reviewed paper

SmartRegio – Employing Spatial Data to Provide Decision Support for SMEs and City Administrations

Martin Memmel, Andreas Abecker, Sebastian Bretthauer, Heinz Kirchmann, Roman Korf, Markus May, Richard Wacker

(Dr. Martin Memmel, German Research Center for Artificial Intelligence (DFKI GmbH), Trippstadter Straße 122, 67663 Kaiserslautern, Germany, martin.memmel@dfki.de)
(Dr. Andreas Abecker, Disy Informationssysteme GmbH, Ludwig-Erhard-Allee 6, 76131 Karlsruhe, Germany, andreas.abecker@disy.net)
(Dr. Sebastian Bretthauer, Goethe University Frankfurt a.M., Theodor-W.-Adorno-Platz 4, 60629 Frankfurt am Main, Germany, bretthauer@jur.uni-frankfurt.de)
(Heinz Kirchmann, German Research Center for Artificial Intelligence (DFKI GmbH), Trippstadter Straße 122, 67663 Kaiserslautern, Germany, heinz.kirchmann@dfki.de)
(Roman Korf, USU Software AG, Rüppurrer Straße 1, 76137 Karlsruhe, Germany, r.korf@usu.de)
(Markus May, Disy Informationssysteme GmbH, Ludwig-Erhard-Allee 6, 76131 Karlsruhe, Germany, markus.may@disy.net)
(Richard Wacker, YellowMap AG, CAS-Weg 1-5, 76131 Karlsruhe, Germany, richard.wacker@yellowmap.de)

1 ABSTRACT

When decisions have to be made which are based on the characteristics and expected developments in specific spatial environments (such as finding the best place for a new production site or for a new shop), geo data and the information that can be derived from it plays a crucial role. While larger companies typically can afford the setup of the required organisational units as well as the access to relevant data from commercial providers, smaller organisations such as SMEs or city administrations are at a disadvantage. The aim of the SmartRegio project was to develop solutions for such organisations that combine freely available (mass) spatial data from many different sources as a decision-making basis focusing on governmental and private actors operating with a focus on a specific region. The data sources include data from infrastructures like energy and mobility, data from public entities, and also data from social media and media channels. The SmartRegio project successfully identified and tackled major technical and legal challenges when aiming to exploit such data, while at the same time realising a generic infrastructure that supports the required processes within the given context.

Keywords: data economy, (big) spatial data architecture, sociodemographics, smart data, spatial data analytics

2 INTRODUCTION

Governmental and private actors operating with a focus on a specific region must constantly adapt to diverse changes in their environment. This concerns a variety of organisations ranging from companies to city administrations. In order to understand the characteristics of the focused environment, the existence of spatial data about it is becoming more and more important, as well as the ability to work with this data.

While large companies are usually able to employ departments that are able to get access to relevant data and to derive strategic measures from them, city administrations, non-profit organizations as well as small and medium-sized enterprises are at a disadvantage. On the one hand, their financial and (IT) technological possibilities are often limited. On the other hand, they are much more rooted in their specific region – and for small-scale areas, data at a high resolution is sometimes more expensive, and the data quality is often not very good.

Consequently, the aim of the project SmartRegio[1] funded by the German Federal Ministry for Economic Affairs and Energy was to develop solutions that combine (mass) spatial data from many different sources – ranging from closed sources only available to the data owners to open data sources – as a decision-making basis focusing on these players. The data sources include data about infrastructures like energy and mobility, data from public administrations (like occurrences of administrative processes involving citizens) and data from social media and media channels. These sources provide some outstanding advantages. Firstly, as a by-product of normal business operation, data collection is relatively cheap. Moreover, they are arising continuously, so it can be assumed that in the near future it will be possible to detect trends at a very early stage – even before people are becoming aware of it. Thirdly, they enable the machine-driven learning of

[1] see https://www.smartregio.org

ISBN 978-3-9504173-2-6 (CD), 978-3-9504173-3-3 (print)
Editors: Manfred SCHRENK, Vasily V. POPOVICH, Peter ZEILE, Pietro ELISEI, Clemens BEYER

patterns and correlations on a low level. The major challenge lies in their interpretation. In order to allow the exploitation of this variety of data sources, reference architecture was developed within SmartRegio that allows for integrating and analysing heterogeneous spatial data in order to better understand the local environment. This includes the characterisation of spatial areas according to several criteria as well as the recognition of developments.

The use of heterogeneous (mass) data poses high technical and legal challenges. Firstly, they are distributed over many different and separated data silos, and the potential data suppliers often have no experience with the involved problems. Secondly, they are available in many different formats and structures, so that many have to be preprocessed comprehensively in order to be able to work with the content itself. Thirdly, the references to spatial and temporal entities often differ in terms of representation means and granularities (e.g., some data sets provide aggregated information from several months about a large spatial entity, while other data sets provide daily information about much smaller regions), which makes their comparison more difficult. While statistics are collected, for example, for administrative areas, infrastructures are technically dependent on their own spatial divisions, and in the case of media or discussions in social networks the spatial reference is often difficult to determine. Fourthly, many of the sources contain personal information and their anonymisation is particularly difficult because of the spatial reference and the combination of many sources.

In the following, the SmartRegio project, the approach and the technical architecture followed within the SmartRegio project, the main challenges when aiming to work with several types of spatial data from different sources, questions regarding security, anonymity, and legal aspects, business models as well as the most important lessons learned will be presented. Furthermore, some specific use cases will be described in more detail.

3 THE SMARTREGIO PROJECT

3.1 Project Context and Partners

SmartRegio is a consortium project sponsored by the German Federal Ministry for Economic Affairs and Energy (BMWi) from 12/2014 until 05/2017. It is part of the technology program "SmartData – Innovations from Data" where a total of 13 selected lighthouse projects were promoted in order to develop innovative services, and to stimulate the broad use of intelligent, data-based technologies.

The SmartRegio consortium comprises partners for technology development and transfer (YellowMap AG, USU Software AG, Disy Informationssysteme GmbH), a partner for the examination of legal aspects (Research Center for Data Protection at the Goethe University Frankfurt a.M.), and a research partner (DFKI GmbH). To realise usage scenarios in the selected pilot region Kaiserslautern, the municipal service provider SWK Kaiserslautern as well as the City Administration of Kaiserslautern joined the project as associated partners.

3.2 General Vision of SmartRegio

Geomarketing and location planning are a profitable business and since decades an oligopoly of few predominant market players. The reason for the high concentration in the market is the exclusive access to socio-economic data. These data are collected, analysed and used to determine, e.g., the spending power, the demands, and the economically relevant behaviour of customers in a given region. Finally, evidence-based recommendations on many essential corporate decisions, e.g., how to create an optimal fit between a branches location, product and service portfolio, marketing strategy, etc. can be given. The outcome is impressive. E.g., given a portfolio change, new product or marketing campaign large retailers receive an estimate of their future revenues with 90 percent accuracy. But there are significant drawbacks:

(1) The services are not affordable for SMEs, creating a competitive disadvantage for those.

(2) The services provide a snap-shot – they do not reflect the continuous development of a region.

(3) The provided data is limited to answer a pre-known set of questions.

SmartRegio, on the one hand, addresses the fact that the more technology enters our everyday life, the more the behaviour of people is reflected in various kinds of data. Moreover, recent advances in the integration and processing of heterogeneous data sources make it possible to assemble those data fragments to an

increasingly precise picture. This picture shows development and includes more facets allowing to ask a variety of questions – including such we do not know yet. And by the fact that this data is emerging as a by-product of conventional business operation, its price can be significantly lower making it available to a wider target group.

On the other hand, the business architecture of SmartRegio stands in contrast to traditional business models in that field. Its main principle is the design of an open platform that supports collaboration between participants of different kind – like data providers, service and software developers.

4 SMARTREGIO ARCHITECTURE AND APPROACH

4.1 Architecture and Components

This section gives a short overview of the SmartRegio platform architecture. SmartRegio intends to offer an open, cloud-based platform for processing and analysing heterogeneous geo-spatial data. The architecture as depicted in Figure 1 follows a standard three-layer approach comprising components for data integration, data processing, and data visualisation (see, e.g., Kahn et al. 2012 for a more detailed discussion of required capabilities for platforms in the given problem context). SmartRegio aimed at using existing and proven tools whenever possible, and the architecture was continually evaluated by the accompanying research.

In the data integration layer, we mainly distinguish between three different kinds of data: (i) geo-spatial data with direct geo-spatial reference, (ii) social media data with indirect or direct geo-spatial reference, and (iii) big-data with usually no or indirect geo-spatial reference. Within the platform these pre-processed data sets are stored in a central data-warehouse. The metadata management system for data and services manages these data sets and provides services for data set retrieval. The service library functions as an equivalent to the data-warehouse, but for the services used to process and analyse the data. Services are also registered at the metadata management for data and services.

Having (i) data sets and (ii) services working on these data sets, the configuration cockpit can be used to implement data pipes, orchestrating several services and data sets to process and analyse the data. The SmartRegio engine executes these service orchestration programs. If such a data pipe generates a new data set, the data set is registered within the metadata management and is published at the Geo-Server. The Geo-Server publishes the data based on OGC[2] standards for visualisation.

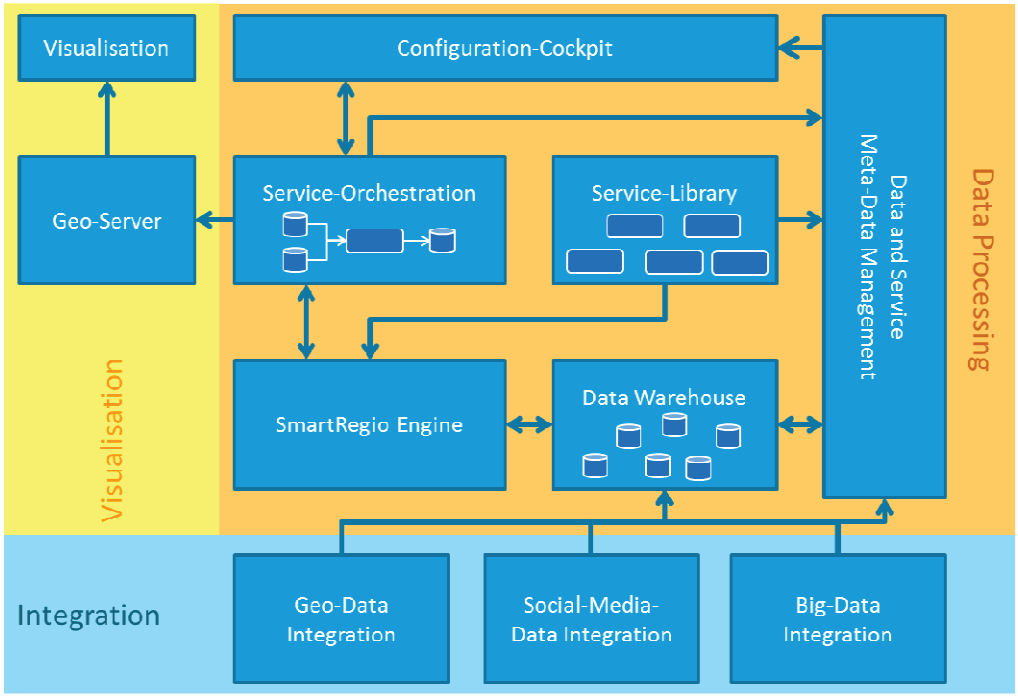

Figure 1: An overview of the SmartRegio System Architecture

[2] http://www.opengeospatial.org

ISBN 978-3-9504173-2-6 (CD), 978-3-9504173-3-3 (print)
Editors: Manfred SCHRENK, Vasily V. POPOVICH, Peter ZEILE, Pietro ELISEI, Clemens BEYER

4.2 Data Acquisition, Data Integration and Processing, and Data Analysis

4.2.1 Data Acquisition

The first step in processing and analysing data is data acquisition. SmartRegio made use of a variety of data sources that provide information on a regional level. Figure 2 shows these data sources ordered according to their characteristics. On the one hand, there are open data sources that can be accessed by anyone; on the other hand, there are closed data sources that, at least, require some administrative steps (e.g., regarding privacy protection or national and international law) before they can be exploited. Furthermore, it is important to make a distinction between public, non-profit data providers and commercial entities.

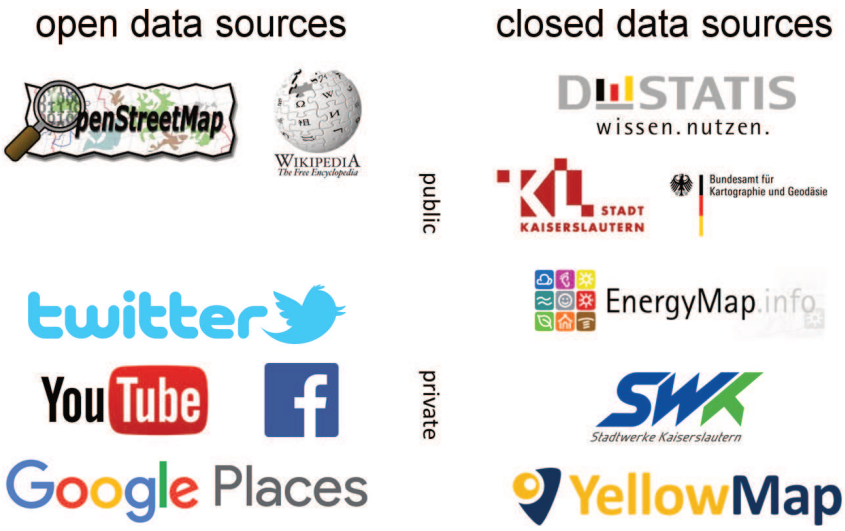

Figure 2: Data types and data sources used within SmartRegio

Depending on the data source type, different challenges have to be tackled when aiming to work with the data. This starts with the problem of finding relevant data sets: E.g., there are still only few open data portals, searching data on these portals is often very time-consuming, and raw data that is crucial for any further processing steps, is often not published. When working with companies and public authorities, there is often a lack of well-structured overviews regarding existing data, and superior contact persons with the required knowledge are often missing. Thus, relevant sources can often only be identified in a complex and lengthy process involving contact persons in different specialist departments that usually have no experience with such requests. Furthermore, raw data is often very expensive in these contexts.

Once a relevant data source is found and access to the data is possible, questions of data characteristics, data quality, and data representation arise. The most common problems here are that we often only find aggregated and preprocessed data (e.g., in annual reports) instead of raw, machine-processable data. Data about smaller regional entities is often completely missing. Furthermore, the data sets are often not up-to-date. A further problem that is of great importance when trying to work with spatial data from heterogeneous sources, is the fact that they usually are not referring to uniform spatial entities, which makes any comparison between data sets difficult. In SmartRegio, this is usually solved by mapping spatial data on a common grid.

4.2.2 Data Integration and Processing

Before using data in a system or platform, it usually first has to be (pre)processed and integrated in a way that allows its usage, e.g., for data analysis, or data visualisation. In the SmartRegio data integration process, different data pools were used for different purposes, depending on the nature of the data (e.g., distinguishing between spatial data and text data). Consequently, different tools such as Talend Open Studio[3], Kibana[4], or

[3] Talend Open Studio (see https://www.talend.com/products/talend-open-studio) is an Open Source ETL & Data Integration tool

[4] Kibana (see https://www.elastic.co/products/kibana) is an Open Source plugin for the Lucene-based search engine Elasticsearch that is used to host large amounts of mostly text-based data in SmartRegio

Katana Platform[5] were used exploratory data visualisation and analysis. In several preprocessing steps, the data is transformed into a generic format that allows conducting further processing steps independent of the format of the original data source. Such transformation steps include, e.g., the mapping of spatial data to a common grid, the mapping of metadata to the schemas used in SmartRegio, or the mapping of terms to semantic concepts in the SmartRegio ontologies. It is very important to note that there are almost no generic approaches for data integration – there are countless different data formats, and each data set requires individual decisions.

As a last step in the SmartRegio integration process, metadata about each data set is published on a metadata management system that allows the publication and description of metadata about data sets, and that offers convenient means to search and access the data. In SmartRegio, the established Open Source data portal platform CKAN[6] was used for this purpose.

4.2.3 Data Analysis

In SmartRegio, four different kinds of analyses have been distinguished:

(1) Aggregated display of different topics: Here, several topics can be displayed together in a way that allows users to draw their own conclusions. This is technically rather simple, yet requires usually a lot of expertise on the user side, as it is often very difficult to recognise connections or dependencies.

(2) Different visualisation means: The same data set can be displayed in many different ways, depending on its characteristics. Examples for such visualisations in the context of SmartRegio are scatter diagrams, heatmaps, or timelines.

(3) Prepared evaluations for a question or issue: A user evaluates a result which has been preprocessed in a way that allows for a quick and easy analysis. Such a prepared evaluation requires some domain knowledge on the side of the technology experts who then can choose and prepare proper data sets and services.

(4) Exploratory data examination: A viewer can filter, highlight, and process interesting records with other records.

For all types of analyses, tools and services have been realised within SmartRegio. Examples and sample use cases will be presented in Section 7.

5 SECURITY, ANONYMITY, AND LEGAL CONSIDERATIONS IN SMARTREGIO

This section provides some insights into the challenges and requirements on security within the SmartRegio platform as well as the technical concept and realisation on legal aspects as discussed in Section 5.3.

5.1 Experiences in Applying Security in SmartRegio

SmartRegio intends to offer an open, cloud-based platform for processing and analysing heterogeneous geo-spatial data. This leads to several challenges in terms of security, some of which are summarised here. Open and cloud-based implies that the platform has open borders. Several stakeholders with different intentions need access to the platform. Data sets from several sources and processing services from different providers need to be run on the platform. Topics like trust, integrity, and authenticity play a major role in using the platform. And within a globally operating platform, local and global legal aspects need to be considered on where data are stored and processed.

Based on the architecture of the SmartRegio platform (cf. Figure 1) we identified several stakeholders required to operate the intended platform. Stakeholders are:

- Platform operator – operating the SmartRegio platform,
- Data provider – providing data sets for the platform,
- Service provider – providing services for processing the data within the platform,
- Application developer – orchestrating services for processing data sets, to provide added values for end users, and

[5] USU Katana Platform (see http://katana.usu.de/) is a platform for data processing on big scale, high velocity and variety

[6] see https://ckan.org

ISBN 978-3-9504173-2-6 (CD), 978-3-9504173-3-3 (print)
Editors: Manfred SCHRENK, Vasily V. POPOVICH, Peter ZEILE, Pietro ELISEI, Clemens BEYER

- End user – consuming processed data.

These stakeholders were used as direct input for developing the role concept within SmartRegio with their respective competencies and access rights within the platform. Another dimension within the platform is the access right on services and data sets we needed to consider. Different tenants, e.g., providing data sets or services, need different access to the resources even within the same role. Application developers need to access data sets and services in order to process data for end users. This requires dynamic access right management for resources based on the licensed resources.

Within SmartRegio we analysed several frameworks for identity and access management, amongst them freeIPA[7], OpenIAM[8], FORGEROCK[9], gluu[10], and Keycloak[11]. Main requirements amongst the security requirements for identity and access management were (i) the support of open standards, especially of the World Wide Web and for extensibility of the security concept, (ii) the ease of integration of legacy systems and third party services as well as (iii) the ease of administration of the system. Based on our own experience with the systems we decided for Keycloak. It provides web-based administration, fulfils the functional as well as non-functional requirements and is part of the RedHat[12] community projects.

In addition to the identity and access management, a major challenge in the security concept of the platform is data security. This includes transport security as well as storage security. Being an open, cloud-based platform requires building on open web-standards. The data upload or provision for the platform is realised by protocols like https[13] or other secure protocols like ftps[14], ssh[15], and ssh file transfer protocol. Platform components and in particular services communicate via https-based on RESTful web-services.

Storage security can be achieved via several mechanisms. For critical data, like data only accessible for one tenant, data need to be stored within a secure storage only accessible for the tenant itself. This can be achieved, e.g., physically or virtually providing appropriate access right mechanisms for these secure areas. Other data sets need to be secured by access right rules within the identity and access management. Providing additional security for preventing security flaws can be achieved by encrypting data. This allows the protection of data where only users and systems with access rights and the key for decryption can use the data. There are several software and hardware options on the market.

Even though we have found solutions in order to implement a security layer within SmartRegio, the complexity still exists in terms of implementing access rules for different tenants and roles.

5.2 Anonymity Approach within SmartRegio

With a rising awareness and discussion on privacy protection within Germany and the EU there is a need on protecting these rights within the SmartRegio platform. Conventional methods on protecting these rights are done via pseudonymisation and anonymisation. While the former only replaces the identifiers, the latter transforms the data set so no single entry causes a risk on identification. Conventional technical solutions are data encryption and data transformation (cf. Ronning et al. 2005, Höhne 2010). However, increasing anonymity usually leads to a reduction of expressivity of the data in terms of the analysis potential (Rosemann 2006, Hochfellner et al. 2012).

Having an anonymised data set, however, does not imply that a de-identification by correlating different data sets is not possible. In (Bender 2015), the author mentions several risks on de-identification:

- Identity disclosure – risk of identifying individuals,
- Attribute disclosure – risk of identifying sensitive attributes of an individual, and

[7] http://www.freeipa.org
[8] http://www.openiam.com/
[9] http://www.forgerock.com
[10] https://www.gluu.org/
[11] http://ww.keycloak.org
[12] https://www.redhat.com/en
[13] https://tools.ietf.org/html/rfc2818
[14] https://tools.ietf.org/html/rfc4217
[15] RFC 4250 – 4256, RFC 4335, RFC 4344 – 4345

- Membership disclosure – risk of identifying sensitive attributes of an individual when proving the individual is likely part of the data set.

The author also lists criteria for anonymity that prevent these risks for de-identification. The criteria are namely:

- k-anonymity – divides the data set into several equivalence classes where each equivalence class has at least k different data set entries,
- l-diversity – requires that each equivalence class has at least l diverse sensitive attributes, and
- t-closeness – requires that the distribution of sensitive attributes value within one equivalence class and the distribution of this attribute value within the overall data set exceed a threshold t.

Within a data processing platform as prototypically realised within SmartRegio, a lot of data sets from third parties are uploaded to the system. This involves the risk that these data might not be conforming to data protection regulations. Therefore, anonymity has to be ensured. The approach chosen within SmartRegio is to measure the above mentioned anonymity criteria in order to evaluate the level of anonymity of data sets uploaded to the platform. Additionally, these KPIs need to be evaluated after the analysis process and before the analysed and combined data are presented to the end user in order to prevent de-identification risks.

In (Bender 2015) the author evaluated several frameworks for anonymising data sets and implemented distributed algorithms for anonymisation. However, the need on efficiently checking anonymity levels for large amounts of data still existed. Our approach is based on platform services utilising Apache Spark data-frames[16] within the USU Katana Platform[17] which provide efficient algorithms on grouping data that we are using to calculate k and l for k-anonymity and l-diversity. To the best of our knowledge we were the first that actually realised the calculation of these KPIs within a cloud based platform to ensure anonymity before data are uploaded and after they were processed and combined with additional data.

5.3 Legal Considerations

5.3.1 In general at European level

After a thorough analysis of the technical part of SmartRegio, we want to have a look at the legal aspects of smart data analysis especially in the field of data protection law. Therefore the SmartRegio approach must be in accordance with legal requirements. So we investigate the technical progress as described above with regard to its compliance with legal requirements, in particular with European Law and the upcoming General Data Protection Regulation (GDPR). At European level, the “Directive 95/46/EC on protection of individuals with regard of the processing of personal data and on the free movement of such data” (Directive 95/46/EC) and the “General Data Protection Regulation” (Regulation (EU) 2016/679) have to be legally assessed. A directive shall be binding, as to the result to be achieved, upon each Member State to which it is addressed, but shall leave to the national authorities the choice of form and methods. In comparison to the directive which has to be implemented by the Member States and is insofar not directly legally binding by itself, the regulation is however directly applicable. The Directive 95/46/EC is repealed with effect from May 25th 2018 (cf. Article 94 GDPR), so that the GDPR will apply from this date on. Therefore subsequently we will only have a look at the GDRP.

5.3.2 General Data Protection Regulation (GDPR)

Before any data protection requirements can be considered at all, the fundamental question about the applicability of data protection law must be clarified as a particular challenge. Data protection law is only relevant in the case of processing personal data (Article 2 GDPR). Therefore, it must always be the first step to check whether the analysis relates to personal data or not.

The legal link for the assessment of the existence of personal data is Article 4 no. 1 GDPR. Personal data in the context of the GDPR means any information relating to an identified or identifiable natural person; an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier, or to one or more

[16] see https://spark.apache.org/docs/latest/sql-programming-guide.html

[17] USU Katana Platform (see http://katana.usu.de/) is a platform for data processing on big scale, high velocity and variaty

ISBN 978-3-9504173-2-6 (CD), 978-3-9504173-3-3 (print)
Editors: Manfred SCHRENK, Vasily V. POPOVICH, Peter ZEILE, Pietro ELISEI, Clemens BEYER

factors specific to the physical, physiological, genetic, mental, economic, cultural, or social identity of that natural person. In the context of SmartRegio for each individual processing step (data acquisition, data processing, data analysis and data visualisation), it is necessary to check whether personal data are concerned or not. It is highly controversial when a person is identifiable. To handle this problem the use of technical methods for anonymisation as described above is useful to exclude the personal data, and then data protection law is not applicable, or to minimise the risk of identification at least. In case of doubt, data protection regulations should be observed. Then, a legal basis is necessary. Processing shall be lawful only if the data subject (Article 4 no. 1 GDPR) has given consent to the processing of his or her personal data for one or more specific purposes (Article 6 no. 1 lit. a GDPR) or a legal provision exists (e.g., Article 6 no. 1 lit. f GDPR). Processing is lawful if processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data (Article 6 no. 1 lit. f GDPR). Therefore a balancing of interests is necessary. The smart data controller can refer to Article 15 (Freedom to choose an occupation and right to engage in work), Article 16 (Freedom to conduct a business), and Article 17 (Right to property) Charter of Fundamental Rights of the European Union. The data subject can refer to Article 7 (Respect for private and family life) and Article 8 (Protection of personal data) Charter of Fundamental Rights of the European Union. Both legal positions must be harmonised. Furthermore smart data collection and analysis techniques are in general in conflict with fundamental data protection requirements like the principle of purpose limitation (Article 5 lit. b GDPR), the principle of necessity, the principle of data minimisation (Article 5 lit. c GDPR), or the principle of transparency (Article 5 lit. a GDPR). In addition, the controller (Article 4 no. 7 GDPR) has in general many legal obligations (Article 24 et seq. GDPR). Also the data subject has various rights (e.g., right of information and access to personal data, right of rectification and erasure, right to object and automated individual decision-making, cf. Article 12 et seq. GDPR). This should also be taken into account when processing, evaluating and transmitting data, because, if necessary, the procedural steps must be verified and the smart data controller must also be able to comply with the claims of the data subject. In the view of the tendency of the data protection law to eliminate existing enforcement deficiencies increasingly and to prove infringe¬ments with considerable fines, the observance of such obligations is already increasingly important in the sense of compliance. Infringements can be subject to administrative fines up to 20 000 000 EUR, or in the case of an undertaking, up to 4 % of the total worldwide annual turnover of the preceding financial year (Article 83 no. 5 GDPR). Thus, companies which collect and analyse personal data based on smart data analyses are subject to numerous legal obligations which they must comply with. All these problems, which illustrate the challenges concerning new technologies and law, must be analysed and require further intensive scientific research.

6 BUSINESS MODEL AND BILLING

In Section 5.1 we identified different stakeholders and their roles within the SmartRegio eco-system. In order to generate revenue these different stakeholders need to have a common business concept, while all want to sell their products and services:

- Platform operator – generate revenue on management and maintenance of a platform for third parties,
- Data provider – generate revenue on data usage,
- Service provider – generate revenue on usage of implemented algorithms and services, and
- Application developer – generate revenue on orchestrated data-pipes providing new insights on the data.

The first question remains on how to operate in terms of who pays whom in this eco-system. Within SmartRegio we identified several possible options. A selection (cf. Figure 3) of these options is shortly discussed here.

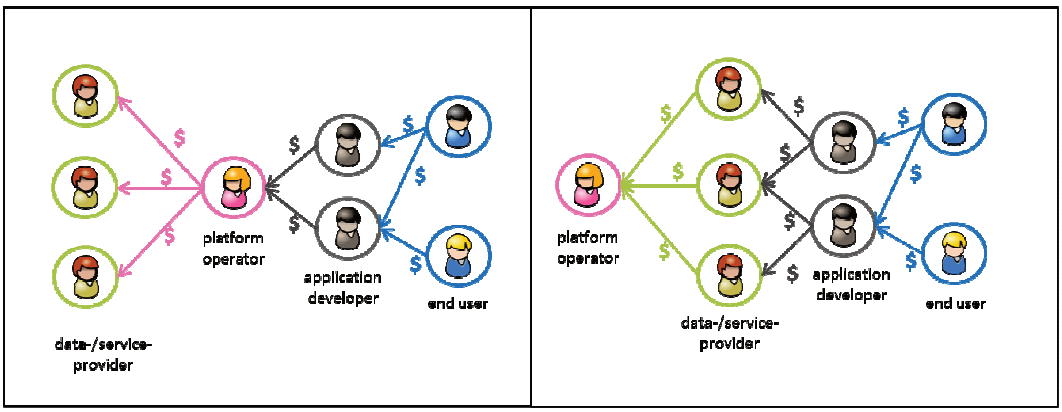

Figure 3: Two possible distribution options within SmartRegio

(1) The end-user pays the application developer as the one that provides the added value to the user.

(a) The application developer pays the platform operator; the platform operator pays data and service providers (cf. Figure 3, left side).

(b) The application developer pays data and service providers, and these pay the application provider (cf. Figure 3, right side).

(2) The end-user pays the platform operator which pays the other stakeholders.

(3) The end-user pays all other stakeholders (which leads to a complex billing for customers).

Having defined the distribution options one needs to define how to evaluate the costs. Here SmartRegio evaluated different options of which a subset is shortly discussed here.

(1) Buying the platform with selected services and data sets, which is not applicable for SMEs due to the large costs this implies.

(2) Pay-per-use based on, e.g., time or data and service usage. Here an appropriate logging mechanism for the platform usage needs to be implemented. There is still variable cost for end users based on the usage. A slight modification would be tiered pricing where the end user gets discounts after a defined amount of usage.

(3) Flat-rate, which allows the usage for a pre-defined time or usage of data and services. This model allows defining fixed costs for end users.

While app stores for several smart phones and operating systems have solved some of the issues arising on the multi-stakeholder scenario the platform for geo-spatial data processing and analysis, SmartRegio still requires some more practical evaluation.

7 SAMPLE USE CASE: THE SPATIAL PROFILER APPLICATION

In this section, the Spatial Profiler application developed within SmartRegio will be presented as a sample use case that illustrates a variety of the challenges and approaches mentioned so far.

The idea behind the Spatial Profiler was to provide an interactive and easy-to-use tool for experts that allows for showing characteristics of given regions and to allow the comparison between regions, given potentially arbitrary many datasets that contain information about them.

Consequently, the Spatial Profiler was developed as a generic, modular web application that offers

- to integrate arbitrary many data sources about a region,
- to work with different spatial granularities, and
- to choose among a variety of different visualisation methods, while at the same time offering to easily integrate new methods.

Furthermore, textual information is supported in a way that allows integrating different mappings for ontologies, so that information from different sources can be easily aggregated or compared for further evaluation.

ISBN 978-3-9504173-2-6 (CD), 978-3-9504173-3-3 (print)
Editors: Manfred SCHRENK, Vasily V. POPOVICH, Peter ZEILE, Pietro ELISEI, Clemens BEYER

The application was developed based on the existing infrastructures ALOE (Memmel & Schirru 2007) and RADAR (Memmel & Gross 2011) developed in DFKI and uses an Elasticsearch index for storing spatial entities as well as the information about them. The visualisations where realised based on the D3 JavaScript library.[18]

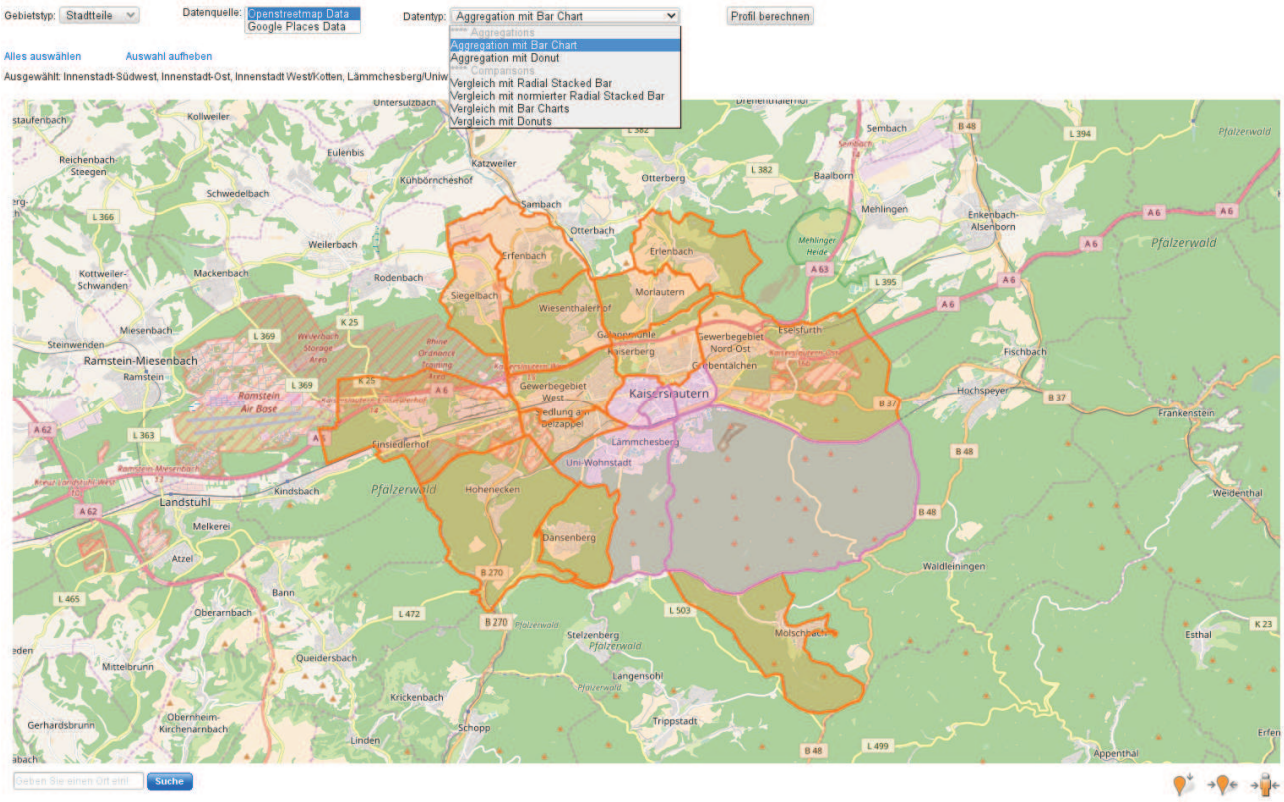

Figure 4: Selection of regions, data sources and visualisation types within the Spatial Profiler application

Figure 4 shows a screenshot of the Sample Profiler with a scenario realised for analysing different quarters in the City of Kaiserslautern. In the top bar, users can choose the following parameters:

- Spatial type: In this scenario the granularities “urban quarters”, “school districts”, and “electoral districts” are offered. The respective shape data required for the display on the map and the analysis of data were provided by the City of Kaiserslautern. Users can select and deselect the regions by simply clicking on them.
- Data source: In our example, two different sources (OpenStreetMaps and GooglePlaces) have been integrated with the focus in distinguishing city parts according to the type of POIs that can be found. They are harvested regularly to ensure that the data sources are up-to-date. Users can choose to use single data sources, but also combinations of them. In order to evaluate the disjunct categories from the sources, a mapping was created to our own ontology that is based on the YellowMap branches categorisation.
- Analysis type: In a first step, users can here decide whether they want to examine all selected regions as one, aggregated region, or if they want to compare them. Depending on this choice and the visualisation means that are integrated, different options are offered then.

[18] see https://d3js.org

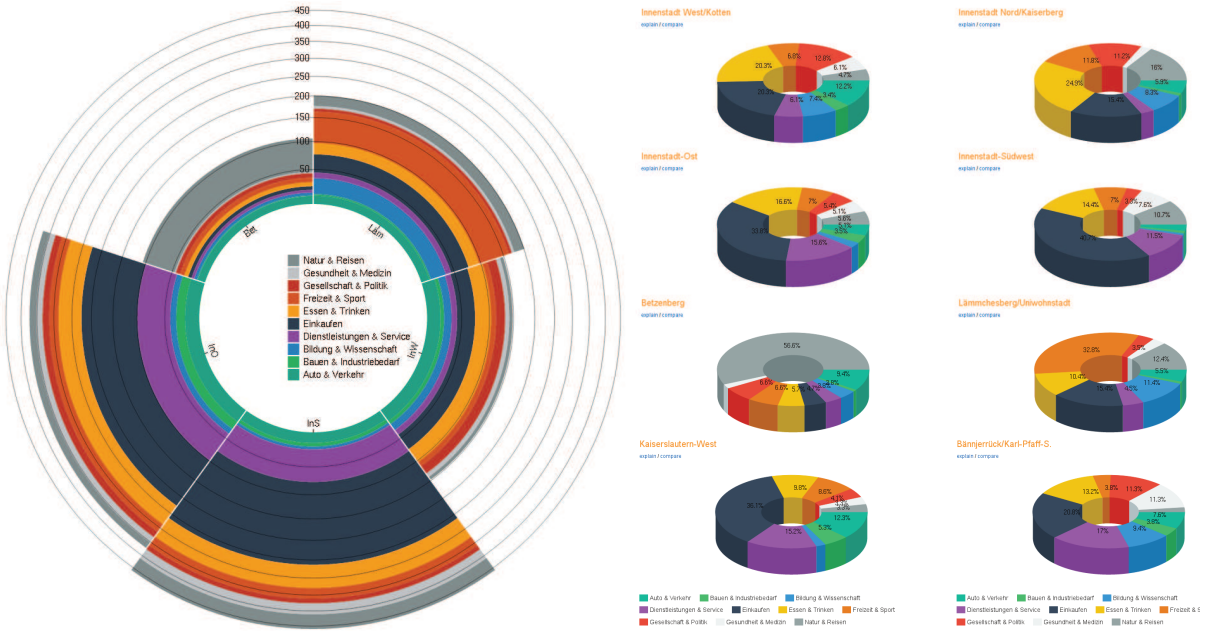

Figure 5: A comparison between the number/distribution of entity types in different city quarters using a non-normalised radial stacked bar (left) and donuts (right) within the Spatial Profiler application

Two different analyses created with the Spatial Profiler are shown in Figure 5. On the left side, five city quarters are compared in a non-normalised radial stacked bar. The size of each sector corresponds to the total number of POIs found in the respective region, and the colour distribution corresponds to the types of POI. This allows, e.g., for characterising quarters as inner-urban or not, while at the same time also providing a lot of information about the structure of the quarter. On the right side, eight city quarters have been chosen for a comparison using donuts to display the distribution of different POI types. Here, it is also possible for each region to display a comparison to the average of all regions in order to find out whether there are interesting specifics.

8 SUMMARY AND CONCLUSIONS

The main hypothesis of the SmartRegio project was that the intelligent analysis of spatial data available already today or in the near future, could help SMEs and local/regional administrations significantly with their strategic decisions regarding local and regional topics – in a great variety of applications such as infrastructure planning, urban planning, geo marketing, all kinds of location planning, to mention just a few. Novel and upcoming Smart Data technologies should be able to integrate heterogeneous kinds of data (open government data, user-generated data, social media content streams, geo-referenced sensor data streams, proprietary data sets like customer and marketing data, etc.), to analyse them intelligently and to visualise the results in user-friendly ways – thus making Smart Data analytics available to domain experts without specific Data Science know-how.

Within the SmartRegio project, we developed and integrated a number of building blocks and technologies for such a solution and showed how they can play together in an integrated architecture. For data acquisition, data ingestion, and data integration, ETL processes have been established for different kinds of data sources (geo data, social media content, non-spatial big-data streams) which are then merged into the SmartRegio Data Warehouse. Central topics were always (i) the creation of spatial references for data without them (geocoding), and (ii) the spatio-temporal integration of data with spatial references in the presence of different spatio-temporal resolutions, granularities and reference systems. For data processing and analytics, a processing-workflow machinery has been realised based on a processing-service library, a metadata management for data sets and processing services, and an orchestration engine; all that managed from the SmartRegio Configuration Cockpit for Data Scientists who are competent with the data and the processing services. For end users, knowledgeable in the application domain and in the analysis goals, but not necessary in the ICT technologies and manipulated data sets behind, a couple of intuitive, expressive visualisations have been created, some of them interactive, in order to give to the end users some freedom for their own experiments and analyses (examples of this interactive visualisation approach have been shown in Section 7).

ISBN 978-3-9504173-2-6 (CD), 978-3-9504173-3-3 (print)
Editors: Manfred SCHRENK, Vasily V. POPOVICH, Peter ZEILE, Pietro ELISEI, Clemens BEYER

Further modules for ensuring data security and for checking privacy requirements have been added to the platform in order to prepare a real-world application.

Business model and billing aspects have been investigated and partially integrated into the software prototype. Challenges in that respect also rely on the acceptance of the stakeholders. Our experience is that specifically SMEs want to have transparent and, ideally, fixed costs so they can plan ahead without taking too much risks. However, this increases the complexity in having a billing model for other stakeholders like data and service providers.

To sum up: Many technical solution elements have been realised at least in a basic form. The level of heterogeneity should not be underestimated, whereas the accessibility and quality of available spatial data should not be overestimated. Often, the required data is not available at the needed level of granularity. Integration of data with different resolution scales and spatial reference systems may require tricky and/or heuristic approaches. This makes spatial data integration still a cumbersome and often manual process. Here, more transformation operators / workflows and a higher degree of automation still provide some interesting future work. Also the realisation of successful business models and billing approaches requires more practical experience and experimentation. Finally, the implemented interactive visualisation methods found great acceptance by test users. But for a broader and easier application of the SmartRegio platform in a wide range of real-world applications, the processing-service library and the toolbox of visualisation methods will still have to be extended if non-technical end users shall be able to work efficiently and intuitively with the SmartRegio platform. Yet, SmartRegio developed prototypical and promising solutions.

So far, SmartRegio delivered a number of technical solution elements for practical employment of Smart Data technologies. But, the biggest problems for project impact were not encountered on the technical, but on the non-technical side.

On the one hand, legal aspects may become a significant hurdle for practical Smart Data applications and prevent businesses to invest in it. We have thoroughly analysed the current state of affairs. In terms of security, trust, and privacy protection there is still some uncertainty for all stakeholders (Kollmann 2016), (QSC 2016), (Pols & Vogel, 2017), (Zacher 2016). And with the European General Data Protection Regulation (GDPR) applicable as of May 25th, 2018 the uncertainty on the provider side will possibly rise (Schonschek 2017). Even though the cloud usage has increased in the last few years, the number of private cloud users is still higher than the number of public cloud users (Pols & Vogel, 2017). Therefore trust in European cloud services needs to be implemented by providing secure services and privacy protection. SmartRegio addressed these topics and builds upon existing standards and state of the art technology.

On the other hand, the availability of useful and usable Open Data may have been the biggest hurdle. Especially, the lack of machine-consumable data was one of the biggest problems. In particular in public administrations, a lack of organisation structures (workflows, defined procedures, responsibilities) and staff that enables an organisation-wide gathering, analysis and publication of data, could be observed. Knowledge about internally available data sets, legal and licensing (!) issues, potentials as well as challenges are often missing, and processes that could help to overcome these problems do not exist. If Open Government Data (OGD) shall act as a fuel for initiating a European Data Economy, much political investments must be spent here. Many concrete activities can be imagined: Concrete large-scale lighthouse projects, more standardisation activities, budgets and binding plans for OGD realization, a closer cooperation of politics, administration, companies, science, and much more. The investments would certainly pay off in better decisions and novel usage ideas in the administration and in the economy.

9 REFERENCES

Andreas Bender: Anwendbarkeit von Anonymisierungstechniken im Bereich Big Data. Master Thesis, KIT, Karlsruhe, Germany. 2015.

Sebastian Bretthauer: Compliance-by-Design-Anforderungen bei Smart Data – Rahmenbedingungen am Beispiel der Datennutzung im Energiesektor, ZD 2016, S. 267 ff.

Daniela Hochfellner, Dana Müller, Alexandra Schmucker and Elisabeth Roß: Datenschutz am Forschungsdatenzentrum. FDZ method reports. 2012. Ref.: http://doku.iab.de/fdz/reporte/2012/MR_06-12.pdf

Jörg Höhne: Verfahren zur Anonymisierung von Einzeldaten. Band 16 der Reihe Statistik und Wissenschaft. Statistisches Bundesamt. 2010.

Zaheer Khan, David Ludlow, Richard McClatchey and Ashiq Anjum: An architecture for integrated intelligence in urban management using cloud computing, in: Journal of Cloud Computing: Advances, Systems and Applications, Volume 1, Number 1. Springer, 2012.

Malte Kollmann: Cloud-Anbieter haben Nachholbedarf beim Thema Sicherheit. Electronic document. Date of publication: November 15, 2016. Retrieved May 30, 2017, from https://www.computerwoche.de/a/cloud-anbieter-haben-nachholbedarf-beim-thema-sicherheit,3326314.

Martin Memmel, Rafael Schirru: ALOE - A Socially Aware Learning Resource and Metadata Hub, in: Martin Wolpers, Ralf Klamma and Erik Duval (Eds.): Proceedings of the EC-TEL 2007 Poster Session. CEUR workshop proceedings, 2007.

Martin Memmel, Florian Groß: RADAR - Potentials for Supporting Urban Development with a Social Geocontent Hub, in: Schrenk, M.; Popovich, Vasily, V.; Zeile, P. (Eds.): Proceedings REAL CORP 2011, p. 777-784, ISBN 978-3-9503110-1-3, Schwechat, (Austria), 2011.

Axel Pols, Marko Vogel: Cloud Monitor 2017 – Eine Studie von Bitkom Research im Auftrag von KPMG. Electronic document. Date of publication: March 14, 2017. Retrieved May 30, 2017, from https://www.bitkom.org/Presse/Anhaenge-an-PIs/2017/03-Maerz/Bitkom-KPMG-Charts-PK-Cloud-Monitor-14032017.pdf.

QSC AG: In die Cloud? Aber sicher! Electronic document. Date of publication: July 15, 2016. Retrieved May 30, 2017, from https://digitales-wirtschaftswunder.de/in-die-cloud-sicher.

Gerd Ronning, Roland Sturm, Jörg Höhne, Rainer Lenz, Martin Rosemann, Michael Scheffler and Daniel Vorgrimler: Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten. Band 4 der Reihe Statistik und Wissenschaft. Statistisches Bundesamt. 2005.

Martin Rosemann: Auswirkungen datenverändernder Anonymisierungsverfahren auf die Analyse von Mikrodaten. Institut f. Angew. Wirtsch.-Forsch. 2006.

Oliver Schonschek: Datenschutz-Grundverordnung - was Cloud-Nutzer wissen müssen. Electronic document. Date of publication: May 04, 2017. Retrieved May 30, 2017, from https://www.computerwoche.de/a/datenschutz-grundverordnung-was-cloud-nutzer-wissen-muessen,3330645.

Matthias Zacher: Unternehmen müssen ihre Vorbehalte abbauen. Electronic document. Date of publication: December 19, 2016. Retrieved May 30, 2017, from https://www.cio.de/a/unternehmen-muessen-ihre-vorbehalte-abbauen,3260926.

ISBN 978-3-9504173-2-6 (CD), 978-3-9504173-3-3 (print)
Editors: Manfred SCHRENK, Vasily V. POPOVICH, Peter ZEILE, Pietro ELISEI, Clemens BEYER