Indexing the Deep Web

Matt Eckert
What is the Deep Web?
Why do we care?

- Estimated to be 400 times the size of surface web
- Wealth of information over a wide range of categories
Finding the hidden data

- 2005 Google – Sitemaps
- DARPA/Google Memex
- Dark web specific search engines
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <loc>http://www.example.com/</loc>
    <lastmod>2005-01-01</lastmod>
    <changefreq>monthly</changefreq>
    <priority>0.8</priority>
  </url>
</urlset>
Finding the hidden data

- 2005 Google – Sitemaps
- DARPA/Google Memex
- Dark web specific search engines
Common Approaches

- Domain specific searches
- Registration of domain
- Machine Learning to mimic human behavior
- Educating website developers how to develop a site easily indexed
- Sample keyword search
Future Work

- Flexible web crawling solutions
- Scalable solution
- Query time searches
Questions?