# Objective

Google Dataset Search is a new service that aims to improve the discoverability of public online datasets [1]. Data publishers need to provide metadata tags in their web pages that describe the datasets. The metadata will then be indexed by Dataset Search.

This note gives recommendations on how to add metadata tags to existing pages in the Open.NRW data portal.

# Approach

Google gives guidelines for publishers so that their datasets can be indexed by Dataset Search [2]. These were used as the basis for this note. In addition, Google created a structured data testing tool [3]. This tool can be used by the authors of web pages to add the metadata tags that Dataset Search requires to index the pages correctly. This tool was used to validate the recommendations in this note.

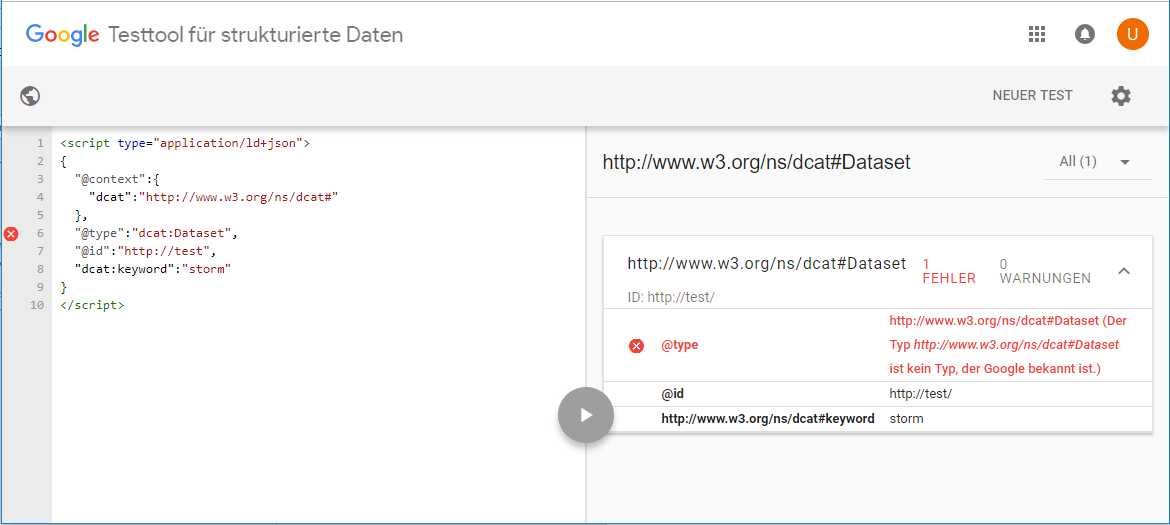
# Adding Dataset Metadata to Web Pages

There are different options to add metadata to portal pages. The first option concerns the metadata syntax or encoding, the second the vocabulary that defines the available tags.

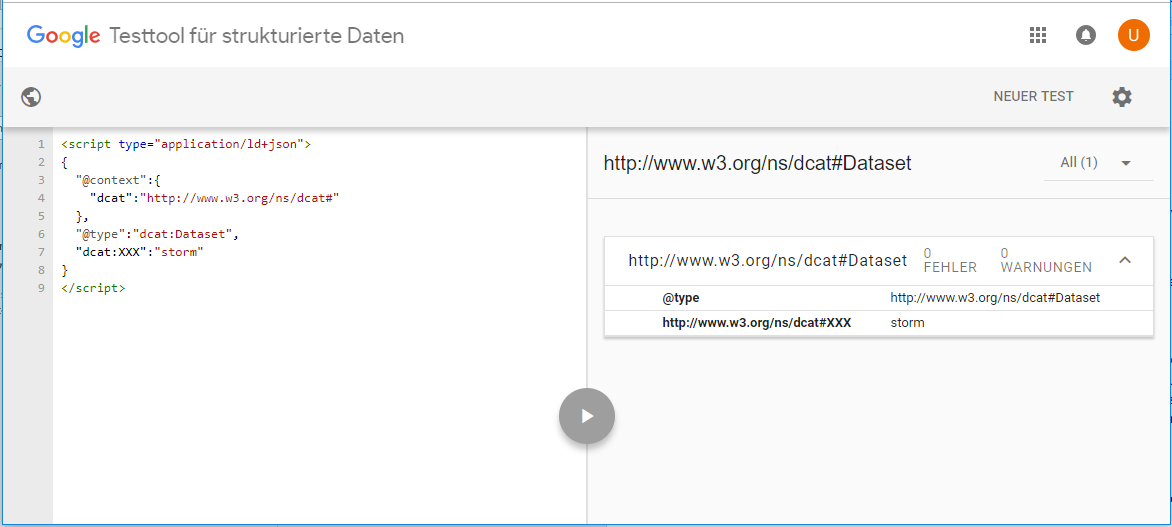
For encoding, Google supports metadata in RDFa 1.1 [4], Microdata [5], or JSON-LD [6]. JSON-LD is the syntax preferred by Google (see [2]). The advantage of JSON-LD is that the full metadata can be embedded in a single script tag in the HTML page. In RDFa or Microdata, the tags are distributed over the HTML page, thus requiring more places where the page that must be modified to embed the metadata. The downside of JSON-LD is that some content must be duplicated. In this note we follow the recommendation of Google to use JSON-LD syntax.

For vocabulary, the Google guide on datasets lists schema.org [7] and Data Catalog Vocabulary (DCAT) [8] as options. DCAT would be a good candidate, because most European Open Data portals already can provide dataset metadata in this vocabulary. However, at least in the structured data testing tool, support for DCAT seems to be not as mature as for schema.org. Some examples illustrate the problem:

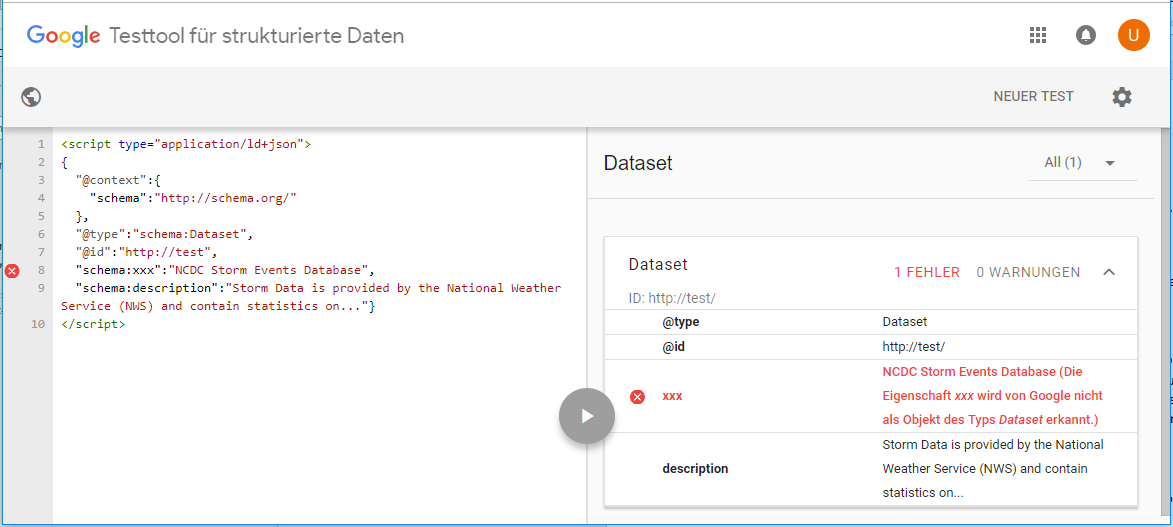
* For metadata in DCAT, adding an @id to the dataset produces an error in the testing tool that DCAT Dataset is an unknown type:



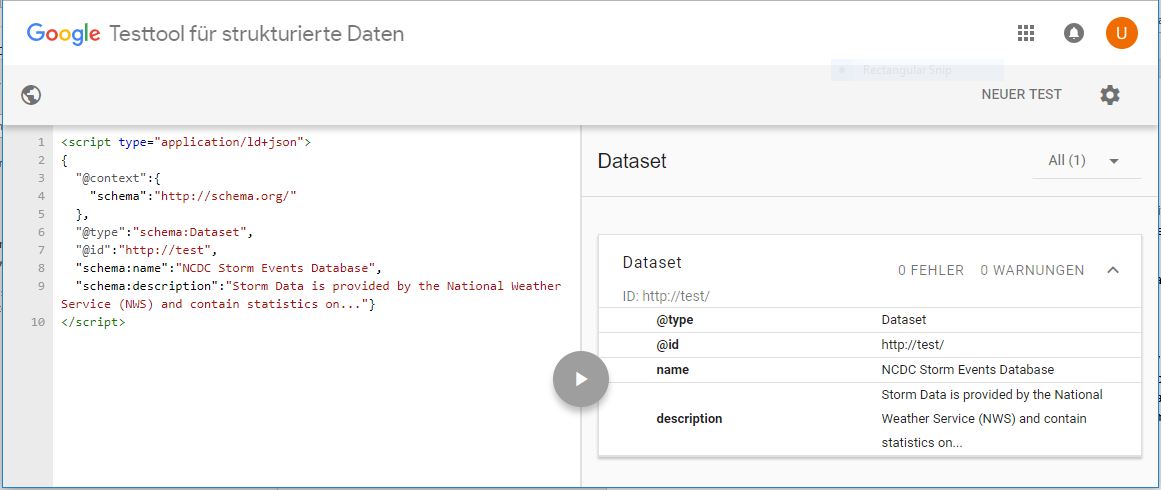
* Also, for metadata in DCAT, the testing tool is unable to validate the metadata and accepts properties that are not in the DCAT vocabulary:



* In contrast, if metadata is using the schemo.org vocabulary, invalid properties are detected:



* Also, with schema.org it is no problem to add an @id:



As it is unclear what effect these errors in the testing tool will have on Dataset Search, in this note we recommend using metadata in schema.org vocabulary for the time being. Once support for DCAT is more mature, it should be straight forward to exchange the schema.org metadata with DCAT-AP.

A dataset can appear on several portal pages, e.g. a details page for the dataset or a result list. Google recommends adding the full metadata for the dataset only to the dataset’s landing page, or “canonical” page. If the dataset appears on another page like result lists, it should make use of the schema.org “sameAs” property to link to the landing page, instead of duplicating the full metadata.

## Landing Page

To create the sample landing page, a random metadata set was chosen from the Open.NRW portal. The details page was downloaded with the required assets and used as the basis: <https://open.nrw/dataset/ldbnrw-service-82711-06izldb>

It should be mentioned that CKAN already has a profile that allows serialization of metadata fields as schema.org in JSON-LD encoding [9] and should also have the capability to embed this metadata in the dataset landing pages [10]. Since this note is a general guideline, we describe the required steps independent of a specific product. But if the portal is based on CKAN, it is recommended to evaluate if the functionality provided by CKAN already fulfills all requirements.

So, for this sample, the DCAT metadata was manually mapped to schema.org and added to the result page. The mapping follows proposals that are currently in the work, like [11]. The JSON-LD encoded DCAT-AP metadata was retrieved from the new portal version, as the DCAT JSON-LD encoding is not available in the old one: <https://ckan.test.open.nrw.de/dataset/a642d7e5-bf1c-57f9-89df-cdad67c7c0fc.jsonld>

Some notes on the mapping result:

* dcat:contactPoint was dropped, because schema.org Dataset has no matching property
* schema:about was mapped to type Class
* dcatde:maintainer was dropped because there was no matching property
* dcatde:politicalGeocodingLevelURI was dropped because there was no matching property
* Type ContactPoint requires either property url or telephone to be present, plus contactType from a reserved list of values, to be validated without errors in the structured data testing tool. This is just to enable the Google corporate contact feature [12]. In the sample we added the URL to make it completely valid in the tool, although it was not contained in the original DCAT metadata. However, it should also be possible to ignore these errors if the corporate contact functionality is not needed.
* dcatde:licenseAttributionByText was dropped because there was no matching property
* DataCatalog currently only contains the name, this could be expanded if required

The resulting JSON-LD document must be added to the HTML page inside a script element with a type of "application/ld+json":

<script type="application/ld+json">

…

</script>

Dataset search will find such tags that are embedded in the HTML page and index the schema.org metadata contained within. This approach is demonstrated in the sample HTML page dataset.embedded.html.

If directly embedding the schema.org metadata is difficult, it is also possible to load it dynamically and inject it in the page, e.g. from some REST endpoint. This approach is demonstrated in the sample HTML page dataset.dynamic.html. Note that this sample only works if loaded from a web server, and not from the local file system, because of security restrictions in the browser.

## Multiple Pages for the same Dataset

Datasets often appear on other pages than their landing page. One example is as part of a list of datasets, like in the results of a search. In such cases, the complete metadata should not be added to such pages. Instead, if the page must contain metadata about the listed datasets, links to the landing pages of the datasets should be added. The link can be defined with the schema.org “sameAs” property. An example of embedded datasets in schema.org JSON-LD encoding could look like this:

<script type="application/ld+json">

{

"@context": "https://schema.org",

"@graph": [{

"@type": "Dataset",

"sameAs": "http://example.com/dataset-1"

},

{

"@type": "Dataset",

"sameAs": "http://example.com/dataset-2"

}]

}

</script>

A similar approach can be used to express provenance information. It is common for open datasets that they are republished or reprocessed or aggregated. In the case of republication of the same dataset, “sameAs” can be used to point to the original source of the dataset. The property “isBasedOn” should be used if the original dataset was changed significantly or aggregated. Finally, the “identifier” property should be used to attach Digital Object Identifiers (DOIs) of the dataset.

## Sitmap

To help Google to better find and index the datasets that are available in the portal, it is recommended to create a sitemap [13]. A sitemap informs search engines about the pages on the site that are available for crawling. Google supports sitemaps in the following formats:

* XML
* RSS, mRSS and Atom
* Text
* Google Sites

Please adhere to the general Google guidelines on sitemaps [14].

# References

[1] <https://toolbox.google.com/datasetsearch>

[2] <https://developers.google.com/search/docs/data-types/dataset>

[3] <https://search.google.com/structured-data/testing-tool>

[4] <https://rdfa.info/>

[5] <https://www.w3.org/TR/microdata/>

[6] <https://w3c.github.io/json-ld-syntax/>

[7] <https://schema.org>

[8] <https://www.w3.org/TR/vocab-dcat/>

[9] <https://extensions.ckan.org/extension/dcat/>

[10] <https://github.com/ckan/ckanext-dcat/issues/75>

[11] <https://github.com/w3c/dxwg/issues/251>

[12] <https://developers.google.com/search/docs/data-types/corporate-contact>

[13] <https://www.sitemaps.org/index.html>

[14] <https://support.google.com/webmasters/answer/183668>