

Web Archives of devastated area sites
&
De-Duplication Project
&
NutchWAX Multilingualization

IIPC General Assembly (May 11th 2011)

Masaki Shibata
National Diet Library, Japan
m-shiba@ndl.go.jp

Japanese Quake and Tsunami on March 11th 2011

- ◆ Thank you for many condolences.
- ◆ Unfortunately we have to cancel IIPC WG with iPRES

Web archives of devastated area sites

- ◆ We are archiving devastated area official sector sites on the law since March 14th.
 - ◆ During March
 - #Dairy:local government
 - #Weekly or Biweekly:government
 - ◆ After April
 - #Basically Weekly:all sites

- 被災地域ウェブサイトの保存
- 海外機関との協力

[被災地域ウェブサイトの保存]

▶ 国立国会図書館は、東日本大震災に関する記録を後世に伝えるため、被災地域の自治体のウェブサイトを重点的に収集しています。ここでは一部のみをご紹介しますが、今後も準備が整い次第、順次公開していく予定です。

画像をクリックすると、収集した時点でのウェブサイトがご覧になれます。掲載されている情報は、収集時点のものであり、現在とは異なる場合がありますのでご注意ください。

岩手県



収集日: 2011/03/14

久慈市 (岩手県)



収集日: 2011/03/22

石巻市 (宮城県)



収集日: 2011/03/24

多賀城市 (宮城県)



収集日: 2011/03/20

仙台市 (宮城県)

南相馬市 (福島県)

福島市 (福島県)

水戸市 (茨城県)

Collaboration

- ◆ Internet Archive

We provide especially private sector sites URL concerning the disaster.

IA has archived and provide open access of this collection through Archive-it.

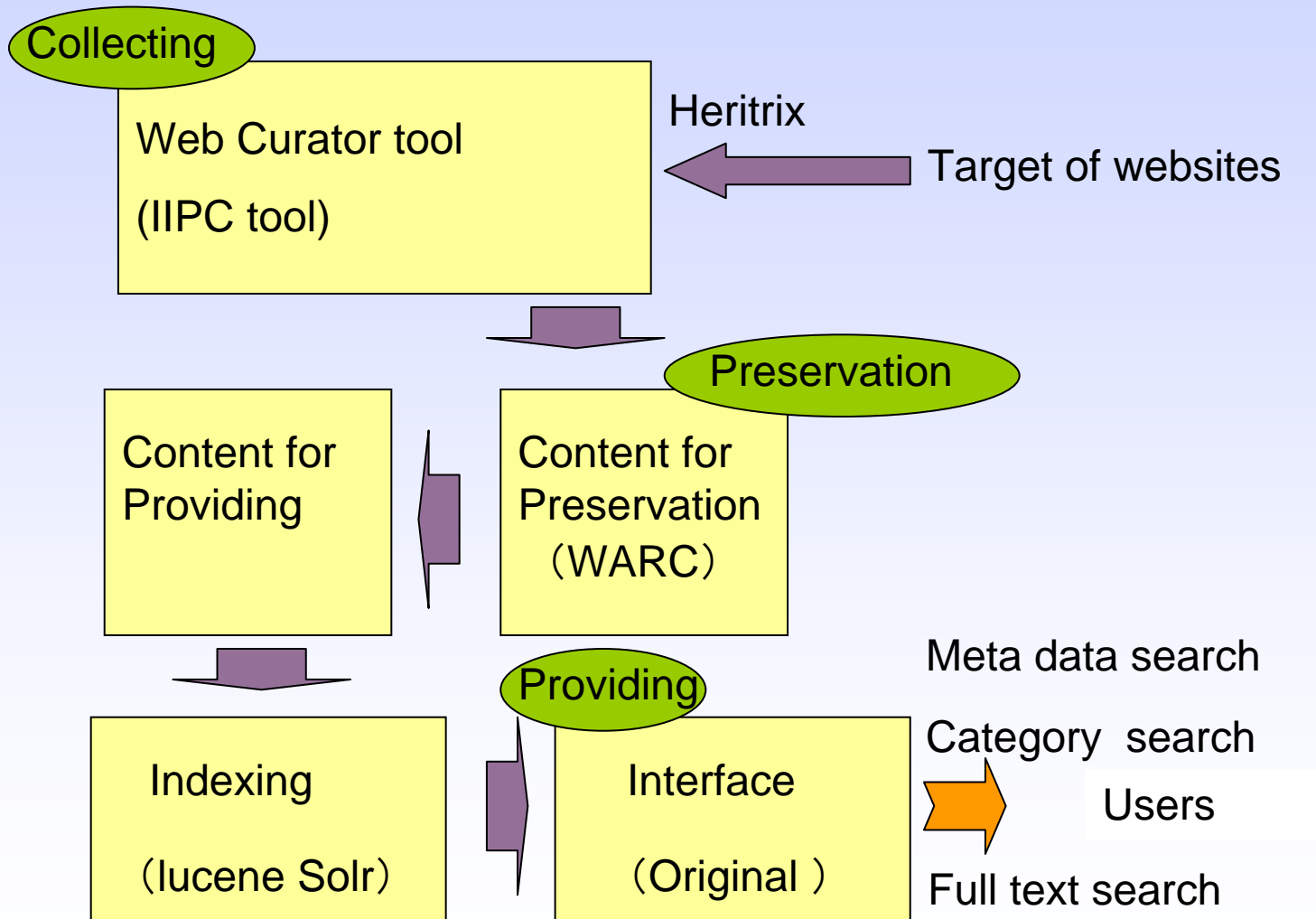
- ◆ Harvard University Reischauer Institute of Japanese Studies

We are exchanging opinion and information for their project “Digital archive of the Japan2011 Earthquake and Aftermath”

De-Duplication Project

- ◆ We has begun archiving official sector sites on the base of the law since April,2010.
 - ◆ #Monthly: National Government
 - ◆ #Quarterly: Local government,
Universities
- ◆ We have archived about 55TB in a year.

Web archiving system



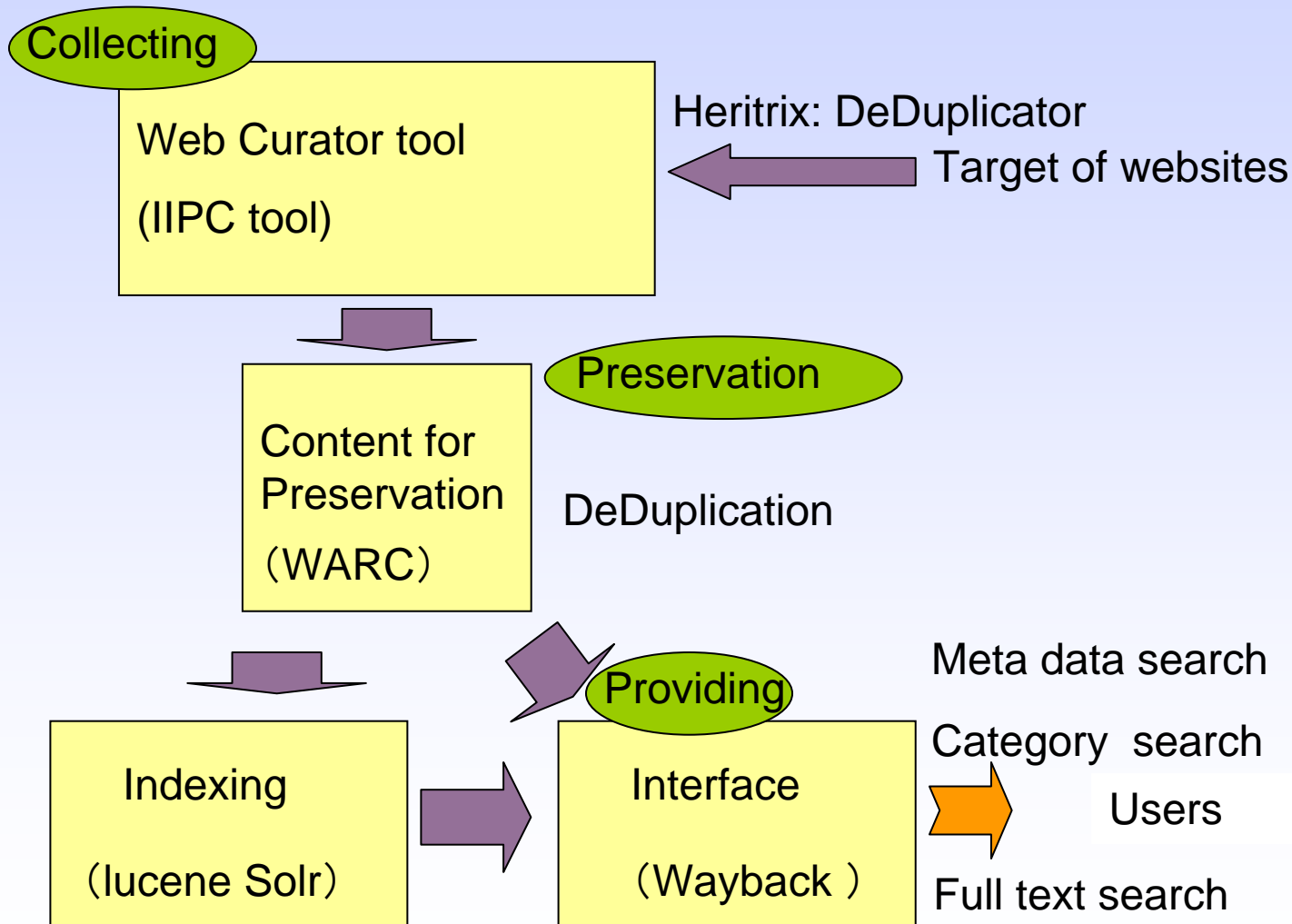
De-Duplication Project

- ◆ Budget is difficult to grow up
But the size of archives is easy to grow up
- ◆ We have to figure out how to handle the growth of archives.
- ◆ We had studied de-duplication in 2 years.

De-Duplication Project

- ◆ Finally we have come to the conclusion we has to implement De-Duplication on our system.
- ◆ We plan to develop some functions from 2011 to 2012 and open de-duplication system in 2013.
- ◆ The rate of redaction(estimation)
 - # Monthly archives:80%
 - #Quarterly archives:45%

Next Web archiving system (2013 plan)



De-Duplication Project

- ◆ For details, please read our uploading document on the netpreserve.org

<http://netpreserve.org/forum/viewtopic.php?f=62&t=513>

NutchWAX Multilingualization

a progress report

Masayuki ASAHARA

Nara Institute of Science and Technology,
Japan

National Diet Library, Japan
masayu-a -at- is.naist.jp

Masaki Shibata

National Diet Library, Japan
m-shiba@ndl.go.jp

Objectives

- ◆ Multi-lingualization of NutchWAX
 - ◆ An extension of a search engine software Nutch
 - ◆ Nutch cannot handle Asian languages (incl. CJK)
 - ◆ Indexing
 - ◆ Caching
 - ◆ Handling UTF-8 (conversion from/to CJK encodings)
 - ◆ Multi-lingualization of Nutch is requisite for NutchWAX

Plan in 2009

The status of (sky)blue parts are completed.

The pink (or red) parts are cancelled due to FOSS based constraints on May 2010.

The status of yellow parts are not completed.

Deliverable ref. & title	Description	Delivery date	Lead
Task 1: WP management			
Objective: Coordinate progress and successful completion of WP			
Task1-1	Work plan	2009/11/02	M. ASAHARA
Task1-2	Quarterly progress reports	¼ of the year	M. ASAHARA
Task 2: Nutch-1.0 for Korean and Chinese			
Objective: Test Nutch-1.0 with KoreanAnalyzer and ChineseAnalyzer			
Task2-1	Nutch-1.0 with Simplified Chinese	2010/01/31	M. ASAHARA
Task2-2	Nutch-1.0 with Korean	2010/04/30	M. ASAHARA
Task2-3	Nutch-1.0 with Traditional Chinese	2010/07/31	M. ASAHARA
Task 3: Nutch-1.0 for Other Asian Languages			
Objective: Test Nutch-1.0 on Other Asian Languages			
Task3-1	Nutch-1.0 with Urdu	2011/01/31	M. ASAHARA
Task3-2	Nutch-1.0 with Other Asian Languages	2011/07/31	M. ASAHARA
Task 4: Multilingualization of NutchWAX			
Objective: System Integration Test of NutchWAX and Multilingualized Nutch-1.0			
Task4-1	NutchWAX with Japanese	2010/7/31	M. ASAHARA
Task4-2	NutchWAX with Chinese (Simplified/Traditional)	2010/10/31	M. ASAHARA
Task4-3	NutchWAX with Korean	2011/1/31	M. ASAHARA
Task4-4	NutchWAX with Urdu	2012/07/31	M. ASAHARA

Progress so far achieved

- ◆ Nutch-1.0 for CJK languages
 - ◆ for Japanese completed with FOSS (-Oct. 2009)
 - ◆ for simplified Chinese completed with FOSS (-Feb. 2010)
 - ◆ for Korean completed with FOSS (-Apr. 2010)
- ◆ NutchWAX for Japanese
 - ◆ completed with FOSS (-Aug. 2010)
- ◆ Language Identification Issue is solved
 - ◆ LanguageIdentifier for 49 languages by Mr. Nakatani
 - ◆ ... actually, this is not our contribution... (-Jan. 2011)
- ◆ NutchWAX for Chinese/Korean
 - ◆ beta release (-May. 2011)

1. LanguageIdentifier

- ◆ Developed by Mr. Shuyo Nakatani
 - ◆ His blog (in English):
<http://shuyo.wordpress.com/2011/01/13/language-detection-plugin-for-apache-nutch/>
 - ◆ code:
<http://code.google.com/u/nakatani.shuyo/>
 - ◆ cover 49 languages:
<http://code.google.com/p/language-detection/wiki/LanguageList>

2. NutchWAX for Chinese/Korean

- ◆ Similar to NutchWAX for Japanese
 - ◆ Bug fix for UTF-8 handling
 - ◆ Incorporate FOSS-based word segmenter
 - ◆ For Chinese, Paoding Chinese Analyzer
 - ◆ For Korean, kspin or LuceneKorean
- ◆ We made beta release (contact masayua@is.naist.jp)
 - ◆ Tester wanted
 - ◆ Crawler test (hosting Chinese/Korean sites)
 - ◆ UI test (Native Chinese or Korean)

Summary

- ◆ LanguageIdentifier for 49 languages by Mr. Nakatani
- ◆ NutchWAX for Chinese and Korean – beta release

Future work

- ◆ NutchWAX for Chinese/Korean
 - ◆ We will make a final release with documentation
- ◆ Next issue:
 - ◆ NutchWAX with Solr for CJK

Thank you for Attention

URLs

- ◆ Nutch Japanization
 - ◆ <http://sites.google.com/site/masayua/m/nutch/nutch-japanization>
- ◆ Nutch Chinezation
 - ◆ <http://sites.google.com/site/masayua/m/nutch/nutch-chinezation>
- ◆ Nutch Koreanization
 - ◆ <http://sites.google.com/site/masayua/m/nutch/nutch-koreanization>
- ◆ NutchWAX Japanization
 - ◆ <https://sites.google.com/site/masayua/m/nutch/nutchwax/nutchwax-0129-ja2>