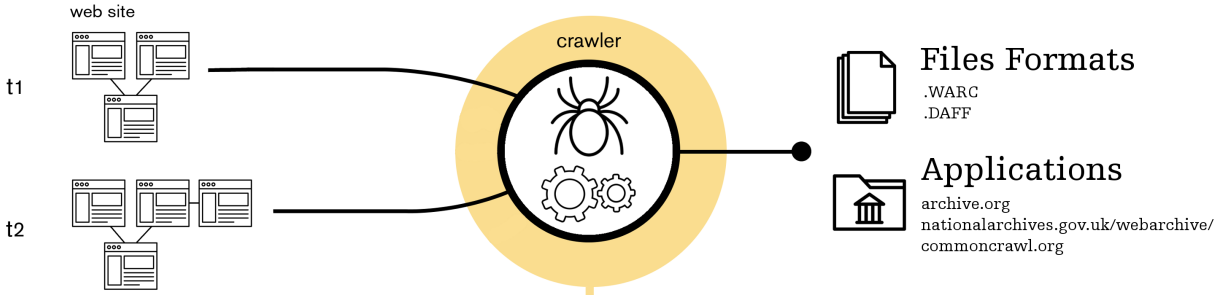
Mining and Searching web archives

Quentin Lobbé, Télécom ParisTech, Université Paris Saclay

Web Archives

"The digital heritage consists of unique resources of human knowledge and expression (...) Digital materials include texts, databases, still and moving images, audio, graphics, software and web pages, among a wide and growing range of formats (...) The world's digital heritage is at risk of being lost to posterity." | Charter on the Preservation of the Digital Heritage, Unesco 2003



1 Automatic Archives Enrichment

Content Extraction

Based on CMS rules & ad hoc assumptions
Going under the page level by mining links, users, comments ...

Focused Object Retrieval by Exploiting Significant Tag Paths, Oita & Senellart, 2015

Events Detection

Archives based model through topic modeling variations
External events correlation by text similarity

A Computational Approach to Understanding Historical Events Using State Department Cables, Allison J.B. Chaney, Hanna Wallach, David M. Blei, 2015

Quality and Consistency

Defining archive quality based on requests
Measuring the probability of archives inconsistency due to unregistered evolutions

Data Quality in Web Archiving, Spaniol & Senellart, 2009

We use the e-diasporas corpus (10000 sites) archived by INA between 2010 and 2015 (70 TO of archives)

Web Sites Atlas

Index of Web Pages

Querying

building a new formalism for querying archives through time, link, page, site, user, keyword ...

Temporal connectives versus explicit timestamps to query temporal databases, Abiteboul, 1995

Visualization

Visualize the exploration of data for an innovative restitution of scientific

Extracting Evolution of Web Communities from a Series of Web Archives, Toyoda & Kitsuregawa, 2006

Methodology

Establishing a socio-technological methodology as a starting point for web archeology studies

E-diasporas Atlas, Diminescu & Jacomy 2012

2 Time Travel Search Engine

Driven by diasporas and migration issues

This upcoming work is an evolution of the e-diasporas project, as he propose an inovative way to understand web interractions and connexions of many migrante populations through time comparisons.

Spark Solr d3.js

quentin.lobbe@gmail.com