Intelligent fire risk monitor based on Linked Open Data

Nicky van Oorschot
Netage
nicky@netage.nl

Bart van Leeuwen
Netage
bart@netage.nl

ABSTRACT

Every day the Fire department of the Netherlands work hard to save people's lives. Therefore, they have been investing in Business Intelligence approaches for several years, to get more information for accident prevention and accident fighting. In this paper, Linked Open Data has been used as a business intelligence approach for the creation of dwelling fire risk profiles based on demographic data. During the research a Proof of Concept showed the appliance of Linked Open Data for this purpose. However the data have some quality mismatches, such as: outdated, accuracy issues and not 100% complete. Evaluation session proofed that the outcomes show similarities with a fire incident map and the gut feeling of several firefighters.

Keywords

Semantic Web, Linked Data, Open Data, Firefighting, Risk Assessment.

INTRODUCTION

The primary objective of the fire department is to save people's lives. Saving lives is about seconds. Complete information could make the difference. For instance, the maximum response time is determined by Dutch law (Binnenlandse Zaken, 2011). The factors which are taken into account to determine the maximum response time for a building are the type of the building and the year of construction. But the risk of a dwelling fire and the development of a dwelling fire depends on multiple factors. Some of these factors are not taken into account. Therefore, the use of "Business Intelligence" within Brandweer Nederland is an upcoming interest. Investments to harvest data and process data into valuable information have been made since the last couple of years. At the moment, Brandweer Nederland is working on business intelligence, loads of data are stored in data warehouses to get information on accident prevention and accident fighting (Nederland Brandweer, 2014). Besides the approach with data warehouses, Brandweer Nederland is highly interested in another approach: linked open data (LOD). "Linked Open Data (LOD) is a growing movement for organisations to make their existing data available in a machine-readable format. This enables users to create and combine datasets and to make their own interpretations of the data available in digestible formats and applications." (Bauer & Kaltenböck, 2012). Open datasets have grown exponentially (Stellato, 2012). Combining various datasets could provide more viable information needed for accident prevention and accident fighting. For example, research showed that residences which are more and better isolated (A+ Energy Label1) develop another type of fire, once the building is on fire (Schaap, 2013). Therefore, the level of isolation could be used to determine the risk of dwelling fire, and maybe even forecast the type of fire that will develop. The results of combining datasets will mainly be used to organise better fire safety education in specific neighbourhoods because of certain demographics. Therefore such system will be most valuable in preparation phase. The Linked Open Data approach differs from the warehouse approach. Data warehouses and their maintenance are expensive. With linked data one can gather the information from the original source when it is needed and combine it instantly (without the need to use various database techniques such as download and exporting data). Less investments are necessary in the technical infrastructure. The datasets are maintained by the source itself and provided by an API or by files. Combining more datasets means that new outcomes and insights are created or established outcomes and insights change. Moreover, change in outcomes could affect the determined response time or other decisions, which provide the extra seconds necessary to save somebody's life. This study examines whether linked open datasets could be used as a dynamic way to provide the fire department with quality intelligent information.

MAIN RESEARCH QUESTION

Does linked open data provide a qualitative and dynamic way to create a dynamic fire risk profile monitor for cities and neighbourhoods?

- Research Sub Questions:

1. Which datasets are necessary and which datasets are available
2. What is the quality of the data and is the quality sufficient?
3. To what extend is it possible to use linked data to get a solid demographic fire risk profile?
4. To what extend is it possible to create a dynamic linked data system, for the creation of fire risk profiles?

RESEARCH DESIGN

The research questions have been used as a guideline for the research. Depending on the identified datasets and the availability (format availability) of these datasets, the research has been affected. Based on statistical relations from the Fire department Amsterdam ("Handreiking sociaal woningbrandrisicoprofiel"), open datasets were gathered from the World Wide Web. These datasets were combined and linked to each other and used to develop a Proof of Concept, the outcomes of the Proof of Concept were validated by four firefighters from the safety regions Rotterdam-Rijmond and Midden- en West Brabant. Validation was based on their gut feelings and based on an older map with fire incidents (2005-2008) and a traditional fire risk map of the fire department.

OPEN DATA & OPEN DATA QUALITY

Linking data on the World Wide Web is important, because of the exponentially growing amount of data and information available on the internet (Stellato, 2012). Besides the term "The Semantic Web" another term has been around for the last decade "Linked (Open) Data". Heath & Bizer (2011) define Linked open data as a recent development in which all datasets that are freely available on the Internet are interrelated by using semantic rules and techniques in an effort to publish these in a machine readable way and to facilitate analysis in human understandable form.

OPEN DATASETS

Linked Data is about using the Web to create typed links between data from different sources. "*Technically, data has to be published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external datasets, and can in turn be linked to from external dataset*". (Bizer & Berlin, 2009). In the past years, organizations have been publishing their data in such a machine-readable way. More organizations are willing to publish their data on the web. Together the published open datasets form the global data space.

Berners-Lee, (2006) outlines a set of 'rules' for the publishing of data on the web. Rules to create a single consistent global data space:

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
4. Include links to other URIs, so that they can discover more things.

These rules have become known as the 'Linked data principles', and provide a fundament for the publishing of data adhering to the World Wide Web architecture and standards.

Furthermore, a network of URIs allows the creation of a network of Linked Data, thus contributing to the construction of a global data model, which is the Linked Open Data. (Bellini, Nesi, & Venturi, 2014)

LINKED DATA TECHNIQUES

Resource Description Framework (RDF) is a technique for data interchanging on the web. RDF is a graph-based data model. The core structure of RDF is a set of triples (Subject -> Predicate -> Object). This triple concept is known as the Triple A principle: "Anybody, can say Anything, about Anything". A concrete example of the triple concept applied on data would be: http://example.org/john --> http://example.org/isFatherOf --> http://example.org/george (Cyganiak, Wood, & Lanthaler, 2014).

In addition to RDF there are more semantic frameworks: RDFS & OWL. RDFS (RDF SCHEMA) is an semantic extension on RDF. The purpose of RDFS is to add more expressivity to RDF graphs by adding more

complex relations like: class. Subclass and resource types (Brickley & Guha, 2014).

In some cases the expressivity of RDF and RDFS are not sufficient enough to describe the complexity of a data model. Therefore, the web ontology language (OWL) can be used (Dean & Schreiber, 2014). For simple data exchange RDF is mostly sufficient. More complexity could be found for example in ontologies etc.

DATASET QUALITY

Data is only as useful as its quality (Zaveri et al., 2012). The challenge is to determine the quality of datasets published on the Web and to make this quality information explicit (Zaveri et al., 2012)

A popular definition of quality is “fitness for use". Therefore, the interpretation of the quality of any data item depends on who will use these data, and for which task the user intends to employ it. While one user may consider the data quality sufficient for a given task, it may not be sufficient for another task or another user. (Mendes, Mühleisen, & Bizer, 2012)

According to Zaveri et al., (2012), data quality is commonly conceived as fitness for use for a certain application or use case. Datasets on the Web of Data already cover a diverse set of domains, therefore data could be used for specific purposes. Whether a dataset fulfils the information need is defined as “Fitness for use”. Moreover, fitness for use is commonly used as a definition of data quality. (Debattista & Lange, 2014b)

Zaveri et al., (2012) define a Data Quality Assessment Method. A data quality assessment methodology is defined as the process of evaluating whether a piece of data meets with the information consumers need in a specific use case. The process involves measuring the quality dimensions that are relevant to the user and comparing the assessment results with the users’ quality requirements.

Both Mendes et al., (2012) and Ruckhaus et al., (2013) maintain three main quality dimensions and the metrics to make the quality explicit. The main quality dimensions are:

- Completeness
- Conciseness
- Consistency

Besides these three main quality dimensions, the role of trust is specifically mentioned by (Bizer & Berlin, 2009) as a quality characteristic.

Completeness

On schema level, a dataset is complete when it contains all of the attributes needed for a given task. On data (instance) level, a dataset is complete when it contains all the objects for a given task. Naturally, the completeness of a dataset can only be judged in the presence of a task of which the ideal set of attributes and objects are known.

Completeness on data level is named “extensional completeness”. Completeness on schema level is called “intensional completeness”. The extensional completeness (data level), can be measured in terms of the proportion of target URIs found in the output (Equation 1), while the intensional completeness can be measured by the proportion of target properties found in the output (Equation 2). Completeness in this study is about missing extensional and intensional data. Completeness is not necessarily an issue, as long as it is possible to get information of nearby objects to fill in the gaps in the dataset.

$$extensional\ completeness = \frac{||uniq.\ obj.\ in\ data\ set||}{||all\ uniq.\ obj.\ in\ universe||} \quad (1)$$

$$intensional\ completeness = \frac{||uniq.\ attr.\ in\ data\ set||}{||all\ uniq.\ attr.\ in\ universe||} \quad (2)$$

Figure 1; Equations completeness

Conciseness

On schema level, a dataset is concise when it does not contain redundant attributes (two equivalent attributes with different names). On data (instance) level, a dataset is concise when it does not contain redundant objects (two equivalent objects with different identifiers). The extensional conciseness measures the number of unique objects in relation to the overall number of object representations in the dataset. Similarly, intensional conciseness measures the number of unique attributes of a dataset in relation to the overall number of attributes in a target schema.

In this specific research case, conciseness means, that when two attributes or even objects are the same, it becomes harder to choose how to use the dataset. It makes the data set less understandable and more data have to be processed in order to get information (influences the performance).

Consistency

A dataset is consistent when it is free of conflicting information. The consistency of a dataset is measured by considering properties with cardinality 1 that contain more than one (distinct) value. Mendes et al., (2012) have defined the consistency of a dataset for a given property p to measure the proportion of objects that do not contain more than one distinct value for p, with regard to the universe of unique property values.

Error of data

An important aspect of the data is the error in the dataset itself. Errors are hard to locate. Mendes et al., (2012) use another dataset to locate errors. In their example, they use the Portuguese and the English version of DBpedia. By doing a consistency test between these two datasets, they find the differences and thus errors in the datasets.

Trust

A less measurable data quality dimension is Trust. Trust is mentioned by Bizer & Berlin (2009) in order for data consumers to assess the quality of data and to determine whether they want to trust the data. Data should be accompanied with meta-information (provenance) about its creator, creation data as well as the creation method. Provenance information together with a certain authority of the publisher of the data, will give the data consumer an impression of the data quality.

Dataset Alignment

Datasets are provided by different sources and in slightly different ways. Using Linked Data the focus is on RDF because it mainly involves simple data sharing, RDFS & OWL are not used because of their unnecessary complexity. RDF can be requested from the source, mostly by using SPARQL. SPARQL is not used by all open data sources since it is not required. Providing a spreadsheet file is the most used option for exchanging open data, which organizations use to provide data in an Open Data approach. Combining datasets which are provided using different methods is something that one has to keep in mind using Linked Open Data (Miličić, 2011). Structure and logic of datasets differ from dataset to dataset. It is important to understand the structure and logic of combining datasets, otherwise: "*you can end up with a dataset that you think is ready for analysis, but is really utter nonsense*" (University of Wisconsin, 2011).

Linked open datasets provide for example: semi structured data, unstructured data, documents in mark-up languages. The use of such a variety of sources could lead to problems such as inconsistencies and incomplete information (Debattista & Lange, 2014a). Kontokostas et al., (2014) maintain that datasets are of varying quality ranging from extensively curated datasets to extracted data of often relatively low quality.

FIRE RISK DEFINITIONS

There are more definitions of "risk", some definitions only mention realization of unwanted consequences to human life or property, while other definitions mention more aspects like for example harm to business continuity. Watts and Hall (as cited in Hadjisophocleous & Fu, 2004) use the following definition of risk: "*Risk is the potential for realization of unwanted, adverse consequences to human life, health, property, or the environment*". Estimation of risk (for an event) is usually based on the expected value of the conditional probability of the event occurring times the consequence of the event, given that it has occurred. According to Meacham (as cited in Hadjisophocleous & Fu, 2004), a comprehensive definition of fire risk is given as follows: "*Fire risk can be viewed as the possibility of an unwanted fire hazard in an uncertain situation, where loss or*

harm may be induced to the valued, typically life, property, business continuity, heritage, and/or environment". The key factors include unwanted outcomes or consequences, uncertainty, valuation, and likelihood of occurrence. Building fire risk analysis can be considered as the process of understanding and characterizing fire hazard in a building, unwanted outcomes that may result from a fire, and the likelihood of fire and unwanted outcomes occurring.

As Bukowski & Safety (1996) mention, it is difficult to express risk to life in a way that can be understood by the public. This leads to the consideration of other metrics for risk. Financial loss is the perfect metric for the property related risk, however, the primary focus of fire codes is life safety, which then requires that risk to life must include a measure of the value of human life. Thus, risk to life is usually assessed separately to avoid the difficulty of assigning a monetary value to human life. (Hadjisophocleous & Fu, 2004)

RELATED WORK

Higgins et al,. (2013), researched the challenges and barriers of safety assessment. Therefore community profiles were created based on open datasets. Over 130 datasets were identified for analysis, whether these datasets were applicable to the community profiles. The community profiles were incorporated in a vulnerability index. The purpose of the vulnerability index is to look at factors present in an individual's lifestyle that are known to increase fire risk. In particular, it is important to understand whether an individual has a number of risk factors present, which could mean the likelihood of injury or fatality was increased should a fire occur. With the community profiles Higgins et al,. (2013) were able to create a city map with heat zones which display the distribution of the community profiles. The research of Higgins et al. (2013), is similar to this study. The Higgins study is broader, since it involves all kind of criminal activities (like burglary) and accident risks.

Jennings (2013) used social and economic information from literature study to assess the risk of residential fire in urban neighbourhoods. Main goal of the research was to use this information to ensure better preventive activities by the government. The risk information has been combined with geographical data. Therefore it became possible to display the information in a geographical map. The research of Jennings (2013) gives an understanding of factors which are related to fire risk in neighbourhoods. Jennings (2013) views factors from different perspectives. However, these factors cannot be copied from this research, since there is a difference among factors between different countries and the Netherlands.

Trinh, Do, Wetz, & Anjomshoaa (2014) tried to develop a method to handle Linked Open Data in a dynamic way. The idea introduced, is to modularize functionalities into blocks. Users are able to combine linked open datasets to dynamically obtain, enrich, transform and visualize data in different ways. One of the research purposes is to study the flexibility in handling datasets during this research.

PROOF OF CONCEPT

Through interviews with several firefighters, it has become clear that in this area little research has been done ("We do not know a lot about which and how demographics affect fire risk, therefore we do not use it in the risk profiles we construct"). The safety region of Amsterdam-Amstelland has completed a statistical research on the relations between demographic factors and the risk of a dwelling fire. With this information, they could conduct precise communication with the appropriate target risk groups.

DWELLING FIRE RISK INDEX

The relations were extracted from "Handreiking social woningbrandrisicoprofiel". Some factors influence the risk positively and some negatively. The risks are based on demographics and involve risks like:

- Western immigrant resident – 15,7% lower risk.
- One parent family – 40% higher risk.
- Singe-person household – 11,5% higher risk.

Since the factors are not independent, it is not possible to do an independent probability calculation and therefore another method had to be used.

Dwelling Risk Index Calculation

All found data (proportions) were multiplied with the probability and added to each other. Therefore, it was possible to create an dwelling risk index for every neighbourhood. The proportion of the factor, times the probability gives a subscore. The subscores of all factors added up, give the risk index. For example, an area with 15% of "one parental families" (40% risk) results in the following sub score: $F_i*P_i = 15*0.40 = 6$. By linking the datasets, in accordance with the statistical relations and formula, the risk indexes have been calculated. Missing results have been filled out with the mean of the dataset. Computing this formula on the data has resulted in the left skewed distribution of risk indexes in Figure 2.

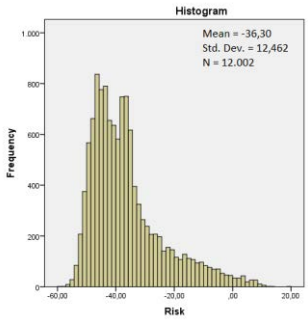

Figure 2. Distribution dwelling fire risks

NECESSARY DATASETS & DATASET APPROACH

Datasets were gathered from the internet. Extra datasets were gathered to create analysis beyond only the dwelling fire risk assessment, such as: response times and fire department locations. Spatial information was necessary to create geographic visualizations. The following datasets which are used in the Proof of Concept:

- Socio-Demographic datasets
- Spatial Data
- Response times and locations of the fire department
- Cadastral and address information

PROOF OF CONCEPT SCENARIOS

Putting all datasets together with the help of C#.NET and the Dotspatial library, has resulted in the Proof of Concept. The Proof of Concept calculates a risk index for all neighbourhoods in the Netherlands and gives a colour to the neighbourhoods corresponding to the risk index.

The Proof of Concept is able to provide insight into three different scenarios. In the future the Proof of Concept should provide more scenarios, or should be able to provide flexible scenarios. In the proof of concept the following scenarios are applied:

- Dwelling fire risks
- Urbanity versus fire department response
- Dwelling fire risk versus fire department response time.

Dwelling fire risk scenario

To determine whether a risk index is high or low, statistical percentiles have been used. The percentiles (and corresponding colours) have been used to determine the ratio in high or low areas. Risk indexes within the lowest percentile (10^{th} percentile) correspond with the lightest green, the highest percentile (100^{th}) corresponds with the darkest red. Figure 3 shows the result of this calculation, important to notice is that the more urbanized parts are the city parts with the highest dwelling risks, although population sizes were not used in the calculation.

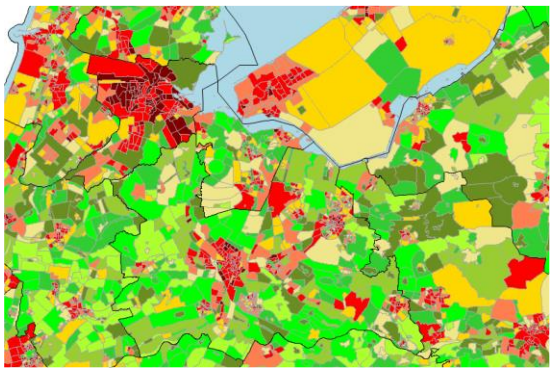

Figure 3. Dwelling fire risk Utrecht (Amsterdam red spot left).

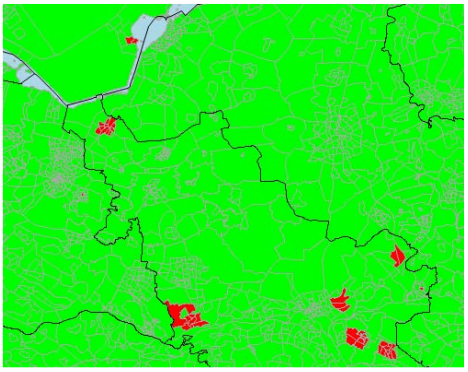

Figure 4. Urbanity vs. Response time Gelderland

Urbanity versus fire department response time scenario

Within the demographic data an attribute which represents the urbanity of the neighbourhood is present. A number between one and five is given. One is very strong urbanity and five is very weak urbanity. From moderate urbanity to very strong urbanity is used in this scenario. The maximum response time for structures with a residence function is 8 minutes. It might be interesting to know, in which regions the urbanity is moderate, strong or very strong with a response time beyond 8 minutes. Several neighbourhoods came up with this scenario. The red areas in Figure 7, have a moderate or strong urbanity (1000 up to 2500 addresses per km2) and the response time of the fire department is over 8 minutes.

Dwelling fire risk index, versus Fire Department Response Time scenario

The risk index could be used for more scenarios beyond the risk index itself. In the third scenario, the dwelling fire risk index has been combined with the fire department response time. Neighbourhoods with a high dwelling fire risk index are reflected in the response time of the fire department. Therefore, the risk indexes from the 90th percentile are shown in red in Figure 5 (high dwelling fire risk) when the response time is beyond the 8 minutes statutory maximum time.

Both scenario`s (Figure 6 & Figure 5) show almost the same areas in red. This is mainly caused by the fire department response time. The areas shown in red in both views are problem areas, because in these areas urbanity is strong, fire risk is high and response time is beyond 8 minutes statutory maximum response time.

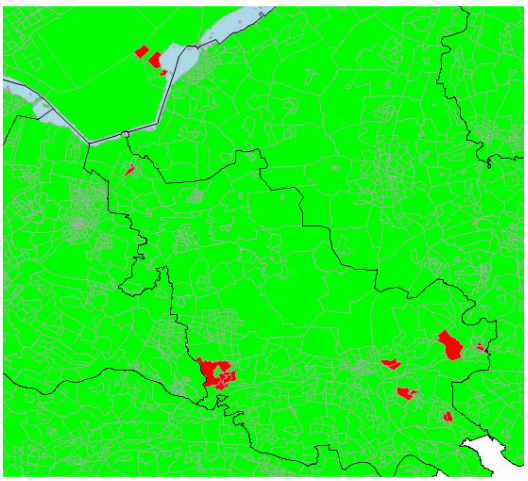

Figure 7. Risk index versus Response time Gelderland-Midden

PROOF OF CONCEPT EVALUATION

The statistical relations used have not been validated. The most important question is: 'How do the outcomes of the Proof of Concept compare with reality?'.

Gut Feel Domain Expert

The three views made with the Proof of Concept have been verified. In these sessions has been shown that the dwelling fire risk index is the most valuable view, since the response time is needed for the other views and these response times are only available on municipality level. In this dataset the median of the response times per municipality has been taken as the mean. Therefore the results are not satisfactory altogether since the interesting results lie in the deviation of response time within a municipality. With a high probability the fire department could tell that the results of the two other views deviate from the reality.

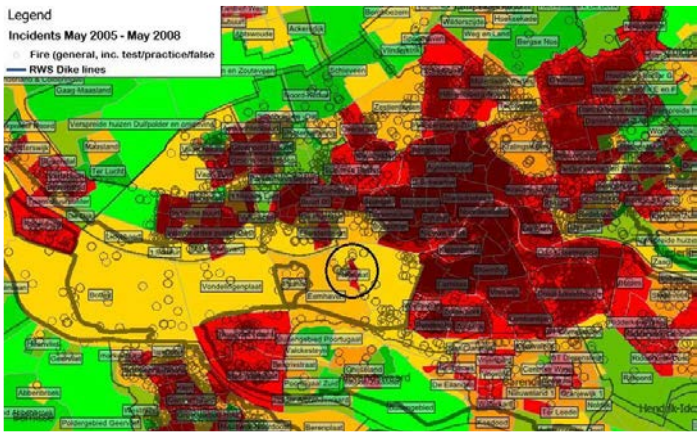

Figure 6. Overlay of fire incidents over dwelling fire risk indexes (Rotterdam-Rijmond.) "Heijplaat" is black circled.

Fire Risk Index

The verification of the dwelling risk indexes are based on the gut feelings of the fire department in their own region. The fire department has compared the outcomes from the Proof of Concept of their own region with the gut feelings of the fire risks in their own region. From this evaluation it turned out that there were similarities between the gut feelings of the fire department and the Proof of Concept: "*It seems to fit well*". The main statement is: "The risk indexes seem to fit well in general, probably there are some errors and deviations". Although as a test during the evaluation a small high risk area "Heijplaat" (Figure 6) was taken which has a high risk, but from their gut feelings the firefighters disagree with this high risk. According to the firefighters this area should have a lower risk.

Fire Incident Map

The Fire Department of Rotterdam-Rijmond compared the outcomes with the incidents from 2005-2008 plotted on a map during evaluation. In Figure 6 the incident map is placed over the outcomes from the Proof of Concept. The high risk neighbourhoods had more fire incidents in the past than the low risk neighbourhoods. This overlay gives a better comparable view.

All participants in the evaluation agreed, that the Proof of Concept provides a good fire risk index. This conclusion has purely been drawn based on the comparison between the incident map and the fire risk map from the Proof of Concept tool. Moreover, "Heijplaat", the area from the gut feeling evaluation part, appears to have had more fire incidents in the past than the average low risk areas. The high risk assigned to "Heijplaat" appears to be connected to reality.

Time needed to add a dataset

Because flexibility and dynamics are part of this study, the time needed to add the datasets was recorded. In this way an overview with the periods of time could be made. Important to note is that two periods (outliers) have a larger duration due to a learning curve needed to use the datasets.

Dataset	Time
Socio-Demographic data	3 Hours
Spatial data neighbourhoods	2 Hours
One parent family	2 Hours
Measured Response Time, Fire department	2 Hours
Locations of all fire departments in the Netherlands	15 Hours
Spatial data Safety regions	2 Hours
Construction years	12 Hours

Table 1. Time needed to add a dataset to the Proof of Concept.

In these results the "Locations of the fire departments" took 15 hours, due to the use of a different coordinate system. It took some time to find out how to align the different coordinate systems. The "constructions years" took more time as well. The dataset with the construction years is approachable with SPARQL. In this 12 hours the time learning about SPARQL is included.

Dynamics of datasets

The dynamics of a dataset are determined by the way of the dataset is provided and how easy it is to add the dataset to the Proof of Concept.

Only one dataset is provided in a dynamic way, which means that the dataset could be requested any time, any place (and monthly updated without the effort of downloading and updating the dataset locally). The other datasets (which are updated regularly) provide some form of dynamics as well, but since individual files should be downloaded from different sources on the Internet the dynamics are affected negatively. Which means these datasets should be managed by the Proof of Concept instead of requested and calculated from the source.

Dataset alignment

The identified datasets were from various sources. However the initial goal of the research was to use rather Linked Data (RDF) datasets, most datasets are in non-linked formats. Aligning(linking) the used datasets has been done upfront. Therefore administrative links were used which are already present in the datasets. Aligning has been done based on national determined neighbourhood codes and by following administrative links to other granularity levels (neighbourhood to postal code or even municipality level). The cadastral data provided as linked data was aligned on a similar way. After harvesting the data via their SPARQL endpoint, the resulting data was stored with the resulting data from the alignment process. Only one dataset was provided as pure Linked Data and therefore the power of "getting data directly from the source and link it" was not possible.

DATA EVALUATION

Quality dimensions

All data could be assessed on data quality. The quality is important to know, since the quality of the resulting information is determined by the quality of data.

Table 2. Quality dimensions. The not shown datasets are 100% on all dimensions.

Dataset	Extensional completeness	Intensional completeness	Consistency	Conciseness
Socio-Demographic data	100%	90.09% - Western foreign resident 90.09% - Dutch resident 93.45% - More residents on one address 93.45% - Single-person households 90.09% - 0-14 age in household 90.09% - 15-24 age in household 90.09%- Main tenant is 65 or older 89.05% - Sale houses 76.56% - Single storey apartment *100%* - Neighbourhood name *100%* - Municipality name *99.15%* - Population	100%	100%
Construction years (BAG)	76.63%	100%	100%	100%

As discussed under chapter Data Quality, completeness has been measured on the intensional and extensional level. Extensional completeness is about whether a dataset contains all objects from the real world (contains the dataset with buildings all buildings from the real world). Extensional completeness is represented in the second column of table 2. Intensional completeness is about whether a dataset contains all used or necessary attributes from the real world (contains the dataset with buildings a construction year, location and coordinates for each building). Intensional completeness is represented in the third column of table 2.

Furthermore, the accuracy levels of the datasets are different. The following datasets provide data on neighbourhood level:

- Socio-Demographic data
- Spatial data neighbourhoods
- Construction years (BAG)

The following datasets, provide data on municipality level:

- One parent family
- Measured Response Time

Error of data

According to the domain experts, the data present in the municipal databases are regularly incorrect. The demographic data used in this research are gathered by the CBS from the municipal databases (Basisregister Personen - BRP). The government conducted 5.000 home visits to research the correctness of BRP in 2014. Their conclusion: "*During the house visits it appeared in almost one third of the cases that there is indeed a difference between the registered persons and the people who actually live there*" (Rijksdienst voor Identiteitsgegevens, 2015). Therefore, there is a certain level of incorrectness in the data.

Completeness

The completeness of datasets is extensional almost 100% complete. The extensional completeness of BAG is not 100%. According to various sources (including the BAG itself) there are no exact numbers public about the completeness and the consistency of the BAG.

The intensional completeness of five datasets is 100%. The intensional completeness of the dataset with the demographic data of all neighbourhoods and the BAG are not 100%. This is not a problem, since the mean of the intensional completeness of both datasets is about 90%. Although, the missing data have been filled out with

the mean, it will always cause errors in the information. In the outcomes, this means that no correct risk calculation can be done for areas with missing data on multiple factors.

Trust

All datasets in the proof of concepts are from government organizations. These organizations have a certain authority, the cadastral institution is responsible for publishing information on buildings and addresses (which are gathered from municipalities). Other datasets are published from the Dutch Health Organization and directly from municipalities. All datasets provide some form of meta-information (provenance). Which means that all datasets mention the creator and creation date, but none of them describes the creation method. The cadastre, does not describe a certain creation date but records a mutation table. Moreover, the cadastre does not record the creator of the data.

However trust is hard to measure, the trust of the used datasets is high. All data is retrieved from authorities and have limited meta-information.

Accuracy

Almost all datasets are on neighbourhood level. Since most datasets do not provide details about individual people, it is not possible to obtain a narrower level (house level e.g.) due to privacy restrictions. Several datasets are on municipality level. During the evaluation sessions the accuracy of the response times was a topic of discussion. The response time is now an average of the response times for a municipality, while it is interesting how this response time fluctuates among the neighbourhoods within a municipality.

The statistical relations used for the Proof of Concept, are about main tenants. The available demographic data contain the proportions of the characteristics in a neighbourhood and do not contain only the proportions of main tenants of the characteristics.

Furthermore, two factors on the age of children (0-11 and 12-17) are not equal to the factors in the data. The demographic data on children is from 0-15 and 15-25. It is unclear how these accuracy problems influence the quality of the outcomes. Probably the proportions of main tenants and the general proportions relatively relate to each other.

DISCUSSION

Necessary data

All necessary datasets were found to conduct a proper risk calculation. However, the statistical relations from the dwelling fire risk profile (“Handreiking sociaal woningbrandrisicoprofiel”) are not validated by other research nor further investigated. Therefore, the statistical relations are not complete (probably there are more factors) and there is no confirmation on the correctness of these factors.

Since the “Handreiking sociaal woningbrandrisicoprofiel” is the only statistical knowledge that is available, no other options were available to use for this research. Improvement of statistical knowledge and improvement of the factors will probably improve the dwelling fire risk calculation in the future.

Individual files

The main idea of this study is to create information by combining open datasets. All datasets used in the Proof of Concept are open datasets available on the internet. Compared to a business intelligence approach in which data warehouses are used, a linked open data approach keeps the data on the side of the provider. In practice it appears that almost all datasets are provided as a spreadsheet file which has to be downloaded from the internet instead of using a SPARQL interface to gather data from the provider. In the Proof of Concept only the data about construction years from BAG were provided through a SPARQL interface. Individual (spreadsheet) files means downloading and storing data yourself, in other words creating a data warehouse. Due to these static datasets, data alignment is harder, since static files contain different file formats which are stored locally. In a dynamic way (i.e. SPARQL) the output will be linked and combined with each other on the spot.

Outdated data

Four out of seven datasets used in the Proof of Concept are from 2013 or older. Important to note is that the dataset which contains the demographic data and the spatial data of all neighbourhoods in the Netherlands date from 2013. During the evaluation sessions the outdated data were discussed. During both sessions the fire department participants stated independently that demographic characteristics do not change in a few years. The outdated demographic dataset would not be a problem. However, the government sometimes invests a lot of money in a neighbourhood which could change the demographic characteristics of a neighbourhood. During the evaluation sessions the participants mentioned that the number of times such investments happen are negligible. The demographic dataset is always one to two years outdated and therefore the obsolescence of this dataset will not cause problems.

Dynamics of datasets

The dynamics of the Proof of Concept would be better, when all data could be requested as linked data, calculated and presented the outcomes, instead of downloaded, managed, calculated and presented the outcomes (which is currently the case).

Proof of Concept Architecture

During the development of the Proof of Concept, all data were added to the main dataset, which is the demographic dataset. In the end this turned out to be a wrong way of handling data, since most datasets are individual files. Updating data is not easy since all data are combined in one file. A better way to manage data, is to keep the datasets separate and link datasets together on data level. Because of this the management of data would be a lot easier.

One of the ways to do this, is to create a SPARQL endpoint for all datasets separately. The datasets could be linked to each other on the spot in the Proof of Concept. Because datasets stay separate, they could easily be updated (which is certainly not the case now). Moreover, adding new datasets would be easier when new datasets could be handled exactly the same way as all other datasets whatever data format they have. Therefore, dynamics and flexibility will be improved by using such an architecture.

CONCLUSION

The necessary datasets to conduct a dwelling fire risk index were found, in some cases the accuracy is a problem. Furthermore, there is a slight accuracy mismatch in the demographic dataset. Also, the meta-information is too limited. From most datasets only the creator and creation date are available. All datasets are provided by authorities which affects the trust positively in the data quality. Most datasets are 2 years old, this is not necessarily a problem. During the research the purpose was to create a risk profile based on particularly demographics. Demographics of a neighbourhood do not change very often. Demographics tend to stay the same over a longer period. However there are exceptions, after a restructuring or neighbourhood renovation demographics can change drastically due to an significant attraction of people with other demographic characteristics.

Based on gut feeling all domain experts could verify that the results seemed to be right on the first impression. The comparison of the outcomes with the fire incidents in the past plotted on a map from the fire department, was surprisingly similar. The fire incident map and the outcomes of the Proof of Concept have been compared successfully by a domain expert as well. With the used datasets and the used statistical relations, it seemed to be possible to do a successful dwelling fire risk assessment. The dynamics of the Proof of Concept was not as expected. Since almost all datasets are provided as individual files and had to be converted to linked data ourselves before we could easily combine, match and calculate data. The evaluation sessions with the domain experts have given an indication of the correctness of the outcomes. However, it would be a good thing to find more ways to verify the outcomes.

ACKNOWLEDGMENTS

I would like to acknowledge those who have helped to achieve the goals of this study. I would like to thank Jan Wielemaker (VU University Amsterdam) for his guidance and support throughout the research. I would like to thank Bart van Leeuwen as well. Bart fulfil the role as domain expert, Bart has helped with his firefight and Linked Open data expertise. I would like to thank the domain experts of the fire department Netherlands which have contributed to this research during the evaluation sessions. Without their help it would be impossible to learn about the correctness of the outcomes.

REFERENCES

Bellini, P., Nesi, P., & Venturi, A. (2014). Linked open graph: Browsing multiple SPARQL entry points to build your own LOD views. Journal of Visual Languages & Computing, 25(6), 703–716. http://doi.org/10.1016/j.jvlc.2014.10.003

Berners-Lee, T. (2006). Tim Berners-Lee (M.I.T.), father of the World Wide Web... Retrieved from https://www.youtube.com/watch?v=Jev2Um-4_TQ

Bizer, C., & Berlin, F. U. (2009). Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems, 5(3), 1–22. Retrieved from http://eprints.soton.ac.uk/271285/1/bizer-heath-berners-lee-ijswis-linked-data.pdf

Brickley, D., & Guha, R. V. (2014). RDF Schema 1.1. Retrieved March 7, 2017, from https://www.w3.org/TR/rdf-schema/

Bukowski, R. W., & Safety, F. R. and. (1996). Risk and Performance Standards. Thirteenth Meeting of the US/Japan Government Cooperative Program on Natural Resources (UJNR) Panel on Fire Research and Safety.

Cyganiak, R., Wood, D., & Lanthaler, M. (2014). RDF 1.1 Concepts and Abstract Syntax. http://doi.org/10.1007/s13398-014-0173-7.2

Dean, M., & Schreiber, G. (2014). OWL Web Ontology Language. Retrieved March 7, 2017, from https://www.w3.org/TR/owl-ref/

Debattista, J., & Lange, C. (2014a). daQ , an Ontology for Dataset Quality Information.

Debattista, J., & Lange, C. (2014b). Representing Dataset Quality Metadata using Multi-Dimensional Views.

Hadjisophocleous, G. V, & Fu, Z. (2004). Literature Review of Fire Risk Assessment Methodologies, 6(1), 28–45.

Handreiking sociaal woningbrandrisicoprofiel. (n.d.).

Higgins, E., Taylor, M., Jones, M., & Lisboa, P. J. G. (2013). Understanding community fire risk—A spatial model for targeting fire prevention activities. Fire Safety Journal, 62, 20–29. http://doi.org/10.1016/j.firesaf.2013.02.006

Jennings, C. R. (2013). Social and economic characteristics as determinants of residential fire risk in urban neighborhoods: A review of the literature. Fire Safety Journal, 62, 13–19. http://doi.org/10.1016/j.firesaf.2013.07.002

Kontokostas, D., Westphal, P., Cornelissen, R., Bibliotheek, S., Hellmann, S., & Lehmann, J. (2014). Test-driven Evaluation of Linked Data Quality Categories and Subject Descriptors. Www2014, 747–757.

Mendes, P. N., Mühleisen, H., & Bizer, C. (2012). Sieve: Linked Data Quality Assessment and Fusion. Proceedings of the 2012 Joint EDBT/ICDT Workshops, 116–123. http://doi.org/10.1145/2320765.2320803

Miličić, V. (2011). Problems of Linked Data (4/4): Consuming data. Retrieved from http://milicicvuk.com/blog/2011/08/04/problems-of-linked-data-44-consuming-data/

Rijksdienst voor Identiteitsgegevens. (2015). Kwaliteit BRP. Retrieved June 26, 2015, from http://www.rijksdienstvooridentiteitsgegevens.nl/BRP/Kwaliteit

Ruckhaus, E., Baldizán, O., & Vidal, M. E. (2013). Analyzing linked data quality with LiQuate. Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8186 LNCS, 629–638. http://doi.org/10.1007/978-3-642-41033-8_80

Trinh, T., Do, B., Wetz, P., & Anjomshoaa, A. (2014). A Drag-and-block Approach for Linked Open Data Exploration.

University of Wisconsin. (2011). Stata for Researchers: Combining Data Sets. Retrieved from http://www.ssc.wisc.edu/sscc/pubs/sfr-combine.htm

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2012). Quality Assessment Methodologies for Linked Open Data: A Systematic Literature Review and Conceptual Framework. Semantic Web – Interoperability, *Usability, Applicability*, *1*, 33. Retrieved from http://www.semantic-web-journal.net/sites/default/files/DQ_Survey.pdf