WIKIMEDIA
CANADA

Final report

"Weather Observations of Environment and Climate Change Canada in Wikimedia Commons”

Report written by: Pierre Choffet and Ha-Loan Phan, Wikimedia Canada; Translated by: Jean-Philippe Béland, Wikimedia Canada

Date: April 26, 2021

WIKIMEDIA

CANADA

Abstract/Abstract

Le projet « Observations météorologiques d'Environnement et Changement climatique Canada dans Wikimedia Commons » est un projet porté par Wikimédia Canada (WMCA) et financé par Environnement et Changement climatique Canada (ECCC) entre le mois de juin 2019 et le 31 mars 2021. Les objectifs sont d'importer 100 ans de données météorologiques des 8756 stations météo disséminées sur le territoire canadien dans Wikimedia Commons et de réutiliser ces données dans les projets Wikimedia. Le présent rapport documente les différentes étapes du projet. Des précisions sont apportées sur les besoins en ressources humaines, le budget, ainsi que tout détail logistique ou en termes de communications utiles à la réalisation du projet pour réplication par d'autres organisations intéressées par la démarche, partout dans le monde. Ce rapport est complémentaire aux notices réalisées dans les projets Wikimedia qui documentent in situ les différents outils créés dans le cadre de ce projet.

The "Weather Observations of Environment and Climate Change Canada in Wikimedia Commons" project is a project led by Wikimedia Canada (WMCA) and funded by Environment and Climate Change Canada (ECCC) between June 2019 and March 31, 2021. The objectives are to import 100 years of weather data from 8756 weather stations across Canada into Wikimedia Commons and to reuse this data in Wikimedia projects. This report documents the different steps of the project. Details are provided on human resources requirements, budget, as well as any logistical or communication details useful to the realization of the project for replication by other interested organizations around the world. This report is complementary to the records made in the Wikimedia projects that document in situ the different tools created in the framework of this project.

Introduction

Wikimedia Canada (WMCA) is an independent non-profit organization dedicated to the growth, development and distribution of free knowledge in Canada. It is an official chapter of the Wikimedia Foundation since 2011.

Wikimedia Canada's mission is to support and educate Canadians to collect, develop and disseminate knowledge and other educational, cultural and historical content in all of Canada's languages, including Indigenous languages, under a free license or in the public domain.

To do this, Wikimedia Canada will:

- Aid and encourage people to collect, develop and disseminate knowledge and other educational, cultural and historic content in the public domain or under a licence that allows everyone to freely use, distribute and modify said content without the payment of royalties.

- Collaborate with public and private Galleries, Librairies, Archives, Museums and Universities (GLAMUs) in making their holdings more broadly and freely available to those interested in an effort to preserve the heritage of Canada.
- Make use of, encourage the use of, and instruct people in the use of free and open source information resources; either electronic or printed for the advancement of education.
- Encourage the development and release of these materials in the languages of Canada including but not limited to English, French, and the Canadian Indigenous languages.

The project "Weather Observations of Environment and Climate Change Canada in Wikimedia Commons" from Wikimedia Canada, has received funding from Environment and Climate Change Canada (2019-2021). The project consists of importing into Wikimedia Commons a century of weather data collected by 8,756 stations across Canada. The objectives of the project are to reuse and valorize our meteorological heritage in Wikimedia projects, to influence the rest of the world to import similar data in open access into Wikimedia Commons, and to try to solve some of the climate change issues that affect us.

The project was divided into two main phases:

- **Phase 1** (2019-2020): Import weather data into Wikimedia Commons from 8,756 Canadian weather stations over the last 100 years.
- **Phase 2** (2020-2021): to valorize these data, in particular, to use and disseminate them in Wikimedia projects, including Wikipedia.

Objectives of the report

The objectives of this final report are to:

- Document the approach used during this project for future reuse by other organizations interested in the project;
- List the tools that have been created in Wikimedia Commons;
- Give directions for continuing the project in the near future.

Phase I: Importing data into Wikimedia Commons

The first phase of the project aims to make Environment and Climate Change Canada's historical data available on Wikimedia platforms. In this ecosystem, two sites were likely to receive the historical weather data: Wikidata or Wikimedia Commons. A first approach would have been to import the entire data into Wikidata. Since each station already has its own elements, it would have been theoretically possible to create a property with a record for each point in time. However, given the volume of data available, this solution would have resulted in large elements and would probably have been rejected by the community.

The alternative of integrating data into **Wikimedia Commons** is then considered. Usually used to distribute images, photos and sounds, recent developments allow to add **raw data, in JSON format**. A precedent exists with the integration of weather data from New York[1]. The **metadata of the stations** will be integrated in **Wikidata**.

The integration of Environment and Climate Change Canada's data into Wikimedia Commons pushes the latter into previously unexplored territory. First of all, the volume of data is unusual: the integration of more than 170 years of pan-Canadian records from several thousand weather stations represents a mass of raw data rarely seen on the platform. The nature of the data is also unusual, as the community is more used to making contributions in the form of photos, sounds or videos.

In order to make the reuse of the data as easy as possible, it was quickly decided to start from the existing one (the New York data) and to extend the model so that it could accommodate the richer Canadian data. By keeping a similar structure for the data, we keep the compatibility with the existing models on the different Wikimedia projects that are already able to reuse the data available in the US. For simplicity, we will keep the organization as **one web resource per station**. The addresses of the resources will be as follows:

https://commons.wikimedia.org/wiki/Data:Weather.gc.ca/Almanac/<cid>.tab

https://commons.wikimedia.org/wiki/Data:Weather.gc.ca/Monthly/<cid>.tab

They have the advantage of aggregating data from Environment and Climate Change Canada. Provinces and territories are not included to avoid problems related to border changes. In order to allow the creation of station lists by province, it is planned to use the Mediawiki *Categories* mechanism.

The work was therefore subdivided into six successive tasks, leading to the import:

1. Adding categories to JSON data in the Mediawiki platform
2. Correcting the disseminated schemes
3. Writing more restrictive validation schemes

[1] https://commons.wikimedia.org/wiki/Data:Ncei.noaa.gov/weather/New_York_City.tab

4. Structured data generation for Wikimedia Commons
5. Community consultation
6. Import

Adding categories to JSON data in the Mediawiki platform

The first step of the project was to write a **patch** for the JSONConfig extension of Mediawiki, used for managing JSON files on Commons. This operation has been prioritized because it depends on the goodwill of third-party contributors to our project to be completed. Delays can be mitigated by doing this step early in the process. Ideally, the patch will be integrated before the import, scheduled for March 2020.

With this in mind, a first version of changes was developed as early as October 2019, and proposed to the merge on the 17th of the same month[2] . After some editorial changes, a first feedback from the Mediawiki team was obtained on December 10, 2019. The extension's maintainer understands that the lack of Commons category management is a problem, but favors another approach to fix it. Our initial implementation proposed to extend the structured data system by adding a dedicated "categories" field in JSON content. In its place, we were asked to use a generic field containing Wikicode, read by Mediawiki to deduce - among other things - the list of categories to which the resource belongs. A new patch taking these requests into account was submitted for comment the following week, on December 17, 2019.

Due to a lack of response, a reminder was sent on February 17, 2020, with no consequences. As the import had to be done in phase 1 ending in March, the data was regenerated to not contain categories. Nevertheless, in the hope of being able to add the categories *a posteriori* during future republishing of the data, reminders to integrate the patch were done since then, without concrete results.

Correcting the disseminated schemes

The **XML data** distributed by Environment and Climate Change Canada is accompanied by **XSD schemes**. A related technology, XSD schemes provide a better understanding of the structure of the data provided by declaring rules that validate the downloaded data by imposing their format. When used with appropriate software, they **ensure that the entire dataset is valid** and meets the standards announced by Environment and Climate Change Canada.

[2] https://gerrit.wikimedia.org/r/c/mediawiki/extensions/JsonConfig/+/543934

These schemes have been coded to be not very restrictive. They impose the general tree structure of XML files and restrict the types of numerical data (integers and floats in particular). They thus leave an important freedom in the values, and the distributed metadata.

However, during their use, it was discovered that the distributed data did not conform to these schemes. We therefore decided to rewrite the schemes so that they would validate the dataset as expected. Although the work in this sense is relatively minor, this new version has been distributed to the public[3] and to Environment and Climate Change Canada with the intent of making their usage broader.

Writing more restrictive validation schemes

All databases contain errors and approximations. It is understood that, among other things, due to their volume and age, it is normal for the data published by Environment and Climate Change to follow this same trend.

As described in the previous point, the distributed schemes only impose criteria chosen by Environment and Climate Change Canada. **Importing data into Wikimedia Commons requires following additional rules**, defined by the community in charge of the development of this media library. For example, a contributor must **act to the best of his or her knowledge, in the best interest of the project**. In the context of a mass import into a community project, it is a prerequisite to **demonstrate that the imported data is cleaned and validated before publication**. Given the amount of data to be processed, it is important to ensure that the reconciliation of the data between the scheme proposed to Wikimedia Commons and the original data is done in good conditions.

For this purpose, it was decided to develop a **second validation scheme, complementary** to that of Environment and Climate Change Canada, which would allow to force the application of new, more restrictive rules. For example, if the temperature data seems, by empirical test, to use Celsius degrees (°C), a single validation by observation cannot be sufficient. At this stage, it is necessary to prove that the processed data do not deviate from what is expected.

The following tests have been added (non-exhaustive list):

- the values accepted for the "province" field are standardized;
- the stations’ geographical coordinates represent a point in a quadrilateral surrounding the Canadian territory;
- the units of the readings are standardized (millimeters, centimeters, Celsius degrees , ...);

[3] https://git.wikimedia.ca/?p=eccc_schema.git

- the allowed values for the "flag" fields are limited so that they can be processed on a case-by-case basis during the conversion;
- the extremes found in the latest version of the data have been hard-coded. The goal is that any value that would exceed them in future versions would require manual approval;

This new validation scheme is distributed **under a free license** with the rest of the project.

Errors found in the database were also reported to Environment and Climate Change Canada (see Appendix C).

Generation of structured data for Wikimedia Commons

A first import test was performed in 2017 with monthly data from one of the Environment and Climate Change Canada stations. As this one was prepared using tools not adapted to a batch operation, it proved necessary to rewrite this step. The original format - which contains all the data present in the monthly files - was nevertheless used as the initial target for this project.

Two XSLT stylesheets have been written for the transformation of the **monthly** and **almanac** data. After validation of the XML statements by the script described in the previous point, this tool allows to transform the original data into the JSON format expected on Wikimedia Commons. These two tools aim at **producing the visible part of the import**, the one that will be directly consultable by the Internet users on Wikimedia Commons. An important effort has been made to keep the tool **simple, easily understandable** and above all to generate logs on the processing it performs. This simplifies **maintenance**, and the reports allow everyone, even non-programmers, to understand the origin of the published data.

The JSON format does not allow the inclusion of metadata as easily as the original XML. The structure used on Wikimedia Commons in particular is designed to serve as a common base for all weather data, regardless of geographic or administrative origin. As such, it cannot represent the specifics of the model used by Environment and Climate Change Canada as accurately as the original data published on the Environment and Climate Change Canada website. **As a result, a number of internal descriptions and subtleties are notably absent from the data published on Wikimedia Commons**.

As an example, non-numeric values are sometimes present in the original data. These are sometimes obvious errors (a value containing only "#" characters) or approximations (a value of "<31") that cannot be represented in the JSONs on Wikimedia Commons. These cases were replaced with absent ("null") values. Some other cases showed the presence of numerical outliers such as negative rainfall totals. These obvious errors were also removed, and their presence is recorded in the processing log.

Community consultation

While massive data imports are a regular occurrence for free community projects and in particular those managed by the Wikimedia Foundation, they are governed by specific rules. For Wikimedia Commons in particular, a help page is available as a reference for GLAMUs (galleries, librairies, archives, museums, universities) and other institutions.

There is no detailed information about structured data yet, but it is possible to find some general rules: propose a license, prepare the files, do a test on a small volume, ask for a bot status, and finally proceed with the import. This penultimate step comes with obligations: it is about **convincing the community that the operation is useful, and desirable** for the future of Wikimedia Commons.

The import of Environment and Climate Change Canada data represents, to our knowledge, a first in terms of volume of raw data in Wikimedia Commons. It is also the **first massive import of historical weather data**. As such, it was important to ensure community support before taking any concrete action. An outcry, even from a handful of contributors, after the import would block the integration of the weather data in the library for a long time.

Following a preliminary study of the weather data used in the different projects (see Appendix F) and in order to minimize possible clashes, **only two of the four datasets were presented: the monthly readings and the almanac data**.

Two calls for comments were sent out:

- March 6, 2020, regarding the very principle of including historical weather data in Wikimedia Commons (see Appendix E)
- March 13, 2020, regarding the Environment and Climate Change Canada data import project (see Appendix D)

Overall, these texts have elicited few reactions. In a community environment, this is a good thing: everyone has had the opportunity to take a position on these issues, and no one has opposed them, which leads to the installation of a default consensus for the project. This does not mean, however, that the meteorological data are definitely accepted: a contrary movement can always take place in the future, but it will then have to be motivated and go through a new community request.

For the future, **the new compromise theoretically opens the door to the integration of the two other datasets (daily and hourly)** from Environment and Climate Change Canada. It should be noted, however, that these eventual imports will probably encounter **additional technical constraints**: the size of the JSON files containing the records will probably exceed the limit allowed on Wikimedia Commons. An increase of this limit will necessarily require the approval of the project administrators, employees of the Wikimedia Foundation.

Import

Any massive import of data into Wikimedia Commons by a robot must be authorized by the administrators. This step is mainly technical, not editorial. Its purpose is to ensure that the data corresponds to what is advertised and that the importer follows the community rules (such as limiting the speed of the import).

The tool developed in Bash for the needs of the project is voluntarily minimalist, based on cURL, using the Mediawiki API. The management of the identification and the sending of the data is done in a hundred lines of code, published under a free license so that its maintenance is possible in the long term.

The import on Wikimedia Commons was followed by the **reference of the data in the corresponding Wikidata elements of the stations, via their P4150 property (weather history)**[4]. The choice to reuse the QuickStatements 2 tool[5] allowed to perform this part of the import without requiring any authorization or additional software development.

Phase II: Reuse of data in Wikimedia projects and communications about the project

Project presentation and brainstorming sessions to a wide audience

Following the import of the data into Wikimedia Commons, Pierre Choffet quickly identified its limits in terms of inventiveness and scientific precision as to the tools to be created to reuse these data. A think tank (WMCA) was set up and the need to involve other partners in a larger scientific network became apparent.

The reflection committee therefore brought together two institutional partners from its network, namely Acfas and IVADO. These two partners were seduced by the project and were easily convinced to participate.

[4] https://www.wikidata.org/wiki/Property:P4150
[5] https://quickstatements.toolforge.org

Thus, two meetings with the general public have been planned:

- **A first meeting (January 27, 2021)** of one hour to **present the project to a large audience** (open invitation made jointly by WMCA, Acfas and IVADO, in their respective networks)
- **A second meeting (February 10, 2021)** lasting two and a half hours, in the form of a **brainstorming session**, to come up with the best ideas for reusing the weather data. A group of about 10 people was selected by WMCA, in collaboration with Acfas and IVADO.

In the **appendices,** you will find **the complete report of these two sessions.**

What we noted:

- **First meeting** (January 27, 2021) to present the project: **about sixty** people connected virtually, coming from **different backgrounds** (school, university, NPO, private institutions, governments...). We received many positive comments. This shows the **strong attractiveness** and the **high potential** of the project. This invitation was also a first starting point to make the project known to a large public.
- **Second meeting** (February 10, 2021) - brainstorming session. **Twelve** people were selected to participate in this structured brainstorming session. They were all present at the first meeting or received the minutes. They came from different backgrounds as for the first meeting (school, university, NPO, private institutions, governments...), which allowed for a rich discussion, as much during the separation of the group into two sub-groups of 6 people (+ observers) as during the plenary meeting. During this session, we retained different ideas for the future of the project - see Appendix B:

    - Data visualizations
    - Descriptive valorization of weather data to "augment" information about events, places or people
    - Prescriptive valorization of weather data to support decision making in different sectors and other countries
    - Creation of Meta-Wiki[6] pages (pages used for the coordination of projects) to invite Wikimedians around the world to reproduce the project

[6] https://meta.wikimedia.org/wiki/Main_Page

Tools developed

Geographic discrimination of stations

The content available on Wikimedia platforms is often organized according to their administrative boundaries: a page dedicated to Manitoba on Wikipedia, an article dedicated to visiting Toronto on Wikivoyage, or the category dedicated to Victoria Island on Wikimedia Commons. As for the data made available by Environment and Climate Change Canada, the stations are accompanied by their geolocation (in WGS84 format), and their respective province of installation. Finding the correspondence between their geolocation and the administrative structure to which they belong is complex for the following reasons:

- the decimals of geographical coordinates are truncated in a manner to generate an imprecision more than a kilometer;
- only the provinces and territories are indicated, not the different administrative levels;
- the indicated provinces and territories in the metadata don't correspond to the existing borders at the time of publication of the data. The evolution of the boundaries between 1840 (start of the survey) and 1999 (date of the last change) is not represented.

The lack of a free dataset from which to derive this information over time prevents us from doing so on a large scale. Nevertheless, it is possible to facilitate the work of contributors to Wikimedia projects by offering them a visual tool to identify the stations present in a given area. To this end, a web interface has been developed and is maintained online[7], allowing to list the stations present in an area drawn by the user, and to extract their identifiers in a customizable format. They are then listed in a string of characters that can be reused in third-party tools.

As an example, its use allows to list the stations located on the Island of Montreal, past or present:

7024400, 7025000, 7026839, 702FHL8, 7026612, 7027280, 7026073, 7026213, 7025250, 7025251, 702S006, 7021945, 7025270, 7024118, 7025228, 7025260, 7025267, 7024745, 7025280, 7024256, 7027440, 7025252, 7025257

Merging of files

In order to be able to **generate files representing the surveys of several stations**, a **merge script** has been written. Based on an XSLT stylesheet, it takes as input the paths of at least two XML files released by Environment and Climate Change Canada and merges them

[7] https://stations.wikimedia.ca/

according to **user-customizable rules**. The tool is designed to work with the tool described in the previous point, and its output is an XML file compatible with the other tools developed for the project. This file can be converted to JSON format in order to be sent to Wikimedia Commons. Using the previous example, it becomes possible to generate a single file containing the recalculated data of the 23 Montreal stations listed above. Once uploaded to Wikimedia Commons, it can be seamlessly reused with all templates that are normally unable to work with multiple input data files.

We learned after writing this tool that Environment and Climate Change Canada had issued rules for merging multiple survey files. These rules are probably not publicly available and as such this tool is probably not compliant. In the event of a future release, it would be **worthwhile to revise these tools in order to maintain compatibility between the data calculated internally by the Department and those made available on Wikimedia Commons**.

The merge tools have been released **under a free license** along with the rest of the code written for the project.

Reuse of data in the Climate table of Wikipedia

Beyond simply making resources available for download to the general public, the integration of Environment and Climate Change Canada's historical data into Wikimedia Commons makes sense in the context of real-time reuse by the various Mediawiki instances. The software allows, once properly configured, to seamlessly download content from Wikimedia Commons and integrate it into its own wiki. In the case of Environment and Climate Change Canada's historical data, this allows users to download, perform calculations, apply formatting, and display the result.

The case of Wikimedia projects has been taken into account in a special way. **A number of templates allow, for example on Wikipedia, to display tables or graphs based on climate data**. Developed several years ago, these representations are present in a large number of pages, including Canadian cities. Historically, they have always required that the people who insert them make their own calculations, and add the data in hard copy in each of the encyclopedia's pages. Importing the data into Wikimedia Commons can improve the experience of contributors. As an example, we have created a new template[8] that can perform calculations on the imported data, and **automatically fill in some rows of the pre-existing "Climate" table**. The implementation allows the data to be kept up to date in a transparent manner, without the need to manually recalculate values based on the latest Environment and Climate Change Canada publications. In the event that the formatting of the original table is modified, it is also automatically reused in the new template.

[8] https://fr.wikipedia.org/wiki/Modèle:ClimatCommons

The source code of the new template is available **under a free license** on Wikipedia.

Other ideas for tools to develop in the future

The brainstorming session and our own reflections allow us to outline some tools that would be useful to create, if we were to continue the project:

- Create interactive graphs of the evolution of the climate, with the possibility of choosing dates;
- Create a data comparison tool;
- Create animated graphics in a more shareable format such as GIF;
- Generate infoboxes[9] highlighting weather data, for example, for historical event pages in Wikipedia, or to describe the climate at the time of a famous person in his or her biography;
- Improve the accuracy of metadata associated with different types of files that can be matched to weather data;
- Link weather data to files describing extreme and non-extreme weather events, including illustrating physical phenomena related to climate and words to describe them in different languages and cultures;
- Integrate other data and templates in Wikimedia projects, for example, data related to agriculture, which could be interesting to cross with weather data.

Meta Page

A Meta-Wiki page for the project has been created. This page can be translated into several languages and this report will be integrated for consultation by Wikimedians: https://meta.wikimedia.org/wiki/Projet_ECCC.

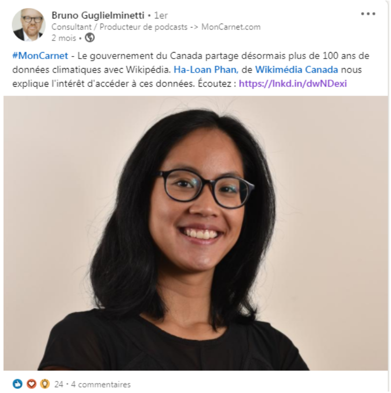
Other benefits of the project

An interview in French by Ha-Loan Phan (WMCA) on the "100 Years of ECCC Weather Data on Wikimedia Commons" project was conducted with Bruno Guglielminetti, in the #MonCarnet podcast of January 22, 2021: https://soundcloud.com/moncarnet/mon-carnet-du-22-janvier-2021#t=22:53.

Mr Guglielminetti, a digital expert, is followed by more than 80,000 people on Twitter and by more than 11,000 people on LinkedIn. This interview provided great exposure for the project,

[9] https://fr.wikipedia.org/wiki/Aide:Infobox

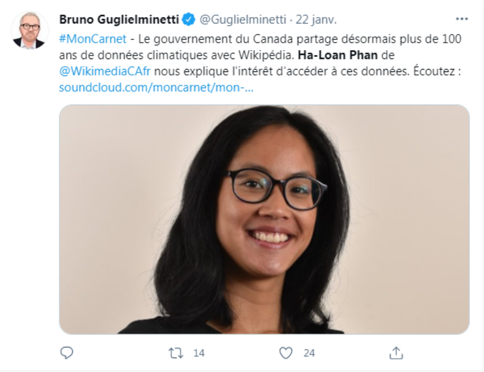
especially in preparation for the two project presentation and brainstorming sessions on January 27 and February 10, 2021.



Bruno Guglielminetti • 1er
Consultant / Producteur de podcasts -> MonCarnet.com
2 mois •

#MonCarnet - Le gouvernement du Canada partage désormais plus de 100 ans de données climatiques avec Wikipédia. Ha-Loan Phan, de Wikimédia Canada nous explique l'intérêt d'accéder à ces données. Écoutez : https://lnkd.in/dwNDexi

24 • 4 commentaires



Bruno Guglielminetti @Guglielminetti · 22 janv.

#MonCarnet - Le gouvernement du Canada partage désormais plus de 100 ans de données climatiques avec Wikipédia. **Ha-Loan Phan** de @WikimediaCAfr nous explique l'intérêt d'accéder à ces données. Écoutez : soundcloud.com/moncarnet/mon-...

14 24

We have submitted an abstract for an oral presentation at the CMOS/SCMO scientific conference - scheduled for late May/June 2021, which has been accepted.

The theme of this conference is "Climate Change: Risk, Resilience and Response". The conference brings together a wide range of scientists and other professionals to discuss topics such as climate, atmosphere, ocean, and Earth sciences. This presentation to a specialized audience allows WMCA to introduce the project to other organizations and individuals directly interested in the field of meteorology. We anticipate that many people in government departments will have access to this science communication.

> Title: 100 years of Weather Observations from the Meteorological Service of Canada in Wikimedia Commons (Wikipedia)
>
> Wikimedia Canada, the Canadian chapter of the Wikimedia Foundation, received a grant in 2019 from Environment and Climate Change Canada. The project consists of uploading 100 years of weather data from 8,756 weather stations across Canada to Wikimedia Commons, a data repository that feeds, among others, the Wikipedia encyclopedia.
>
> The uploading and sharing of governmental and institutional big data in Commons is a world premiere. This provides a collective workspace with features and computing capabilities for different organizations. This project aims for the free access to data for as many people as possible, their reuse and their potential crossing with other data.

We will explain how we adapted the Meteorological Service of Canada data model to the Commons model, and what new features are now available, thanks to this data import. We will explore some of the tools developed and how to improve Wikipedia articles from Canadian locations, in all languages. We want to demonstrate the strong potential of such an approach that could inspire new projects by governments in the open data field.

WMCA will supervise an internship in partnership with IVADO and in collaboration with Acfas: "Des données pour raconter" (Data Storytelling) - planned for summer 2021: https://ivado.ca/bourses-et-subventions/bourses-de-stage-des-donnees-pour-raconter/.

This competition was born in 2019 from a collaboration between Le Devoir, Polytechnique Montreal's JData lab, and IVADO. The program aims to develop the skills of graduates:

- on problems directly related to society;
- in data journalism / "data storytelling" in the broadest sense, on subjects requiring the application of innovative techniques and approaches;
- in teamwork and multidisciplinary contexts.

Following a successful partnership for the two brainstorming sessions in this project, IVADO offered to WMCA to host two data science and data journalism interns. WMCA has offered to work in collaboration with Acfas to continue this project, since Acfas has expertise in science communication. We anticipate that this collaboration will be very fruitful in showcasing and spreading the word about this project, and in reaching out to others interested in creating tools using 100 years of weather data in Wikimedia projects.

Project schedule

PHASE I: DATA IMPORT	
October 2019	Project planning Adding categories to Wikimedia Commons Developing a tool to download the data from Environment and Climate Change Canada Rewriting the XML data validation scheme Writing data validation schemes seen from Wikimedia Commons
November 2019	Writing stylesheets to transform XML data into JSON
December 2019	Second version of adding categories to Wikimedia Commons

January - February 2020	Writing import scripts in Wikimedia Commons
February - March 2020	Proposal of the import to the Wikimedia Commons community
March 2020	Final import of data into Wikimedia Commons Linking data in Wikidata
PHASE II: DATA REUSE	
October - November 2020	Station data fusion tools
November - December 2020	Map allowing to discriminate the stations by geographical location
January 27, 2021	Session 1: Presentation of the project by Pierre Choffet and Miguel Tremblay (see Appendix A)
February 10, 2021	Session 2: Brainstorming (see Appendix B)
February - April 2021	Realization of the tools Writing of this report Final accountability

Meeting schedule

- **May 28, 2019** - First mention of the possibility of a project related to Environment and Climate Change Canada data in Wikimedia Commons.

- **July 24, 2019** - Reflection on the possibilities of the project, the team, its planning and its budget.

- **December 4, 2019** - Ongoing monitoring of the project

- **October 2, 2020** - Follow-up of the work done, planning of phase 2

- **November 7, 2020** - Discussions around the possibility of organizing events with other Canadian institutions

- **November 8, 2020** - Work on a presentation document to approach other partner institutions

- **November 10, 2020** - Discussions on the integrity of imported data
- **November 20, 2020** - Presentation of the project to IVADO
- **December 4, 2020** - Presentation of the project between partners (WMCA and Acfas and IVADO - partners in the organization and dissemination of brainstorming sessions)
- **December 8, 2020** - Preparation of brainstorming sessions and consideration of who to invite
- **January 6, 2021** - Contact and explanations of the reporting mandate for the January 27 and February 10, 2021 brainstorming sessions
- **January 7, 2021** - Discussions on the content of the January 27 presentation, following the proposal of a first draft of the slide show
- **January 13, 2021** - Discussions on the content of the January 27, 2021 presentation, following the second draft of the presentation slideshow
- **January 15, 2021** - Follow-up meeting for the January 27 and February 10, 2021 brainstorming sessions.
- **January 19, 2021** - Rehearsal for the January 27 presentation
- **January 27, 2021** - Presentation of the project to the general public (see Appendix A)
- **February 2, 2021** - Planning meeting for the February 10, 2021 brainstorming session
- **February 10, 2021** - Structured brainstorming session, WMCA in collaboration with Acfas and IVADO. Facilitation: Sébastien Paquet and Ha-Loan Phan (see Appendix B)
- **February 10, 2021** - Follow-up meeting for the reporting mandate
- **February 12, 2021** - Follow-up to the February 10, 2021 brainstorming session and CMOS conference discussions.
- **February 22, 2021** - Ongoing monitoring of the project.
- **February 23, 2021** - Planning the end of the project

- **April 12, 2021** - Follow-up of the end of the project
- **April 16, 2021** - Follow-up of the end of the project
- **May 7, 2021** - End of project meeting, debriefing

Administrative management and implementation of the project

Human resources

Name, title and institution	**Role**	**Tasks**	**Paid, pro-bono or partnership**
Jean-Philippe Béland, Wikimedia Canada Institutional Advancement Manager, self-employed	Administrative coordination of the project	Management of administrative tasks (administrative reports, budget tracking, administrative links with ECCC)	Paid by the grant
Guillaume Chicoisne, Scientific Advisor, IVADO	Advising brainstorming sessions	Reflection on the direction of the discussions, reflections on the guests and on the project, mobilization of the participants in these sessions	Wikimedia Canada-IVADO partnership
Pierre Choffet, consultant, self-employed	Developer	Project conceptualization and development of tools, project presentation, accountability (writing of this report)	Paid by the grant
Johanne Lebel,	Advisor and	Organization of	Service agreement

Editor and Project Manager, Acfas	implementer of the brainstorming sessions (project presentation and brainstorming)	online brainstorming sessions (mobilization and communications and registration management, logistical management of the sessions)	paid by Wikimedia Canada-Acfas grant + partnership
Sébastien Paquet, Scientific Affairs Manager, ServiceNow	Advisor and brainstorming facilitator	Reflection on the direction of the discussions, reflections on the guests and on the project, mobilization of the participants in these sessions	Pro-bono in a personal capacity
Ha-Loan Phan, Administrator, Wikimedia Canada	Administration, management, consulting and coordination of the project	Management and coordination of the various stages of the project, reflection and quality assurance of deliverables, mobilization of partners and implementers, accountability (writing of this report), communication of project impacts	Pro-bono
Benoit Rochon, Administrator, Wikimedia Canada	Project administration, management and consulting	Management and coordination of the different stages of the project, reflection on the project	Pro-bono
Camille Vézy, PhD student, Université de Montréal	Reflection session reporter (project presentation and brainstorming)	Written report of the sessions	Wikimedia Canada-Camille Vézy Grant Service Agreement

When the grant application was submitted, only two paid staff were planned (developer and administrative coordinator). Finally, as is often the case in Wikimedia projects, WMCA volunteers (including Board members) joined the project to improve it and take it further, for a more ambitious vision. The brainstorming sessions were born from these meetings, and WMCA formed a partnership with Acfas and IVADO, two non-profit organizations working in the scientific field, with a large network of researchers in Quebec.

One of the difficulties observed in terms of human resources was especially the voluntary time of "scientific" coordination/orientation of the project (see estimated time - budget) that the people planned at the beginning could not assume alone. In the future, if such an ambitious project is carried out, it is advisable to foresee a specific budget for the "scientific" coordination which requires particular competences and a network (other than the administrative follow-up of the project). In fact, relying on volunteers for an important part of the project can be a major weakness, putting the realization at risk.

Consolidated budget

Item	Funded by ECCC	Funded by WMCA	Total
Subcontractor for development	$47,200		$47,200
Computer equipment	$713		$713
Travel expenses	24 $		24 $
Overhead (project management, coordination, marketing, communications, accounting, translation and reporting)	$6,831	$20,000 (in kind[1])	$26,831
TOTAL	**$54,768**	**$20,000**	**$74,768**

[1] Wikimedia Canada's "in-kind" contributions include pro bono management, coordination, translation and consulting services.

Conclusion

The excitement around the “Weather Observations of Environment and Climate Change Canada in Wikimedia Commons" project demonstrates its strong potential. We have been able to attract people from different backgrounds around common themes, generating curiosity and interest. This project could be replicated at different scales: 1. with other Canadian organizations that have datasets that can be cross-referenced with ECCC datasets and that are interested in making their datasets freely available on Wikimedia projects, 2. with governmental organizations equivalent to ECCC in other countries. Since the Wikimedia movement has the advantage of being international, partnerships with other chapters could be considered to foster the local aspect of project management, while allowing regular exchanges between chapters.

As with all projects in the Wikimedia sphere, and in the open source field in general, volunteers have joined the projects. Of the people who have heard about the project, most have shown interest in continuing with WMCA in some way. The discussion, the generation of ideas, and the search for solutions are rich when you succeed in gathering a diversity of people around the table.

This final report was designed to be reused by anyone who would like to replicate the project. It is therefore to be placed in all hands and a translation into several languages would be beneficial.

Contact

Wikimedia Canada website: wikimedia.ca, email: info@wikimedia.ca.

Appendices

Appendices A and B: Project Presentation (January 27, 2021) and Brainstorming Session Reports (February 10, 2021)

Appendix C: Reporting Problems to Environment and Climate Change Canada

This is a translation in English of the original message in French.

Hi Miguel,

The import of the data into Commons is now finished, I focused on the error report in question. This only concerns the English XML files released by ECCC, I did not work with their CSV files.

Here are the problems I encountered:
- The XML files are all malformed in the sense that the distributed schema[10] does not validate them. I won't go into technical details here (I can do that with the developer responsible for their generation), but I have written an alternative and very incomplete version that is available on the Wikimedia Canada server[11];
- Regarding flags: the process of generating XML should be better described to avoid the user trying to interpret their meaning. The "I" flag is present in the almanac files and is not documented;
- latitude-longitude are sometimes 0. One station (8104200) is located in Russia;
- "totrain" sometimes contains "#" characters instead of a numeric value (1010720). Also, its value seems to be limited to 1000 mm - if this is a limit of the measuring tools, it should be indicated;
- "totsnow" can have negative values (6032119);
- totprecip" can contain "#";
- for "dirmaxgust", the difference between 0 and 36 should be documented;
- speedmaxgust" has an obviously aberrant maximum value (467
km/h) and may contain a non-numeric value "<31" ;
- the element "precipitation" also has a maximum value that seems arbitrary (999).

[10] http://climate.weather.gc.ca/climate_data/bulkxml/bulkschema.xsd
[11] https://git.wikimedia.ca/?p=eccc_schema.git;a=summary

Overall, a few simple improvements could be made to improve the reliability and comprehensibility of the data:
- the scheme would benefit from being widely completed in order to ensure the reliability of the data disseminated;
- the flags should be better documented, that would allow to better appreciate the quality and the limitation of the distributed data;
- beyond just the province, more specific information on the administrative location such as the municipality where the station is located could be included. If the information is missing, one or two significant digits could be added to the geographic coordinates, so that they can be cross-referenced with other databases.

I attach the log that led to the generation of the files actually sent to Commons. It's verbose, but you'll get the details of the problematic files.

Good afternoon, see you soon.

Appendix D: Project Announcement and Call for Comments

Hello Commons,

At the end of last year, Environment and Climate Change Canada (ECCC) and Wikimedia Canada decided to explore possible synergies between Wikimedia projects and the federal agency. A first reflection crystallized around the data produced by the latter and distributed under open license on their website.

As a result, we have created tools to convert official historical data to the JSON format required for integration into Wikimedia Commons. The inclusion of weather data in Commons is not a first, a similar work has already been done in the United States using data from nceii.noaa.gov. The basic principle of this import is to reuse the same data structure, which allows to keep the compatibility with the existing models in the different Wikimedia projects. Importantly, these tools are available to all and are designed to keep the data up to date in the future.

The data proposed for import has been published in a Git repository and I am now calling for comments on the integration of this data in Commons. The tree structure should be understood as follows:

- the file weather.gc.ca/Monthly/4026175.tab would be imported into Commons as Data:Weather.gc.ca/Monthly/4026175.tab. The "4026175" identifier in the resource refers to a stable code given by ECCC for each of the weather stations.

Some notes on the conversion process and the final result:

- ECCC distributes its data in four levels of detail: hourly, daily, monthly and an almanac [1]. For the moment, only the last two are offered for import, because they generally fit the needs of the models of the different Wikimedia projects. However, if the community decides that other levels of granularity are relevant, it is perfectly possible to add them before publication;
- In order to ensure a high level of quality in the data published on Commons, all values are tested before being added. In particular, obviously outliers (unrealistic temperatures, negative precipitation, ...) are eliminated. As such, the data on Commons would be a subset of the data published by ECCC;
- Contributors used to structured data on Commons will have noticed the presence of an unusual "wikicode" field. Its presence is linked to a proposed change to the JsonConfig extension that will allow, among other things, to add Categories (in the Mediawiki sense of the term) to the data. If its integration is refused by the development team, this property will simply be removed.

I would like to take this opportunity to point out that not all column descriptions are currently translated into languages other than English and French. If you are able to make adaptations in other languages, I would be happy to integrate your work before the final data import.

(the same topic was opened today on the *Village Pump* - if you are familiar with English you can follow the conversation there too)

Appendix E: Historical Climate Data on Commons

Last week, I've posted a request for comments concerning import of Canadian structured weather data into Commons. Donald Trung advised us to post here so I can get the confirmation this type of files can be distributed through Wikimedia Commons. Please note this is only about *historical* weather data - forecasts are specifically excluded from this discussion.

My understanding of Commons:Project_scope leads me to think that hosting that kind of data is in the scope of Commons, because every requirement is checked:

- it's educational content that "provides knowledge". This is especially true since the climate change research has become high priority in many countries during previous years.
- shaped like what was proposed by Yurik in 2017 it's media files because JSONs is *structured* data that can easily be used to automatically generate graphs. As a matter of fact, it is already done on Wikimedia platform. Commons also now has a nice table

based visualisation interface that allows everyone to directly read content in an intelligible way.

- it uses free file format - a lot of tools can read/write JSON data, especially on the web.

Any thoughts?

Appendix F: Correspondence of data used in Wikimedia project templates

Data	**ECCC (per station)**
Template:Climate (French Wikipedia)	
Minimum temperature	✓
Average temperature	✓
Maximum temperature	✓
Duration of sunshine	✗
Relative humidity	Calculable (hourly)
Amount of precipitation	✓
Amount of rain	✓
Amount of snow	✓
Number of days with storms	✗
Number of days with hail	✗
Number of days with snowfall	✗

Number of frost days	Calculable (daily)
Number of days with fog	✗
Number of days with zero sunshine duration, greater than or equal to one hour, at 5h	✗
Number of days with rainfall greater than or equal to 1mm, 5mm, 10mm	Calculable (daily)
Number of days with gusts greater than or equal to 16 $m.s^{-1}$ or 28 $m.s^{-1}$	Calculable (daily)
Minimum temperature of the month with its month-year date	✓
Minimum temperature of the highest month, with its month-year date	✓
Minimum temperature of the day with its day-month-year date	✓ (almanac)
Highest minimum temperature of the day, with its day-month-year date	Calculable (daily)
Maximum temperature of the month with its month-year date	✓
Highest temperature of the month, with its month-year date	✓
Maximum temperature of the month with its month-year date	✓

Lowest daytime maximum temperature, with its day-month-year date	Calculable (daily)
Highest temperature of the day, with its day-month-year date	✓ (almanac)
Lowest monthly precipitation, with its month-year date	✓
Highest monthly precipitation, with its month-year date	✓
Extreme daily rainfall, with its day-month-year date	✓ (almanac)
The most days with precipitation, with its month-year date	Calculable (daily)
Extreme snowfall, with its day-month-year date	✓ (almanac)
Sunshine duration of the lowest month, with its month-year date	✗
Sunshine duration of the highest month, with its month-year date	✗
Highest sunshine duration of the day, with its date	✗
Instantaneous extreme wind of the day, with its day-month-year date	Calculable (daily)
Lowest pressure of the day, with its day-month-year date	Calculable (hourly)

Highest pressure of the day, with its day-month-year date	Calculable (hourly)
Pressure at sea level	✗
Number of days with minimum temperature less than or equal to -10°C, -5°C	Calculable (daily)
Number of days with maximum temperature greater than or equal to 35°C, 30°C, 25°C and less than or equal to 0°C	Calculable (daily)
Solar radiation in MJ/m².	✗
Potential evapotranspiration for the month	✗
Wind over 10 minutes	✗
Monthly rainfall, over 4 years	✓
Template:Graph:Weather_monthly_history (English and French Wikipedia)	
Minimum temperature	✓
Maximum temperature	✓
Average minimum temperature	✓
Average maximum temperature	✓
Number of days with precipitation	Calculable (daily)
Total precipitation	✓

Number of snow days	Calculable (daily)
Total snow	✓
Template:Climate (Wikivoyage)	
Minimum temperature	✓
Maximum temperature	✓
Amount of precipitation	✓
Number of days with precipitation	Calculable (daily)
Amount of snow	✓