Building Curated Collections:
A Brief Illustrated Guide

Abstract – A brief illustrated guide for building curated collections with Minerazzi.

Published: 04-13-2016; Updated: 10-06-2016

Keywords: minerazzi, curated collections, recrawls, fqu bot

Introduction

A three-step solution for creating curated collections with Minerazzi is given below.

Step 1. Start with an initial URL found through Minerazzi.

Step 2. Click the icon to discover new linked URLs so you are doing a recrawl.

Step 3. Copy/Paste URLs as you would or by clicking the top-right **{S}** of a results page.

To build your own collection, repeat steps on a result as many times as you need to. If necessary, Minerazzi will update its own collection by indexing recrawled URLs.

Visualization

A visualization is provided in Figure 1.

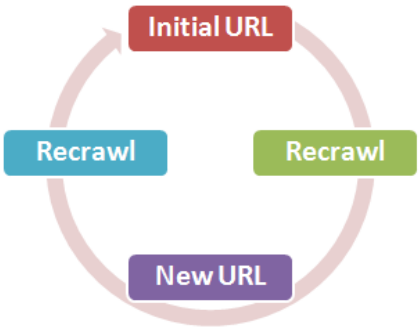

Figure 1. Recrawling cycle of a miner.

Example: The Panama Papers

The Panama Papers mega collection is, at the time of writing, the largest data leak of deception and corruption documented at https://panamapapers.icij.org

In that sea of information, building subcollections driven by user's judgments is a time consuming process. Minerazzi helps users to easily do this as illustrated with the following example.

On 4-13-2016, the query [icij] in The Panama Papers miner (http://www.minerazzi.com/pp) found the mentioned link (Minerazzi, 2016a). See Figure 2.

Figure 2. Panama Papers record.

Clicking the icon returns the internal and external URLs linked from the initial one. Partial results are given in Figure 3.

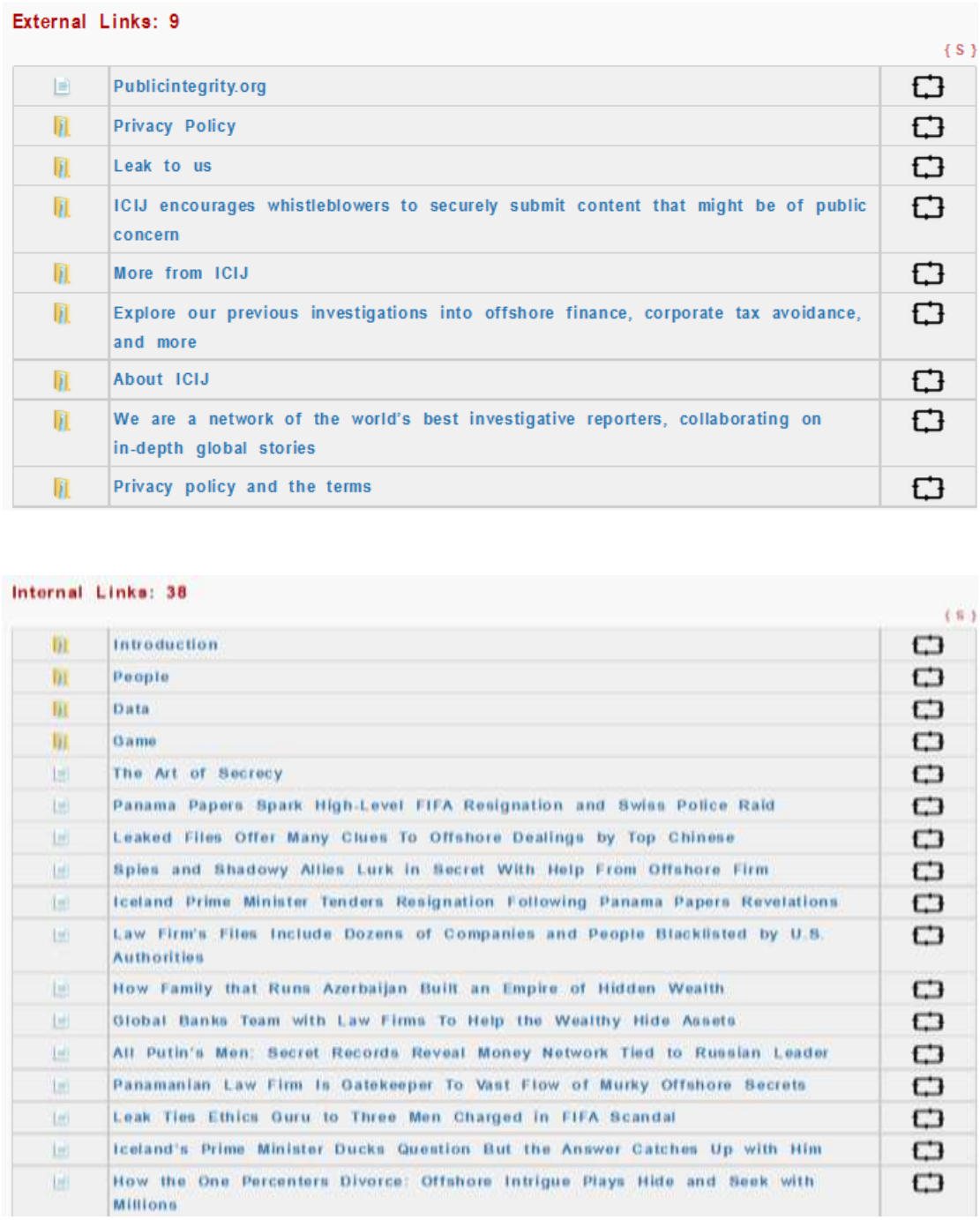

Figure 3. Recrawling results obtained from record shown in Figure 2.

Clicking again the ⌑ icon next to a result retrieves new linked URLs.

Simplifying Discovery of New URLs

We have built the FQU Bot that, as its name implies, extracts Fully Qualified URLS from an input URL. This tool is available from the Tools section of Minerazzi at http://www.minerazzi.com/tools or by just going to http://www.minerazzi.com/tools/fqu/fqu.php (Minerazzi, 2016b).

To use the tool, just input an initial URL from a search result or recrawl. The tool also accepts a piece of text or source code as the initial input. Many users find that it greatly simplifies the creation of curated collections when the input is topic-specific.

Conclusion

Minerazzi simplifies discovery and classification of relevant linked URLs. A single URL can be used as the starting material for building collections and subcollections. The recrawling process is driven by the relevance judgments made by a user.

References

Minerazzi (2016a). The Panama Papers. Retrieved from http://www.minerazzi.com/pp

Minerazzi (2016b). FQU Bot. Retrieved from http://www.minerazzi.com/tools/fqu/fqu.php