Overview

Neontribe is a small web development house with some experience of working with data to produce digital products and services. Our processes will be useful for some of a project of this type, or for all of a smaller piece of work. We would typically work with partners with specific domain knowledge for a project of this scale, concentrating our efforts on those areas where our expertise is strongest.

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?

Neontribe's user-centred design process is designed to develop prototypes for digital products and services. It's part of our usual work during a discovery phase of a project. It's centred on the people whose needs we are are looking to meet. We'd run a series of workshops with DEFRA staff, and other stakeholders representing specialist and non-specialist users. These workshops:

- Explore people and their situations to produce ideas for digital products and services.
- Expand ideas into stories that articulate the benefits and outline the features of a concept.
- Iterate lo-fi prototypes to gain a deeper understanding of how we'll meet people's needs.

People, situations and ideas

This workshop starts by outlining the objectives of the project to participants. That done, we think about people, situations and ideas.

We know we are talking about 3 audiences: citizens, scientific data re-users and other data re-users. We'll articulate segments of those three audiences as personas: fictitious characters designed to advocate for their audience. We carry on by thinking about the situations these people might find themselves in. Something like "Listening to a radio programme about chemical contamination" is enough. That character, in that situation, leads to ideas. Something like "Somewhere I can propose a product for testing next year" would be an example of an idea. An idea might be a digital product, or an interface to the data, or a feature for a web property etc. At the end of the day, we'll roughly prioritise the ideas for further work. We've blogged about one method of doing this here: http://www.innovationlabs.org.uk/2013/08/11/innovation-cool-wall/

We've run these workshops for a variety of projects: there's a description of one run for a very different audience here: http://www.innovationlabs.org.uk/2011/12/20/on-spending-a-day-innovating/

Ideas to concepts

This workshop starts with an introduction of the relevant outcomes of "People, situations and ideas" day. What's on the list of ideas we intend to take further, and how did we get there?

Then, small groups develop one or more linked ideas into concepts for taking a stage further. During the day, each group will present their concept to the rest and then iterate on the basis of feedback. You'll find a

descripion of a workshop we helped run which had a similar objective here: http://www.innovationlabs.org.uk/2012/02/19/lab-2-the-end-of-the-beginning/

Paper prototyping

After an “ideas to concepts” day, we’d decide which should be taken further. We’d produce a paper prototype of each of these for examination. In a series of test-and-iterate sessions, a pair of people will test each idea while another pair observes. That done, the four of them will work with us to make changes to the prototype. That might be changing something that already exists, or developing a new features. We’d end the day with a tested prototype of each concept, that could be taken to alpha development.

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?

The needs of the end users of the data will help us understand how clean or untouched the data must be for it to be of use, and how frequently it must be updated.

Three basic techniques are helpful to us.

1. run code that automatically collects, cleans and then presents data to a regular data import service. This works well when the availability of incoming data is variable, but the format is consistent. It is likely to be useful for updating some reference data and aggregate statistical data.
2. create a series of tools that allow an administrator to easily view and choose what data to import and when, and/or to correctly format the data. This is useful when data sources and quality vary to the extent that only a human can make a decision, automation is too costly or data needs to be added which is not otherwise available. This is likely to be needed on data held in Word and Excel. While an Excel spreadsheet *may* be well-structured and understandable simply saved as a .csv, it is likely to be otherwise.
3. develop a service that can collect data directly from 3rd parties. This is useful when there are reliable, consistent, data sets. Much of the raw data could qualify here, especially if the database exposes an API which makes it available to us for aggregation and statistical analysis, or through a data transform for consumption by other data re-users.

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’d be looking to our user-centred design process to inform what we’d build here. Matching people to situations helps us frame meaningful questions, and prototyping helps us refine them before development.

On the past, we’ve used simple Javascript libraries to consume data from an API. An example would be a small hackday project that has grown up a little, Career Trax: http://career-trax.herokuapp.com/#search. This tool uses data from http://api.lmiforall.org.uk to answer the question “What might the job market be like for a career I’m interested in?” More at https://github.com/neontribe/lmi-everywhere

Another example would be our work for the Young Lives project, where we used Ontowiki to manage selected data and metadata, and then produce some configurable graphs that showed that data. http://data.younglives.org.uk/view/?r=ylp%3AEthiopia Those graphs can answer questions like “How has access to clean drinking water changed in the regions of Peru?” from the data that is available to them. This project also attached narrative written by the data owners to sub-sets of the data. This approach is likely to be useful for the needs of citizen-users. More about the project can be found at: http://data.younglives.org.uk/view/?r=ylp%3Aabout

Legal issues

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

Open data must be available for anyone to use, reuse, and redistribute. Our first instinct in these projects is to rely on a partner’s knowledge of legal issues in data publishing, as we do not have specific expertise in this area. However for the purposes of this exercise, we observe that FERA’s website is © Crown Copyright, and licensed under the Open Government License: http://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/ so it seems likely that this will apply to the data they publish also. They helpfully provide an email address to check. The OGL specifically grants the right to “exploit the Information commercially” with attribution, which we believe is sufficient as long as all data is so licensed. Case law is sparse in the area of derived data from research, however, we’d apply the pragmatic approach of abiding by the conditions of the most restrictive licence for any of the data sets. In terms of privacy, it is likely that data concerning usage of the data would be covered by the FOI, and we would need to check that.

Practical issues

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

We’d seek to add the data to such catalogues as exist: data.gov.uk is an obvious starting point. However, the best way of making the data findable is getting it used and therefore talked about, and suggesting searchable citations for data during the download process.

We believe it is also important to document in a simple human readable form what the plan for data processing is, where the data endpoints are and what, if any, transforms have been needed on the data. We are not aware of any standards work into documenting data quality. We’d investigate if formally adding such issues as metadata to a dataset would be of use to any of our users.

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?

Our instinct is to create a data model first, by drawing diagrams and writing human-readable labels to them,

and then think about different serialisation options later on. Those options would be driven by those which made most sense to particular user communities, but include meta-languages such as XML, JSON, CSV, or some member of the RDF family. During this process, we'd be aware of existing work such as:

- schema.org
- W3C standards: eg. SKOS and data cubes
- Other working groups; DCMI for example.

That data model would define the entities we'd expose at persistent identifiers, and the metadata we'd require. At a minimum, we'd expect that metadata to provide dates of data import and processing, and organisational data. We'd expect to be spending the smallest possible proportion of the project budget on an early delivery of a simple model and single serialisation option, and retaining as much resource as possible to iterate that delivery and support any developer community that grew up around the data. This might mean delivering additional serialisation options, or it might simply mean supporting and engaging with those who are actually using the data. It is extremely unlikely that any one format will satisfy all data re-users. In the past, we have started with JSON: http://linked-development.org/ and at the moment we find developments in JSON-LD and Linked CSV particularly useful in this area.

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 suggest documentation, source code, and examples would live on a wiki, with technical advice available from the project team to registered users. To us, a key issue around publishing open data is metrics for usage. Only with demonstrable use can you secure the business process changes needed to sustainably feed open data in the long-run. We'd examine support through this lens, and:

- offer people the choice of registering, which we'd do that after any necessary download had been triggered, and provide a benefit for registration: e-mail updates when dataset is updated etc.,
- deliver some sort of benefit to using API keys, without requiring them; perhaps the ability to track data re-user's own usage of the API.

Team

What team would you put together to provide this service?

Harry Harrold (5 years experience in UX and facilitating workshops) eg. http://data.younglives.org.uk
Heydon Pickering (8 years experience in HTML/CSS/design) eg. http://career-trax.herokuapp.com/ http://greatbritishpublictoiletmap.rca.ac.uk/
Rupert Redington (5 years experience in UX and facilitating workshops, 8 years experience in front-end development) eg. http://career-trax.herokuapp.com/ http://data.younglives.org.uk http://greatbritishpublictoiletmap.rca.ac.uk/
Neil Dabson (3 years experience in data modelling, 15 years experience in back-end development) eg. http://data.younglives.org.uk http://linked-development.org/
Katja Mordaunt (5 years experience in back-end development) eg. http://career-trax.herokuapp.com/