WFD data for Wikidata

Final report
André Costa
2017-06-28

Background

The EU Water Framework Directive (WFD) requires reporting of water data, in a standarised format, from all member states at the end of each 6-year water cycle. The data is stored in Reportnet at the European Environmental Agency (EEA) and exposed to the public as XML and RDF.

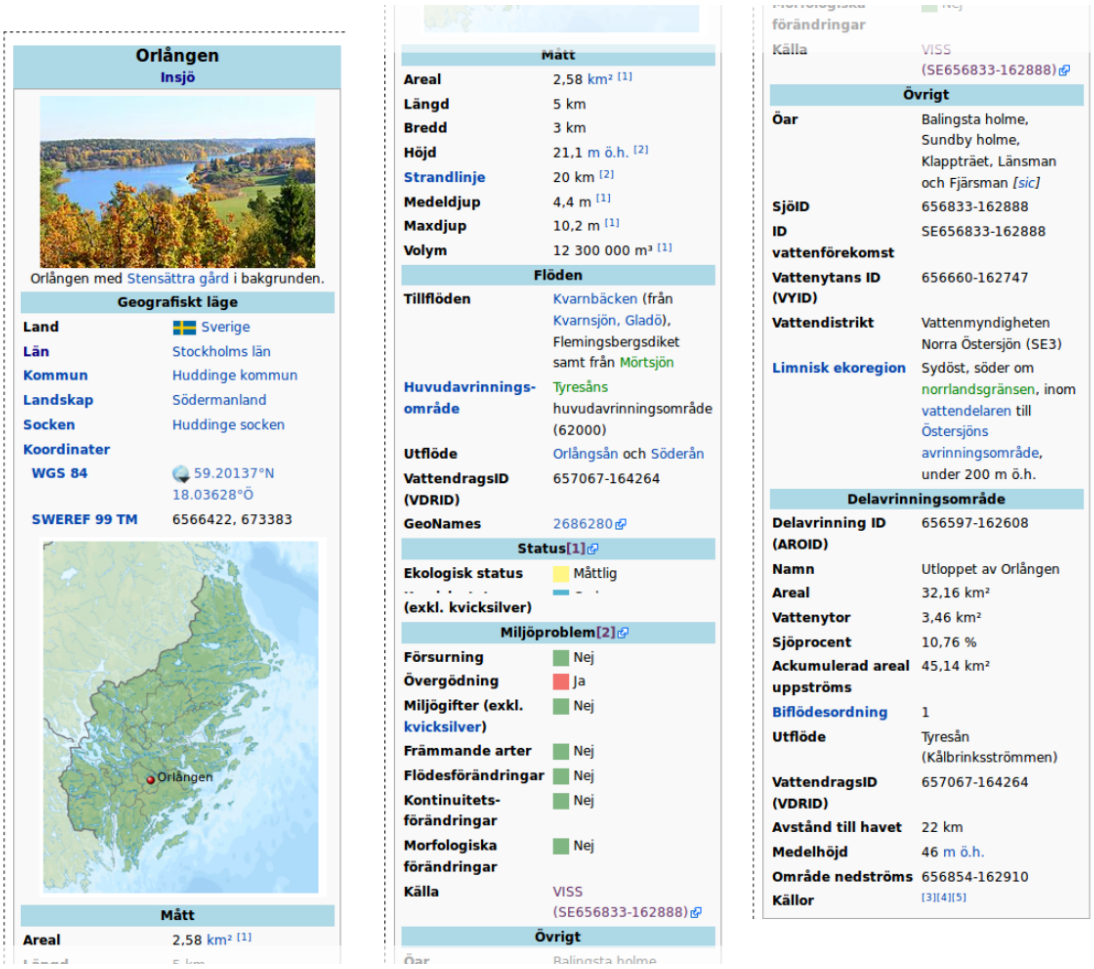
A lot of public spending is used to classify the water bodies of Europe. This knowledge could be exposed in Wikipedia articles throughout Europe for all waters (lakes, rivers, groundwater, coastal and transitional waters). Wikipedia articles are ranked high in all internet search engines and there is a huge possibility of highlighting WFD data to the general public, stakeholders, NGO:s, politicians, students and more.

In 2013 a contributor to Swedish Wikipedia generated articles[1] for all larger water bodies in Sweden using WFD reporting data as exposed through Water Information System Sweden (WISS)[2]. This illustrates what reporting data might look like when exposed through Wikipedia. Maintaining these articles is however done manually and updating them whenever new data is made available is time consuming and error prone.

At about the same time Wikidata[3] was created to be the shared knowledge base of the Wikimedia projects. Data stored in Wikidata is, as opposed to WIkipedia, structured, and multilingual. As facts from Wikidata can be exposed in Wikipedia articles it provides a simplified method for updating, maintaining and translating facts for use across all language versions of Wikipedia. Since the WFD reporting is done in a structured and uniform format across the member states it provides an excellent resource for systematically importing data into Wikidata.

With this as a background the Kalmar County Administrative Board (***Länsstyrelsen i Kalmar län***) and Wikimedia Sverige decided to start a joint project, ***WFD-data till Wikidata***, in the first half of 2016. The aim was to take the initial steps needed in order to make reported water data in Europe publicly available though Wikidata and exposed through Wikipedia.

[1] For an example article see https://sv.wikipedia.org/w/index.php?title=Orlången&oldid=23812764
[2] http://viss.lansstyrelsen.se/
[3] https://www.wikidata.org/



The infobox for Orlången, populated by data from WISS.
Corner image: Orlången från Stensätra / Johan Fredriksson / CC BY-SA 3.0

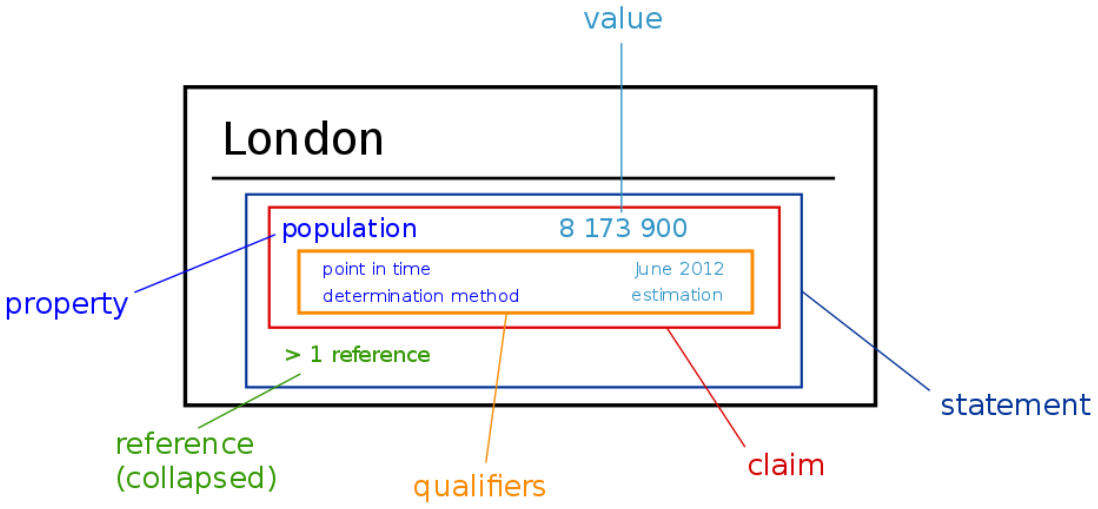
Initial analysis

To limit the scope of the project it was early on established that the initial focus would be on data for River Basin Districts (RBDs) and Lake Surface Water Bodies (Lake SWBs) while keeping other SWB types and Ground Water Bodies (GWBs) in mind for the future. The limitation was in large part based on the information believed to be of most interest for Wikipedia as well as being conceptually more easy to understand.

At the first initial meeting the WFD Reporting Guidance was gone through and for each statement it was decided if it was suitable/relevant for Wikidata, how prioritised it should be, and a layman's explanation for what the statement meant was constructed. This provided the basis for a living documentation[4] and served as the starting point for the process of matching these statements to Wikidata.

[4] https://se.wikimedia.org/wiki/Projekt:WFD-data_till_Wikidata_2016/Mappings

Structure on Wikidata



An illustration of the structure of a statement on Wikidata.
Wikidata statement / Kaganer, Kolja21, Bjankuloski06en, Lydia Pintscher, Addshore / CC BY-SA 3.0

An item on Wikidata contains multiple statements each being constructed from a claim and references. Each claim is constructed from a property, a value and optional qualifiers. Values can be either simple data types (such as a date or string) or references to other items on Wikidata. Qualifiers are property-value pairs which serve to clarify the scope of a claim or the conditions under which they are valid. Multiple claims using the same property are allowed.

When matching the WFD statements to Wikidata the first step is determining whether a property already existed which fills the need or if a new one needs to be constructed. The second step is identifying if the WFD structure can be kept or if it needs to be adapted to fit Wikidata.

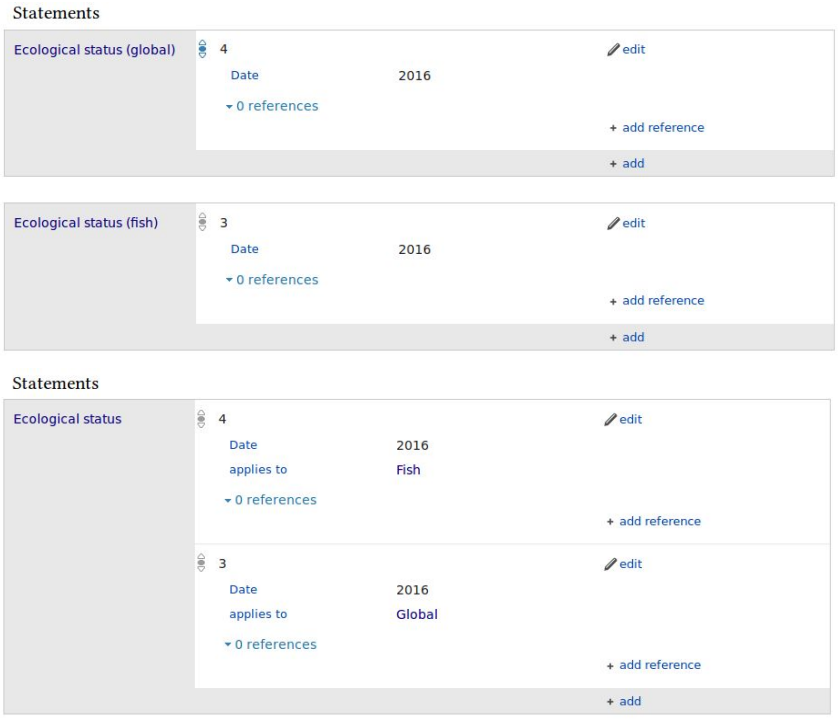
Structure adaptation

Wikidata has a rather flat structure in that while you can add multiple qualifiers to a claim a qualifier can itself not be further qualified and since the order of qualifiers are not guaranteed this cannot be used to encode further information either. By comparison the WFD data contains separate statements for e.g. the confidence or assessment methods of other statements. It is also not the custom of Wikidata to create items for aspects of a claim if they are not themselves concepts, by comparison WFD statements, e.g. Priority Substances (**`370jagjalq3mZklYf[]`**) can include classes and structures which are up to three levels deep. Notable is further that there is no boolean data type on Wikidata requiring a restructure of any such statements.

As Wikidata can contain both current and historical data it also becomes important that any implicit assumptions, such as year for the reporting cycle, is included in any claims. This also requires strategies for how to deal with situations where a statement is present in one

reporting cycle but not in the next, e.g. a given Significant Impact type (**`ko3a_fa^a[Yfl)ehY[l4qh]`**).

Property construction



Illustrating how qualifiers are used to reduce the number of needed properties, here for the modeling of Ecological Status (**`ko%[gdg_a[Yd3lYlmk/j0gl]flaYd6Ydm]`**).

The need to adapt structures affects which properties can be proposed, which can be combined and any rules they need to include with respect to allowed values or required qualifiers.

New properties are proposed through a formal process and need the approval of the community before they are introduced. The recommendation on Wikidata, in most cases, is to avoid the creation of multiple similar properties favoring instead to leverage qualifiers to distinguish use cases. As such the workflow for proposing a new property consists of:

- Ensuring there are no similar properties which could be used (e.g. Competent Authority was deemed to be covered under the definition of Operator[5]).
- Ensuring that the proposal is not too narrow or overly broad.
- Designing rules for allowed values.
- Designing rules for required qualifiers or the interpretation of optional qualifiers.

The rules for values and qualifiers should also be encoded so as to allow for automatic validation which, when implemented, will create constraint reports.

Determining when a proposed property is too narrow is a careful balance between ensuring an accurate interpretation of the values and being able to compare more items on the same property. To give an example Ecological Status can be either introduced as property which is limited to the WFD definition or as a more general property. In the narrow implementation it

[5] P137. See also the discussion in relation to the rejection of a more specific property.

can only be used for water bodies in the membership countries but we can apply strict rules on the allowed values. In the broader case we can use one property to compare the status across water bodies all over the world, but we sacrifice the accuracy of the values since the scale used in WFD might not correspond to that used by e.g. China.

It is important to also acknowledge that we only have control of what the proposal looks like initially, if it gets implemented, and how it gets implemented, is in the hands of the community. That said a sound proposal is unlikely to change drastically and the feedback from the discussion often serves to clarify details which were inadequately explained in the proposal.

Reference group meetings

In addition to the introductory meeting in Kalmar there were two meetings with the broader reference group. The first in Copenhagen in June 2016, and the second in Brussels in October of the same year[6]. The reference group consisted of representatives from ministries and water authorities in 11 different countries as well as representatives from the EEA and the European Commision.

At the meetings the first draft of new properties were presented and revised after discussions. The Wikidata test environment[7] was used which allowed new structures and properties to be instantly created and explored during the meetings. Using an environment which mirrored that on Wikidata made it easy to recognise any constraints in the models, explain the Wikidata structure to participants new tp the project and ensured all proposed designs were implementable.

The meeting also served to validate the choice of which statements from the WFD reporting to include on Wikidata. It was also a natural environment to address any worries or questions related to the data becoming accessible through Wikidata including, but not limited to, the fact that the data remains open to further editing once it is there. Also raised were identified needs with respect to visibility and accessibility that making the data available through Wikidata and by extension Wikipedia would fill. Finally the meetings were an opportunity to investigate the technical aspects of extracting data from Reportnet to Wikidata.

Copyright, using CC0 for data

A separate questions which was raised during both of the reference group meetings was the requirement that the WFD reporting data be explicitly licensed under CC0 in order to be eligible for inclusion on Wikidata. All Reportnet data is CC BY licensed unless otherwise indicated,[8] but for the purpose of this project that is not sufficient. While a few of the reported datasets include explicit licensing info the majority don't. This is partly due to technical limitations and the fact that there was no natural place to report this in the data. For the next

[6] See meeting notes below for links to the notes incl. participants lists.
[7] See e.g. the demo item for a lake.
[8] CC BY 2.5-DK per http://cdr.eionet.europa.eu/legalnotice.

reporting cycle EEA are recommending that a licensing attribute be added to the schema and looking into how this could be surfaced in association with the dataset in Reportnet.

Most participants ate the meetings were of the opinion that the (non-spatial) data is not copyrightable to begin with hence the question of licensing had never arisen. The requirement from Wikidata on explicit licensing, in part due to the EU database protection directive[9], requires that each member state looks into releasing the data if they wish to have it included.

In addition to the above the explicit outcomes, with regards to copyright, from the meetings were:

- EEA proposed CC0 in Reportnet for the data created by them (RBD and lakes regarding: ID, name and relations).
- A proposal was drafted to the Water Directors to suggest that parts of the publically reported data be made available under CC0.[10]
- A survey was designed by EEA to gauge the licensing situations among the member countries.
- The project was presented to the Data & Information Sharing working group of WFD making the case that the produced data be made as accessible as possible whenever there aren't any security implications to doing this.[11]

In addition to Sweden[12] the Netherlands were found to have already CC0 licensed (the spatial) parts of their data[13] and it was indicated by Finland and Poland that their data is also available under CC0.

Work on Wikidata

The work done on Wikidata was all aimed at laying the groundwork needed for the later imports. Apart from designing and proposing properties it included:

- Starting discussions about best practices (e.g. how to deal with boolean values or how to best describe data on measurement methods associated with a value).
- Creating basic concepts needed to describe the nature of the WFD data and the proposed properties.
- Mapping allowed property values to Wikidata items and creating new items when needed.
- Handling pre-existing value on Wikidata to ensure these are compatible with the data being imported.

It is worth noting that not all statements in the WFD reporting data needs its own property. As an example SWB category (**`kmj^Y[]7Yl]j#Yl]_gjq`**) was solved by using the **instance**

[9] http://ec.europa.eu/internal_market/copyright/prot-databases/index_en.htm
[10] https://upload.wikimedia.org/wikimedia/se/a/a5/Free_use_of_reported_WFD_data.pdf
[11] https://circabc.europa.eu/sd/a/420decec-d487-4d7b-84d3-b9eb8af281f6/Note_publication%20of%20data%20reported%20in%20WISE_final.docx
[12] http://viss.lansstyrelsen.se/About.aspx?aboutPageID=5
[13] http://cdr.eionet.europa.eu/nl/eu/wfd2016/spatial/envv9wchq under ***access constraints***.

of[14] property together with a created structure of items[15] relating the various SWB categories to the concept of an SWB.

Proposed properties

As part of the project 5 separate properties were proposed and accepted, one further property proposal was rejected, one is still under discussion and one pre-existing property was updated to clarify its relation to one of the new properties. The properties were proposed in sequence so that the feedback from one could be applied in the next. The time it took for a proposal to conclude varied greatly from 9 to 57 days with the average being about one month.

The first properties to be proposed were for **SWB code**[16] and **RBD code**[17]. Both are examples of unique external identifiers, a category of properties which are fairly common on Wikidata and thus were accepted fairly quickly. SWB code had a fairly big overlap with the pre-existing ***sjö id***[18] property used to identify lakes in Sweden. This meant that the subset of Swedish lakes which were also included in the WFD reporting (i.e. all with a surface area >1km²) could automatically get an SWB code added to them. There was some initial confusion about the scope of the pre-existing property which were clarified thanks to the efforts of this project.

Focusing on RBDs the next property to be proposed was one for **Competent Authority**[19]. After some discussion it was decided that this relationship could be covered by the pre-existing Operator property[20]. It should be noted however that this was by no means a unanimous decision and it highlighted the gap between the property proposal process, which recommended a broadening of a pre-existing property and the part of the community which maintain existing properties. The final resolution was non-ideal but the experience serves to illustrates the non-permanence of implemented properties, something which is important to be aware of when proposing broader properties.

The next proposal was a property for **Significant environmental impacts**[21] in relation to impact types in SWBs. Here there was added complexity due to this being a boolean datatype in the WFD data. The discussion focused on how you can show that an SWB is no longer suffering from a particular environmental impact (i.e. when importing data from two reporting cycles). This property was on purpose designed to be broader than just the WFD scope to allow it to be used for water bodies outside of the EU where the classification of

[14] https://www.wikidata.org/wiki/Property:P31
[15] https://www.wikidata.org/wiki/User:AndreCostaWMSE-bot/WFD/surfaceWaterBodyCategory
[16] https://www.wikidata.org/wiki/Wikidata:Property_proposal/EU_Surface_Water_Body_Code
[17] https://www.wikidata.org/wiki/Wikidata:Property_proposal/RBDcode
[18] https://www.wikidata.org/wiki/Property:P761
[19] https://www.wikidata.org/wiki/Wikidata:Property_proposal/Competent_authority
[20] https://www.wikidata.org/wiki/Property:P137
[21] https://www.wikidata.org/wiki/Wikidata:Property_proposal/Significant_environmental_impact_types with boolean values being discussed at https://www.wikidata.org/wiki/Wikidata:Project_chat/Archive/2016/10#How_to_model_properties_with_yes.2Fno_values

impact types might be different. The allowed impact types under WFD were then mapped to Wikidata items[22] for those values which were relevant to SWBs.

The final property to be accepted was that for (overall) **Ecological status**[23]. By contrast with the previous property this was specifically tailored to the WFD usage. The type was however changed from numeric (1-5) to taking items as a value. This allowed the scale used in WFD to be mapped to unique items which could be given clear multilingual labels (such as Poor status) making it more readily understandable. The allowed values were all created and collected together with their official definitions[24].

The final property to be proposed was that for (overall) **Chemical status**[25]. Being very similar to Ecological status in structure the same design was used. The allowed values were all created[26] and the proposal is currently being discussed by the community.

Prepared properties

In addition to the proposed properties one more was prepared but not proposed within the time frame of the project. This was a property for the ***Ecological status or potential per quality element (QE)***. This is similar to the accepted (overall) Ecological status property in its structure. The QE property was decided to be split from (overall) Ecological status during one of the reference group meetings the motivation being that the Ecological status value is an overall value supported by the values for the QEs done in such a way that any failing QE results in a failing status overall. The complication here will be to clearly communicate why one cannot use the same property with no qualifier indicating the overall status. The second complication is the sheer number of quality elements[27] which could be applied to a single SWB (each reporting cycle) which might be a deterrent for getting it accepted.

Imported data

Two datasets were imported to Wikidata within the scope of the project and a third is ready to go, awaiting clarifications about copyright. We started with RBDs as: there are less of them; they rarely existed on Wikidata (just six prior to the import); they contain less complex data; and they fill a hierarchical need in the later import of SWBs.

The code used in the imports is available under a free MIT licence on Github[28]. It was designed to make it easy to use to import WFD data from other countries (without having to edit the code) and for use with future reporting cycles (assuming the xml structure remains stable). Every statement that it adds comes with a reference indicating both the dataset (WFD reporting cycle and member country) and the .xml file from which the data was gotten.

---

[22] https://www.wikidata.org/wiki/User:AndreCostaWMSE-bot/WFD/swSignificantImpactType
[23] https://www.wikidata.org/wiki/Wikidata:Property_proposal/WFD_Ecological_status
[24] https://www.wikidata.org/wiki/User:AndreCostaWMSE-bot/WFD/swEcologicalStatusOrPotentialValue
[25] https://www.wikidata.org/wiki/Wikidata:Property_proposal/WFD_Chemical_status
[26] https://www.wikidata.org/wiki/User:AndreCostaWMSE-bot/WFD/swChemicalStatusValue
[27] https://www.wikidata.org/wiki/User:AndreCostaWMSE-bot/WFD/QualityElement
[28] https://github.com/lokal-profil/WFD_import/

The code respects pre-existing data meaning it won't overwrite user contributions and it can be run twice in a row without risking duplicate imports. This also allows it to be developed incrementally with more properties being supported over time. Additionally the code comes with preview functionality making it possible to do a trial run where the data is processed but outputted to a table rather than added to Wikidata.[29]

Swedish RBDs

The first import was done using a one-off script with the the primary purpose being to illustrate that the mapping process used for RBDs (and Competent Authorities[30]) worked. The 10 RBDs of Sweden were created together with their 10 Competent Authorities. The reason for doing this using a one-off script was that the official data was not yet available through Reportnet and thus the provided source data was slightly differently structured.

Finnish RBDs (prepared)

As the Finnish RBD data was already available this was selected as the trial dataset for the country neutral code. It requires that all Competent Authorities are first mapped to Wikidata items (or that such items are created) after which it can be feed the url to the RBDSUCA .xml file (and associated .gml file).

Imported data are:

- The English name of the RBD
- The local name of the RBD
- The RBD code
- The fact that it is an RBD
- The country
- The prime Competent Authority
- The surface area

The gml file is only used to access the name of the RBD in the local language.

Due to uncertainty with regards to the license of the Finnish data the import had to be paused after the first couple of items. Once this issue is resolved it can be restarted.

The script was also used to re-run the import of Swedish RBDs in order to ensure all statements were properly sourced with references to the WFD data.

Swedish SWBs

The import of Swedish SWBs is of particular interest since it largely overlaps the set of lakes created on Wikipedia back in 2013. As much of the information in those articles became outdated with the finalization of the 2016 WFD reporting cycle there is now a golden opportunity to update them by importing the data to Wikidata.

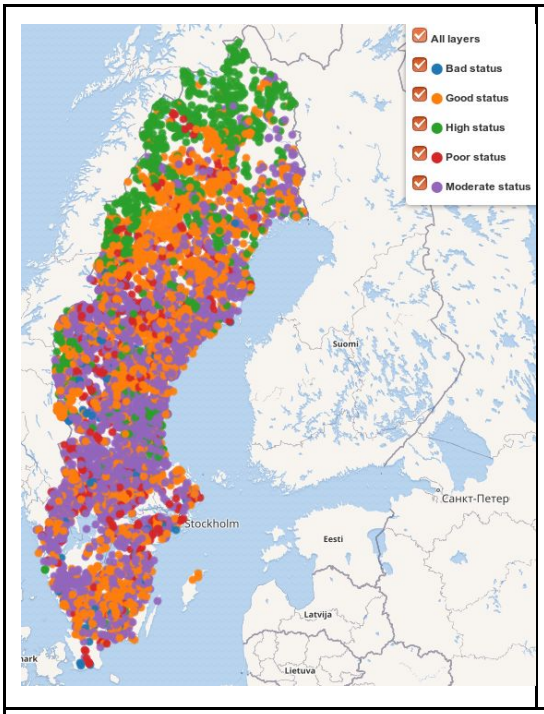
---

[29] An example of the preview data can be found at https://www.wikidata.org/wiki/User:AndreCostaWMSE-bot/WFD/preview.
[30] https://se.wikimedia.org/wiki/Projekt:WFD-data_till_Wikidata_2016/Mappings#CompetentAuthority

With this as a focus the first run only updated pre-existing items containing an SWB code, no new items were created. For Sweden these are about 6800 entries. When creating new items care must be taken to ensure that suitable connections are made between the administrative SWB item and the geographical body of water, especially in the cases where the relation is not 1:1.

Imported data are:

- The English name of the SWB (if any)
- The local name of the SWB
- The SWB code
- The SWB category of the item
- The country
- The RBD to which it belongs
- The ecological status
- Any significant impact types



The imported SWB data is here used to create two interactive maps. Left: SWBs coloured according to their ecological status. Right: SWBs coloured based on their river basin district.

Future use of the data

Now that the Swedish SWBs have been imported it is appropriate to start a discussion on the Swedish language Wikipedia about updating the infoboxes in the lake articles so that they make use of the up-to-date facts on Wikidata. There are plenty of similar examples of

Wikidata connected infoboxes around meaning implementing most of these facts should be fairly easy for any lake which has a 1:1 relationship to an SWB.

Certain statements, significant impact type in particular, have a slightly more advanced structure and may require the assistance of volunteers who are experienced in building these types of templates. Similarly for the situation where one lake consists of more than one SWB, or vice versa, additional assistance might be needed.

As part of the project a few example visualisations of the imported SWB data were produced. These did however not explore how the imported data can be connected to other datasets on Wikidata. Reaching out to the larger open data community it would be interesting to see what correlations they can find in the imported data and which new visualisations they can come up with when it is connected to the rest of Wikidata.

Conclusion

Due to unforeseen and external circumstances the project ended up taking longer (calendar time) than expected. This also included the delayed finalization of the Swedish reporting data which prevented early test imports. The delays have meant that some aspects of the project have not had time to be fully finalised. In particular there are more properties ready to be proposed,a few more facts which can readily be added to the import framework and a bit more work that can be done around communicating the need for CC0 licensing of WFD reporting data from more countries. In one area in particular there is plenty left to do and that is the preparation and prototyping of infoboxes on Wikipedia which make use of the imported data.

That said the project should still be considered a success. We created a framework which can be used for importing RBDs and SWBs from any member country that reports under the WFD. The framework can also easily be extended to coastal waters and GWBs as needed. We showed how the reported facts can be approached in a systematic manner to map these to the appropriate Wikidata properties and how property proposals can be constructed so as to be successful.

We got clear and positive feedback around the need to clarify the licensing of reported data in order to ensure that it can be as widely used as possible without the fear of violating database rights or copyright. There was also a clear positive response to the idea of making the reported data available through Wikidata and by extension on Wikipedia.

In short we have fulfilled our goal of laying the groundwork needed for anyone who wishes to use WFD data on Wikidata.

Links and additional materials

Created properties

- EU Surface Water Body Code (P2856)
- EU River Basin District code (P2965)

- Significant environmental impact (P3643)
- WFD Ecological status (P4002)

Presentations and letters

- Wikidata and WFD reporting - Given at the first reference group meeting, Copenhagen 2016-06-09
- Wikidata and WFD reporting - Given at the second reference group meeting, Brussels 2016-10-17
- Wikidata och WFD rapportering - Given to VISS applikationsråd, Online 2017-06-08
- Draft letter proposing that all reported data be made available under CC0

Meeting notes

- Start-up meeting 2016-04-12
- First reference group meeting, Copenhagen 2016-06-09
- Second reference group meeting, Brussels 2016-10-17