ODI: Assessment proposal

BUSINESS PROPOSAL

Author(s):	Chris Palk, Jamie Matthews
Issue date:	21st March 2014
Version:	1

Contact Information:
Chris Palk (Managing Director)
E: chris@dabapps.com
T: 01273 208222

Table of Contents

Introduction *3*

Understanding Data Needs *3*

Data processing *3*

Data exploration *4*

Presentational view *4*

Underlying data sets *4*

Legal Issues *4*

Practical issues *5*

Technical issues *5*

Social issues *5*

Team *6*

Introduction

The below proposal is a response by DabApps to the Open Data Institute's assessment of Suppliers of Open Data Services. DabApps are a small development agency based in Brighton. We specialise in API and data-driven projects.

Understanding Data Needs

The DabApps team will work with the team at DEFRA to identify who the users of the data will be. We would work together to come up with a set of personas (http://www.romanpichler.com/blog/agile-product-innovation/persona-template-for-agile-product-management/) for each type of user we identified. This would be a interactive session where together we discuss and define each type of user and try to identify what their key requirements for the data will be. The personas will be documented and published so they can be reviewed by DEFRA.

From here we will build a set of user requirements - both for the data and for the way that users would like to access it.

If time and budget allows we would also conduct interviews with some key users of the data. We can use this data to validate the personas we have identified and to add to the user requirements

Finally we would produce a matrix detailing the data and access needed and the relative priority of each component.

Data processing

Data processing is the task of taking data from a variety of sources and applying transformations to unify them in a common format, so that they can be stored, integrated, and more easily made available to consumers. Each data source comes with its own set of technical challenges. The DEFRA scenario specifies three different source data formats: semi-structured Word documents, a relational database, and Excel spreadsheets.

The simplest of these to work with is the relational database, as these are explicitly designed for the purpose of making data available to other systems in a predictable and flexible manner. Once the schema is understood, it's a relatively simple task of connecting to the database via an adapter for the chosen programming language (in Python, we use SQLAlchemy, which provides connectors for all popular open-source and commercial database engines) and extracting the necessary data.

Excel and Word are designed to allow humans (rather than other programs) to work with and visualise data, and so tend to be far less predictable in their structure. That said, it is possible to read and manipulate these documents programmatically using open-source libraries. There is often a manual "sanitisation" step, where a human intervenes to clean up and reformat the data in a way that would be prohibitively complex (or even impossible) to automate. As the data must be regularly refreshed and kept up-to-date, it is vital that the transformation workflow is thoroughly documented, and the import process is kept as foolproof as possible.

In dealing with external data sources, we find it is important to strike a balance between the robustness principle - "be liberal in what you accept" - and a guideline from The Zen of Python: "errors should never pass silently". In other words, when dealing with messy data, attempt to automate as much of the workflow as possible, but ensure any problems or failures are escalated for analysis by a human where necessary.

Data exploration

Presentational view

The landing point for users would be a presentational view onto the data.

This would include textual content giving an overview of the motivation of the study, and an overview of the data being presented.

The central element of the presentational view would be an interactive graphic allowing casual users to view and navigate around the data. The style of presentation we would provide is best demonstrated through the D3 gallery (https://github.com/mbostock/d3/wiki/Gallery#visual-index) and examples such as The Guardian's usage of open data sets (http://www.theguardian.com/news/datablog/interactive/2012/dec/04/public-spending-uk-2011-12-interactive).

The aim of the landing page is to present enough information for non-technical users to explore the data, and to present a initial point from which users can further access the data.

* Linking to the underlying data sets. See below.
* Linking to a GitHub page containing the source and documentation for the interactive demo. This resource would be aimed at helping developers wishing to create their own visualisations onto the data sets.
* Linking to any particularly relevant third party resources for users wanting to educate themselves about the domain.

Underlying data sets

The data sets would be exposed via a web browsable API (http://www.django-rest-framework.org/) in a way that is accessible both to casual users and to more technical users who need data export or programmatic access to the data.

* When navigating the data sets in a browser the user would be presented with a tabular representations of the data sets, with control for searching and ordering the data.
* Each page would also be available as a JSON based Web API for programmatic access to the data. Brief inline examples of how to access the data programmatically would be included on each page.
* Each page would include a link for export of data into CSV format for use with spreadsheet type applications.
* The unedited source data files for each data set would linked to from each page.
* For admin users the interface would also provide for import, editing and flagging data as ready-to-publish.

Legal Issues

When making data available we need to be aware of the restrictions by a number of laws and acts. Some examples are Copyright Law, Competition Law, Freedom Of Information Law, Privacy Law as well as the data protection act. We would need to review the data against the main points of these acts and if necessary we would consult legal experts to avoid any doubt. Our recommendation would be use a existing licenses such as the Creative Commons Zero (CCO) ,Open Data Commons Attribution or the Open Government license, as a standard to work from.

Practical issues

The stated aim of the scenario is to make data available to a number of different users. If data is not discoverable and accessible by its "target audience", then it might as well not exist.

The most common path for any user trying to research a given topic is the Web. For this reason, it is vital that the "entry point" to the published data sets is set of web pages, which describe and document the available data and make it available for download (or access via APIs etc). If appropriate, these pages should include statements on the quality of the data, details of any missing data, license information, and provide links to further topics of interest.

A useful side-effect of creating a "home page" for the dataset(s) is that it can be easily indexed by search engines, improving discoverability for casual researchers. The URL may also be linked to from other dataset aggregation services such as data.gov.uk.

Technical issues

Any decisions on formats for data publishing must take into account both the data itself, and the intended audience.

It's important that datasets aimed at relatively non-technical users are published in a simple, universal format (ideally an open, text-based format such as CSV, or Excel if absolutely necessary). However, many complex datasets may not lend themselves to such a flat, two-dimensional structure. In some cases, it may be practical to introduce simple relationships between such flat datasets, perhaps by assigning each data point an identifier such as a UUID (universally unique identifier), or identifying a naturally unique identifier in the source data.

If the consumers of the data are programmers, then it makes sense to publish in a widely-accepted data interchange format such as JSON or XML. Again, the exact format(s) chosen depends on the nature of the data. A decision will need to be made about whether to provide only the "raw" data (in the form of downloadable dump files) or whether to allow the user to interrogate the database programmatically via one or more APIs. The data dump approach is usually less work, and is sometimes more appropriate for very large, low-level datasets. The API approach is very useful if the intention is to allow applications to be built on top of the data. This is described in more detail in the Data Exploration section above.

Finally, if the audience are in a specific field (scientific, technical, etc), it is useful to research and conform to data publishing standards that are already established within that discipline. This may include specific formats, schemas or metadata that make it easier for data to be integrated into an existing workflow or toolkit.

Social issues

The presentational view would include a link to a GitHub (http://github.com/) page with its source code and documentation. This would be presented primarily from the viewpoint of a developer or technical user interested in presenting their own visualisations based on the data sets.

The page should include not just the source code, but tutorial style examples of how to create alternative interactive examples.

A prominently linked FAQ page would be provided, including a contact form. Comments made through the contact form would be processed in such a way as to allow the FAQ to be easily updated as part of a response to user questions.

The site owner would also administer a Twitter feed and Google Group. The Twitter feed would be intended primarily for notifying interested parties to new updates to the data sets, and for answering casual questions. The Google Group would be intended for in-depth communication between interested parties, including researchers, developers, and actively involved citizens.

Team

The Core Team would be made of

Chris Palk (MD): Chris's background is in the financial services sector, managing large IT projects for city banks and financial institutions such as Morgan Stanley and Lehman Brothers. He has experience of large database systems.

Jamie Matthews (Technical Director): Jamie is in charge of our technologies, standards and development strategies. He is an accomplished systems designer and developer with particular expertise in programming frameworks. Jamie has a keen interest in Big Data and is a founder of Big Data Brighton.

Caroline Pickering (Creative Director): Caroline is an experienced User Experience (UX) and visual designer, as well as being a front end engineer, who has worked on a wide variety of large projects for clients across the public, private and “third” sector (Charities and Not for Profit Organisations). She would be responsible for the user requirement gathering.

Tom Christie (Senior Developer). Tom is a an API expert and author of popular Django Rest Framework open source project. He has experience of building many APIs on top of that framework and would architect the DEFRA system.

We also have a number of other experienced developers in-house who would be part of the team if required.