Automated Semantic Validation of Crowdsourced Local Information – The Case of the Web Application "Climate Twins"

Alexander Kaufmann, Jan Peters-Anders, Sinan Yurtsever, Luca Petronzio

Automated semantic validation of crowdsourced local information - the case of the Web application "Climate Twins"

Alexander Kaufmann[1], Jan Peters-Anders[1], and Sinan Yurtsever[2]

[1] AIT Austrian Institute of Technology GmbH, Vienna, Austria
{alexander.kaufmann, jan.peters-anders}@ait.ac.at
[2] ATOS, Madrid, Spain
sinan.yurtsever@atosresearch.eu

Abstract. Climate Twins is a freely accessible Web application which provides easily comprehensible information on the projected climate developments in Europe at a regional level. Climate Twins is based on the idea that an impression of the future climate of a specific region can be received by indicating regions which have a similar present climate. Within the 7th EU FP project "TaToo" ontology-based functionalities have been added to the basic Climate Twins Web application in order to allow users to enrich the original information base - restricted to only two climate parameters ('temperature' and 'precipitation') - by themselves by adding any further geo-located information which they consider relevant in the context of a specific local climate (e.g. information on local vegetation and fauna). This kind of crowdsourcing information raises the problem of safeguarding the quality of the added information. In this paper the authors show how the domain ontology of Climate Twins can be used to semantically validate the coherence of new entries in order to prevent incorrect links.

Keywords: semantics, information enrichment, climate change

1 Introduction

Climate change is often discussed on a rather abstract level. Raising awareness for what climate change actually means to people is important to bring the discussion from an academic level to the level of political action to induce necessary changes in the way we produce the goods and the way we behave as consumers. In order to visualize climate change on the regional level, the AIT Foresight & Policy Development Department has developed a Web application called "Climate Twins" which makes the projected changes of the climate manifest on a high spatial resolution by showing regions with a comparable climate today or in the recent past [5]. The similarity of climate conditions, however, is based on temperature and precipitation only. The picture of the climate in a region is much richer. Therefore it has been decided by the Climate Twins publisher to to involve the user community of the Climate Twins Web

application to enrich the regional climate information in a self-organized way. This required additional functionalities which have been developed in the 7th EU FP-project "TaToo" for which Climate Twins was one of three validation scenarios. An important issue of community-based information enrichment is the problem of safe-guarding quality when users can freely upload any information they want. In this paper we describe how we use semantics – the Climate Twins domain ontology – for the automated check of user entries to block incorrect links.

2 Climate Twins - a Web tool for visualizing climate change

Climate Twins is a regional climate simulation application, developed by the AIT Foresight & Policy Development Department, that is freely accessible on the Web [1]. The objective of Climate Twins is to offer a simple to use tool to people to get a vivid impression of the future development of the climate conditions in certain areas of Europe they are interested in. 'Simple to use' means that people do not have to be climate research experts for being able to get information from Climate Twins.

The basic idea of Climate Twins is to show people what the climate in their respective regions will look like in the future by showing other regions which have similar climate conditions today. For example, a person living in Austria's capital Vienna wants to know how the local climate will be two decades from now. The Climate Twins application returns those areas in Europe that have today similar climate conditions that are projected to be prevalent in the Vienna area in 20 years. The comparison of climate conditions in Climate Twins is based on two parameters: 'temperature' and 'precipitation'.

Climate Twins uses the temperature and precipitation data from the regional climate projection model COSMO-CLM (COnsortium for Small-scale MOdeling – Climate Local Model) at a 18 x 18 km resolution [2]. Projected data are available until 2100. Historical data are available back until 1961. Finding comparable areas is done by comparing the projected data of the climate parameters 'temperature' and 'precipitation' in the future with the respective present data. Those areas where future and present climate parameters are similar are then called "climate twin regions" [8].

For a particular Climate Twins query the user locates the grid cell he/she is interested in – the so-called "point of interest" (POI) on the map "Selection" in the Climate Twins Web application. Then he/she selects the future time period for the POI and the reference time period for all other grid cells (available time periods are the decades between 1961 and 2100). Further choices available to the user concern the climate parameters (temperature or precipitation or both), the similarity measures (proportional similarity or Hellinger coefficient), which part of the year to be compared (the whole year or one of the four seasons), the weighting of the climate parameters (if both are selected) and the similarity measure thresholds (the closer to 1, the less grid cells will be found with a similar climate). The results are then displayed in the "Results" map. There the user can easily see which places in Europe show similar climate conditions today as they are projected to be in the future at the POI [9].

3 Involving users to enrich the information of Climate Twins

The project TaToo has provided several Web services that could be user to improve the functionality of Climate Twins with regard to information enrichment by involving its users. For this purpose three TaToo public services have been integrated: The tagging service enables users to add additional information on a particular Climate Twin region for the benefit of other users. The discovery service enables the search of additional information on a particular Climate Twin region. The evaluation service gives the users a possibility to evaluate the community-added resources. The evaluation functionality helps the users to distinguish valuable from weak resources added by other users. Originally, this has been the only way of pointing out good as well as weak resources and relied completely on self-regulation within the user-community.

If a user wants to upload the link to a certain resource he/she considers to contain useful climate-related information to be shared with other users the Add Info-tab within the Climate Twins interface has to be used (see Fig. 1). There the user has to select a particular location that fits best to the spatial relevance of the additional information. The user does this by clicking on an adequate Climate Twins raster cell. For this cell he/she gets a list of GeoNames [3] located in this cell from which a particular GeoNames location has to be selected. Then he/she has to fill in the link to the resource (i.e. the URL) and the title of the resource. He/she can also add a brief description and select an adequate topic from the Climate Twins domain ontology as a keyword of the resource. With pressing the submit-button the user uploads the URL of this additional resource and adds the corresponding meta-information to the TaToo knowledge base. It is from then on retrievable by other users of Climate Twins.

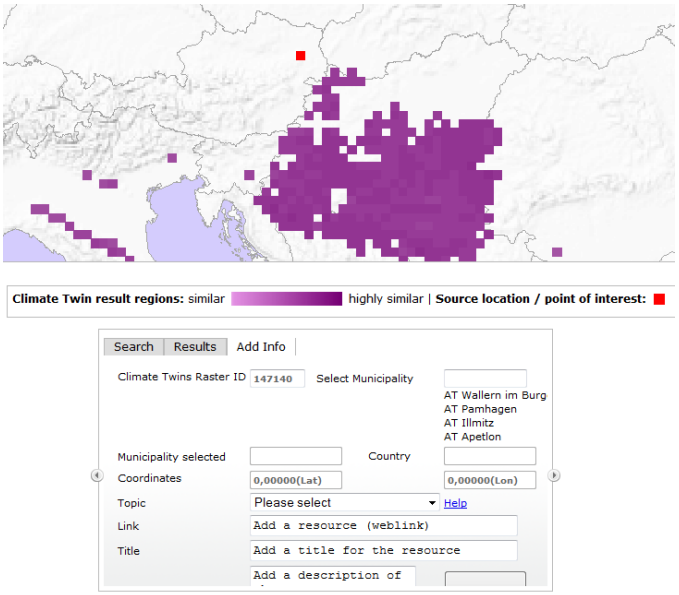

Fig. 1. Screenshot of the Climate Twins interface: part of the Results map and the Add Info tab

Fig. 1 shows the Add Info tab which the user has to fill to upload a new information to the knowledge base. The Add Info tab is right below the Results map showing the similar regions for a certain Climate Twins query. In this example the user has selected a Climate Twins raster cell in Burgenland at the eastern border of Austria. In this cell four GeoNames are present (Wallern, Pamhagen, Illmitz, Apetlon) from which the user has to select one.

Since there was no a priori-restriction of eligible resources in the original version of the Web application there was no guarantee that an added link points to a valuable information. There was no check whether an information is climate-related in any meaningful way or whether it has been geo-located correctly or whether a link has been correctly entered. The only process of quality assurance foreseen was user evaluation, i.e. self-regulation by the user-community itself.

Evaluating information can be done by either ranking its usefulness on a scale ranging from one to five stars or writing a comment or using both options. Other users see the average rating in the list of retrieved results (having used the TaToo-discovery function of Climate Twins). Furthermore if a user opens a link from the list of results in the dedicated viewer he/she finds more detailed evaluation information in the upper part of the viewer: the thread of comments and the individual ratings of the users before. It is also the place where the user enters his/her evaluation.

Of course, by relying on this kind of evaluation, it is impossible to block incorrect information. Other users can only rank it low and write critical comments, but the information remains in the knowledge base and cannot be removed.

4 Safeguarding quality of user entries by means of semantics

After having integrated the three TaToo services the new Climate Twins application with its added functionalities has been evaluated by test users from several research fields with experience in environmental issues (natural sciences, economics, geography, spatial planning, nature protection and engineering). Many comments expressed serious doubts whether self-regulation by rating and commenting could ever be sufficient to maintain an adequate quality level of the resources added by the user community. As a consequence, the idea was born to use the domain ontology for automated quality checks. Before the Climate Twins domain ontology was only used as a means to refine searches by ontology topics, so this opportunity to make more use of semantics seemed very attractive to the provider of Climate Twins.

The Climate Twins domain ontology has been constructed from two ontological resources: SWEET and reegle. The "Semantic Web for Earth and Environmental Terminology" (SWEET) is a large ontology developed by the Jet Propulsion Laboratory [4]. It covers almost any field of environmental science. So far, however, it lacks sophisticated relations between the concepts. "Reegle – the Clean Energy Info Portal" is an information portal for clean energy information [7]. It consists of country energy profiles including energy statistics, a clean energy search tool, a directory of relevant actors and stakeholders, a reegle blog and an extensive glossary of energy-related terms and concepts. Both sources could not be used directly as the Climate Twins

domain ontology. SWEET is too broad, reegle is too detailed for the purpose of Climate Twins. As a consequence, a new much more focused preliminary domain ontology has been derived from these two sources (using only parts of the namespaces of the two ontologies), aligned to the TaToo bridge ontology and applied in the background of the Climate Twins application. The bridge ontology includes also GeoNames representing geographical space in the TaToo ontology framework [6].

The Climate Twins domain ontology is still preliminary. It is not yet covering all important concepts related to climate change, nor does it include GeoNames for all of Europe. Still it is sufficiently comprehensive to demonstrate the semantics-based automated quality check of user entries. In the following we describe the two cases where we have so far applied this method: geographically misplaced resources and resources where title and content description are inconsistent.

4.1 Misplacing a resource

In this case a user uploads a link to a reasonable resource but geo-locates it incorrectly. Due to the fact that Climate Twins requires the selection of a particular GeoName, errors should occur less often than in the case of Web applications where resources can be located freely in geographical space (e.g. Google Earth). Nevertheless, it can still happen. A frequent source of mistakes are distinct places which share the same or a very similar name. Misplacing a resource in Climate Twins means that the user has linked it to a wrong GeoName. The logic of the automated check procedure in this case consists in the comparison of the GeoName selected for geo-locating the resource and any GeoName appearing in the title of the resource. If both are not identical, there needs to be an error and the entry is blocked.

In the following flowchart (see Fig. 2) this procedure is demonstrated with the example of the small village 'Andau' mixed up with the similarly sounding but distinct village 'Antau'. Both names are not only very similar, they are even located in the same Austrian province of Burgenland which makes such a mistake not unlikely.

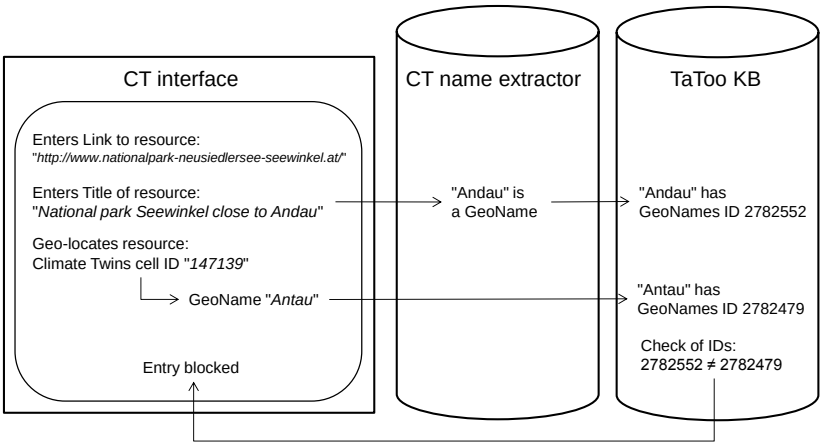

Fig. 2. Flowchart of the checking process for correct geo-location

In this example the user adds information to the national park "Seewinkel" in north-eastern Burgenland, close to Lake Neusiedl. The URL of this web resource is correct but not relevant here, because it does not contain a GeoName. He also adds a title. This entry matters, because there is a GeoName in it. Within the Climate Twins application there is a procedure to extract names from any text the user enters into the Add Info tab. One of these text elements in the title is the GeoName 'Andau'. Then the user makes a mistake while choosing the corresponding Climate Twins raster cell. He ends up in a cell containing the GeoName 'Antau' and selects it for geo-locating the information. Both GeoNames are sent to TaToo where the GeoNames IDs of both places are compared. Since they are not the same, the entry is blocked. Andau and Antau are distinct GeoNames, both cannot be true for the same information, so the entry must contain an error.

4.2 Selecting a topic that is incorrectly related

In this case there occurs an inconsistency between two terms. Technically speaking the user chooses two disjoint ontology concepts for the same entry. This means there are two ontology individuals describing the same resource but are not accordingly related. Of course, for being able to run such a check procedure the underlying ontology has to contain all the required relations between the concepts as object properties.

The following example demonstrates this type of check: A Climate Twins user wants to add an information on animals typical of a peculiar Austrian region, the steppe region at its border to Hungary. What is wrong is that she links it to a wrong climate zone, i.e. a climate where a steppe cannot exist.

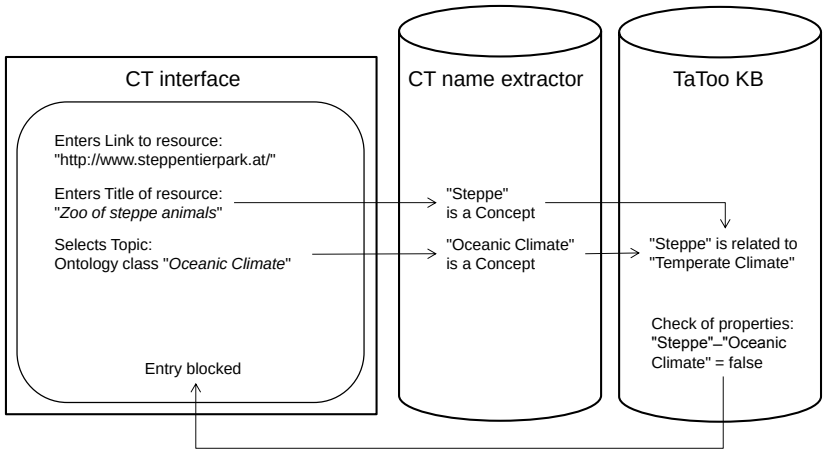

Fig. 3. Flowchart of the checking process for consistent terms

The user enters the URL of the respective zoo and the title. In both entries the word 'steppe' appears. In the title it is a distinct word, so it can easily be extracted and sent to TaToo as an ontology concept. Furthermore, the user selects one of the ontology individuals as a topic, i.e. a keyword indicating the content of the added information.

Here she chooses 'oceanic climate'. TaToo checks the object properties. In the ontology 'steppe' cannot be related to 'oceanic climate'. They are disjoint classes. This is due to the fact that a steppe can only exist in temperate and continental climates. With typing 'steppe' in the title and selecting 'oceanic climate' as topic the user has wrongly established a de facto-relation between the ecosystem 'steppe' and an incompatible climate zone. Since this relation cannot be true, the entry is blocked.

5 Conclusions

Semantics have the potential to provide a good means to check automatically the quality of freely uploaded information by users of the Climate Twins application. The Climate Twins publisher has no chance to do this personally, so an automatic procedure was essential. As the two examples described above show, the approach works basically. It is a first step, however, and the number of possible checks has to be extended a lot and some technical issues have to be tackled in order to be of real practical value in safeguarding the quality of user entries. In the following we list some major issues:

The name extraction procedure used in the Climate Twins application is still very simple. In the present version it cannot handle phrases. Regarding the first example of misplacing a resource this means that a lot of place names cannot be extracted, because they consist of more than one word. For example, there are two towns in Austria with the name 'Waidhofen', one with the extension 'an der Thaya', the other with 'an der Ybbs'. The name of the river is crucial for identifying the place unambiguously. The present name extractor cannot distinguish such places and hand over the correct GeoName to the TaToo system. A more powerful name extraction procedure is therefore needed.

The domain ontology must get more comprehensive. This is not only about adding concepts, which, of course, is important for its own sake. It is more important, however, to increase the internal complexity of the ontology. Formulating new specific relations between ontology concepts will allow for more types of checks. In the second example above, we only dealt with the relation between climate and vegetation. No other relations between ontology classes have been defined so far. Of course, this is an almost open-ended task with a lot of tricky issues to be tackled regarding the definition of relevant and correct properties which correctly link the concepts under all circumstances. This is more difficult than it seems at first glance, because what is obviously a correct relation in one context can be wrong in another.

Climate Twins is still indicated as beta-version. The main reason is a still weak usability. Much of it concerns deficiencies regarding its user interface: More intuitive titles and labels are needed, a more detailed map to make it easier for Climate Twins users to orient themselves. Furthermore the two maps now should be collapsed into one map for both selection of the POI and viewing the results. But usability concerns also the new automated checks. User-friendliness in this context means that a user who has made one of the errors for which an automatic check exists receives meaningful feedback. For now a wrong entry is simply blocked. In the future it is planned

that there are more meaningful error messages which contain the reason why an entry has been blocked. In the first case of a misplaced resource this could be something like: "Your information is probably misplaced. Please, check the location again." Overall, usability has to be improved a lot, before we can expect that self-organized information enrichment by Climate Twins users will take off.

Acknowledgements. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement nr. 247893.

References

1. Climate Twins Viewer: AIT Climate Twins Viewer v3.0, http://foresight.ait.ac.at/projects/tatoo/
2. COSMO-CLM: COnsortium for Small-scale MOdeling – Climate Local Model, http://www.cosmo-model.org/
3. GeoNames: GeoNames geographical database, http://www.geonames.org
4. Jet Propulsion Laboratory, California Institute of Technology: SWEET - Semantic Web for Earth and Environmental Terminology, http://sweet.jpl.nasa.gov
5. Loibl, W., Peters-Anders, J., Züger, J.: Climate Twins – a Tool to Explore Future Climate Impacts by Assessing Real World Conditions: Exploration Principles, Underlying Data, Similarity Conditions and Uncertainty Ranges. Geophysical Research Abstracts 12, EGU2010-12149 (2010)
6. Pariente T., Fuentes J.M., Sanguino M.A., Yurtsever S., Avellino G., Rizzoli A.E., Nešić S.: A Model for Semantic Annotation of Environmental Resources: The TaToo Semantic Framework. In: Hřebíček J., Schimak G., Denzer R. (eds.) Proceedings of ISESS 2011 Brno, pp. 419–427. Springer, Heidelberg (2011)
7. Renewable Energy & Energy Efficiency Partnership (REEEP): reegle - Clean Energy Info Portal, http://www.reegle.info
8. Ungar J.: A Comparative Analysis of Region Pairs Matching Current and Future Climate Conditions. Diploma thesis, University of Vienna, Department of Geography and Regional Research (2010)
9. Ungar J., Peters-Anders J., Loibl W.: Climate Twins – An Attempt to Quantify Climatological Similarities. In: Hřebíček J., Schimak G., Denzer R. (eds.) Proceedings of ISESS 2011 Brno, pp. 428–436. Springer, Heidelberg (2011)