Experimenting in techniques and approaches in search on structured georeferenced bibliographic data sources

STSM Reference Code: **COST-STSM-ECOST-STSM-IC1302-220815-060715**

STSM Applicant: Dragan Ivanović, dragan.ivanovic@uns.ac.rs, University of Novi Sad, Serbia

Host: University of Ljubljana, Slovenia, contact person: Marjan Čeh, Marjan.Ceh@fgg.uni-lj.si

Type of STSM: Regular (from Serbia to Slovenia)

Dates and duration: 22.08.2015.-29.08.2015. (a week)

1. Purposes of the STSM:

GEOPOLO is a software system for georeferencing scientific-research publications, i.e. it is a system which enables joining geospatial data to scientific-research publications. This system is implemented for theses and dissertations defended at the University of Ljubljana (http://geopolo.fgg.uni-lj.si/). Currently, librarians manually add geospatial data to publications. The mainly purpose of this short term scientific mission is to research literature about using NLP techniques in order to find out toponyms (geospatial keywords) from the text in order to enable automatic georeferencing of publications. Besides that, the aims of the STSM are developing of new research ideas involving researchers from University of Novi Sad and University of Ljubljana and knowledge transfer in order to enable implementation of similar system as GEOPOLO is at University of Novi Sad.

2. Activities

- In the first part of the STSM, Marjan Čeh introduced Dragan Ivanović with his colleagues and their research fields.
- After that, we started discussion about GEOPOLO and its weak points. Furthermore, we defined the main direction of improving the GEOPOLO system which should enable automatic georeferencing of publications.
- In order to do that, we made research methodology which is going to achieve that goal.
- Also, we reviewed and collected the available Slovenian geospatial data catalogs, as well as Slovenian natural language processing tools.
- As the last, but not at least important, we created the data set which we are going to use as training and validation data set for machine learning based system for automatic georeferencing data set.

3. Results

- The research methodology which is going to achieve automatic georeferencing of publication has been defined. The methodology is based on machine learning techniques, Slovenian natural language processing tools and Slovenian geospatial catalogs. There will be a few challenges in this research which are going to be discussed at The first International KEYSTONE conference (IKC 2015) in Coimbra.

- Slovenian geospatial catalogs has been analyzed and following catalogs are going to be used for this research:
  - RPE (register of spatial units)
  - EHIŠ (register of house numbers)
  - REZI (register of geographical names)
- Slovenian natural language processing tools has been analyzed and Slovene parser (http://eng.slovenscina.eu/tehnologije/razclenjevalnik) and Slovene tagger (http://eng.slovenscina.eu/tehnologije/oznacevalnik) have been selected for this research.
- Training and validation data set has been created by exporting the GEOPOLO database and adopting it for this purpose.

4. Further collaboration

Further collaboration includes realization of the research which has been started within this STSM. The result of this research will be published as scientific paper in some journal of conference proceedings. Besides the scientific paper, the implemented system for automatic georeferencing will be the result of this research and it is going to be integrated within GEOPOLO. The further research and collaboration follows two main directions. The first direction is creation of a strategy for developing the similar system to georeference theses and dissertations defended at the University of Novi Sad (http://dosird.uns.ac.rs/phd-uns-digital-library-phd-dissertations). Preconditions for this are well-developed natural language processing tools for Serbian language and existence of Serbian geospatial data. Those preconditions (especially the second one) are not meet in this moment, but it should be in the future. The second direction is research of using Google Maps JavaScript API (https://developers.google.com/maps/documentation/javascript/examples/) for automatic georeferencing keywords from a query in order to enable keywords-based search on georeferenced publications even though the certain keyword from the query is not mentioned in the certain publication, but the publication's geospatial data are close to or are in relation superset-subset with geospatial data of the certain keyword.

Further collaboration could be achieved using the contemporary ICT technologies such as email and Skype, or through some funding program for supporting joint research activities including this COST action. There is a special program supporting joint research activities of Serbian and Slovenian researchers, but the deadline for the call for two year period 2016-2017 was expired in June of 2015 (http://www.mpn.gov.rs/medjunarodna-saradnja/naucna-saradnja/bilaterala/523-slovenija) and the next call probable will be in June of 2017.

Ljubljana, Slovenia, 29/08/2015

Dragan Ivanović