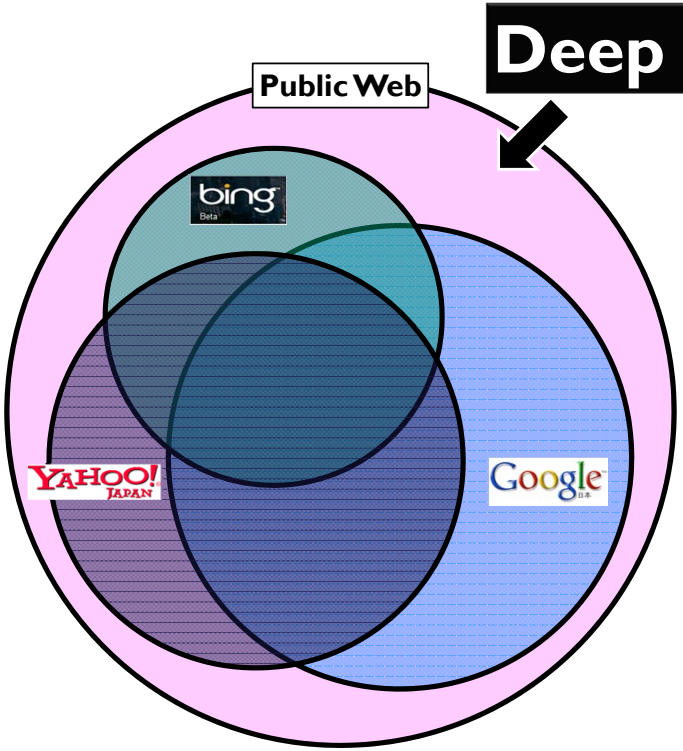
SE

- Search 404,431 URLs as queries between September 6 and September 8, 2009

Search Engine Coverage

N=404,431

	Google	Yahoo! Japan	bing Beta	Google + Yahoo! Japan + bing Beta
No. of searchable URLs	215,259	174,805	115,679	291,024
Coverage Rate	53.2%	43.2%	28.6%	72.0%



Coverage and Overlaps of Search Engines

$$\text{coverage rate} = \frac{\text{No. of searchable URLs}}{\text{No. of investigated URLs}}$$

Overlap of Search Engines

	Google	Yahoo! Japan	bing Beta
Google		54.2%	37.9%
Yahoo! Japan	66.7%		39.1%
bing Beta	70.5%	59.2%	

Conclusion

- The deep web is assumed to be roughly 30%.
  - This study investigated the current status of the deep web using a full-text URL collection harvested from 92 IRs in Japan.
  - The coverage rate was highest with Google at 50%, and when used in conjunction with two other major search engines, coverage increased to about 70%.

ASIST 2010, October 22-27, Pittsburgh, Pennsylvania, USA