Open Source CMS SaaS for Government
10 John St, London, WC1N2E, UK
W: www.pretagov.co.uk
P: +44 (0) 87 0392 7071
E: contact@pretagov.co.uk
Company number: 8408608
VAT No: GB 166569956

Open Data Digital Service Delivery Proposal

Project: To publish research data from the Food and Environment Research Agency (FERA)

PretaGov's Approach

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 first step would be ensure the identify the different users groups and create a profile of who they are what their needs are. In this case three user groups have already been identified, Citizens, Scientific data re-users and other data re-users. Our experience is that further analysis can often result in uncovering additional user groups or important sub-divisions within identified groups that could have important implications later on. The first step would be interviews with the client to gain from their knowledge of the different user types.

We would then try to build up an concrete profile of the user needs of each user group. This might involve identifying representatives of each group and interviewing them in person. It might involve producing a survey. It may involve interviews with experts within the client organisation or related parts of government who are familiar with the users groups. The choice will be dependent on the time, budget and specifics of the user groups. The aim of the profile itself will be to create as concrete a description as possible of who these users are, for what reasons they need the data, what users they have for the data and what mechanisms they would prefer to access with. Where possible we ensure that representatives on these user groups can be available on an ongoing basis to answer questions during development and help during the user acceptance testing.

The solution developed that meets the needs of specialist and non-specialist users would change based on the needs of those user groups. We would apply techniques we would normally apply while determining a site Information architecture that has to work for audiences with different needs. One such technique we use when we have enough user representatives available is card sorting. This and other techniques helps us identify the words and concepts that we can use to lead users to the areas of the site that best match what they are looking for. It is likely in this scenario we may end up with the main part of the site handling data visualisation for citizens with a separate dedicated section for those wishing to reuse data.

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?
Because messy data can cause a variety of unexpected issues and delays when not anticipated it is very important to try as much as possible to identify sources of inconsistent or missing data early on. There are a variety of methods that can be used to achieve this. Trying to obtain samples of all the different kinds of

data early in the process can be very helpful. Prototyping the import process to determine data import issues is helpful.

We also use extensive use-case analysis during the website design to ensure that we cover all the ways the data will need to be accessed. Working backwards from this we can often further determine which parts of the data are important to get the desired end result. For example, if data needs to be visualised as a map, will geographic coordinates per provided with the data?, and if not, how accurate will the geocoding need to be?, and how clean is the address data that will need to be geocoded?, etc.

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?
Designing a useful and intuitive user experience (UX) is an iterative process. Once we have profiled the different user groups to get a concrete picture of their needs we start creating use-cases which document concrete scenarios of interaction with the system. A use-case consists of a description of the need, a main scenario detailing step by step interaction and many alternative scenarios outlining what happens if things don't go to plan. Each use-case allows us to determine what visualisations or interfaces give us simpler scenarios. We aim to produce where possible an exploratory UI where a single visualisation can be used to answer many questions that we have anticipated and also those we haven't.

Wireframes would be constructed based on the scenarios. We would then organise a workshop with user representatives and step through the scenarios and wireframes to try and determine the suitability of the designed interfaces for the users. This process is designed to maximise feedback and to gain feedback earlier rather than later. This enables us to iterate our designs at each stage cost effectively so the final design maximises the goals of the users.

Once wireframes and use-cases (Discovery Phase) have been signed off we progress to developing a prototype (Alpha Phase). Our cloud based coding environment allows collaboration with the client and users, allowing us to quickly iterate the solution as we gain more feedback. Built into visualisation tools make certain graphs very quick to produce which can be tested to see if they are suitable before most custom solutions are produced. To procure

Legal issues

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

The following is our generic approach to dealing with legal issues which would apply to the DEFRA Project.

Licencing

In discussion with the client, a licence would be selected under which the data is to be released, ie the UK Open Government Licence. We would also discuss the attribution to be applied, typically the name of the organisation and a link to either the organisation's home page or a page about the data or content being licenced. We would include copy on how the attribution is to be presented, preferably on a separate page which space is limited such as mobile apps.

We would ensure in working with the client (unless specifically requested to not implement) that the license covering the content or data made available is:

- in a human-readable description as content on the website
- implemented as computer-readable metadata on the website

The human-readable descriptions and marks as defined on the following websites:

- Creative Commons licence chooser
- Open Data Commons licences

The metadata would comply with the correct vocabulary to publish rights metadata in a number of different ways as detailed at: https://github.com/theodi/open-data-licensing/blob/master/guides/publisher-guide.md. This includes Publishing Copyright Notices, Licences and Rights Statements.
We would discuss how they would like the data to be accessible via scraping, web APIs, JSON etc, then implement the appropriate rights statement format.

Protection of Derived Data

In consultation with the client, the risk of derived data containing personal information needs to be assessed and if necessary solutions to protect personal data such as the use of Multi-Level Security (MLS) labels to associate data with its protection level and apply usage functions (*transformations*) to ensure protection requirements.

For some datasets, it may not prove possible to transform them sufficiently such that a reasonable level of privacy can be maintained for citizens; these datasets should not be opened to ensure data protection.

Copyright and Database Rights

To protect the copyright and database rights of the government department and local government authority, PretaGov would implement technical solutions such as robots.txt or other methods such as obfuscation and use of javascript to help prevent scraping. In addition we would advise our client to specify that data can not be scraped without permission in the Terms and Conditions.

Practical issues

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

Making information easy to find is standard information architecture. It involves using different forms of navigation and working with the user profiles to identify the key phrases and paths users would identify when trying to find that data. This normally involves dropdown menus, free text, faceted and metadata search interfaces as well as highlight particularly popular or topical data sets where needed. Our CMS gives the publisher the tools to modify the navigation as the relevance or importance of various data sets or visualizations change over time.

Quality issues will be able to be documented by the publisher via accompanying documentation and also in the metadata for the datasets, documents or visualisations.

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?

The methods for download and access of the data depend on the profile of the users who want to gain access to the data, the nature of the data itself as well as performance and cost considerations. In general our goal would be to support as many methods as possible as it's not always possible to predict which format all users would prefer to use. Our most common forms of data publishing is CSV download as it is supported by both technical and non technical users. This can be provided to both authenticated or unauthenticated users.

Sometimes data is provided in unstructured forms such as Word documents , spreadsheets or PDF's which we make available can be free text search, metadata search or downloadable.

We can provide a REST API utilising the HTTP open standard to publish data via JSON or XML. Normally when importing data we assign our own UUID style identifiers which can then be used in APIs where a suitable persistent ID is not present in the original data.

All webpages including those containing downloadable datasets and visualisations of open data will be accompanied by eGMS metadata information embedded in the page. This includes fields dealing with the sources and allowed uses of the data.

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?

We think it is important to provide both reference documentation for any APIs as well as tutorial style and worked examples as people learn differently. Our architecture is based on an open source content management system (CMS) which allows for documentation and frequently asked questions to be updated by the publisher over time as more feedback is received.

Our CMS allows forms to be created and embedded in other pages. Forms can be created and adjusted by the publisher themselves so they can be adjusted by the publisher themselves to hone the feedback they are receiving.

Team

What team would you put together to provide this service?

The Team for DEFRA project as defined by the Digital Roles in the Digital Services is as follows:

Senior Business Analyst (1)
Product Manager (1)
Senior Developers (1)
Junior Developer (1)
Technical Architect (1)

Please refer to our daily rates for these roles in the Digtial Services Store at https://www.digitalservicesstore.service.gov.uk/

About PretaGov

PretaGov focuses on reducing the risk to government when entering the cloud to procure CMS Software as a Service. They do this by ensuring the Government procure cloud based software as a service with the lowest risk possible as their services comply with government standards such as the Security Policy Framework (SPF), accessibility standard WC3 WCAG L2 for the Disability Act , the Data protection Act, data sovereignty, Public Records Act, government branding and information architecture guidelines, as defined by the Government Digital Service manual. We have extensive experience in high availability solutions. Our PretaGov SaaS guarantees 99.95% availability using geo-redundant data centres with automatic failover.

PretaGov specialise in open source technology and 'Agile' delivery of secure, fully supported intranets, web applications to digitise government services, open data digital services and websites to the public sector. The advent of the G-Cloud and Digital Services Frameworks have enabled us to cost-effectively deliver fit for purpose web solutions that the previous procurement environment made cumbersome, time-consuming and expensive for government.