APEX 数据和知识管理实验室

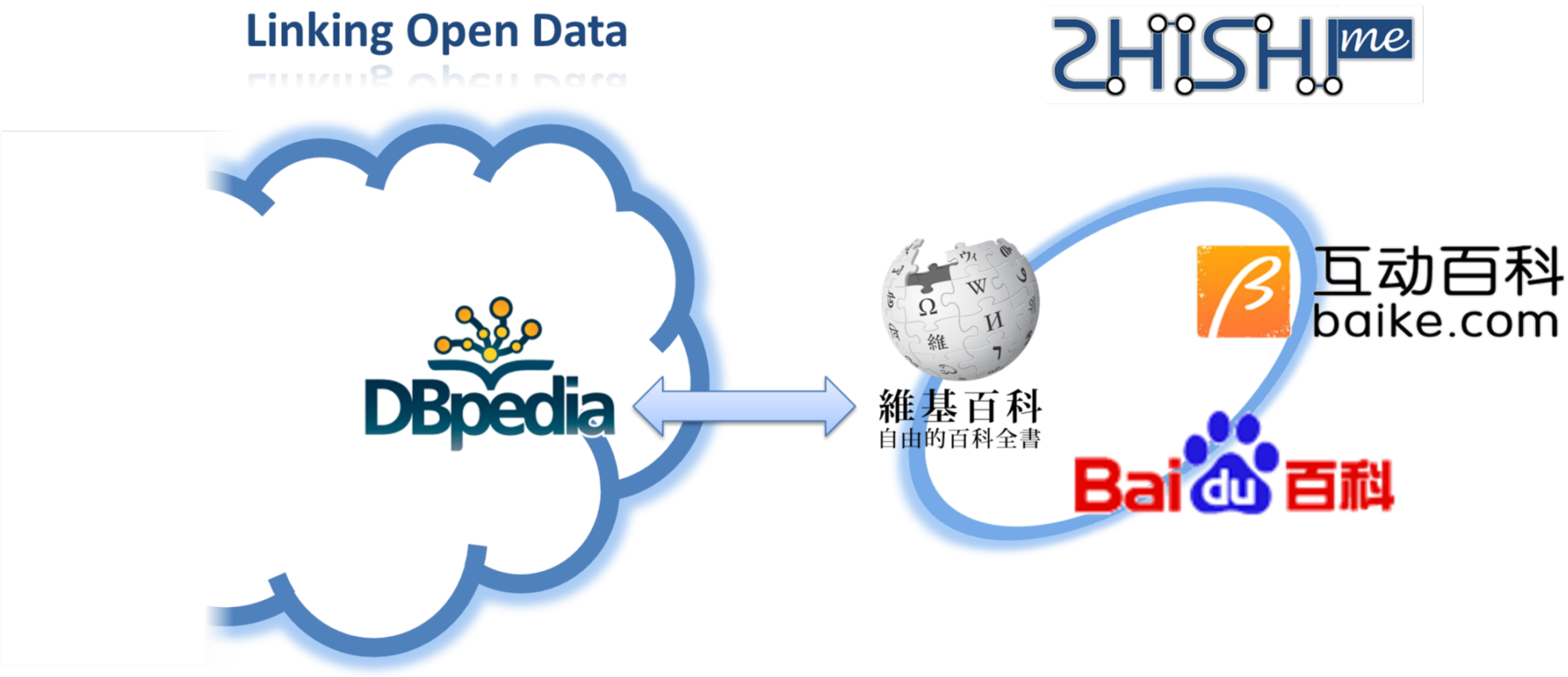
DATA & KNOWLEDGE MANAGEMENT LAB

The state of the art of Chinese LOD development

Haofen Wang

Overview of Zhishi.me

- Zhishi.me (**http://zhishi.me**) is the first effort to publish large scale Chinese semantic data and link them together as a Chinese Linking Open Data (CLOD). The statistics are collected by Feb, 2013
  - Over 8 million distinct instances
  - Over 1 billion RDF triples



Multilingual Issues

- There are no one-size-fit-all mechanisms for providing resource identifiers.
- Using IRIs?
  - Chinese characters are non-ASCII, un-encoded words are reader-friendly.
  - XML specification: encoded URIs are not allowed to act as XML properties.
- Using URIs?
  - IRIs are incompatible with HTML 4, non-ASCII characters should be encoded with the URI escaping mechanism to generate legal URIs as “href” values.
  - Most Web browsers automatically encode IRIs which users entered into the address bar. Servers always receive URIs.
- A compromise
  - Storing IRIs in databases and transforming received URIs into IRIs. Trying our best to provide IRI embedded resource files when response users’ data request.

Future Work

- Assigning uniform resource identifiers for matched instances.
  - Naming resources properly
  - Uniting traditional Chinese names and simplified Chinese names
- Developing multilingual instance matching algorithms to discover more links between Zhishi.me and LOD.
- Integrating the e-commerce Web sites (360Buy and Taobao) and social Web sites (e.g., Weibo, Dianping) as the first effort of Chinese Linked Open Stream Data
- Extract ontologies from CLOD and linked with well-known thesaurus and taxonomies like schema.org
- Providing more APIs (like entity linking, complex relation finding and allowing users to upload their own data and return the links with CLOD for federated querying purpose)

Thanks!