ScraperWiki

- **Understanding data-user needs**

How would you go about understanding the needs of different types of data user to develop a solution that meets the needs of specialist and non-specialist users alike?

Our approach is to apply an agile framework methodology. The first phase is the *discovery phase* - where user requirements for the internal users are defined and analysed. This process will include conferencing, meeting and interviewing the key stakeholders and documenting the results. The output will include 'user stories' which we define, prioritise and estimate.

A system is developed in sprints of 10 days and released often and iteratively to ensure that the internal users needs are being met. The ScraperWiki data science team introduce clients to a rigorous extreme programming form of agile development often referred to as XP. Our teams test first using Behaviour Driven Development this process enables fast reaction to feature changes. We prioritise customer interactions. By following this approach it enables our team to deliver working software frequently and welcome changing requirements even late in development cycle. This approach removes many of the risks associated with waterfall development and substantially improves speed of development and reduces costs. All the tools used are open source which reduces lock-in.

Example: Government Performance Platform Team

In January the UK Government Performance Team (part of GDS) approached ScraperWiki to help collect data from Gov.UK, Google Analytics and myriad other sources. The objective is to build 24 departmental dashboards to feed the https://www.gov.uk/performance . The project started in ernest at the beginning of March. A business analyst and senior developer spent a week onsite with the performance team, meeting the stakeholders, documenting the requirements, creating and prioritising user stories. Sprints are a week long and iteration meetings are happening every week to ensure that the users see progress.

- **Data processing**

How would you take messy data from a variety of sources and in a variety of formats and import it into a backend system that can drive a website?

ScraperWiki's platform and services are designed and fit for this purpose We collect data from a range of diverse sources such as websites, PDF files, and online data services. We clean and normalise it or conjoin it as required and output it in different formats. These output formats include simple one-off static files supplied as plain text or spreadsheets, but can also be live and continually updated data hosted on our data hub, which can be accessible as a resource across any organisation to common analysis and visualisation software, such as Excel, R, Tableau and Matlab. Alternatively as a data feed which can be consumed by any other application which can source data from a URL. We are regularly asked to provide data in JSON.

Example: UNOCHA

The UN Office for the Coordination of Humanitarian Affairs (UN-OCHA) is responsible delivering relief in parts of the world which have experienced natural or human disasters.

This usually requires provision of data to a wide range of governmental and non-governmental organisations, charities and commercial providers in countries around the world.

ScraperWiki supported UN-OCHA in building a data hub containing detailed country indicators like child mortality, infrastructure, and health care, sourced from 20 different websites including the World Bank and UNICEF. The HDX data hub has become a one-stop-shop for analysts and suppliers to the UN, and will, in future, be supplemented by internal data from OCHA such as staff disposition and donor information, and ultimately operational data from the field.

- **Data exploration**

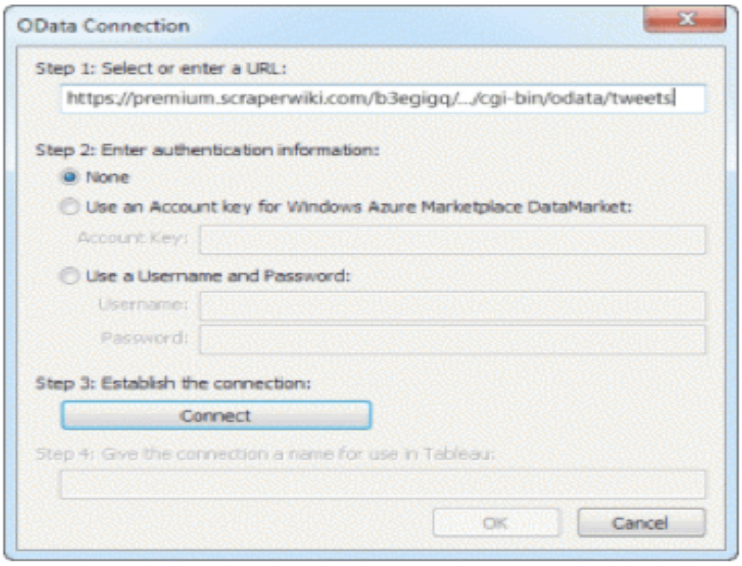
How would you create a useful and meaningful interface and/or visualisations so users can pose and answer simple questions from the data?

We will do a 'discovery', define the broad scope, speak to the stakeholders, ask questions, and pose example hypotheses. A prototype of a simple version is built and we make iterative changes.

ScraperWiki's platform also provides some *standard* visualisation tools including the 'Summarise my data' tool, 'Plot a graph' tool and 'View on a Map' Tool.

Many organisations have a preference for using 3rd *party visualisation tools* and ScraperWiki offers support to a number of these and which is why we built the **"Connect with Odata'** tool. This is a hassle-free way to get ScraperWiki data into analysis tools like Tableau, QlikView and Excel Power Query.

Lastly some organisations want customised visualisations for a specific purpose and we are able to create and have created these to meet the needs of our customers.



OData Connection

Step 1: Select or enter a URL:

https://premium.scraperwiki.com/b3egigq/.../cgi-bin/odata/tweets

Step 2: Enter authentication information:

● None

○ Use an Account key for Windows Azure Marketplace DataMarket:

Account Key:

○ Use a Username and Password:

Username:

Password:

Step 3: Establish the connection:

Connect

Step 4: Give the connection a name for use in Tableau:

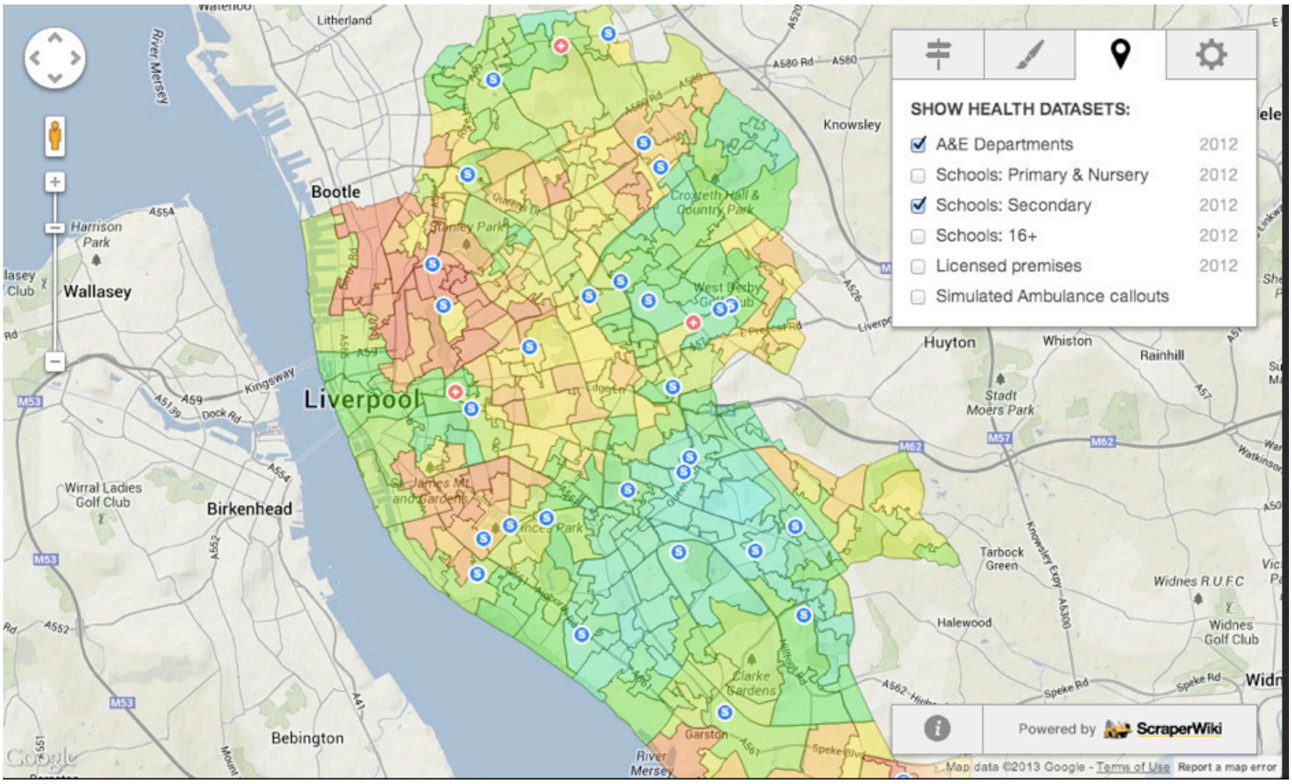
OK

Cancel

Example: Health Research Map – Liverpool John Moore University

A research department at the University wanted to provide data to the emergency services to support their work. For example, identifying factors such as numbers of assaults near licensed premises and schools, or numbers of ambulance call outs to falls by local demographics.

For this work we matched up public data from the Office of National Statistics and local authorities with internal data from the emergency services. This data was then displayed on an interactive map which enabled researchers to make sense of the complex interactions in the data.



SHOW HEALTH DATASETS:

☑ A&E Departments 2012

☐ Schools: Primary & Nursery 2012

☑ Schools: Secondary 2012

☐ Schools: 16+ 2012

☐ Licensed premises 2012

☐ Simulated Ambulance callouts

We created a short video to explain how the map works http://youtu.be/HPTOnUZxxII

- **Legal issues**

What activities would you undertake to tackle the legal issues around data publishing, including licensing, intellectual property rights, derived data and privacy?

We're from an open source background, and members of our team have been scraping Government websites and republishing the data for 11 years.

We would advise the client on legal issues around scraping, which are complex and involve multiple areas of the law. In short, obeying robots.txt and sensible resource use covers most cases.

Publishing the data (rather than just analysing it) raises a further set of issues. It is possible for organizations to get paralysed worrying about this. Private data, and commercially sold data, should not be published - they aren't meant to be open data.

There is, however, a wide range of Government data that is not explicitly licensed as open data, and should be. Depending on the clients media and legal risk needs, we may advise them for practical
purposes to republish such data as if it were licensed as open data.

If they are themselves a Government agency, we would work with them to properly release the data with an open license.

- **Practical issues**

How would you ensure that the data is easy to find, and that any issues about its quality are documented?

This depends on the clients needs. If they are using an open data publishing platform, like CKAN, we would use it for discovery and quality meta-data.
Otherwise, we'd start with good SEO for the pages publishing the data.
Most people find most data via search engines.

- **Technical issues**

How would you publish different types of data (reference data, raw data, aggregate statistical data) as open data in ways that meets a variety of different data users needs? What formats and open standards would you use? How would you approach creating persistent identifiers? What additional metadata, such as provenance metadata, would you provide?

We approach all data processing and conversion tasks based on the needs of the end user. In this case there are two different kinds of user - the agency publishing, and others who then use the data.

The agency may have existing standards and systems for data publishing, and methods they use to internally make them publish more data. We would follow those, or advise on improvements to them.

Otherwise, we would publish in the appropriate simple format. Typically this is CSV files, or for more complex/technical datasets JSON.

We reuse existing identifiers where possible, otherwise we would create new ones in simple situations just short strings of text, or in more complex ones as permanent URLs.

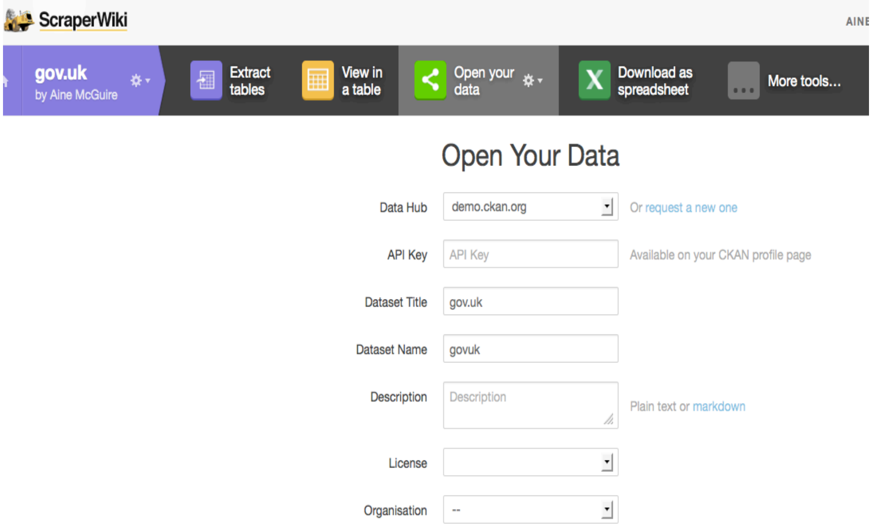
We would make sure new and more complex formats are added if there is a sign that end users could benefit from them (e.g. publishing to Tableau).

All data is useless unless you know how it is corrected, so we would also publish links to the original source and methodology, and source
code that processed it.

- **Social issues**

How would you support users of open data with relevant documentation, source code snippets, example queries and technical advice? How would you provide feedback routes to the publisher?

Since its first launch ScraperWiki has worked closely with the Open Data community. We have built a relationship with a worldwide community of data activists and data hackers. We released an **'Open Your Data'** tool to make it easier for data geeks to publish data that they liberate on ScraperWiki. We also provide free premium datahub accounts for open data projects.



Open Your Data

Data Hub	demo.ckan.org	Or request a new one
API Key	API Key	Available on your CKAN profile page
Dataset Title	gov.uk	
Dataset Name	govuk	
Description	Description	Plain text or markdown
License		
Organisation	--	

The ScraperWiki website is Open Source and licensed under the Affero GPL. Check us out on Github.

ScraperWiki has also cultivated a relationship with a worldwide community of journalists and journalism schools by giving them premium accounts with the objective of helping them to exploit and make use of open data.

- **Team**

What team would you put together to provide this service?

ScraperWiki will put together all or some of the team that is suitable for agile development and the number of roles will be depend on the size of the project and whether we are taking complete ownership for the whole project delivery. For many of our projects **Team Lead** (Business Analyst/Program Manager) / **Team Members** (Architect/ Developer/ UI Specialist/). We are also accustomed to working with organisations where the Product and Delivery Managers are supplied by the customer.