Collaborative user oriented metadata production on EuroSDR Geometadatalabs platform

Bénédicte Bucher, Marie-Dominique Van Damme
LASTIG, Univ Gustave Eiffel, ENSG, IGN, Saint Mande, F 94 160, France

Geographical metadata have initially been designed for the management of geographical data in a production environment and for the exchange of geographical data between a provider environment and a user application. These metadata were to be processed in contexts belonging to the general domain of expertise of geographic information (GI). In the past twenty years, geographical metadata standards stemming mainly from the GI community have been elaborated, in particular ISO19115 to describe geographical data series. Metadata datasets have been produced to describe authoritative databases, in particular in the context of the European spatial information infrastructure, INSPIRE. On-line catalogue services were developed to exploit these metadata and exchange them metadata conforming to the CSW standard.

Geographical metadata may also be needed in contexts not belonging to the general field of GI expertise. In the early century, the Web has been widely adopted as an open distributed architecture. In an open distributed architecture, no restriction should be set on the potential users of a given data service and metadata should support not only data exchange but also data discovery and reuse in domains possibly far from the data specific domain. Users who may not have been trained to use complex geographical data could benefit from spatial data infrastructure like INSPIRE. To achieve this, current GI catalogues must be improved to be usable not only from GI specialists but also from novice users. Since ten years, efforts have been devoted by the GI community to reach new users by adopting new formalism and standards outside the GI field of expertise. For example the DCAT metadata standard is used to produce geographical metadata more legible outside the GI community.

The work presented here targets the identification and production of geographical metadata to improve the capacities of existing GI catalogues and make them more usable for novice users of geographical data. We wish to remove from the user perspective existing silos between data technologies and funding programs that he currently has to cope with when he searches for geographical data. Our research hypothesis is that this identification and this production can be organized through a collaborative platform connecting representative users and experts of the different relevant GI components, geographical data and geographical software or services to pre-process the raw geographical data for the user application. Our approach is to foster the analysis by the requirements of a chosen application. For this application, we analyze what metadata are required for users to retrieve and use geographical data and are currently missing. Then we experiment a process to produce these missing metadata on a collaborative platform, in a way compatible with standard models so that the produced metadata content can eventually feed existing catalogues.

To implement our approach and evaluate our research hypothesis, a first step was to design a collaborative platform accessible from representative users of an application domain and from experts in geodata or geosoftware.
This platform is the EuroSDR Geometadatalabs platform. Geometadatalabs was designed to support the identification of missing metadata by eliciting user requirements and connecting them with existing metadata, more precisely querying existing catalogues based on their requirements and presenting them the results. A strong constraint from our perspective was to support user-centered access to data and to mobilize geodata experts who may be reluctant to interact with

different platforms. We implemented Geomatadatalabs as a unique platform that will host different projects depending on the user community. These use-oriented projects are called infolabs, as illustrated on Figure 1. The user sees the specific infolab dedicated to his usage whereas the geodata provider can have an access centered on his geodata and transversal to all thematic infolabs.
Geometadatalabs was also designed to support the production of missing metadata. With this respect, it must be capable of supporting collaborative production of whatever constitutes metadata in today's context: textual comments, structured data more and more to be organized as linked data and images. Mediawiki engine was chosen mainly because it powers the successful collaborative project Wikipedia which demonstrates that its editing interface can be learnt by anyone. Besides, it supports the edition of textual and semi-structured information. It can integrate RDF data and hence interact with more structured and GI oriented catalogues powered by Geonetwork for example as well as yield structured data for these catalogues.

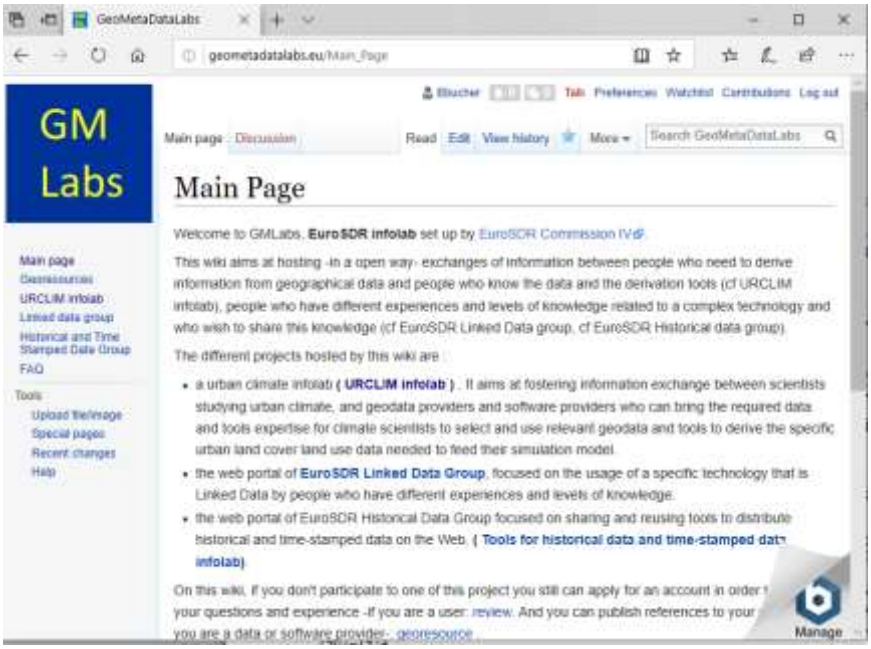

Figure 1. Geometadatalabs, a collaborative platform hosting specific infolabs dedicated to exchanges between GI experts and a specific users community, on this page URCLIM infolab for the urban climate community to connect with experts in geodata and geosoftware relevant for their needs.

The next step is to select an application domain for this research and experiment our approach on this domain, i.e. our capacity to identify missing metadata and to collaboratively document them. The criterion to select an application domain is the fact that representative users are motivated to engage in our experiment.

Our first proof of concept is developed within the application domain of urban climate modelling. Scientists studying urban climate design canopy models to simulate interactions between meteorological phenomena (wind, moisture, temperature) described at a given scale and the surface of earth described at a finer scale in order to calculate finer meteorological phenomena. To obtain land data required to feed these canopy models, this community has developed a common strategy: 1) agree on common formal specifications of such land models (a.k.a Local climate Zones) (Bechter et al. 2015), 2) design production procedures of such land data affordable by the community itself (Ching et al. 2018). This strategy has been successfully applied to produce low resolution land model out of Landsat imagery on the World Urban Databased and Access Portal Tools project (http://www.wudapt.org/). In order to account for local phenomena like for instance

urban heat islands or air pollution, more resolute climate models are needed. To design these models, scientists need more resolute land data describing the city morphology and land use (Masson et al. 2019). The purpose of the URCLIM project is to design local urban climate models on a set of European cities –Paris, Toulouse, Ghent, Brussels, Helsinki, The Randstadt, Bucarest- reusing local open data, and local data which falls into the scope of the Public Sector Information Directive, i.e. which may not be open at the start of the project but that have a good probability to become open in the near future. The URCLIM infolab has been developed on Geometadatalabs platform to connect climate scientists and data specialist to identify and reuse relevant geo-resources to produce land models to feed urban canopy models at a high resolution (Bucher and Van Damme 2018). Its main page is visible on Figure 2.

Figure 2. Infolab dedicated to urban climate research hosted on Geometadatalabs.

The identification of missing metadata was performed by decomposing the retrieval process step by step, using the platform to share the expression of user requirements and the presentation of metadata results and discussions on data relevance. We identified the following missing metadata.

- To improve the expression of a user need, more alignments between existing ontologies must be established and maintained to connect user domain with geo-data domains. INSPIRE schemas are a very useful asset to perform these alignments. Similarity measures are needed, firstly to assess if comparable data are available on a different city, and second to support the query extension.
- To improve the retrieval of resources and evaluate their relevance, multi-lingual metadata are needed as well as data samples. Similarity measures are needed to compare datasets. More metadata about derivation processes are needed than metadata about the different software projects –users do not want to contribute to the software but rather to preprocess their data-. In this specific step, Geospatial user feedback from similar users is especially relevant. Cross-references between similar data and similar software or services are missing.
- During the exploitation of the data, for users to be able to question the results, they need legible provenance documentation (the entities involved, their expertise, the technologies). Besides, there is a need for a technology neutral way to describe uncertainty.

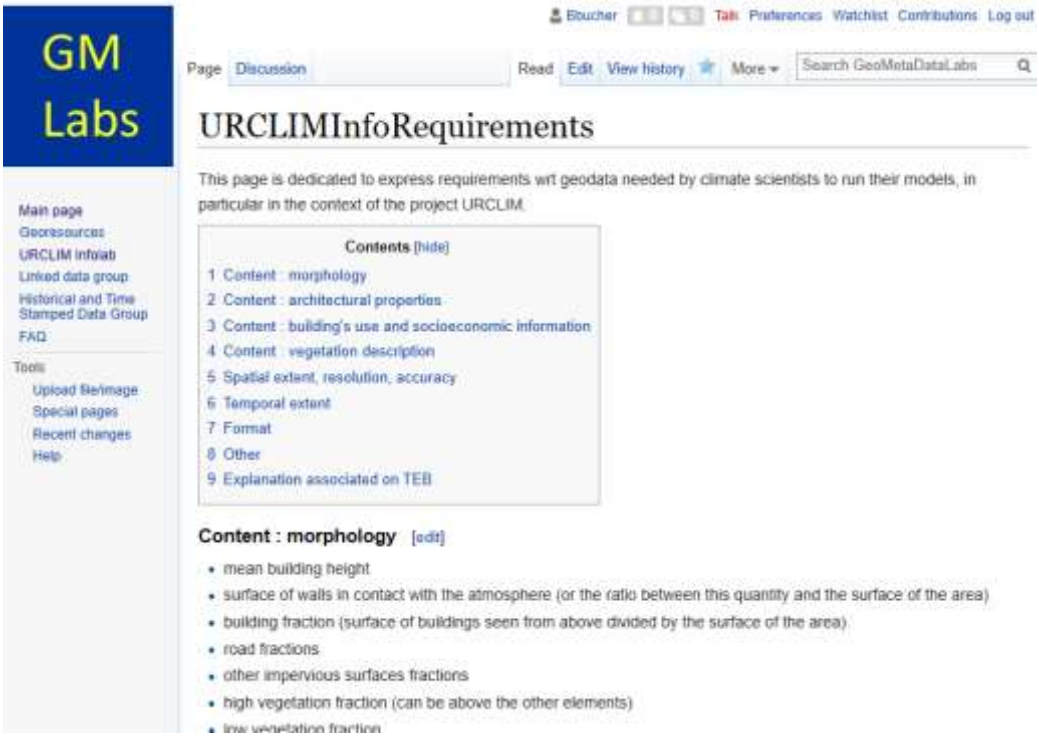

Figure 3. Elicitation of user requirements for geographical information on URCLIM infolab.

Further work must address the production of cross references between data set in the different cities of the project and n the documentation of derivation process to process the raw datasets and yield the required information layer.

This work is developed in the context of EuroSDR commission on information usage and has been funded by EuroSDR. It also benefited from fruitful exchanges with Eurogeographics Quality KEN and INSPIRE KEN. It has been partly funded by ERA4CS, an ERA-NET initiated by JPI Climate with co-funding from the European Union (Grant n° 690462) for the URCLIM project (www.urclim.eu).

Bechtel, B., Alexander, P., Böhner, J., Ching, J., Conrad, O., Feddema, J., Mills, G., See, L. and Stewart, I. 2015. Mapping local climate zones for a worldwide database of form and function of cities. *International Journal of Geographic Information*, 4(1), 199-219

Ching, J., Mills, G., Bechtel, B., See, L., Feddema, J., Wang, X., Ren, C., Brousse, O., Martilli, A., Neophytou, M., Mouzourides, P., Stewart, I., Hanna, A., Ng, E., Foley, M., Alexander, P., Aliaga, D., Niyogi, D., Shreevastava, A., Bhalachandran, P., Masson, V., Hidalgo, J., Fung, J., Andrade, M., Baklanov, A., Dai, W., Milcinski, G., Demuzere, M., Brunsell, N., Pesaresi, M., Miao, S., Mu, Q., Chen, F., Theeuwes, N., 2018. WUDAPT: An Urban Weather, Climate, and Environmental Modeling Infrastructure for the Anthropocene. Bull. Amer. Meteor. Soc. 99, 1907–1924.

Masson V., W. Heldens, E. Bocher, M. Bonhomme, B. Bucher, C. Burmeister, C. de Munck, T. Esch, J. Hidalgo, F. Kanani-Sühring, Y-T Kwok, A. Lemonsu, J.-P. Lévy, B. Maronga, D. Pavlik, G. Petit, L. See, R. Schoetter, N. Tornay, A. Votsis, J. Zeidler, 2020 : City-descriptive input data for urban climate models : Model requirements, data sources and challenges, Urban Climate, Elsevier, Vol 31, https://doi.org/10.1016/j.uclim.2019.100536

Bucher B., Van Damme M.-D., 2018, URCLIM Deliverable D2.1-1, URCLIM Infolab, Paris