Geonovum - *Spatial data on the web*

Topic #2
A usable spatial data publication platform

Spotzi
Veilingdreef 17
4614 RX, Bergen op Zoom
Nederland

Authors:
Dhr. T. van Sprundel – Lead Developer
teun@spotzi.com
Tel.: 0164-240000

Dhr. M. Pegels – Senior data analyst
matthew@spotzi.com
Tel.: 0164-240000

Dhr. ing. R. Dolman – CEO Spotzi
remco@spotzi.com
Tel.: 0164-240000

Date:
March 17th 2016

Index

1. Introduction

Geonovum has been working for years towards the realization and success of the public sector spatial data infrastructure. The intended users of the SDI however are experts in the geospatial domain. The OGC standards cover the full range of geospatial use cases – some of which are unavoidably complex. Seen from outside the geospatial domain, the data behind the OGC services is part of the "Deep Web", because the data is published behind specialized web services and not readily available for the majority of web developers. In this testbed Geonovum asks the following question: "In which way can, in an evolutionary way, the current SDI be leveraged and the majority of web developers be reached as well." (*Geonovum invitation to tender – Spatial data on the web, 2015*)

This research is split into different topics that all answer their own questions. Spotzi is researching the second topic of the research: "**A usable spatial data platform**".

The leading perspective of this research topic is using the data, i.e. the user leading perspective. Its goal is to find out how to make spatial data easy to find and more specifically to explore the idea of 'government as a platform' i.e. make data easier to use by providing not only the data itself but also a community surrounding it. Given the fact that now there are disparate data sources full of opaque data, what needs to be done differently and by whom?

An important platform to compare with is OpenStreetMap (OSM). In recent years OpenStreetMap has become an ubiquitous platform for storage, management and extraction of geographical data. The quality of the data has increased considerably. An ever growing number of companies is using it to power their innovative geo-services. OSM's already rich tooling ecosystem is expanding at a considerable rate as more and more companies and organizations are contributing resources. Finding, querying and downloading data from OpenStreetMap has become straightforward and can be done in many different ways. Plugins that read data from OSM exist for many popular GIS, data visualization tools and client side frameworks.

A big part of the research done by Spotzi for this topic is answering the questing; "Is a platform like OpenStreetMap suitable for use as a modern dissemination platform for open data?"

2. Spotzi

Big Data. Open Data. You've probably heard of these terms. Every day new data sources see the day of light. Information that is extremely valuable for analyzing, marketing or for private use. All this information is published in different formats and on different websites. Companies and individuals mostly have insufficient technical knowledge to gather the data that might be valuable to them.

With a young and dynamic team in the Netherlands and Canada, Spotzi is continuously working on solving data issues. To ensure that we do not get lost in the huge amount of data we focus on gathering geographical information. We offer an intelligent and rich web- and mapservice via which our clients can retrieve their data or push their data.

Next to offering data the visualization is even more important. Then you will recognize patterns you didn't know existed. Our clients however are no Data or GIS specialists. We therefore entirely focus on making mapping easy.

Our clients are diverse and develop innovative solutions around our data. Real Estate website JAAP.NL shows our extensive neighborhood information next to each real estate listing. Insurance companies use our data to create insights into the risks they take.

3. Motivation

We have chosen to tender for research topic 2, because the product Spotzi already build matches the goal of Geonovum to deliver data instead of plain WMS and other image server. As the product we build is based on Open Source software, Spotzi is willing to share its knowledge with the public domain. Geonovum is the perfect platform for Spotzi to achieve this.

Spotzi has chosen to use Open Source Software and Open Data to build the current platform: a datashop and data visualization platform. The goal of Spotzi is the same as the goal stated in this research topic: let the community add data but in a controlled matter. Spotzi promotes to share as much data as possible. Spotzi is offering a free account to parties who are willing to share their data.

Currently we have chosen to review the uploaded data by our members and only accept new datasets and not updates. Geonovum's proposal of using a two-tier tagging system is an interesting one as this addresses the issue of how to handle updates on existing datasets without the datasets being affected.

Secondly Geonovum's approach of using OpenStreetMap as a "platform" attracted out attention as Spotzi was thinking of methods how to return data into the community instead of keeping it within the Spotzi platform.

4. Approach

In this research Spotzi tries to identify the components that constitute a user- and publisher friendly data dissemination platform. To do this Spotzi is using their own data publishing platform as a basis for this study. The Spotzi Mapbuilder platform is a datashop and visualization platform based on Open Source Software and Open data.

In the tender invitation for this research topic Geonovum describes 9 tasks that need to be completed in this research topic. Spotzi has divided these tasks into four themes:

- User management
- Workflows
- Searchable data
- Linking data

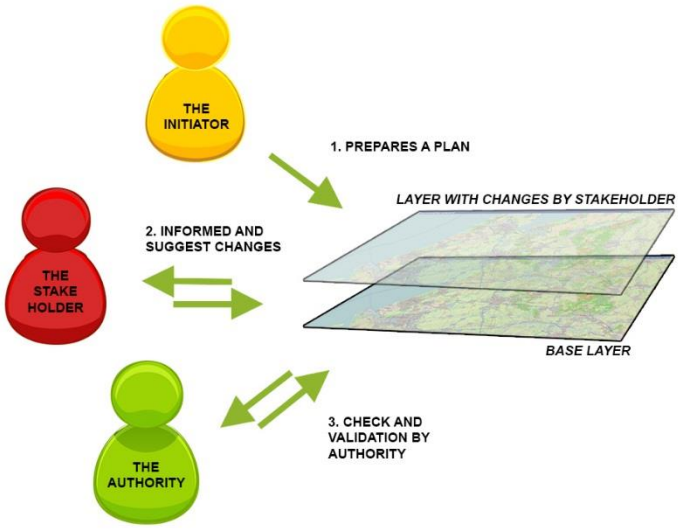
In the next chapter we explain what we've done for each theme to find the best solutions for a usable spatial data platform.

5. Research

5.1. User Management

User management is an important part of a usable spatial data platform. The first 2 tasks of this topic are trying to find a solution how to work with different users and different datasets. Whereas the strength of a publication platform like OpenStreetMap is its crowdsourcing capabilities, authoritative data should not (always) be directly editable by the public. Governments want to use feedback from citizens about their data, but this should always be checked and approved prior to inclusion.

Spotzi investigated the use of a two-layer system around each dataset. The first layer is the main layer of data which can be edited by the authorized party. The second layer is a data layer on top of the main layer where the community can edit or propose changes.



5.1.1. Different kind of users

One of the issues we've encountered in our research is the different kind of users that participate in this platform. If we take a public notification as example, we have the government on one end and different end users on the other end. The government has GIS experts that fully understand all the GIS functionality, but also have officials that have to examine incoming applications without any GIS knowledge. Conversations with different parties already revealed that this is an issue that governments already encounter with their own systems. A lot of time is spent in explaining how the systems work to non-GIS experts.

The same difference in expertise can be found on the other end of the table. The different end users may have (some) GIS knowledge, but this cannot be guaranteed.

This leads to an important conclusion. If we want a spatial data platform that can be used by different kind of users with different knowledge levels, we need to keep working on a platform which is as simple as possible. Spotzi encountered this same issue when it was sharing its own data with the community. From following users that use the system, Spotzi tried to improve its platform step by step. The same applies for this spatial data platform.

5.1.2. Different authorization levels

The next issue in this theme is the difference in authorization levels. As we've stated before, governments don't want citizens to edit data that they are responsible for. For some datasets the authority does not even want it to be visible to citizens.

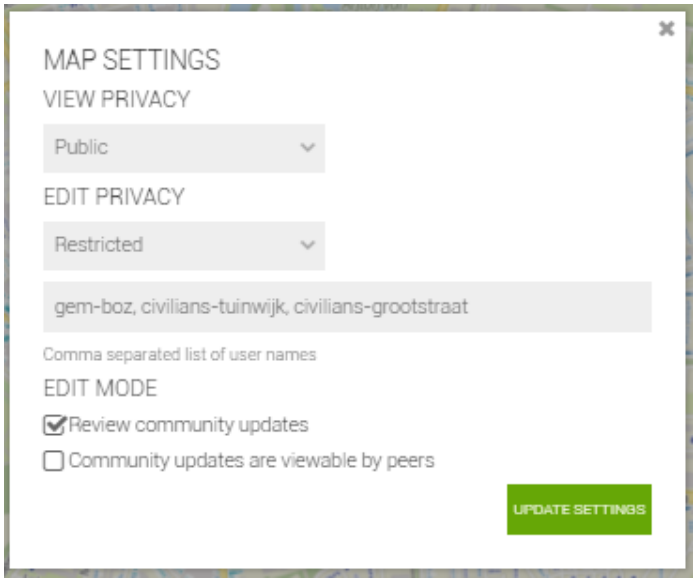
To address this issue, Spotzi created different authority levels in the platform. We've started with the following levels:

Visiblity:

- The data is visible to everyone
- The data is visible to a selected group
- The data is visible to no one

Editability:

- The data is editable by everyone
- The data is editable by a selected group
- The data is editable by no one



Furthermore, we've added the possibility to work with proposed edits. The authority of a specific data set can allow others to make proposed edits, but not edit the data source directly. We've also added an option to control if users are allowed to see propositions from other users. Again it's the authority that has control over these settings.

There are however more levels of authorization that could be implemented in the platform. An example of this is allowing an end user to only see certain records of a dataset, like their own house.

An other issue was matching the platform to the authorization levels that already exist in the current systems of governments. In the current systems, authority levels are already set for each department/user. If a government is to work with the spatial data publication platform, they do not want to set these authority levels again by hand. As each government has his own system, Spotzi could not create a solution for all systems. To address this issue we've tried to make it as generic as possible and Spotzi made it possible to set the authority levels by webservice. This way governments can set their authority levels with the use of a small bit of programming.

5.1.3. Data quality

Next is the issue of data quality. A government is obligated to deliver data with a very high data quality standard. This certainly applies to key registrations. Various organizations with public functions are obliged to use this data in different processes. If there are errors in these datasets problems could arise for civilians, businesses and governments.

According to the Dutch government *(Kwaliteit van basisregistraties, Digitale Overheid 2014*), providing good data quality not only means an authority should deliver good data, but they should also provide good reporting about their data quality levels. In a transparent way there should be explained how the quality is measured and how civilians can get more information about the data quality of a dataset.

Furthermore, authorities should deliver insights in how they are improving the quality of their data. All of these points create a rather complex situation when you're trying to use input from other sources where data quality is not guaranteed.

Spotzi proposes two solutions for this problem, depending on the amount of user input. When you have a large quantity of user input, you could say that this should improve the data quality by itself. An example of this is Wikipedia. There are errors in the data, but with a big community these errors get sorted out really quickly.

However, in practise authorities like the Dutch municipalities do not get that much user input. We could improve the quantity of user input by connecting to existing platforms. We could also improve the reliability from certain users by looking at their status or previous input.

5.1.4. Understandable data

The last issue we've addressed in this theme is making data understandable. As we've stated before in 5.1.1 a lot of different users with different knowledge levels can use the platform and work with the data. It is therefor very important that we make sure the users understand the data they are working with. This understanding reaches out over different levels.

At first we must make sure the end user understands the dataset that they are working with. During our research we've encountered different experiences on this topic. Authorities can struggle a lot with explaining their datasets, to either civilians or to colleagues from other departments. Some authorities spend days explaining their data, while others just assume it's clear. Other authorities cannot identify their selves with this problem. Anyhow, this remains an issue that we need to prevent.

Different datasets can look the same, but have different meanings. For example, a dataset with sewerage can look the same as a dataset with cables and pipelines. Therefor at least a title and description are needed to give a short explanation about the dataset. In addition to that we've showed that you can also use links to non-geo environments to give more information about datasets. We've created a wiki that has links to formal definitions that explain a data set.

Besides explaining what a dataset is, it is also important to make sure end users understand the data itself. The government often uses different descriptions then it's end users. I.e. whereas the Dutch governments describes a certain feature as a "lichtmast", most civilians would describe this as a "lantaarnpaal". Formal definitions are not the most used names for all objects. If authorities want to explain their data, they could add aliases to their data with popular definitions.

Adding titles, descriptions, links to non-geo environments and using aliases are ways Spotzi proposes to make sure that the end-user understands the data. If authorities believe the use of these methods are not sufficient to explain the data, their data should probably not be shared with this group of end-users. In the end an authority has the responsibility of choosing the audience which he shares his data with.

5.2. Workflows

An other important theme in our research was to investigate how suggested changes/updates could be used as input for existing workflows. And if it is possible to setup a toolchain between non-geo administrators and data owners to publish and maintain information.

5.2.1. How to connect with civilians/other users

If we want a spatial data publication platform to work, we need to make sure not only governments are able to use it. We need to connect to civilians and other users and if possible even to a whole community. We've investigated the positive aspects of the OpenStreetMap platform which is a big success.

The first and most important point that needs to be adapted to connect to end-users is to keep the platform simple. Spotzi has a lot of experience with end users of a geo-platform and most of them are not experts. If they struggle using the platform they will definitely stop using the platform and therefor undermine the success of the platform. Keeping the usage of a platform like this user friendly is an ongoing process. From the start we will need to ask for feedback from end users and investigate their patterns while working with the platform.

The second way to connect to end-users is to connect the platform to existing platforms that already are a success. To attract users, we will need to offer them something. Connections with platforms like "verbeterdebuurt" or creating a collaboration with local news stations can make end-users use the platform.

Lastly we advise to find new ways of data input. If the end-users don't come to you, you will need to go to them. Connect to social media sources could increase the input you get from users. Users don't need to log in to the platform to share their opinion. I.e. a user can tweet about something his municipality needs to do and the platform can be used to geocode this message and send a notification to the responsible government.

5.2.2. Use platform to initiate processes

To make a success out of the geospatial data platform, the platform should initiate other processes and not only be a marketplace for data. This initiation could be anything from sending an email to the responsible authority to making a call to a well defined web service.

The platform we've created is capable of doing all of this. When an end-user makes a proposal on a dataset both the user and the authority get an notification by mail or text message on their phone. This message describes the actions that have been made and offers the reader a direct link to the proposal. The link is also shareable with co-workers or friends, so it's easy to pass it on.

The same happens when a proposition is accepted or declined. Not only do the maker of the proposal and the authority receive a notification, it is also possible to send a notification to every user that has shown its interest in the selected object.

It is also possible to initiate automated processes. As an example we've created a simple webpage where the accepted proposals result in public notifications. This way a municipality can easily share its public notifications with the public without having to do any extra actions.

Authorities like municipalities can easily adopt this spatial publication platform in their existing workflows. With the automation of geo-processes, they can save time and share their geo information with their civilians.

Spotzi Geonovum <visualization@spotzi.com> | ldegroot@spotzi.com

Spotzi Geonovum - your map update proposition

Dear map owner,

You proposed a map update on Spotzi Geonovum. Please wait for the map owner to review your proposition. Below you will find details of the update.

Map	Permits (Rebuilding - bicycle stand)
Update type	Geometry edit
Update ID	c9a1aa50-e5f4-11e5-b7f2-00265522ea30-28
Description	Bicycle stand

You can track the update by clicking here. The update will be listed in the map updates window. Be aware: you may need to log in to your Geonovum account again in order to view the update.

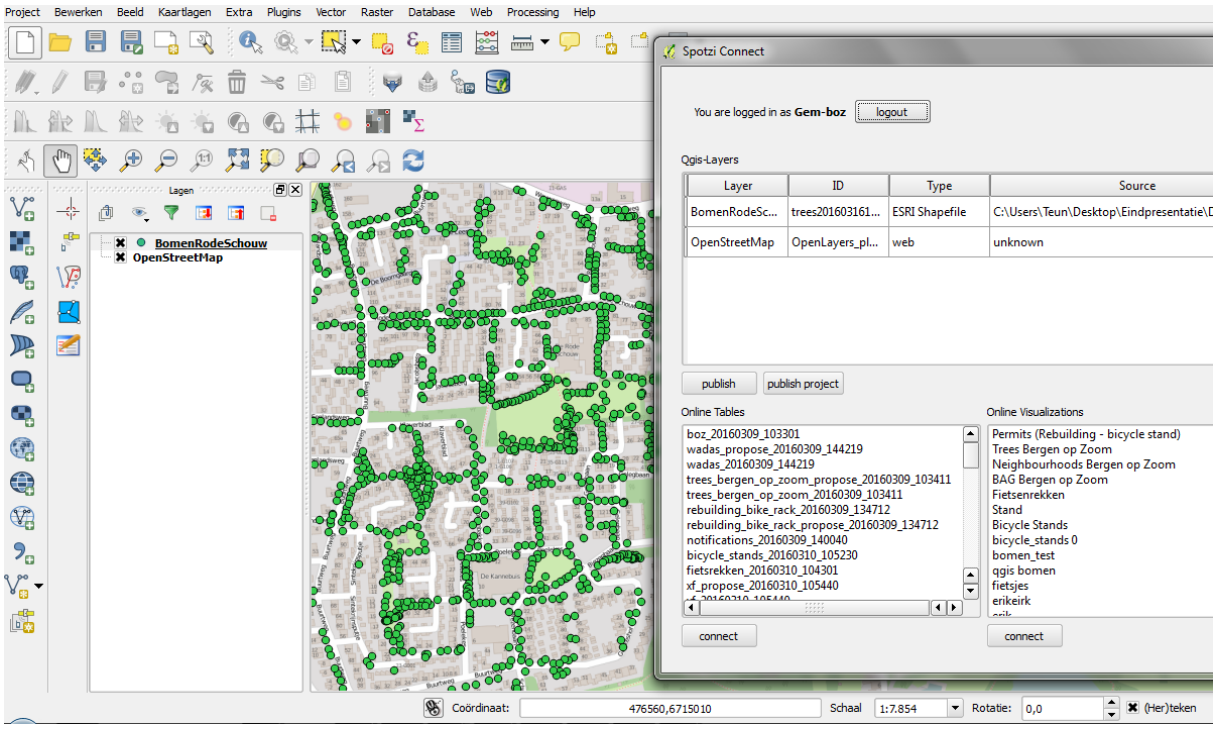
Best regards,
Spotzi

Veilingdreef 17
4614RX, Bergen op Zoom
The Netherlands

5.2.3. Different ways to connect to the platform

Because most authorities already have their own GIS systems the platform needs a way to connect to existing systems. There are a lot of different systems governments use, so we could not develop a one fits all solution. Instead we've developed the platform to be as generic as possible so it would not take much time to connect to the system.

As an example we've created a simple plugin for the open source GIS software QGIS. With this plugin we're demonstrating the possibilities of connecting to the platform. A QGIS user can login the platform within the plugin and can view his available datasets and maps.
He can download data for use in QGIS and can upload data from QGIS to the platform.



5.2.1. Coordinate systems

One of the tasks from the testbed topic was to investigate what has to be done regarding the coordinate reference system. Different authorities and end users all use different coordinate systems. In the Netherlands the three most used coordinate systems are:

Amersfoort / RD New	(http://spatialreference.org/ref/epsg/28992/)
WGS 84	(http://spatialreference.org/ref/epsg/wgs-84/)
WGS 84 Web Mercator (Auxiliary Sphere)	(http://spatialreference.org/ref/sr-org/7483/)

From our research we can conclude that it is very important to keep the platform as simple as possible for the end user. Concerning coordinate systems this means that the end user, both authority as end user, should not need to worry about spatial reference systems. Most of the users are not GIS experts and don't have any experience with spatial reference systems. The users should be able to import any dataset without any indication of the spatial reference system.

To make this possible Spotzi implemented a coordinate system recognition system which recognises the coordinate system a certain dataset is in when it's imported. The system is storing this information in the metadata of the dataset as its "imported SRID". The recognition of the coordinate systems only works for the three coordinate systems described above. It is possible to extend this list of coordinate systems.

After the data import the system projects and transforms the dataset to a single reference system. In our platform we've chosen to use WGS 84, so the platform can handle more than just Dutch datasets.

When the user exports the data, the system projects and transforms the data back to the original reference system. This way all the data in the platform will be shown in the same way, and end-users do not loose their original data while working with the platform.

5.3. Searchable data

Existing geo-data platforms like OpenStreetMap allow users to search in the data with the use of API's, like the Overpass API. In this research we've investigated several ways to make the data in the platform searchable. Our goal was to make the searching user friendly, so every end-user is able to use the search functionality, not only experts.

5.3.1. Search by address

The first and most used way to search data on a map is searching by address. From the experience we've already had with our own platform we can conclude that over 90% of the search requests are search requests based on an address.

In the Netherlands, we have got the BAG that can be used to find the right geo location that matches an address. Here it is fairly easy to find the location you're looking for. However, this does not apply to every country/region. The search by address function we've added to the platform therefor starts by requesting location data from Nominatim, an geolocator based on OpenStreetMap.

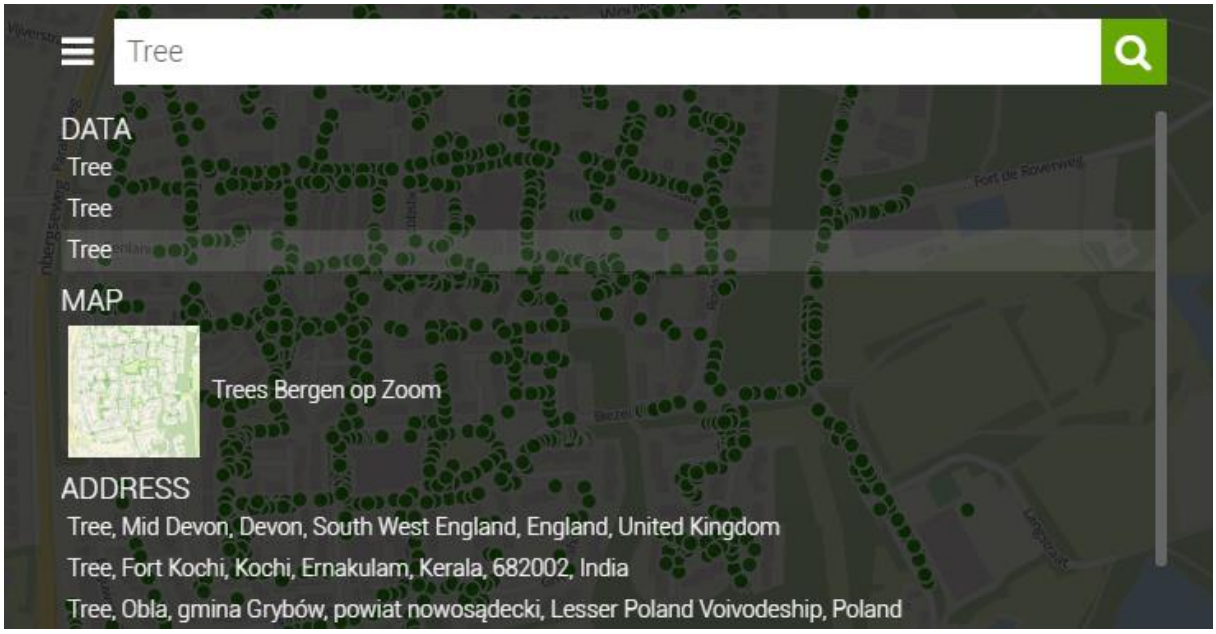
However, we're trying to increase the search results by handling other sentences that do not correspond to an address. For example, if an user searches for "the bakery between 5th an 7th street" we're trying to also return the location of this specific bakery. In co-operation with universities such as the University of Waterloo in Canada, we're striving to improve our geo-location possibilities. This is however a complicated topic and a research on its own.

5.3.2. Search by dataset

The next possibility to search in the platform is to search by dataset. We've allowed users to use the same search bar to find datasets in all of the datasets that they have access to. This not only includes their own datasets, but also community datasets shared by governments. By using this functionality it's easy for an end user to find the dataset they are looking for.

5.3.3. Search by record

That last and most interesting way to search in the platform is to search on the data itself. If an end-user types in a request in the search bar, the system fires queries to an API that sends on its turn requests to all the available datasets. If a hit is found, the search bar will show this result next to the possible datasets and addresses. If the end user clicks the result it will open the corresponding map with a marker on the requested record. This way the user can easily find the data it's looking for.



5.4. Linking data

The last subject of our research concerns linking data. We have researched what opportunities and what difficulties we would encounter when we would try to use Linked Data in the platform. As Spotzi we were not familiar with Linked Data at the start of the project. Our approach was to develop a generic way to create linked data for every dataset that was used in the platform. Furthermore, we wanted to use this linked data to create new functionalities for the platform. However, we have learned that this is very complicated.

5.4.1. Opportunities

The strength in using Linked Data for this platform is that you can combine different datasets to create new information. We've found three opportunities to use linked data in the platform.

At first, we've shown that it is possible to link to non-geo data and its formal definitions with datasets to explain the contents of a dataset. We could link records from a local tree dataset to the formal BGT descriptions. This way even if the formal definitions would change, the dataset would still be accurate.

Secondly, we've tried to use linked data to link the proposition layer to its corresponding data layer. Every proposition has its own unique ID and therefor this was an ideal opportunity.

Our last idea was to use linked data to link all available datasets for a user i.e. if the user looks at a property on the BAG dataset the platform should notify the user that there are also other datasets that have information about this subject. Although we've not yet figured out how to properly do this, this idea is something we will keep working on in the future.

5.4.2. Difficulties

The main difficulty we've encountered was the big variety in datasets that are used by the platform. And more specifically the unknown factor about this. A platform suitable for government, business and civilians can be used for so many different possibilities that it's impossible to create a linked data solution to all of these datasets at once.

It is certainly possible to adopt linked data for one or two datasets in this platform. However, if we want to develop a generic solution that fits all datasets we could not discover more similarities then a name, type and description that most datasets would use. We could link these objects to formal definitions described on websites like schema.org. However, most of the linking should still be done by hand.

6. Challenges

Spotzi has taken a lot of effort into the development of this application for the Geonovum Geo4Web testbed. It took us a lot of preparation and research before we could actually start the development of a tool, which both citizens and municipalities could understand.

This is a good starting point for the description of the first of three challenges, which Spotzi has encountered during the development of this particular tool for the testbed.

Challenge 1: The data and the usability of the tool should be as understandable as possible for both citizens and municipal / government.

In short, the usability of an application must be instantly understandable for the citizens, but especially for a community / government. For example, the BGT makes extensive use of a particular jargon to generalize the appointment of object types to a greater or lesser extent. This jargon should make geographical data at least understandable for a municipality and government. However, we cannot assume that a standard citizen is also aware of this jargon. This can be confusing and will, ultimately, also have influence on the operation of this application tool where communication between municipalities and citizens is a key feature.

However, Spotzi does have some ideas to minimize confusion in the communication between municipalities and citizens. Therefore, Spotzi will try to match each map to a wiki that explains each object on the map. This wiki contains explanations and various national geo-standards that are published by the government. With this wiki, object types on the map can be looked up by municipalities and citizens at any time.

Additionally, Spotzi advises to use aliases of object types provided by the government (SKOS), but also by the community itself. For example, In the case of changing a road section, a citizen may have different names for this object type. ‘ROAD' does also have different names in certain languages. Synonyms for "ROAD" are for example: "LANE", "AVENUE", "STREET" etc., This challenge is already discussed several times within the Geonovum GitHub community with ways to streamline this information. On the one hand we can look at the more frequently visited pages to extract the general terms used in the map.

If these names are added in the application, we can better streamline the general name for an object. Citizens will know what names are normally used, and the municipality has more insight into the names that could use a civilian.

These are just some solutions that are created to make the public data understandable among both municipality and citizens. However, this is a challenge that is large enough for a dedicated research.

Challenge 2: Retention of data quality if the citizen is going to interfere with the public data.

Maintaining the quality of this public data is an important task for local authorities and governments. Maintaining the quality of data requires a great amount of energy and time. With this tool, we want to make it as easy as possible for municipalities and governments to maintain the quality of the public data. In addition, the sharing and editing of this public data by the community also plays an important role in this application.

Within the community, on the one side we have the municipalities / governments that want to maintain the quality of the data. And on the other side, we have the citizens who can edit the public data at all time. This may cause friction and impact the quality of the public data.

To maintain the quality, we want to introduce more roles and authority levels to use this application. So the community can change the public data at all times, but in a more controlled way for the municipalities and the government. The community adds this data to the maps, but we can't deny that there would still be issues, due to incorrect filling of new data by the standards of an example BGT.

How this works in practice, should be investigated. This challenge not only raises a lot of questions, but it is also important to note that this tool should be tested in a practical. It must ultimately be clear who has the final responsibility for the changed data.

Challenge 3: Linking data

During this testbed research we've learned a lot about Linked Data. Even though we've concluded that it is not possible to create a fully automated way to create linked data for every dataset, we could still see many opportunities. Especially if web crawlers like Google's can automatically understand this information. This is something that is being researched by the other testbed topics.

If the above is working it would certainly be worth to further investigate the possibilities that Linked Data could offer. Not only would Google be able to inform you about for example houses that are for sale, but it could also give you insights on other datasets that could be linked to that location. Like a felling permit that has been granted for the tree in front of the house or the public location plans on the opposite of the street.

The platform could be the basis where all the data could be linked and this would generate endless possibilities. This challenge is certainly one that needs further investigation. Hopefully this is something that could have a main role in the final testbed topic.

7. Conclusion and recommendations

We've started our research with the search for a usable spatial data platform to make (sharing) spatial data possible for everyone. After some months of research and developing we now present a platform that contains many of the success elements that made big spatial data platforms like OpenStreetMap successful.

With this platform we're allowing all kind of different users to work with and share spatial data in an easy way. There are still many point we can improve, but the basis is set. With the platform we've created we want to encourage both authorities as civilians to join the ride to improve this platform and find out the possibilities of geo-data.

The platform allows users to import or connect their data sources with and easily visualize their data. Users can search through their data and provide other users, both authorities as non-authorities, with their input and thoughts. Users can share their maps and data and show others what they are working on.

Also we encourage the participants of testbed 5 to always keep the end-user in mind. In our experience and opinion, it's not important to investigate what's possible, but to investigate what's needed and what will be usable.

We recommend the use of the basis we've founded here for further research and we're definitely willing to help where needed. We've put all the different components of our research on the geo4web testbed GitHub. We also encourage everyone to try the platform and to share their opinions and ideas for the platform.

Finally, we want to thank Geonovum for making this research possible and the participating authorities for all their help and input.