Appendix A - Weather Observations of Environment and Climate Change Canada in Wikimedia Commons

Report of the meeting of January 27, 2021

Reporter: Camille Vézy for Wikimedia Canada
Proofreading: Ha-Loan Phan, Wikimedia Canada; Pierre Choffet for Wikimedia Canada; Miguel Tremblay, ECCC. Translated by: Jean-Philippe Béland, Wikimedia Canada

January 27, 2021, 12:30-1:30 p.m.: Project presentation

Attendance

- Pierre Choffet and Ha-Loan Phan (for Wikimedia Canada)
- Miguel Tremblay (ECCC)
- Dissemination partners: Acfas and IVADO
- Home institutions of the people present (+60 people):
    - Acfas
    - Agence science-presse
    - Agriculture Agroalimentaire Canada

WIKIMEDIA
CANADA

    - Anagraph
    - Apple
    - National Archives of France
    - Bibliothèque et Archives nationales du Québec (BAnQ)
    - Office of Ecological Transition and Resilience, City of Montreal
    - Cégep de Chicoutimi
    - Center for Research in Massive Data of Laval University (CRDM)
    - Research and Development Centre, Saint-Jean-sur-Richelieu/AAC
    - Centre de Recherche Informatique de Montréal: CRIM
    - Climate Data Online / Meteorological Service of Canada
    - Environment and Climate Change Canada (ECCC)
    - Element AI (Service Now)
    - Scholar
    - HEC Montreal
    - Snow Info
    - ImpactBee
    - Institut de la statistique du Québec
    - Institut de recherche et de développement en agroalimentaire
    - Institut de valorisation des données (IVADO)
    - La Financière agricole du Qc
    - Ministère de l'Agriculture, des Pêcheries et de l'Alimentation du Québec (MAPAQ) - Direction régionale de la Montérégie Est
    - Ministère de l'Environnement et de la Lutte contre les changements climatiques (MELCC), Gouvernement du Québec
    - Météo-France
    - Mila, University of Montreal
    - My Intelligent Machines (MIMs)
    - Observatoire global du Saint-Laurent (SLGO)
    - Ouranos
    - Polytechnique Montreal
    - LaCogency Digital Productions
    - Natural Resources Canada
    - University of Sherbrooke
    - University of Montreal
    - Université du Québec à Montréal
    - City of Montreal
    - Wikimedia Canada

- Home institutions of those registered but not attending (approximately 20 people):
    - Apple
    - Forestier en chef, Gouvernement du Québec
    - Groupe d'Intervention pour le Développement Durable (GIDD)
    - Groupe Pousse-Vert

- Intelligence and Data Institute, Laval University
- Shopify
- University of Port-au-Prince
- WSP Canada Inc.

The presentation is given in French.

Opening of the session

Ha-Loan Phan, administrator of Wikimedia Canada, presents the "Weather Observations of Environment and Climate Change Canada in Wikimedia Commons" project. The project concerns **100 years of weather data from 8,756 weather stations across Canada.** In 2019, Wikimedia Canada obtained a grant from Environment and Climate Change Canada for this project which will end in March 2021. To our knowledge, this is a world first, importing and sharing massive government and institutional data in Wikimedia Commons.

Availability, access, reuse, cross-referencing with other data, republication and universal participation of this data are the underlying values of this project. Those interested in this data import include those with an interest in massive data, which can be cross-referenced with other datasets in Wikimedia Commons. This project provides a collective workspace, offering functionality and computational capabilities, for different organizations. Wikimedia Commons can also be used as a data repository for scientific publications[1].

The January 27, 2021 meeting is intended to showcase the team's accomplishments and to share some of the future capabilities that come from importing data in Wikimedia Commons. This is a preamble presentation to the February 10, 2021 brainstorming session.

Ha-Loan Phan presents Miguel Tremblay and Pierre Choffet.

Miguel Tremblay is a senior advisor at the Canadian Meteorological Centre. He has been specializing in open data access at the Meteorological Service of Canada (MSC), a branch of Environment and Climate Change Canada (ECCC), for the past ten years. He has been a Wikimedian for about 15 years. He first proposed this project, and it is Wikimedia Canada that is executing it, thanks to the contribution of ECCC.

[1] See podcast "Mon Carnet, le podcast de Bruno Guglielminetti", Friday, January 22, 2021 (22:53) https://soundcloud.com/moncarnet/mon-carnet-du-22-janvier-2021

Pierre Choffet is a free software developer, a long-time Wikimedian and former manager of OpenStreetMap Montreal. He is the one who performed the massive import of the project's data.

Miguel Tremblay and Pierre Choffet take over the presentation.

1. The organizations involved

- The **Wikimedia Foundation**, whose flagship project is Wikipedia
    - Warning: Wikimedia (the Foundation) is different from Mediawiki (the free software that hosts Wikipedia)
- **Wikimedia Commons**, which is the project first created to host the photos and videos used in Wikipedia and to which data can now be added in JSON. This is where the **weather observation data** was imported.
- **Wikidata**, a database created in 2014 by the German national chapter of the Wikimedia Foundation. This is where the **metadata** of individual weather stations are put.
- **Environment and Climate Change Canada (ECCC)**: This department of the Government of Canada is dedicated to weather forecasting to ensure the safety of Canadians and to ensure that economic activities can take place despite the vagaries of the weather. Forecasting requires observations such as those made at weather stations (with several standardized measurement tools) located across Canada. These data are housed at the Meteorological Services of Canada where a climate archive contains all the data from observations made by weather stations since 1840.

2. The data

The data are from about **8,700 stations** (some open, some closed) and were collected **from 1840 to 2018**. These are observations every hour, every day, every month. Those put on Wikimedia Commons so far correspond to the almanac and monthly data (daily and hourly data have not been uploaded).
→ This represents a total of 4.5GB of data converted and then uploaded to Wikimedia Commons, for a total of 26 million values.

Almanac	Monthly data
Temperatures - Extreme maximum	Temperatures - Maximum average

- Extreme minimum - Maximum normal - Minimum normal - Normal average	- Minimum average - Average - Extreme maximum - Extreme minimum
Precipitation - Extreme rainfall - Extreme snow depth - Maximum snow depth on the ground - Percentage of precipitation	Precipitation - Total rainfall - Total snow depth - Total height - Snow depth on the ground on the last day
	Wind - Maximum burst speed - Direction of maximum burst

3. The project

The project has 2 main phases:

- From 2019 to October 2020: importing data into Wikimedia Commons to match the needs of ECCC and to meet the standards of operation of the various Wikimedia project communities.
- Until March 2021: explore how to enhance data, reuse it within Wikimedia projects and beyond.

4. Data models

<u>The ECCC model:</u>

The data are from ECCC and were retrieved according to the Canadian Manual of Surface Weather Observations (ManObs) standards which define the frequency, the manner to write readings, and record data. This data is stored on internal ECCC servers in the archives of the Meteorological Services of Canada. This database is private, but ECCC makes available daily exports of this data in XML format on its own website. The hourly, daily, monthly and almanac data mentioned above can be retrieved and used for projects of our choice.

WIKIMEDIA
CANADA

The Wikimedia Commons model:

The weather data model was developed by the community between 2017-2018 and integrated to Wikimedia Commons in JSON format. This model is thought to be global to suit all countries in the world.

The model is built in a community way, usually by experts from the community (there are gaps, a scientist would probably not find enough information for his own needs). It is evolving: the community can decide to improve the structure of these data to be able to integrate more information than there is currently.

Wikimedia's extended model for ECCC:

How to merge the 2 models to make a coherent model?

The solution chosen was to start from the Wikimedia Commons structure and extend it to include additional data, add fields and give a representation as faithful as possible to the ECCC data.

The operation requires switching from XML to JSON.

The data was matched as follows:

XML (ECCC)	⇒	**JSON (Wikimedia Commons)**
No entry	⇒	No entry
Data not available	⇒	Data not available
Outlier* (if applicable)	⇒	Data not available
Numerical value	⇒	Numerical value
Metadata	⇒	Missing metadata**

Outliers are errors in the data (e.g., -15 mm of rainfall over one month). These outliers have been reported to ECCC and will probably be corrected in their database. On the Wikimedia Commons side, these outliers are not present.

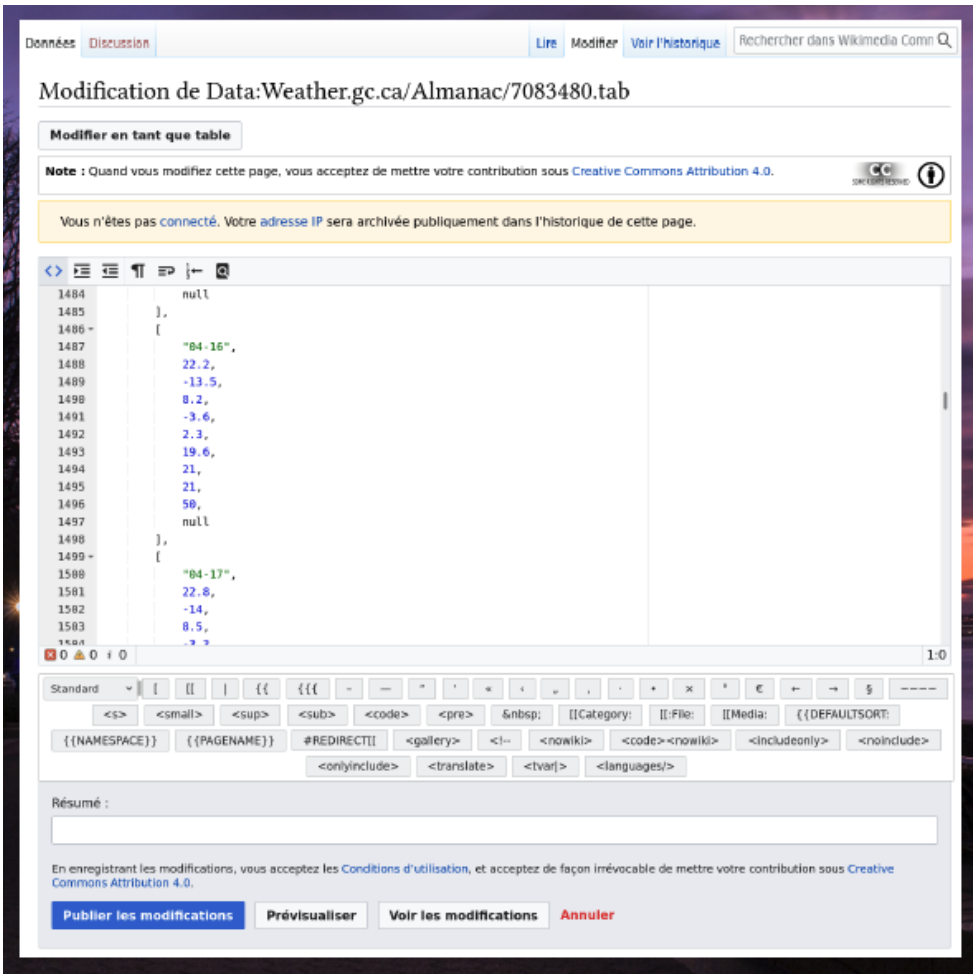
**The JSON format is less descriptive than the XML format, so it is more difficult to add metadata. For example, it is not possible to transpose in Wikimedia Commons information that specifies whether the data is an estimate or a single value from multiple surveys (e.g., snow depth rounded to 0 when there was snow on the ground at the time of the survey). This is important information for scientists, but the project is intended to be used by as many

people as possible. Therefore, it has been chosen not to include this information to simplify the reuse of this data by any contributor.
However, some of this metadata has been transferred to Wikidata: for example, the location of the stations is not present in Wikimedia Commons, but it is now present in the record of each station in Wikidata.

5. The new tools

Several tools were created to carry out this project:

- A **download** tool that allows to retrieve in batch the information that is available on the ECCC website.
- A tool that **merges** XML data from multiple selected stations. It is typically used to combine values from stations located in the same geographical area.
- A tool that allows to **discriminate** the stations by geographical proximity: https://stations.wikimedia.ca. This selection can then be analyzed by the merging tool which will then give the values in this area.
- A **conversion** tool to the Wikimedia Commons JSON model. It is useful mainly for Wikimedia projects contributors but can be reused and adapted for other JSON structures.



Another example of a possible query: from information present in Wikimedia Commons but also in Wikidata, one can cross meteorological information and other information such as those concerning the building, the infrastructures...
For example, the following map shows the weather stations in Quebec that are located within a 1 km radius of a bridge over 100 m long.

6. The added value for the various stakeholders

Through this project, it is possible to disseminate ECCC data in different ecosystems for different future reuses.

This project benefits **Wikimedia projects** by helping to disseminate quality data and facilitating contributions in different languages when updating data.

Example: table from https://fr.wikipedia.org/wiki/Montr%C3%A9al

Relevé météorologique de Montréal (période : 1981-2010)

Mois	jan.	fév.	mars	avril	mai	juin	jui.	août	sep.	oct.	nov.	déc.	année
Température minimale moyenne (°C)	−14	−12,2	−6,5	1,2	7,9	13,2	16,1	14,8	10,3	3,9	−1,7	−9,3	2
Température moyenne (°C)	−9,7	−7,7	−2	6,4	13,4	18,6	21,2	20,1	15,5	8,5	2,1	−5,4	6,8
Température maximale moyenne (°C)	−5,3	−3,2	2,5	11,6	18,9	23,9	26,3	25,3	20,6	13	5,9	−1,4	11,5
Record de froid (°C)	−37,8	−37	−29,4	−15	−4,4	0	6,1	3,3	−2,2	−7,2	−19,4	−32,4	−37,8
date du record	1957	1934	1950	1954	1947	1995	1982	1957	1951	1972	1949	1980	1957
Record de chaleur (°C)	13,9	15	25,6	30	34,7	35	35,6	37,6	33,5	28,3	21,7	18	37,6
date du record	1950	1981	1945	1990	2010	1964	1953	1975	1999	1968	1948	2001	1975
Ensoleillement (h)	101,2	127,8	164,3	178,3	228,9	240,3	271,5	246,3	182,2	143,5	83,6	83,6	2 051,3
Précipitations (mm)	77,2	62,7	69,1	82,2	81,2	87	89,3	94,1	83,1	91,3	96,4	86,8	1 000,3

Source : Environnement Canada[44]

This table could be updated automatically in all languages rather than being filled in by hand as is currently the case.

→ **Wikipedia users** have access to more up-to-date data.
Users of the MediaWiki software can seamlessly retrieve all information from Wikimedia Commons.

If **other countries** follow suit, this project could allow data to be cross-referenced for more global calculations, on an international or even global scale.

7. The future of the project

By March 2021, for the last two months of the project, the aim is to build templates for the different projects of the Wikimedia Foundation - Wikipedia in the first place. Wikinews could also take advantage of this information to present new visualizations of this data. The interest is to have more attractive and more easily accessible pages.

→ On **February 10, a brainstorming session** will be held to **generate ideas for projects to pursue**. The 3 main questions driving it are:

- How do we want to reuse our meteorological heritage?
- How can we influence the rest of the world to import similar open access data into Wikimedia Commons?
- How can this data help us, collectively, to address some of the climate change issues that affect us all?

Questions from the audience

Licensing and compatibility in Wikimedia projects: The data itself is licensed under a Government 2.0 license on the Canadian Open Data website which is compatible with what is eligible on Wikimedia Commons (licensed under Creative Commons, CC BY-SA).
As for the metadata of the weather stations (e.g. their location), according to the jurisprudence in Canada, these data qualify as *facts*, and are therefore not subject to intellectual property rights. On the other hand, if we want to put the data as such in Wikidata, which is technically feasible, there is a licensing problem: we would have to ask the minister to put the observation data in the public domain.

In addition, there is a worldwide movement towards the opening of data (observations and forecasts) which is overseen by the World Meteorological Organization. This would allow countries that do not have the required infrastructure to have quick, high resolution forecasts.

Why not draw from OpenStreetMap to get the polygons of the municipalities (for station discrimination)? In OpenStreetMap, many municipal boundaries are missing, they are not available in open data.

Why have hourly and daily data not been considered in the import yet?
This is mainly because of the very large volume of data. This is not a technical problem (as Wikimedia's servers are capable of handling this load), but it is more of a "community" issue in Wikimedia projects: although the interest in historical weather data is not problematic, there was the fear of "cooling" the community if the imported dataset was too big. In Wikipedia, we need the monthly data to make the tables that summarize this information. But the interest of the Wikimedia community might be weaker towards the hourly data, so there might be a refusal of the community for this dataset which would have been considered not very useful for the Wikimedia projects. But now that the monthly and almanac data is entered, it would be possible to push the granularity even further by adding the daily and hourly data.

One participant mentioned working in **applied research in agriculture**. A team from Natural Resources Canada has developed a 10 km grid with daily data. It could be interesting to consult these people to cross-reference the data.

A participant mentioned in the chat that **ECCC has also been working on rules for reconciling climate data when there are multiple stations in a region and/or over time**. Pierre Choffet is interested in having access to them to adapt the merging tool.

Question about the robustness of Wikidata: is there any way to monitor changes in information on a particular station?
On Wikidata there is a way to monitor every change with the tracking tool, and you get an email immediately. Miguel Tremblay monitors all changes made to weather stations on Wikidata!

According to the World Meteorological Organization, the data is standardized, so if someone in another country wants to do a similar project, will they be able to use the tools that Pierre developed to convert the XML format to JSON?
The ECCC XML format is a proprietary format, so other weather services will probably have their proprietary format as well. That said, once rendered in Wikimedia Commons, homogenization of methods can occur. The calculations are also done on the MediaWiki server side.