Beta guidance on assessing files for the NII

A key aspect of the National Information Infrastructure is the curation of a data list that documents the most significant datasets for the running of the nation. Although this goal may seem rather abstract, potentially descending into a philosophical discussion about ‘what significant means’, we seek to outline below a straightforward approach to achieving just this.

What is significant?

For the purposes of the NII, ‘significant’ applies to those core data assets that are essential in understanding and analysing a field or area of work. Further to that definition, the NII, for the purposes of facilitating assessment, sees those significant datasets in two ways:

Subject data
This is the data for a given subject area that is most critical to its functioning.

It provides validated operational information about subjects and informs on things that affect or shape the subject.

Example: Opening times, timetables, eligibility, financial data, availability of services or commodity, inspections, metrics, caseloads, outcomes.

Reference data
This is the data that connects different datasets together. It provides points of interconnection between public data for example temporal and geographical data (maps or primarily geocoded data), as well as definitions and code lists, including vocabularies.

Reference data uniquely identifies real-world subjects such as organisations, sites, assets, publications (eg legislation) or services by describing their characteristics.

Example: Types of school, health conditions, a council's boundaries, the geographical extent of a ward, a base geospatial file, vocabularies.

Identifying NII data:

What you need to do

STEP 1: Ask the right questions

Start by asking the following questions:

- For the work we do, as an organisation, what are the most fundamental datasets which define (tells you what something is or where it is) key aspects of our domain (the areas of society and the functioning of government your department/agency is responsible for).

These will be datasets without which you could not gain an understanding of an area of work or a domain

Some datasets may be completely obvious (where hospitals or GP surgeries are in the case of health) but others may be less so yet fundamental.

Then ask yourself the question:

- For the work we do, as an organisation, what are the most fundamental datasets which provide information about our activities and that are critical to the functioning of our domain?

These will be datasets without which you could not gain an understanding of an area of work or a domain

Notice the clear distinction between something that identifies or defines like ‘type of schools’ or ‘location of schools’ and things that affect the subject such as ‘test results’.

While data on the location of schools defines both their geographical position and what they are (e.g. New Town Primary School, School Terrace, Reading. RG1 3LS), school results data provides information about the subject (in this case both the state of education in the UK and of a very specific item e.g New Town Primary School).

Using this approach, start by developing a list of datasets that will qualify under either reference or subject types and that can be considered by you, experts in that domain, to be fundamental.

In the education example above, the final questions would be:

- *if I did not have this data, could we function?*
- *If we did not have the location of schools could we function?*
- *If we did not have a classification for types of school, could we function?*
- *Can you try and analyse/understand the education domain in the UK without this?*

If the answer to any of those questions is 'no, we could not' or 'not really, it would be an inaccurate view' then you have found an NII candidate.

To help you in your thinking, you can use the list of existing datasets on data.gov.uk which you can obtain by logging into the site, going to your organisation's page and clicking on 'manage unpublished datasets' on the administrative tools box. The second step on the management page will allow you to download all of your datasets (published and unpublished) currently listed on data.gov.uk.

REMEMBER

This exercise is not limited to what you have on data.gov.uk, data that can be published or just open data. This is about identifying the most significant datasets in a given domain based on an organisation's portfolio and to document them based on the NII conditions and principles.

The fact that a dataset cannot be released (e.g data protection issues, legacy licensing, etc.) does not preclude it from the NII if you consider it should be there. You will document the data by following this guidance and its existence will be registered on data.gov.uk the same as any other NII dataset, though the data itself will not be available. Think of your entire domain.

STEP 2: Document the components

The NII has a clear set of components that must be documented. We have provided you with a spreadsheet and the extra metadata required for the NII. This will allow you to gather the information on the spreadsheet so we can ingest it on data.gov.uk when the time comes, making it very easy for you.

Use the explanations below to help you gather the information you need and populate the NII spreadsheet accordingly.

Standards

Does the data follow any standards e.g. ISO or BSI standard or any other industry standard? Write down the actual standard name and provide a URL, there may be more than one, so you can copy and paste another instance of the standard column, as long as they are next to each other, otherwise we cannot ingest them onto data.gov.uk.

REMEMBER

In some instances regulations prescribe standardised ways of collecting and managing data, you can refer to that as a standard.

Vocabularies and code lists

Here, identify any formal vocabularies or code lists that define the things in the data.

For example, if you have a file that uses country codes, then provide a reference to the file that contains each code and what it stands for. The same applies to a vocabulary, for example, a file that provides an actual definition of terms used in the data (could be financial terms, scientific terms or commercial terms).

Check the current format of those files, they must be in CSV and we will provide you with space to upload them on data.gov.uk. If they are already publicly available, then provide the URL for the file or page (we know sometimes it won't be a file but a site).

As before, just copy and paste an extra vocabulary column next to the first one until completed.

If you have a large amount of vocabularies and code files, we can harvest those from your servers in one go, please have a chat with team@data.gov.uk or notify the person that sent you this guidance.

Licencing

Note the licenses the file is bound by. This could be OGL if the file is public and open, or a combination of licences. Provide the name of the licence and in the column next to it provide a link to a licence terms txt file or a page containing that information, again copying and pasting the licence field and the licence URL field next to each other for as many as you need.

REMEMBER

Although all open data should be published under the Open Government License, we understand that other data that has not been published, or may not be published for valid reasons, will have other types of licenses; we want you to document them.

Statutory relationship

Is the data being captured because a regulation or law specifies it? Is there any other legislation that for other reasons may limit or put conditions on the data?

Again, note which legislation and get the URL for the legislation from legislation.gov.uk. Paste it in the legislation column.

As before, just copy and paste an extra legislation column next to the first one until completed.

Service Levels

All data ingested for this exercise will be granted a service level statement. This is a text which declares an agreement to continue publishing the data and maintain its quality as well as publicly notifying ahead of time if there will be any major changes to the data or any potential cessation of publication. In the case of cessation of publication, you will be agreeing to keep data already available via data.gov.uk. Just put a yes in this column. Next to this there is a “reason for non compliance” box - in the rare instance that you cannot agree to the service level for a given dataset, explain why here.

REMEMBER

Although the expectation is that the data will continue to be published, certain datasets may cease publication because the regulation that requires that data to be captured has been amended; or the regulation has been changed in a way that affects the collection (or removes the requirement for collection); or because collection has been superseded by another data collection exercise with a different name but similar scope.

Structure

For each dataset, provide a csv that documents each item in a dataset in the order they show on the data and provide a brief definition of what each header means.

For a dataset with the following headers:

Price	Company	Admin cost	Date

A csv file would document as follows:

- **price:** the total cost of the contract including VAT
- **company:** The company or legal entity that delivered the work
- **admin cost:** The total cost of internal administration of the project excluding VAT
- **date:** the date the latest payment was made (i.e. not the date of delivery)

It could also be the case that it looks as follows:

- **price:** the total cost of the contract including VAT
- **company:** The company or legal entity contracted for the work
- **admin cost:** The total cost of internal administration of the project excluding VAT
- **date:** the date the latest payment was made (i.e. not the date of delivery)

Notice the importance of documenting this. In one instance, the company heading refers to the company that delivered the work, which could be a subcontractor, while in the second instance it refers to the company that won the contract. The user of the data will not be able to know that and may derive the wrong conclusions from the data unless it is made clear.

Although you probably would not show cost against a subcontractor from a head contract (money would go to the contracted party always) it serves to show that documenting the headers this way gives context to the data.

Give the name of this file. We will provide you with a way to upload each file to data.gov.uk.

Unpublished

Is it published or will it go as unpublished? You will have to note it with a ‘yes’ or a ‘no’ on the exemplar spreadsheet. You still need to provide all the pertinent details and an inventory entry will be created for this file.

STEP 3: Non-compliance

There will be things that your data may not be able to comply with yet. For each component and quality aspect that you cannot meet, provide a reason why (there is a non-compliance column), including details on what you plan to do about it.

We will explore the commitments to remedy non-compliance later in the process, but for now simply capture the basics of it in the knowledge that we will come back to you on it.