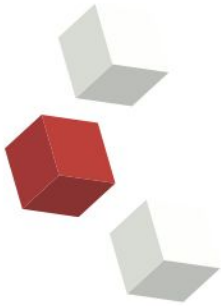

Amsterdam
Data Science

The RDF Unified Migration Portal

Applicants

Supervisor Name	Department/Section	University/Industry
Dr. Rinke Hoekstra and Albert Meroño Peñuela	Knowledge Representation and Reasoning group	Vrije Universiteit Amsterdam
Dr. Auke Rijpma	Research	International Institute of Social History[1] (IISG Amsterdam)

Co-Applicants / supervisors:

Dr. Kathrin Dentler (Vrije Universiteit Amsterdam)

Dr. Richard Zijdeman and Ashkan Ashkpour (IISG Amsterdam)

Rutger Rienks (DataLab Amsterdam, Gemeente Amsterdam)

Annemieke van Velzen (DataLab Amsterdam, Gemeente Amsterdam)

Project Description

Migration has always been a very important topic[2], but with the flow of refugees from crisis areas such as Syria, it is higher on the agenda than ever before. However, data on this topic covers a wide range of topics (migrants, asylum seekers, refugees, countries, municipalities, reception locations (*opvangcentra*), incidents), and is thus highly scattered and fragmented. Datasets on these topics are being published on the Web in different sources, locations, formats, and time periods in a decentralized and unrelated manner. Certainly, these datasets describe real world related phenomena, but they don't reference each other and operate only at a very local scope. This hampers a combined use of these datasets, making a consistent overview of the situation difficult. Consequently, stakeholders such as policy makers, researchers, people involved in the debate, the public, volunteers, and refugees themselves need to spend valuable time and labour in combining the contents of these datasets manually to get meaningful information.

[1] http://www.iisg.nl/

[2] Especially in Amsterdam; see Erika Kuijpers' book on this topic *https://books.google.nl/books?id=fZdyuGV6fFsC*

Our project applies the methodology of Linked Data to publish and link migration related datasets on the Web. Relevant datasets for key use cases on migration, with a focus on the Amsterdam area, will be identified in official and unofficial channels, institutions, media, and social networks (Twitter feeds, Facebook pages, etc.). These datasets will be curated, transformed to Linked Data, described with appropriate vocabularies, and linked to other related datasets and concepts on the Web, to facilitate their automated integration.

The resulting migration knowledge graph will have three valuable applications: First, stakeholders will be saved manual labour when querying across disparate migration data sources in an easily replicable way. Second, intelligent data-driven applications will be possible with minimal effort. Third, the integrated, linked migration data will allow for statistical models to better manage future flows, by potentially enabling the prediction of forthcoming refugee waves.

Project Organisation

The student assistants will work in an interdisciplinary group and collaborate with our stakeholders (see below). We devise two different roles for two assistants:

- *Linked Data expert (0.2 fte).* The first assistant will have experience in the Linked Data publication methodology and strong programming skills. She will be skilled in RDF, Linked Data vocabularies, SPARQL, and will be familiar with typical data integration issues. The assistant is also required to have experience in the whole cycle of software development, with skills on requirements gathering, data modelling, use-case definitions, specification, design, programming and testing. Knowledge on ontology engineering and knowledge modelling, and a record with organisations dealing with the refugee crisis, will be valuable additions. ***This is the assistant that we would like to request via this proposal.***

- *Digital humanities student on migration (0.2 fte)*. The second assistant will be versed in the Digital Humanities, and have interest and expertise in the domain of migration. She will be well aware of the problems that both institutions and refugees face during migration episodes, and suggest where data and tools are needed, thus facilitating provision of use cases. She will be resourceful, and keen on investigating a variety of channels, institutions, media, and social networks to find relevant data, even if poorly structured. ***This assistant will be funded by the IISG.***

Both roles require an interdisciplinary focus, and enthusiasm and curiosity towards both sides of the project, namely on current social phenomena such as migration, and also on technology and computer science, their interaction, and their practical application.

Collaboration

The project will form a strategic partnership between the 'structured data' work package of the NWO funded CLARIAH[3] project, based in Amsterdam (VU, IISG), and with the DataLab of the Municipality of Amsterdam[4], part of the Knowledge Mile. In order to enhance potential synergy amongst the collaborating institutions, the two assistants will spend 0.04 fte at the DataLab Amsterdam facilities. The rest of the time, they will be embedded within the CLARIAH team at VU and IISG. The supervisors will attend weekly meetings at the DataLab once per month.

The data collected and enriched will be published as an official Amsterdam dataset through the DataPunt of the municipality. The use case and study will show the immediate value of linked open data for answering complex societal, and politically relevant questions. This collaboration enables knowledge and technology transfer along three dimensions: 1) Linked Data technology expertise from the VU to the municipality, 2) Valorisation of city data from IISG/UU to the municipality, and 3) concrete datasets, use cases and exposure from the municipality to the academic partners.

[3] CLARIAH is a NWO funded Dutch project for developing national-wise humanities IT infrastructure: http://clariah.nl/
[4] https://www.amsterdam.nl/gemeente/organisatie/overige/datalab-amsterdam/

Concretely, the collaborating institutions are:

- the Knowledge Representation and Reasoning group, department of Computer Science, in the Vrije Universiteit Amsterdam; a world-class research group on Semantic Web technologies and Linked Data publication methodologies.
- the International Institute of Social History[5] (IISH), the world-leading research institute on social history.
- the Municipality of Amsterdam[6], concretely the DataLab Amsterdam[7], a workshop, expertise and partner for urban issues in the municipality.

Deliverables

1. a *study of migration data*, containing a thorough inventory of migration related datasets, especially those related to the Amsterdam city area. It must include the datasets technical particularities (formats, locations, availability, licensing, etc.), and important integration problems that hamper their combination.
2. a *Linked Dataset*, in RDF format, containing a *schema* and *instances* with the contents of the integrated data sources on refugees and their political, geographical, and social context.
3. a *"link set"* dataset, providing links between resources of the linked migration dataset, and resources in the LOD Cloud[8], which will connect the new dataset to a data-Web of 100 billion statements.
4. a *scientific paper*, describing
   (a) the selected data sources and their selection criteria;
   (b) the workflow to generate their Linked Dataset representation;
   (c) modelling decisions behind its schema;
   (d) documentation on the "link set";
   (e) usage instructions, e.g. by providing example SPARQL queries;
   (f) proof that the labour on combining datasets is reduced, the application development is easier, and the migration data gets augmented; and
   (g) prototype and preliminary results of statistical models that can be useful for predictive tasks.
5. a Web-based *demonstration application*, showing the use cases fulfilled.

[5] https://socialhistory.org/
[6] https://www.amsterdam.nl/
[7] https://www.amsterdam.nl/gemeente/organisatie/overige/datalab-amsterdam/
[8] http://lod-cloud.net/

Planning

The project will run for 6 months, with 2 months of summer vacation, from June 2016 - January 2017. The exact plan will however depend on the availability of our assistants. The work plan for these 6 month is as follows:

06 2016	requirements gathering and use cases; dataset study; inventarisation and curation; demo mockup
09 2016	dataset study; inventarisation and curation; vocabulary choices; design choices; application specification
10 2016	dataset study; inventarisation and curation; dataset conversion to RDF; demonstration application development
11 2016	paper writing; dataset conversion to RDF; demonstration application development
12 2016	paper writing; presentation preparation
01 2017	paper submission; presentation on the project results and demonstration