A linked open data based system for flexible delineation of geographic areas

Ali Khalili*, Peter van den Besselaar**, and Klaas Andries de Graaf*

* Department of Computer Science, Vrije Universiteit Amsterdam, NL
a.khalili@vu.nl , ka.de.graaf@vu.nl
** Department of Organization Sciences, Vrije Universiteit Amsterdm, NL
p.a.a.vanden.besselaar@vu.nl

In order to investigate the dynamics urban systems function, the OECD (Organization for Economic Co-operation and Development) in collaboration with EC (European Commission) and Eurostat have introduced the concept of Functional Urban Areas (FUAs). FUAs are based on a fixed set of socio-economic and environmental factors and provide a basis for an agreed definition for measuring development of urban and metropolitan areas. However, as FUAs are predefined in terms of specific characteristics, FUAs will not fit for researchers who conceptualize areas in another way.

Therefore, we developed an adaptive approach for dynamic and multi-faceted delineation of FUAs, rather than merely relying on a rigid schema with a fixed list of FUAs per country. This adaptive definition of FUAs demands integration of data from multiple up-to-date linked data sources. In this paper, we describe an approach and implementation for a Linked Open Geo-Data space, which combines openly available spatial and non-spatial resources on the Web to classify urban areas with the aim to more flexibly monitor and research urban development.

The approach is part of the SMS platform, a data-platform for science and innovation studies (sms.risis.eu), especially the 'geo-services' within SMS. These geo-services are based on a series of open geo-resources, such as GADM[1], OpenStreetMap[2] and Flickr geotagged data[3]. By integrating these geo-resources, the SMS platform can give for an entity's address the geo-location up to 11 different levels (Figure 1).

We illustrate this with an example of a service to determine the geographical location if one knows an address (or even only an organization name). As shown in Figure 2, in the top left part of the screen the address "Vrije Universiteit Amsterdam" is inserted, and the application has as output various maps and, in the bottom right, the geo-characterization of the inserted address at eleven levels.

Figure 2 shows the various administrative boundaries for the geocoded address. in this case, level 8 represents LAU 2 (Local Administrative Unit). The platform can be used to do this for large(r) amounts of addresses, and the output then is not on the screen, but in a tabular form. For this purpose, the SMS platform provides a *spreadsheet add-on* where users can enter a (long) list of addresses to be geocoded. In the future we aim at adding different distance concepts, such as travel distance (time, frequency, price, etc.) as part of the innovative SMS geo-services.

[1] Database of Global Administrative Areas: http://www.gadm.org

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

[3] http://www.flickr.com/services/shapefiles/2.0/

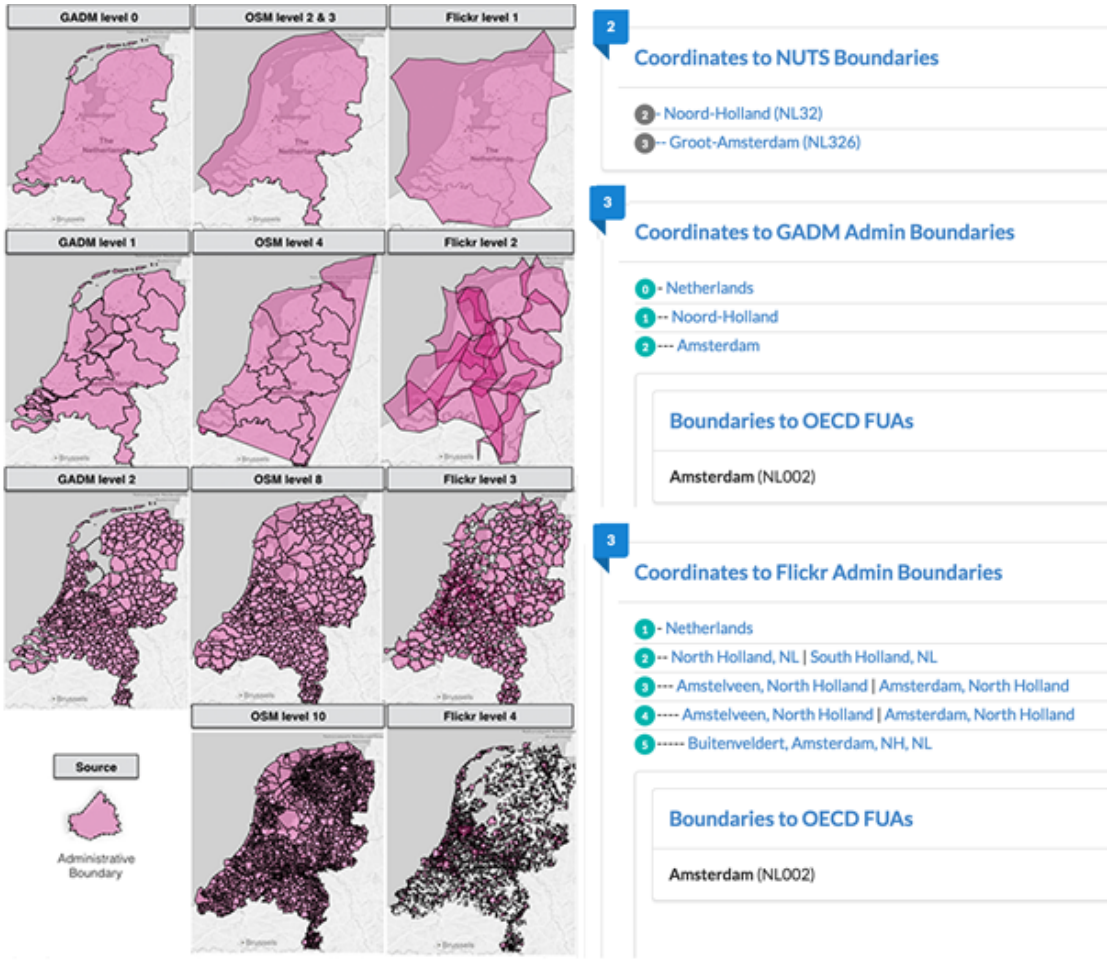

Figure 1: Mapping administrative boundaries using open geo-resources

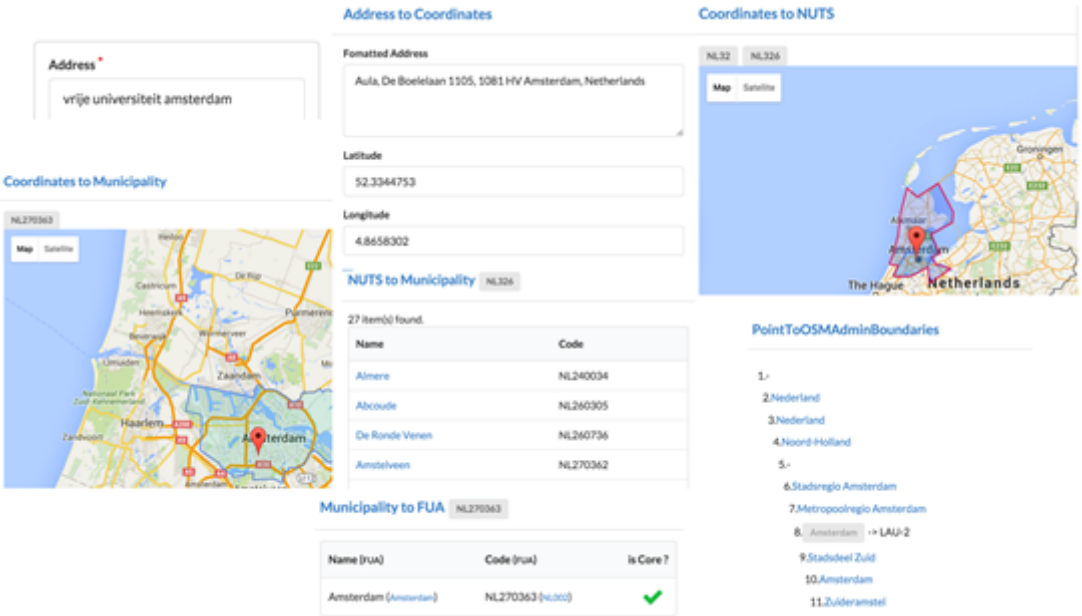

Figure 2: Geo-locating services

The advantage of this service is twofold. Firstly, through location of an entity (e.g., a university) within a *given administrative boundary* (e.g., a municipality), one can link the entity to other (e.g., statistical) data available at that geographical level. For example, National Statistical Agencies like Statistics Netherlands (CBS)[4], and international organizations like OECD and Eurostat, publish many socio-economic, population, transport, environmental, and other data at several spatial levels such as regions, provinces, municipalities and neighborhoods. These data then can be used as context information for the entities.

Secondly, interlinking these statistical data with geographical data from administrative boundaries, *alternative geographical schemas* similar to OECD's Functional Urban Areas can be provided. For example, indicators such as population, living costs, income, facilities, power consumption, can be taken into account for defining the geographical boundaries, specifically adapted to the theoretical or practical perspective of a study. This enables a researcher to create tailored geographical boundaries defined in terms of the characteristics he/she considers relevant. Different levels and aggregates of boundaries can then be mapped based on the statistical data provided for those boundaries.

As a use case, we investigate the regional socio-economic properties of innovative activities, as stimulated by recent research and innovation policies in the Netherlands. The new policy in the Netherlands is oriented at the 'top sectors' of the economy, which were selected after consultation of policy makers, representatives of the research system, and entrepreneurs in the country. After selecting these 'top sectors', a substantial part of public research funding was devoted to this new policy. Consortia can apply for funding, and they should exist of companies and research organizations (such as universities) with a company as main applicant. Because of this context, the funded projects can be considered as a useful representation of RTD (Research Technology Development) collaboration for innovation. In this case, we are interested in the geographical properties of these consortia. In order to investigate this, one needs data about the projects and data about the characteristics of the relevant geographical units. These data are available as open data. In this case, the following open datasets are deployed:

- RVO dataset[5] providing a list of research and innovation projects that have received subsidies and financial support from the Netherlands Enterprise Agency (Rijksdienst voor Ondernemend Nederland). Projects information includes companies and research institutes that are collaborating in the project, but also the geographical coordinates of the project main applicant.
- CBS dataset[6] published by the Statistics Netherlands provides different types of statistical information on dimensions such as employment and income, economy, society and regional aspects of municipalities and regions in the Netherlands.

As one does not know *ex ante* what the level of geographical organization of the consortia is, we need to define the 'geographical containers' in different ways. That would enable us to find out at what geo-level these consortia are in fact organized. Then we can identify the

[4] https://www.cbs.nl

[5] http://www.rvo.nl/open-data-van-rvonl

[6] https://www.cbs.nl/en-gb/our-services/open-data

relevant characteristics of these geographical 'containers' of the projects. As an example, we first determined different sets of Urban Areas based on different statistical properties provided by the CBS dataset and different levels of open administrative boundaries. Figure 3 shows the delineation of these Urban Areas through (i) the population size, (ii) the number business establishment and through (iii) combinations of these two indicators at the municipal level. The geo-location service enables the user to put different weights on different indicators when delineating the boundaries, which could be used for an analysis of the role of specific factors. Boundaries typically differ when defined by different characteristics. When compared with the OECD FUAs[7], the adaptive Urban Areas take into account additional regions (administrative boundaries). Figure 3 show the different maps of urban areas that result from different definitions. The darker the area is, the higher the score on the variables used to define the Urban Areas.

In the second step, geographical coordinates of projects are mapped to these Urban Areas. Fig. 8 shows the result of the mapping where frequency of the projects is highlighted: (again) the darker the color, the higher the number of awarded projects in that area.

As can be seen when comparing Figure 3 and Figure 4, by far not all (Functional) Urban Areas have projects. But more importantly, the different ways the Urban Areas are defined, lead to different outcomes. Using the OECD FUAs (bottom-right map in Figure 4), or the population density based FUA (top-left map), one would miss some of the relevant areas.

We plan to improve the underlying algorithm to be able to closely reproduce the OECD FUA. Moreover, we plan to use geographical data from GADM, instead of OpenStreetMap, which we used for this case study, to improve the mapping of projects (cf. see Figure 4). In the case study we identified hybrid FUAs by selecting the 35 areas that contain most population and businesses. This decision was made in order to compare to the 35 OECD FUA areas.

After identifying the spatial dimension of the project consortia, one may aim to find out whether the ‘geographical containers’ share socio-economic or other factors, in which they differ from less innovative areas. In this example, we only mapped the geo-location of the main applicant in each of the consortia. In a next step, we will also include the geo-location of the other partners in the projects.

[7] We used the shapefile from the Eurostat used in the large study named urban audit: http://ec.europa.eu/eurostat/web/cities/data/database

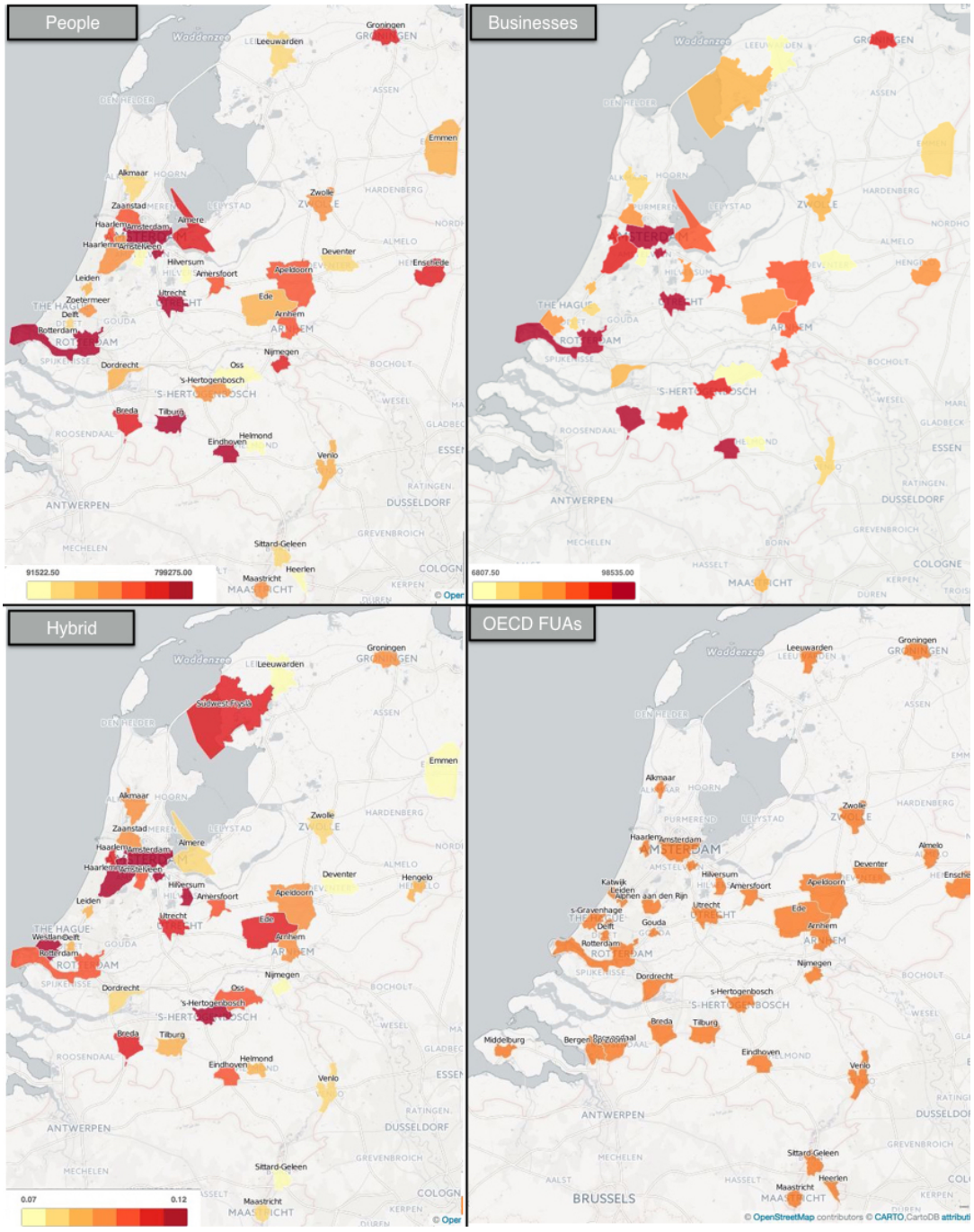

Figure 3. An example of the adaptive delineation of FUAs for the Netherlands based on the open statistical data (populations, business establishments, hybrid and OECD).

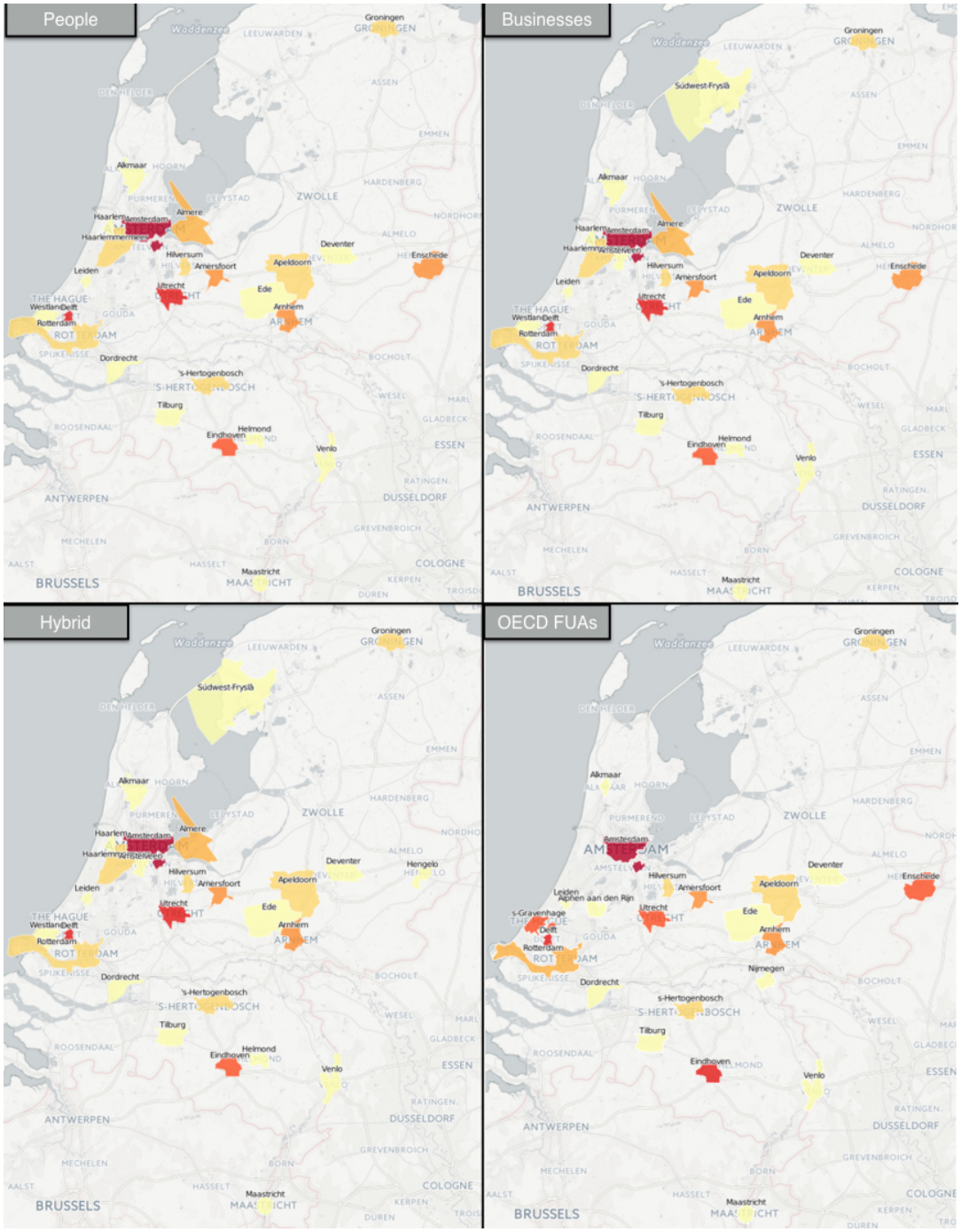

Figure 4. Amount of RVO project subsidies mapped to the dynamically delineated FUAs defined based on the CBS open statistical data and OpenStreetMap boundaries.