Universal Distant Reading through Metadata Proxies with ArchiveSpark

Helge Holzmann (L3S), Vinay Goel (IA) and Emily Novak Gustainis (MHL)

IEEE BigData 2017, Boston, MA, USA

https://github.com/helgeho/ArchiveSpark
ArchiveSpark

- **Expressive** and efficient Web archives data access / processing
- Joint work with the Internet Archive
- **Open source**
  - Fork us on GitHub: [https://github.com/helgeho/ArchiveSpark](https://github.com/helgeho/ArchiveSpark)
  - Star, contribute, fix, spread, **get involved**!
  - **Modular**, easily **extensible**
- **More details in:**
Core Concepts

- Efficient seamless **two step loading** approach:
  - **Filter** as much as possible on **metadata** before touching the archive
  - **Enrich** records with data from payload **instead of mapping / transforming**
    - enrich functions, dependent on each other, document lineage
Benchmarks

• Three scenarios, from basic to more sophisticated:
  a) Select one particular URL
  b) Select all pages (MIME type text/html) under a specific domain
  c) Select the latest successful capture (HTTP status 200) in a specific month

• Benchmarks do not include derivations
  • Those are applied on top of all three methods and involve third-party libraries
Generalization

• Generalize the concept to any metadata / data type and source
  • Allow for remote sources / databases / search queries
  • Abstracted away by Data Specifications

• Common enrichment interface, functions to be used across sources
Recap: Original Objectives

1. A simple and expressive interface
   • powered by Spark / Scala with a declarative enrichment API

2. Compliance to and reuse of standard formats
   • CDX / (W)ARC for Web archives, now any metadata + data

3. An efficient selection and filtering process

4. An easily extensible architecture

5. Lineage support to comprehend and reconstruct derivations

6. Output in a standard, readable and reusable format
   • all carried over to the new generic version
Example Study (Polio / MHL)

• Medical Heritage Library (MHL)
  • “MHL is a digital curation collaborative effort among some of the world’s leading medical libraries. It promotes free and open access to quality historical resources in medicine.”

• Study Polio symptoms mentioned in medical journals
  • Using MHL’s full-text search system as an entry point into their archive

---

**Source - Records Accessed - # Here**

<table>
<thead>
<tr>
<th>Source</th>
<th>Records Accessed</th>
<th># Here</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw data</td>
<td>all</td>
<td>&gt; 170,000</td>
</tr>
<tr>
<td>Meta files</td>
<td>collection</td>
<td>3,148</td>
</tr>
<tr>
<td>Full-text</td>
<td>matches</td>
<td>2,027</td>
</tr>
</tbody>
</table>
Interoperability

- High reusability of code among similar jobs with different data
- On average, external added complexity only 4.8% of internal instructions

<table>
<thead>
<tr>
<th>intern. / extern.</th>
<th>MHL Polio</th>
<th>MHL E. Search</th>
<th>MHL E. Local</th>
<th>MHL E. Remote</th>
<th>Web E. Local</th>
<th>Web E. Wayb.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MHL Smoke</td>
<td>67 / 14 (21%)</td>
<td>136 / 14 (10%)</td>
<td>195 / 16 (8%)</td>
<td>182 / 16 (9%)</td>
<td>225 / 17 (8%)</td>
<td>194 / 17 (9%)</td>
</tr>
<tr>
<td>MHL Polio</td>
<td>71 / 3 (4%)</td>
<td>217 / 7 (3%)</td>
<td>199 / 7 (4%)</td>
<td>195 / 7 (4%)</td>
<td>206 / 7 (3%)</td>
<td></td>
</tr>
<tr>
<td>MHL E. Search</td>
<td></td>
<td>247 / 4 (2%)</td>
<td>226 / 4 (2%)</td>
<td>214 / 5 (2%)</td>
<td>224 / 5 (2%)</td>
<td></td>
</tr>
<tr>
<td>MHL E. Local</td>
<td></td>
<td></td>
<td>24 / 0 (0%)</td>
<td>158 / 3 (2%)</td>
<td>129 / 3 (2%)</td>
<td></td>
</tr>
<tr>
<td>MHL E. Remote</td>
<td></td>
<td></td>
<td></td>
<td>148 / 3 (2%)</td>
<td>122 / 3 (2%)</td>
<td></td>
</tr>
<tr>
<td>Web E. Local</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>87 / 1 (1%)</td>
</tr>
</tbody>
</table>

- Large parts extracted away from the user by DataSpecs / EnrichFuncs
  - To be shared as modules, entire workflows to be shared as recipes
Thank you!

https://github.com/helgeho/ArchiveSpark
https://github.com/helgeho/MHLonArchiveSpark
http://www.medicalheritage.org
http://alexandria-project.eu