

Geo-Semantic Labelling of Open Data



Sebastian Neumaier, Axel Polleres,
Vadim Savenkov

Open Data Search is hard...

Why is Search in Open Data a problem?

- No natural language cues, little context
- Specific terminology
- Existing knowledge graphs don't cover the domain of CSV Open Data well
- **Open Data is not properly geo-referenced**

Our contribution:

Hierarchical labelling of spatial entities in tabular data

Example Table

<i>federal state</i>	<i>district</i>	<i>year</i>	<i>sex</i>	<i>population</i>
Upper Austria	Linz	2013	male	98157
Upper Austria	Steyr	2013	male	18763
Upper Austria	Wels	2013	male	29730
...

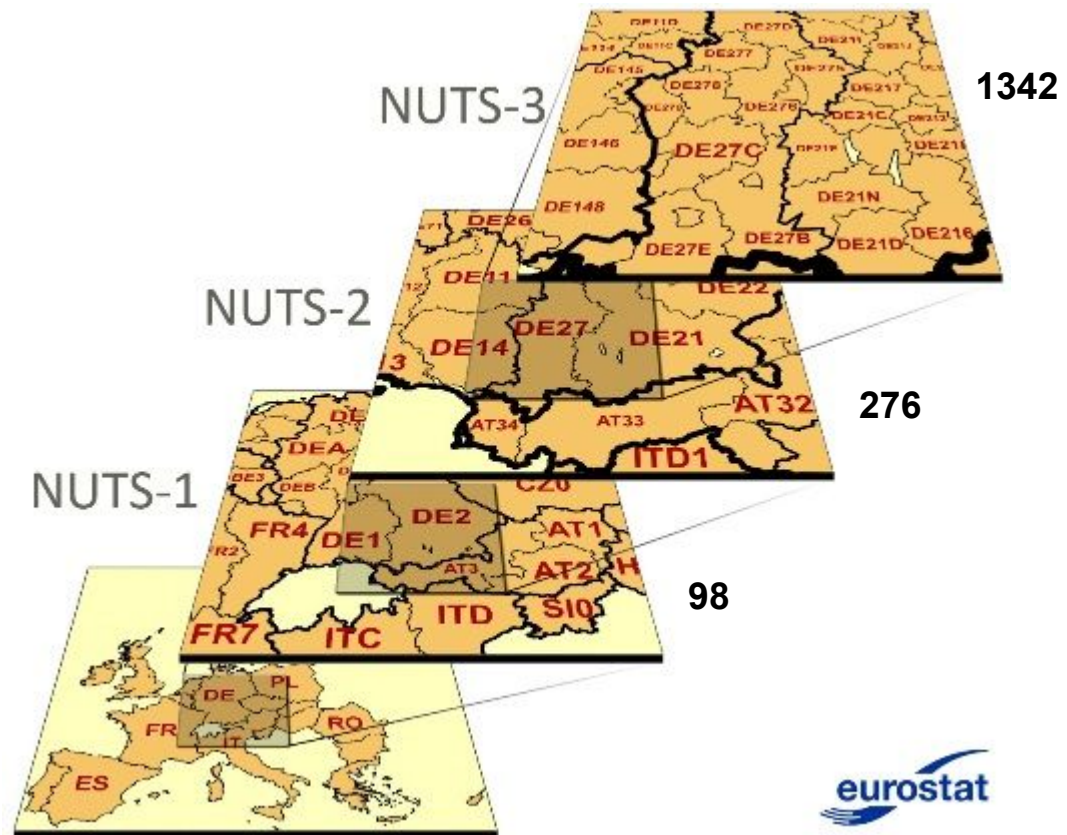
<i>NUTS2</i>	<i>LAU2_NAME</i>	<i>YEAR</i>	<i>SEX</i>	<i>AGE_TOTAL</i>
AT31	Linz	2013	1	98157
AT31	Steyr	2013	1	18763
AT31	Wels	2013	1	29730
...

NUTS: Nomenclature of Territorial Units

Nomenclature des unités territoriales statistiques (fr)

Geocode standard for referencing the subdivisions of countries for **statistical purposes**

Used in many open government datasets, e.g., on data.gv.at



- **Over 11 million** geographical names of entities such as countries, cities, regions, villages, etc.
 - unique identifiers
 - detailed hierarchical description including countries, federal states, regions, cities, etc.
- Also available as RDF Ontology
- **+ additional datasets (e.g. Postal Codes)**

gn:A	gn:A.ADM2	gn:A.ADM3	gn:A.ADM4	gn:A.ADM5
Germany	Bavaria	Upper Bavaria	Munich; Urban District	Munich



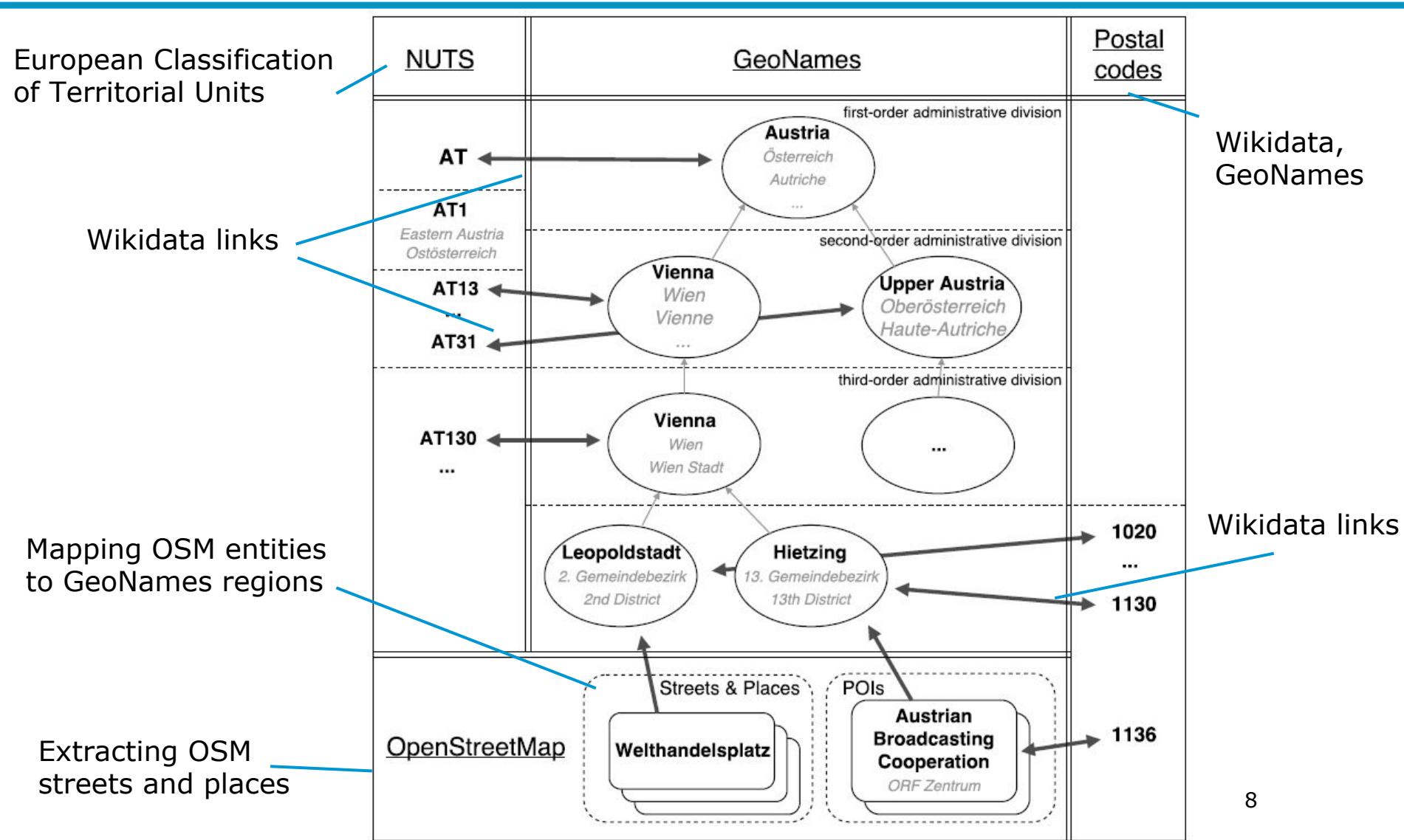
OpenStreetMap

Largest **community-maintained wiki of geo-entities and open map data**. Combines data produced by volunteers using GPS devices with contributions by agencies and companies.

- **Nodes**: specific points defined by a latitude and longitude.
- **Ways** are ordered lists of nodes that define lines or polygons.
- **Relations** between OSM elements.
 - Route is defined as a relation of multiple ways.
- **Tags** are key-value pairs used to describe the meaning of elements,
 - E.g., “highway: residential” on a way element can indicate a road within settlement.

Building a Knowledge Graph of Geo-Entities

Geo-Knowledge Graph Construction



GeoNames postal codes database

- 84 countries and a total of 1.1M entries.
- For each code: place name, and (depending on the country) several parent region/subdivision names.

Heuristic to map postal codes to  GeoNames entities:

- Split the place name of the postal code (from the postal codes database) on delimiters and search for matching GeoNames entries (from the main database) in the respective country.
- Use code's ancestor regions to **disambiguate (resolve multiple matches)**



Mapping NUTS Codes to GeoNames Entities



- **WIKIDATA** maps **1316** out of **1716** NUTS codes to GeoNames entities.
- Remaining 400 codes are typically NUTS regions where no Wikidata and/or GeoNames entry exists
 - In most cases: no corresponding administrative region.
 - For instance: NUTS **AT126 Wiener Umland/Nordteil** and **AT127 Wiener Umland/Südteil** are no administrative districts, but groups of neighboring districts (Northern resp. southern neighborhood of Vienna), and therefore have no separate Wikidata/GeoNames entity to map.



Open Street Maps Integration

1. Use  **GEOFABRIK** to download an OSM snapshot on a country level.
2. With the OSM  **Nominatim** service, get potential OSM entities for the NUTS 1, 2 and 3 identifiers.
3. If multiple matches: choose the OSM entity (e.g., region) at the same NUTS level as the corresponding GeoNames region.
4. Query **Nominatim** again to get polygons of the disambiguated regions.
5. Extract street names, places etc. based on polygons from the OSM snapshot, and add them to the hierarchy under the corresponding GeoNames entity.

Dataset Labeling Algorithm

The CSVs are processed column by column. The following algorithm is used to detect and label geo-entities in CSV columns:

Three cases:

1. NUTS or Postal Codes Column
2. Textual Geo-references
3. Best effort mappings (Address Parsing & OSM Mappings)

Labeling Case 1: NUTS or Postal Code column

- Pattern based preprocessing via regular expressions
- Threshold (currently 90% of values) to classify the column as NUTS or Postal Code
- Values in classified columns are mapped to GeoNames entities
- **NUTS codes** are easy to detect with a regex
 - two letter country code followed by zero to three numbers.
- **Postal codes** are preprocessed with a permissive regex filter
 - Attempt to match with known postal codes of the dataset's country of origin.

Labeling Case 2: (Short) string values

Word(s) or text, try to map the column values to GeoNames:

- Collect all possible entity mappings for all column values, including any ambiguous labels.
 - E.g., all GeoNames entries for “Linz” and likewise for the other values (Steyr, Wels, ...).
- For these, compute the score as a **number occurrences of ancestors**.
 - E.g., Linz as an Austrian city would have score 7 while as a city in Saxony - only score 2.

<i>federal state</i>	<i>district</i>	<i>year</i>	<i>sex</i>	<i>population</i>
Upper Austria	Linz	2013	male	98157
Upper Austria	Steyr	2013	male	18763
Upper Austria	Wels	2013	male	29730

Labeling Case 3: Best Effort Match

If Cases 1 and 2 do not apply:

- Use the **Libpostal** Python library to extract streets and place names from strings as a preprocessing
- Try to match with the known OSM entries
 - Disambiguate multi-matches (same technique as in Case 2)
- Use the entities found in the metadata to reduce the set of multi-matches.
 - Prefer OSM candidate mappings which are within one of the regions detected in the metadata.

Extract geo-entities from the titles, descriptions and publishers of the datasets:

1. Tokenize input fields and remove stopwords.
2. We group the input by word sequences of up to four words, i.e. all single words, groups of two words, ..., and run the labeling algorithm for mapping a set of values to the GeoNames labels (including the disambiguation step).

Dataset Labelling

Metadata descriptions

- Geo-entities in titles, descriptions, organizations
- Restricted to origin country

CSV cell value disambiguation

- Row context:
 - Filter candidates by potential parents (if available)
- Column context:
 - Least common ancestor of the spatial entities

Metadata

Tourismus - Ankünfte und Nächtigungen in Oberösterreich

Ankünfte und Nächtigungen in den oberösterreichischen Meldegemeinden ab dem Jahr 2000

Daten und Ressourcen

Ankünfte und Nächtigungen in OÖ seit dem Jahr 2000 Entdecke

Veröffentlichende Stelle: Land Oberösterreich

Datenverantwortliche Stelle: Land Oberösterreich, Abteilung Statistik

Lizenz: Creative Commons Namensnennung 3.0 Österreich

Link zur Lizenz: <https://creativecommons.org/licenses/by/3.0/at/deed.de>

Attributbeschreibung: NUTS2 => Bundesland Oberösterreich Gemeindenummer bzw. Gemeinename => Erhebungsjahr => Kalenderjahr Erhebungsjahr lt. Tourismus-Statistik-Verordnung 2002 §2 Abs.7: Städte und Gemeinden mit mehr als 1.000 Gästenächtigungen im Kalenderjahr

CSV

NUTS2	Gemeinename	Jahr	Ankuenfte	Naechtigungen
AT31	Linz	2000	340880	579683
AT31	Steyr	2000	38726	78644
AT31	Wels	2000	84370	150417
AT31	Altheim	2000	4989	10744
AT31	Aspach	2000	2637	21316
AT31	Auerbach	2000	484	3541
AT31	Braunau a. Inn	2000	15748	33911

Diagram illustrating disambiguation of the 'Linz' entity:

- Upper Austria (Oberösterreich, Haute-Autriche)
- Linz (multiple instances)
- Saxony
- Germany

Disambiguate

Showcase: Search Interface

data.wu.ac.at/odgraphsearch

Faceted query interface:

- Geo-entities
- Full-text queries

Back end:

- **MongoDB** for efficient key look-ups
- **ElasticSearch** for indexing and full-text queries

Dataset

Search Index



Web Interface

Geo-Entity

Knowledge Graph

Temporal filters

Leopoldstadt

Republic of Austria > Wien > Wien Stadt > Gemeindebezirk Leopoldstadt Spatial entity or Full-text results

Publizistikförderung - Publizistikförderung [RTR-GmbH](#)
<http://data.gvat/>
Im Rahmen der Publizistikförderung werden periodische Druckschriften gefördert, die sich mit politischen, kulturellen oder weltanschaulichen Themen befassen. Beschreibungen der Daten sowie der Möglichkeit der Datenabfrage über eine REST-Schnittstelle siehe <https://data.rtr.at/Publizistik>.

gesetzlichegrundlage	zeitschrift	foerderungswerber	strasse	plz	ort	foerderbetrag	jahr	status
Abschnitt II PubFG 1...	TARANTEL	Werkkreis Literatur ...	Vivariumstraße 8/4/1...	1020	Wien	5742.22	2017	abgeschlossen

Hunde pro Bezirk Wien - Anzahl der Hunde pro Bezirk [Stadt Wien](#)
<http://data.gvat/>
Anzahl der Hunde pro Bezirk und Rasse

NUTS1	NUTS2	NUTS3	DISTRICT_CODE	SUB_DISTRICT_CODE	Postal_CODE	Dog Breed	Anzahl	Ref_Date
AT1	AT13	AT113	90200	..	1020	Zwergspitz / Mischl...	10	20180601

Rechtsträger die der Kontrolle des Rechnungshofes unterliegen [Rechnungshof](#)
<http://data.gvat/>
Der Rechnungshof ist als unabhängiges Organ der externen öffentlichen Finanzkontrolle für Bund, Länder und Gemeinden eingerichtet und dazu berufen, die gesamte Staatswirtschaft zu überprüfen. Dem verfassungsrechtlichen Auftrag folgend prüft der Rechnungshof daher sowohl die Gebarung öffentlicher Einrichtungen als auch privater Rechtsträger, an denen Bund, Länder oder Gemeinden mit öffentlichen Mitteln beteiligt sind. Hier finden Sie eine Liste jener Rechtsträger, für die der Rechnungshof prüfzuständig ist (Stand: 1. Juli 2018; letzte Aktualisierung: 21. Juli 2018).

Bezeichnung des Rec...	Straße und Hausnumme...	PLZ	Ort	Land
Zbp, Zentrum für Ber...	Welthandelsplatz 1 G...	1020	Wien	Österreich

All Organizations at WU Vienna - All Organization Units [WU Wien Open Data](#)
<https://www.opendataportal.at/>
A hierarchical list of all organization-units at the Vienna University of Economics and Business.

org_id	name_en	name_de	location_id	floorname	street	zip	city	otype	url
3979	Universitätslehrgäng...	Universitätslehrgäng...	002_00_OG05_011600	Gebäude EA	Welthandelsplatz 1	1020	Wien	RF	https://execu

Top Locations Wien - top-locations-wien.csv [Stadt Wien](#)
<http://data.gvat/>
Touristische Auswahl der wichtigsten POIs in Wien, Ca 140 POIs in den Kategorien Sightseeing, Museen, Gastronomie, Nightlife, Musik, Shopping, Cafés und Restaurants. Jede Location enthält allgemeine Infos wie z.B. Adresse, Telefonnummer sowie eine Kurzbeschreibung und die Geodaten.

title	category	Beschreibung	address	zip	city	geo_latitude	geo_longitude	tel_1	tel_1_comment	tel_2	tel_2
Wiener Sängerknaben	musicstage	Die "jüngsten musika...	Obere Augartenstraße...	1020	Wien	48,224458	16,3734017				

Die 1001 größten IT-Unternehmen Österreichs - Top 1001 IT-Unternehmen Österreichs [Computerwelt, CW Fachverlag GmbH](#)
<https://www.opendataportal.at/>
Die größten Unternehmen der Informationstechnologie in Österreich, gereiht nach Umsatz und Anzahl der Mitarbeiter. Die Liste wird laufend aktualisiert und Ende August im Sonderheft "Top 1001" der IT-Zeitung "Computerwelt" veröffentlicht.

Name	Kurzbezeichnung	PLZ	Ort	Telefon	Website	Umsatz 2014	Mitarbeiter 2014	Umsatz Info	Schätzung 2014	Umsatz 2013	Mitar 2013

Index Size

Portal	Datasets	Resources	thereof CSVs	Indexed
data.gv.at	2399	9091	2794	2427
opendataportal.at	414	1061	473	442
govdata.de	19711	56584	14542	5396
offenedaten.de	10902	24247	4408	3308

	<u>total</u>	data.gv.at		opendataportal.at		govdata.de		offenedaten.de	
<u>Columns</u>	3054	717	(30%)	7	(2%)	1126	(21%)	1204	(36%)
GeoNames	2850	587		5		1105		1153	
OSM	306	185		3		75		43	
<u>Metadata</u>	11 028	2391	(99%)	441	(99%)	4895	(91%)	3301	(99%)

40 random datasets, manually evaluated

- 16 contain no georeferences in columns.
- **17 correctly labeled cell values** (of remaining 24)
- **4 have inaccuracies in OSM labels and 4 in GeoNames labels**
- **Incomplete labels: 7 datasets**
- Metadata labels: 33 datasets metadata based labels complete, but in most cases (32) extraneous labels added:
 - E.g. data publisher “Stadt Wien” was linked to
 - the city of Vienna, Austria, and
 - **“Stadt Wehlen”**, a city in Saxony mentioned as “Stadt”
 - ***Solution: take the country of origin into account as a workaround, need confidence estimation for mappings***

Evaluation (2): Potential false Negatives

Three sets of datasets picked:

- **20 with metadata labels and no column labels assigned**
- **20 with column labels and no metadata mappings**
- **20 without column or metadata labels**
- Geo-labelling based on title and publisher metadata: 100% complete
- 40 datasets without any assigned metadata labels do not provide any geo-information cues in their metadata descriptions.
- **For 9 of the 60 datasets we identified columns with potential geo-data which remained unlabeled.**
 - Particularly in the set of 20 datasets without any assigned metadata and column labels we found 7 candidates with missing labels.

Evaluation (2): Error classes

1. Corresponding entities are missing in the base knowledge graph.
Solution: integrate more entities and/or find alternative names (e.g., by using the multi-lingual labels from Wikidata/DBpedia).
2. City/region names are embedded in text, or combined with other content in the cell, e.g., the region type.
Solution: improved pre-processing for the CSV cell values.
3. The table consists of several sub-tables, where each sub-table has a geo-label as "title". The column contains very few labels, below the threshold.
Solution: improved parsing algorithm for better understanding of the table's structure, e.g., if the table is horizontally or vertically oriented.

Present Shortcomings of Open Dataset Search

- Most of Open Data it is not (yet) Linked.
- Geo-spatial information is hidden in datasets.

Our contribution:

- Hierarchical knowledge graph of spatial entities
- Algorithms to annotate CSV tables and their metadata descriptions

Current / future work:

- temporal dimensions
- Enable GeoSPARQL (or an alternative geospatial-query language)
- Parsing coordinates in datasets
- Parse other file fomats, e.g., XML, PDF, ...
- Test other domains such as tweets or web pages

Thank you!
Questions?



VIENNA UNIVERSITY OF
ECONOMICS AND BUSINESS

Sebastian Neumaier

sebastian.neumaier@wu.ac.at

twitter: @sebneum

sebneumaier.wordpress.com

Axel Polleres

axel.polleres@wu.ac.at

polleres.net

Vadim Savenkov

vadim.savenkov@wu.ac.at