Scope

This document describes a proposal for the spatial data on the web testbed, research topic 4 “Spatial data on the web using the current SDI”.

The proposal is submitted by the following parties:

- interactive instruments GmbH, Bonn, Germany (contact point)
- GeoCat BV, Bennekom, The Netherlands
- Linked Data Factory, Rozendaal, The Netherlands

Table of Contents

SCOPE **1**

TABLE OF CONTENTS **1**

MOTIVATION **2**

WHY TOPIC 4? 2

ARCHITECTURE CONSIDERATIONS 2

Discovery and integration *2*

URI strategy *3*

Metadata *3*

Spatial functions and indexes *3*

Links *3*

PLAN OF APPROACH **5**

OVERVIEW 5

DESIGN 5

Resources *5*

Representations *6*

DISCUSSION 7

IMPLEMENTATION 7

VALIDATION 8

REPORTING 8

MAPPING OF REQUIREMENTS **9**

TASK 1 9

TASK 2 11

Motivation

Why topic 4?

Research topic 4 is of great interest for us as it combines two aspects that are both important to us. First of all, it allows to think ahead, investigate and develop a proposal for making feature-based spatial data sets available on the web. At the same time, it does not simply start from a clean sheet, but takes the existing infrastructure for spatial data sets and the related workflows for data management and dissemination into account.

We believe that it is important to avoid being constrained by the technical details of WFS, GML, CSW and ISO 19139 when exploring how the data should be made available on the web, i.e. when focusing on the first aspect. In the end, it is essential that the data is useful for those that want to use it, e.g. for implementing web or mobile applications for the new environmental act.

At the same time, there is a lot of utility for building a prototype on the basis of the current infrastructure. This approach provides an indication in how far the current spatial data infrastructures are or can be made compatible with providing spatial data on the web and which conceptual issues need to be addressed. If the approach proves to be feasible then it provides a migration path with a potentially low effort for a data provider - or even a central operator of a spatial data infrastructure - to make his spatial data set(s) also available “on the web” as the prototype software will be made available under an open source license.

Of course, if it turns out that there is strong demand for the data “on the web”, then it may make sense to consider to streamline the data dissemination flow for this channel - up to directly maintaining the data set as resources on the web. Research topic 1 and 2 will be looking at this (and to some extent also topic 3, but that will depend on the approach taken). It will be interesting to work in parallel on related topics and benefit from the collaboration with the other participants and discussions with interested third parties. The testbed is promising to provide a good environment to experiment with different approaches and learn from each other.

Since time and funding is limited, it is also helpful that the scope of the research topic does not cover everything related to spatial data on the web, but is focused on a few aspects that are particularly important:

- crawlability and linkability, i.e. making each resource available via a persistent URI and ensure that all resources can be reached via links from a “landing page” for a data set
- classification of the resources using vocabularies supported by the main search engines on the web
- representations of data for consumption by humans (HTML), web-developers (JSON) and search engine crawlers (HTML or JSON with annotations)
- establishing and maintaining links between data
- discovery of both spatial and non-spatial data by the same search engine

I.e., when looking at research topic 4 with the broader architectural goals of the testbed in mind - see, for example, also appendix B of the “invitation to tender” - it is obvious that topic 4 covers only aspects of the bigger picture. At the same time it is important that the design does not place obstacles on the intended workflows for the environmental act.

Finally, we like about this research topic that it is not mainly a desk study, but the ideas are driven by and validated by implementations and demonstrators.

Architecture considerations

The goal of this research is to integrate the existing separate mechanisms for discovery and use of both spatial and non-spatial data. Before we discuss our plan of approach, we explain the architectural view that underpins our approach. This section describes the current mechanisms and highlights where the main issues are, and how we will address these. Details are provided in the following section, the plan of approach.

Discovery and integration

The standard approach in spatial data infrastructures for discovery of and access to spatial data on the web is via web services. There are relevant services for resource discovery (CSW) and for data access (WFS/WCS/SOS). These services are without modifications not accessible by search engines (Google, Bing, Yahoo). They also have a steep learning curve for developers that want to start working with them.

Alternative sources of spatial data on the web that are published as Linked Data share with the spatial data infrastructures that they, too, are not (yet) accessible by the search engines without modification (this challenge is faced in topic 3). These data sets use ontologies like GeoDCAT-AP or GeoSPARQL and are accessible via SPARQL endpoints. However the linked data community is a lot closer to search engine crawlable data then the geo community. That's why we should verify, if the geo community can get closer to "data on the web" by adopting principles from the linked data community.

The only way to make content crawable by search engines is by adding structured markup in website content (RDFa, Microdata or JSON-LD) defined by the schema.org ontology. Schema.org includes (limited) support for representing geometries[1]. Of these three formats only JSON-LD is both a serialisation for RDF *and* a JSON document [2]. So, if a geo service response is transformed into JSON-LD based on schema.org we cover three objectives:

1. the data is crawlable by search engines if we use it for marking up HTML pages
2. the geo data has the same structure and format as plain Linked Data, and can thus easily be integrated (linked) with other "plain" Linked Data.
3. the "plain" JSON data is available through an API, facilitating it to be programmable for web-developers

URI strategy

Because this work involves proxies that rewrite URL's while using linked data concepts like persistent identifiers and links between resources, it is important we define a clear URI strategy before exposing crawlable content to the search engines. A WFS GetFeature query or CSW GetRecordById request, for example, are in general not suitable as a URI for a resource, because such a request can have multiple forms (POST/GET, order of the parameters). They are also tied to a technology and are likely to change with time.

Metadata

It is interesting to see that the concept of metadata as we know it will probably change on the web of data. Currently structured markup (amongst them geo items) is considered metadata to a web document. In a Linked Data set metadata is recorded at the graph level via schemas such as VoID[3], DCAT[4] and for geo GeoDCAT-AP[5]. Both perspectives of metadata are present in the tasks.

Today metadata records are typically maintained in separated processes and use a different type of web service. When eventually the web of data has evolved to a true, crawlable web of linked resources, "metadata resources" will become first class citizens as much as other resources.

Spatial functions and indexes

Spatial web services are capable of executing spatial queries, such as "give me all buildings *located in* Valkenswaard". These queries use a spatial index to retrieve the results on the topological function "within" the border of the polygon that is called Valkenswaard. The technology behind this is based on a spatial data model and a repository supporting spatial functions that work on coordinates. Common client implementations, like search engines, using the Schema.org vocabulary currently do hardly support spatial functions. In order to make a search engine answer a query like "give me all buildings *located in* Valkenswaard" it is necessary that in the data schema a property representing "located in" is used. In this way, the query is not executed spatially, but just as a normal query. The data must therefore be annotated with the property representing "located in" and the coordinates cannot be used to calculate. It is therefore a good idea to create indexes that simulate spatial similarity, such as aggregations based on location or key attributes as the tender document suggests.

Links

The tender document describes tasks that are targeted to creating and maintaining links. We want to address below the different topics we think are relevant to RDF linking, and how we will incorporate this in our work.

Link discovery

To discover and establish links between data resources and keep them actual is a subject of many (academic) studies. Over the years open source tools have been developed to create links based on resource similarity (examples of these tools are RDF Refine[6], SILK[7] and LIMES[8]). The link discovery

[1] http://schema.org/GeoShape
[2] http://www.w3.org/TR/json-ld/#relationship-to-rdf
[3] http://www.w3.org/TR/void/
[4] http://www.w3.org/TR/vocab-dcat/
[5] https://joinup.ec.europa.eu/node/139283/
[6] http://refine.deri.ie/
[7] http://silk-framework.com/

process, often called reconciliation, is based on crawling linked data and finding similarities in property values and qnames. The similarity that has been found between resources must always be validated (manually approved) by human action to guarantee the quality of links. It is therefore a time consuming process. It is not realistic to assume that link creation can be done in real time while querying and selecting features in the data.

Link maintenance

There are tools developed (e.g. semantic pingback[9]) to overcome problems with links. Link issues are not Linked Data issues, they are web issues. With the web of data we are at a new level due to the fine-grained nature of the data and the number of new resources compared to the web of documents, but the fundamentals are unchanged. We have solved broken links in the WWW for many years now, and we are used to it. It is the responsibility of the data provider to maintain, as much as possible, redirects to deprecated links. The user that downloaded the data must however understand that his version is detached from the data web ecosystem and might become orphaned over time. After all, a broken link in the web of data is nothing different than a broken link in the web of documents. Nevertheless in our solution we will provide links to metadata, and we will record provenance information in a dataset when it is downloaded.

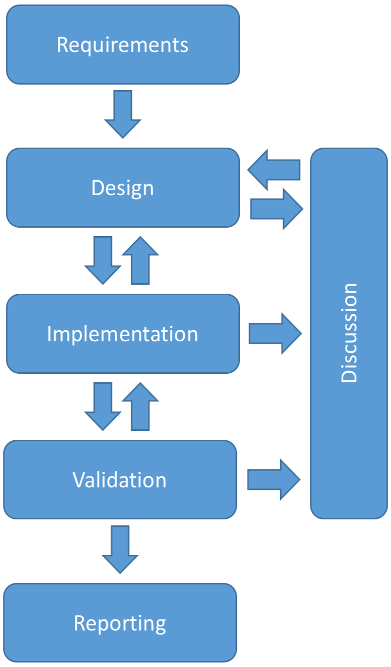
The meaning and value of links

The principle of linking resources is the cornerstone of the Linked Data architecture. It is therefore a powerful instrument for enrichment. Both the data model (ontology) and the data itself (instances) can be linked. The effect of creating a link has an immediate and profound consequence on the data value and quality, hence the usability of the data. Thus it is evident that there are pros and cons to making links between resources. Adding links between vocabularies changes the overall meaning of the now connected vocabularies. Adding links between instances implies that a new relationship between the instances is established. Usually links between instances are of the type that expresses some kind of similarity. The effect of this is that a combination of properties for a “same” resource is established. Other types of relationships are equally allowed. The tender seems to focus primarily on the technical feasibility of link creation and link maintenance (how to retain the link, where is the link stored). In our research we will also address the issue of data usability, value and quality. Meaning that we will ask ourselves questions like: "what is the use case for which we are creating this link?" , "what is the appropriate linking property for this use case?" , "will the general audience understand the context, meaning and intention of this link?"

[8] http://aksw.org/Projects/LIMES.html
[9] http://aksw.org/Projects/SemanticPingback.html

Plan of approach

Overview



The requirements have already been captured in the preparation of the tender.

Based on the requirements and the architectural view described in the previous section, the starting point will be an initial design of the resources that need to be available on the web, their representations and links. The usability for human users, crawlers and web developers as well as the maintainability will be taken into account.

This design will be detailed and updated as needed during implementation, validation and discussion.

Finally, the results will be documented in reports.

Each activity covers both tasks and in task 1 both the data (WFS) and the metadata (CSW) aspects.

The design should be mostly complete in 2015, but we expect some changes due to new knowledge from implementation and validation. Implementation and validation will be executed in close connection over the length of the testbed and start in parallel with the Design. The reporting phase at the end of the testbed will consolidate all the information collected during the testbed.

Design

Resources

Data

For the data sets in task 1 there are at a minimum the following resources that need to be exposed: the data set (the landing page for the data set / service), metadata describing the data set, each feature type (the "layer" in the traditional GIS/map view) and each feature instance.

As there are typically a large number of features for each feature type in a spatial data set and because it will typically be beneficial to group features by aspects that relate to terms / keywords that are used by used in searches, it should be useful to not simply link from each feature type to all its feature instances, but via intermediary resources. We will explore the feasibility of such an approach in a 'crawlifier' proxy[10].

It should also be avoided to return a long list of features or feature references in a single resource, but a paging approach will be used[11]. This is supported by the WFS 2.0 specification and reflects also how search engines present results of a query. One item of discussion should be how much information of a feature is to be included in the list representation and how many features to show per page. In the paging approach, the features must be ordered consistently and this needs to be taken into account in the design and implementation. To facilitate pagination, it can be helpful to follow the google sitemap pagination approach.

We will also specify the links that will be provided for each resource, in particular each feature instance. We will use the principles discussed in the section "The meaning and value of links".

The links will include links to metadata, and we will foresee to record provenance information in a data set when it is downloaded. The provenance information is added to the downloaded data set referencing a selection of properties of the vocabularies VoID, DCAT or PROV.

Metadata

A related metadata for each dataset is referenced from the GetCapabilities response. To link the metadata of the dataset to the individual dataset records in a crawlable way presents some challenges:

[10] For example, the crawlifier proxy could also (be configured to) provide the woonplaats and the postcodes as resources and link to the address instances from the postcodes. Although not exposed by the WFS, additional information could be used to configure provinces as resources, too, which might make it more useful to a human user browsing through the data set.

[11] Note that a paging approach will also be necessary for a general solution as many WFS instances limit the number of features returned by a query. For example, in the ELF project we see limits between 500 and 10000 features per query.

- the result of a CSW request as presented in the capabilities should be converted to JSON-LD annotated HTML,
- the schema.org vocabulary currently doesn't model "Things being part of a Dataset", we need to extend the schema.org vocabulary to model the relation between Thing and Dataset, then we can transform the dataset metadata to the schema.org/Dataset verb.

An alternative use case for the metadata is the opposite route. The catalogue in which the metadata is stored can be registered at google (using a sitemap linking to HTML representations of each metadata). Each representation of a dataset metadata can provide crawlable links to the datasets via the WFS-proxy, so a search engine will automatically crawl the WFS content. In such a scenario there is no need to register individual datasets on search engines.

Representations

JSON-LD for schema.org, RDF/XML for GeoDCAT-AP

The requirement is that the JSON-LD representation shall be "programmable", i.e. be structured data that can be used in applications, for example, for selection, navigation and processing purposes.

All resources will use the schema.org vocabulary. The geospatial metadata resources will also use the GeoDCAT-AP vocabulary. The GeoDCAT-AP representations will not use JSON-LD to encode triples, but RDF/XML.

Task 1: In order to meet this requirement, we will define a general mapping from the GML application schema definitions to RDF - using the schema.org vocabulary as far as possible. JSON-LD will be used as the serialisation of the triples.

As we cannot determine the mapping to the schema.org vocabulary automatically, this will have to be configured when setting up the 'crawlifier' proxy for each service. An idea we have and hope to realize in this project is a panel in GeoNetwork where the data provider can assign schema.org annotations to dataset attributes and store this as a ISO 19110 feature catalogue metadata. The proxy can use this information to transform the WFS data to schema.org.

Regarding the use of the schema.org vocabulary: The main search engines and other sites provide information in how far they support the vocabulary[12]. In addition, there is research available that provides additional insights[13]. We will take this information into account when identifying the mapping of the data to the schema.org vocabulary. We will also work together with the topic 3 contractors regarding their results for an extension of schema.org.

For all resources and properties that do not map to schema.org, we will analyse and decide in the design phase how we extend schema.org[14] (for the testbed, we will use external extensions).

We will reference the data set metadata resources from the spatial objects using a link. We will also have to specify a mechanism to include additional links to other resources that cannot be derived from the source data alone. We will examine mechanisms for this as part of our work on the testbed.

As a result, all the feature properties, schema.org information and additional links will be included in the JSON-LD representation of each feature.

Using a separate Linked Geo Metadata service, we will create resources that contain geospatial metadata every time a CSW service is called. The resources are instances using the schema.org or GeoDCAT-AP vocabulary.

Task 2: We will convert a popular/basic registry to Linked Data. Which registry to use is open for discussion[15]. The vocabulary that we will use for this registry will be schema.org and if needed extensions that need to be determined.

We will define the links that can be created between the spatial objects of INSPIRE BAG and Farmland and the registry. We will investigate to what extent this can be automatized. A common approach is to model both datasets using the schema.org ontology and model the link between the two feature-types as schema.org/location. To detect which location to use for a registry record should be managed by finding similarities in the labels from the spatial objects and the labels of the resources in the registry (for example by address and/or postal code).

In addition, we will investigate to what extent the link creation process can be part of the data conversion as performed by the 'crawlifier' proxy. As already outlined above, link creation is a time consuming and often manual process.

[12] e.g. http://microformats.org/wiki/search-engines, https://developers.google.com/structured-data/, http://www.bing.com/webmaster/help/marking-up-your-site-with-structured-data-3a93e731

[13] e.g. http://dws.informatik.uni-mannheim.de/fileadmin/lehrstuehle/ki/pub/Meusel-etal-Schema-org-Adoption-WIMS2015.pdf

[14] https://schema.org/docs/extension.html

[15] We noticed most of the basic registries (we expect the dutch term "basisregistratie" is meant here) are either not available as open data or they already have geometry included. Potential candidates are handelsregister (overheidsdiensten) or alliander energy usage data.

HTML (with embedded annotations)

There are two key requirements:

1. The HTML representations shall be semantically enriched with schema.org / JSON-LD annotations in order to improve their indexing.

We will implement the recommended approach of embedding JSON-LD and WebComponents. The HTML contains the embedded JSON-LD and includes a WebComponent for the feature type that takes care of the visualisation of the data. Search engine crawlers may parse the embedded JSON-LD with schema.org information. The JSON-LD will be the same representation as discussed above.

2. The HTML representations shall be informative and pleasant to consume.

We will be using commonly used libraries to generate HTML representations that are understandable, can be consumed also on mobile devices and include all the information and links to other resources in a clear, structured way. Resources with a geometry will also contain a context map.

We need to explore, how much information should be included in ‘index pages’ that link to multiple resources, e.g. from a postcode to the addresses.

Additional Representations

GeoJSON and GML representations for features and feature collections will be supported[16], too. GeoJSON may be easier to handle for some developers expecting more “plain” JSON than JSON-LD offers.

Discussion

As discussed in the motivation, there is overlap between the four research topics with regard to design decisions on how spatial data is made available, and also maintained, on the web. Therefore, it will be beneficial to discuss thoughts and design ideas with Geonovum / the sponsors, the other participants and other interested parties. The work plan foresees this interaction and we welcome this.

This is in particular relevant when we consider that research topic 4 covers only aspects of the bigger picture. A key aspect for us will therefore be, that the design does not place obstacles on the intended workflows for the environmental act.

At the same time, we have to consider that the testbed is executed with limited time and we reserve the right to make design decisions for our topic in order to be able to deliver the results in time.

In the interest of keeping costs low, we would welcome, if web-meetings and other web resources are used as much as possible. However, if there is a need for physical meetings we are prepared to attend them.

Implementation

interactive instruments has significant experience with WFS proxies, including a product (XtraProxy for WFS, see references) that implements a WFS client. A key difference is that XtraProxy provides access to data in a WFS via a GeoServices REST Feature Service supported in particular by all ArcGIS clients, while the ‘crawlifier’ proxy will provide a simple, read-only REST interface for all the resources specified in the design. The interface will support content negotiation for determining the representation for the response. Feature data will not be cached in the proxy.

If we win the assignment for topic 4, interactive instruments will make core modules used in XtraProxy available in public GitHub repositories under an open source licence. This includes the WFS client module, which is able to access WFS 1.0, 1.1 and 2.0 as well as GML 2.1, 3.1 and 3.2.

This will be the basis for the implementation of the ‘crawlifier’ proxy on the WFSs. The necessary implementation tasks are mainly determined by the design, i.e.:

- configurable mapping from the GML application schema to JSON-LD / schema.org
- configuration of additional, intermediate resources
- configuration of rule-based additional links to external resources
- the JSON-LD and HTML generation
- the simple, read-only REST interface

Instances of the ‘crawlifier’ proxy demonstrator implementation will be set up for the INSPIRE BAG[17] and Farmland WFSs[18].

[16] There may be limitations to GeoJSON output, if the source data is not compatible with the limitations of the GeoJSON format.

[17] Note that the INSPIRE BAG WFS may be an INSPIRE compliant direct access download service, but the data is not yet in the interoperable representation according to the INSPIRE data specification for addresses.

As stated in the invitation to tender, there are often small issues with WFS deployments that may make them hard to use. We have significant experience about this as a result of our work with WFSs. The WFS client in XtraProxy / the 'crawlifier' proxy is able to "handle" a number of typical issues.

For facilitating metadata crawling we suggest to research two scenarios:

- In the first scenario we create a proxy that forwards the CSW request to a CSW server and transforms the result into either schema.org as JSON or HTML (with embedded JSON-LD) or GeoDCAT-AP as RDF/XML.
- In the second scenario we'll set up an instance of GeoNetwork (as used in nationaalgeoregister.nl) which will offer capabilities to expose ISO 19139 data as schema.org in HTML (with embedded JSON-LD) or JSON and GeoDCAT-AP in RDF/XML. Content negotiation or the output schema parameter of the request is used to determine which ontology and format should be returned.

We'll use a reconciliation service such as RDF Refine or other (see "Link discovery" in the section "Motivation") to establish links between the data sets served via the WFS proxy of task 1 and the basic registry published as linked data of task 2.

Currently we see that in addition to links between resources in the three datasets of task 1 and 2, we will provide example links to third party data sets not listed in the description of the research topic: links to Geonames, DBpedia resources and existing Linked Data sources from Dutch government.

Validation

In this activity we will test the demonstrators against the requirements:

- review them from the view of the relevant environmental act use cases
- review the HTML representation ("informative and pleasant to consume")
- verify that all resources are indexed by Google and one other search engine
- verify that the presentation of the search results is so that non-expert users can understand and use the resources referenced
- investigate, if and in how far the annotations improve the ranking in searches in search engines as well as help answering more structured questions
- analyse the performance of the 'crawlifier' proxies

One expectation is to analyse and demonstrate in how far search engines can respond properly to structured questions[19]. In general, Google/Yahoo/Bing/etc. support structured queries only to a limited extent[20] or via additional APIs[21]. We therefore see this as a two-part activity.

First, we plan to investigate the responses of Google, Yahoo and Bing as the most commonly used search engines to such 'structured questions' on the data and if we can optimise responses by adapting the design or by providing recommendations on how to ask questions[22]. The main result of this activity is to get a better understanding what can be done today and what limitations exist. This is clearly also a topic that is relevant across the different research topics.

Second, we may extend the 'crawlifier' proxy, for example with additional resources for such structured queries, mapping them to WFS queries, in order to demonstrate how processing structured queries could be supported.

We will analyse in how far WebComponents are supported by current browsers (desktop and mobile).

As needed and if time and resources allow this, we will improve the design and implementation based on the validation results.

Reporting

We will summarize all findings as stated in the invitation to tender and this proposal.

In addition, we will also document and explain common WFS issues that we have seen from our experience with WFS. This will include cases which the 'crawlifier' proxy can deal with and other that will require a change in the WFS configuration. We will also include information how these and other issues can be identified by providers, e.g. using validation tools like OGC CITE WFS 2.0 and ETF INSPIRE WFS 2.0.

[18] As a test, we have set up proxies for the two WFSs using XtraProxy (BAG and Farmland), i.e. making the WFSs available to the ArcGIS platform. Here is a web map in the ArcGIS Online map viewer with the proxies dynamically accessing the WFSs. Note that these proxies run on a demo server and may not be available.

[19] Note that the 'structured questions' in the invitation to tender do not fit with the data in the two WFSs as these do not provide information about building types or dates. Therefore, other structured questions will be used.

[20] For example, https://www.google.com/advanced_search

[21] For example, https://developer.yahoo.com/boss/

[22] We will also try to contact authors of research papers and contacts in the companies as needed to clarify questions.

Mapping of requirements

In addition to the list of sections mentioned in the invitation to tender, we have also this section to provide a cross-reference how and where we have addressed and will address each requirement.

Task 1

Requirement	Comment
Design and implement a 'crawlifier' proxy that wraps a WFS that conforms to the requirements of OGC WFS 2.0 and the INSPIRE technical guidance and provides access to its data and metadata in a format fit for consumption and indexation by crawlers/spiders and non-expert users.	Fully covered, see "Design".
Such access should also be provided for dataset metadata in CSW ISO AP format that is linked from the capabilities document - and typically stored in a catalog service. The research shall assess how much of that data can be used or is useful for indexing and information needs of non-expert users.	Fully covered, see "Design" & "Implementation".
The INSPIRE BAG and Farmland WFSs referenced below should be used for this activity.	Fully covered, see "Implementation".
The data shall be accessed live from the WFSs / CSWs and in general do not cache data.	Fully covered, see "Implementation".
The proxy should not interfere with the existing OGC services, i.e. these shall continue to be available as-is.	The WFSs and CSWs may be used as they are. The proxies sit "on top" of the OGC services.
Make individual spatial objects 'crawlable' by serving at least HTML and JSON-LD representations of each objects as well as an index of all data (for an example of an index of an object type, see e.g. http://bag.kadaster.nl/verblijfsobject/). These resources shall include links whenever they are reasonable and can be derived from the WFS and CSW responses.	Fully covered, see "Design".
The metadata records provided by catalog services shall be made available in HTML and JSON-LD representations, too. In addition they shall be made available as RDF conforming to DCAT and the draft GeoDCAT AP (see references for a script that converts ISO metadata to GeoDCAT AP).	Fully covered, see "Design, JSON-LD for schema.org, RDF/XML for GeoDCAT-AP".
Additional representations beside HTML, JSON-LD and RDF (RDF only for metadata records) are welcome, but not a requirement.	We also include GeoJSON and GML representations for features and feature collections, see "Design, Representations".
The 'crawlifier' proxy shall support HTTP content-negotiation.	Fully covered, see "Implementation".

The resource representations provided by the ‘crawlifier’ proxy should provide reasonable links to other resources provided by the proxy including metadata resources.	Fully covered, see “Design, Links”.
Indexes of spatial objects, and maybe metadata records, may also be needed based on other criteria (e.g. aggregations based on location or key attributes) in order to improve findability. The research should provide recommendations regarding the resource structure and the linking between the resources.	Fully covered, see “Design” and “Reporting”
The HTML representations shall be semantically enriched with schema.org / JSON-LD annotations in order to improve their indexing as described by Google.	Fully covered, see “Design” and “Validation”.
The HTML representations shall be informative and pleasant to consume.	This is considered in the “Design” and will be validated and, if necessary, improved. See “Validation” and “Discussion”.
The JSON-LD representation shall be “programmable”, i.e. be structured data that can be used in applications, for example, for selection, navigation and processing purposes.	Fully covered, see “Design”
Investigate, if and in how far the above measures can improve the ranking in searches as well as help answering more structured questions like the examples above. Identify limitations.	This is investigated in the “Validation” activity.
Assess if and in how far DCAT and Geo-DCAT AP may help in realising the goals of this assignment. Provide suggestions for improving Geo-DCAT AP, if applicable.	This will be investigated in the “Validation” activity
The ‘crawlifier’ proxy should be easy to deploy and have minimal impact on existing infrastructure.	XtraProxy is easy to deploy and as the proxy for the WFSs is built on the same core, it will be easy to deploy, too. GeoNetwork is easy to deploy on tomcat or self contained. There is no impact on existing infrastructure.
The ‘crawlifier’ should use common web-technologies to simplify adaptations by web developers and server administrators to meet their needs.	XtraProxy is a modular Java application based on OSGi that was designed with extensibility and adaptability in mind. It uses common web-technologies like Jersey and Jackson to provide JSON/REST services or Bootstrap and AngularJS for client components. The same approach would be used for the ‘crawlifier’ proxy for WFS. GeoNetwork uses either XSLT or Groovy scripting to configure the schema transformations. Adaptations are done in “schema-plugins” that are maintained outside the core product. The skin of the application is based on AngularJS/Bootstrap, the template system in AngularJS facilitates easy customisations of the skins.
The performance of the ‘crawlifier’ proxy should	XtraProxy basically offers the same performance

be close to the WFS and CSW performance.	as the WFS and we see no reason why this would be different in the crawlifier proxy prototype. Performance aspects will be analysed in the “Validation” activity.
Provide input to research topic 1 regarding the experiences from executing the task.	We confirm that we will provide input to research topic 1. This is part of the “Discussion” activity.
Many possible issues with WFS could possibly be solved with good examples, documentation, fixing small bugs, removing ‘special cases’ and exceptions from the WFS specification, and so on. Provide such solutions when possible.	The WFS client in XtraProxy is able to “fix” a number of common issues in WFS deployments. These will also be available in the crawlifier proxy. In the report we will list and explain the issues. We will also include information how these and other issues can be identified by providers, e.g. using validation tools like OGC CITE WFS 2.0 and ETF INSPIRE WFS 2.0.
Verify in how far users can make better use of popular search engines for questions like: ● “give me all buildings located in Valkenswaard” ● “give me all schools located in Valkenswaard” ● “give me all buildings built before 1960”	This is investigated in the “Validation” activity.
Verify that spatial data published by the Dutch government is included in search results of popular search engines and in a way so that non-expert users can understand and use the resources referenced from the search results.	This is investigated in the “Validation” activity.

Task 2

Requirement	**Comment**
Implement and release a Linked Data version of a popular/basic registry.	Fully covered, see “Design, JSON-LD, Task 2”.
Establish links between spatial objects (from BAG and/or the Farmland data in task 1) and non-spatial data from the registry.	Fully covered, see “Design, Links”.
Discuss approaches for establishing, maintaining, representing links in such a set-up with obstacles and/or benefits from both the point of view of providers (of the spatial as well as the non-spatial data) and users.	Fully covered, see “Design, Links”.
The different approaches should be documented with worked examples and may require extensions to the ‘crawlifier’ proxy from task 1.	Fully covered, see “Design, Links”.
Investigate what is required to make linking spatial data usable and programmable for web-developers and non-expert users.	Investigating this, i.e. how programmers and end-users can add/delete links, is a very open topic and could be a research topic on its own. We will include in the report considerations on how this might be supported in an appropriate way in extensions to the architecture used in our

	implementations. In general, we feel that this is a subject that fits mostly to research topic 1.
Identify requirements so that linked properties are retained after user downloads the data (include provenance information, link to metadata, covered by the metadata requirement from above).	Fully covered, see “Design, Links”.