Site scraping techniques to identify and showcase information in closed formats - How do organisations find out what they already publish?

Proposal for the Share-PSI 2.0 Timisoara Workshop: Open Data Priorities and Engagement.

Peter Winstanley

1. Focus

This session addresses the question of how organisations that already publish considerable amounts of information on their website but in non-interoperable formats such as Excel and PDF might 'discover' what they are publishing and present it in various helpful ways to end users (including the organisation's own staff) as part of the engagement to discover priorities for open data publication.

Illustrations from site scraping of Scottish Government and NHS Scotland will be presented.

2. Rationale

Many government and public sector bodies already publish a considerable amount of information including data and reports on their websites. However, this is frequently done under a distributed management process and using content management systems, both of which tend to militate against being able to present in a quick and flexible way the assets of any particular publication format. As a consequence, organisations might find it challenging to know where to start when establishing a programme of converting existing information resources that are not in open formats (1-2 stars on the 5 star model[i]) to more open formats.

The Scottish Government has been experimenting with site scraping using the Python Scrapy library[ii] and Exhibit[iii] to periodically gather from the website[iv] all the links to Excel and PDF documents together with a set of metadata and other information from the web page containing the link. The metadata are then used as facets for sorting and grouping the links in an Exhibit faceted browsing web page[v]. Similar work has been done with the websites of NHS Scotland Information and Statistics Division[vi] and NHS Health Protection Scotland[vii]

3. Issues for Discussion

3.1 Is scraping of websites in this manner a generalizable and useful technique to help organisations know better what assets they currently publish on their websites in specific formats such as Excel and PDF?

3.2 What approaches can be used to improve the involvement of the general public in examining these information sources as a way of assisting organisations with prioritising work?

3.3 Are techniques for automated document conversion (e.g. from Excel to an XHTML representation using Apache Tika Server[viii]) an adequate way to provide an interim and generalisable solution to the publication of information in open formats?

---

[i] http://5stardata.info/
[ii] http://scrapy.org/
[iii] http://www.simile-widgets.org/exhibit/
[iv] http://www.scotland.gov.uk
[v] http://labs.data.scotland.gov.uk/
[vi] http://www.isdscotland.org/
[vii] http://www.hps.scot.nhs.uk/
[viii] http://wiki.apache.org/tika/TikaJAXRS