Tabellen im Web mit Koordinaten in CSV-Dateien konvertieren

Ein Arbeitsblatt für Selbstlernende und Interessierte.

1. Einleitung

Voraussetzungen

Um dieses Arbeitsblatt durcharbeiten zu können, werden folgende Dinge benötigt:

- Notepad++
- Eine Internetanbindung (um auf Wikipedia zugreifen zu können)

Über Wikipedia als Datenquelle

Wikipedia ist eine freie Datenquelle, mit Seiten, die oft auch Tabellen mit Koordinaten enthalten, wie beispielsweise eine Liste von Burgen im Kanton Zürich. Auch im Web gibt es immer wieder Tabellen - mit HTML-Tag table gekennzeichnet - mit nützlichen halbstrukturierten Daten.

Diese Listen könnten möglicherweise Koordinaten als eine Zeile, anstatt standartgemäss zwei (lat und lon) darstellen oder sie könnten unter anderem ihre Koordinaten als String (mit Anführungszeichen vor und nach der Zahl) angeben.

Man beachte dabei aber immer auch die Lizenz (beispielsweise dürfen Koordinaten aus Wikipedia nicht systematisch in OpenStreetMap importiert werden - höchstens für einen Abgleich).

Diese Anleitung zeigt anhand drei Beispielen, wie man auf Wikipedia und allgemein im Web verfügbare Tabellen zu sich holt und in eine CSV-Datei umwandelt und anschliessend "verschönert"

2. Beispiele und Aufgaben

Beispiel 1: Wikimedia-Liste konvertieren (Fehlende Fotos)

Wikimedia bietet eine Liste von Burgen in der Schweiz an. Die Liste zeigt jedoch die Koordinaten als eine Zeile an; dies aus dem Grund, damit ein Link möglich ist, der automatisch auf GeoHack verweist. Da aber der Standard die Koordinaten in zwei Spalten anzeigt, sollte dies geändert werden. Die Liste hat aber noch eine weitere Problematik, eine die erst sichtbar wird, wenn man die Liste in eine CSV-Datei konvertiert hat. Dann werden nämlich alle Koordinaten als String eingetragen, was daran

sichtbar ist, dass Daten in der Koordinatenspalte von Anführungszeichen benachbart werden, was völlig unkonventionell ist und ebenso behoben werden muss.

Link zur Tabelle: Fehlende Fotos in Commons

Aufgabe 1:

1. Die Wiki-Tabelle in eine CSV-Datei konvertieren

Um die Tabelle in im CSV-Format zu erhalten wird Wikitable2CSV genutzt.

Beispieldaten:

```
Location,Upload,Name / Wikidata-Object,Coordinates (GeoHack),Coordinates
(Map),Cat.,KGS-Nr
Reichenbach im Kandertal,,Burg Aris ob Kien,"46.6086,7.69556","https://castle-
map.infs.ch/#46.6086,7.69556,17z,N6074854753",-,-
Oberdiessbach,,Burg Diessenberg,"46.8289,7.63714","https://castle-
map.infs.ch/#46.8289,7.63714,17z,N6757108494",-,-
Wynigen,,Burg Friesenberg (BE),"47.1017,7.72889","https://castle-
map.infs.ch/#47.1017,7.72889,17z,N5629122641",-,-
Sennwald,,Burg Frischenberg,"47.2317,9.44917","https://castle-
map.infs.ch/#47.2317,9.44917,17z,N6020806251",-,-
Madiswil,,Burg Gutenburg (Gutenburg BE),"47.1833,7.79556","https://castle-
map.infs.ch/#47.1833,7.79556,17z,N3414232868",-,-
Gossau (SG),,Burg Helfenberg (St. Gallen),"47.4014,9.22278","https://castle-
map.infs.ch/#47.4014,9.22278,17z,W445209863",-,-
```

1. Per Notepad++ die Daten mit RegEx ausbessern.

   Wie in der Einleitung angemerkt, werden in dieser Tabelle die Koordinaten als Strings angegeben, ergo mit Anführungszeichen. Jedoch sind das nicht die einzigen Angaben mit Anführungszeichen, denn direkt danach folgt ein weiterer String, nämlich ein Link. Diesen wollen wir natürlich als String erhalten und müssen so einen Weg finden, die "falschen" Anführungszeichen zu löschen, während die "korrekten" Anführungszeichen unversehrt bleiben.

2. CSV-Datei in Notepad++ öffnen
3. Ersetzfunktion öffnen (Strg + H)
4. Regular expression aktivieren
5. Die Regular Expressions eingeben

   Dies ist der schwierigste Teil der Aufgabe, denn wir müssen nun eine RegEx bilden, die die Anführungszeichen, welche um jede einzelne Koordinate stehen, auswählt, aber die Anführungszeichen um den Link ignoriert. Hier empfehle ich RegExr, da es mit dieser WebApp recht einfach ist, eine passende Regular Expression zu kreieren. Meine RegEx `"(\.[0-9]),(\.[0-9])"` sollte für unseren Zweck genügen. Diese fügen wir einfach in das obere Feld ein. (Find what :) <<<

6. Das Ersetzende eingeben

   Da wir in der RegEx zwei Gruppen haben, können wir einfach alle Koordinaten mit den zwei Gruppen ersetzen, welche von einem Komma getrennt werden. Die RegEx sieht dann so aus: `$1,$2`, diese geben wir nun in das untere Feld ein (Replace with :)

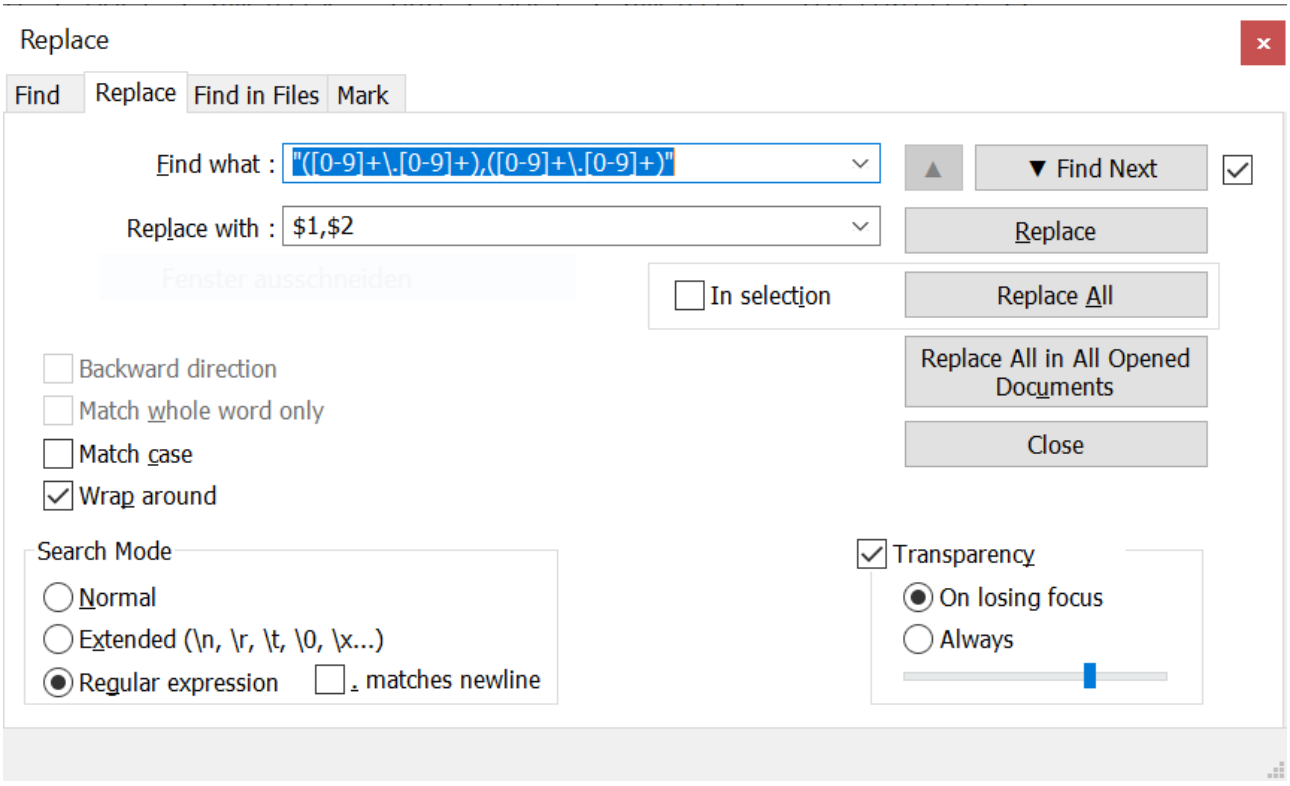

Abbildung 1. So sollte es dann in Notepad++ aussehen

7. Alle Vorkommen ersetzen (Replace All)
8. Spalte hinzufügen Nun haben wir eine Spalte mehr, als in der ersten Zeile vorgesehen, also wählen wir einfach `Coordinates (GeoHack)` aus und tauschen es mit `lat,lon` aus
9. Leerzeichen entfernen

   Wir sind aber noch nicht ganz fertig mit den Feldnamen, denn diese enthalten noch Leerzeichen und andere Spezialzeichen, dies lässt sich aber einfach bereinigen: Siehe "Schritt-für-Schritt Anleitung um Feldnamen zu säubern mit Notepad++" im Kapitel Umgang mit CSV.

10. Die CSV Datei in ein GeoJSON konvertieren mit convertcsv.com
11. Überprüfen mit geojson.io

Nun kann die GeoJSON-Datei direkt in QGIS importiert werden.

Beispiel 2: Wikipedia-Liste konvertieren (Burgen Kanton ZH)

In diesem Kapitel geht es um die konvertierung einer Wikipedia-Liste. Als Daten werden Burgen im Kanton Zürich verwendet.

Hier der Link zur Liste

Aufgabe 2:

1. Die Wiki-Tabelle in eine CSV-Datei konvertieren mit diesem Tool
2. Per Notepad++ die Daten ausbessern.

   In diesem Beispiel müssen wir uns nicht mit einer RegEx rumschlagen, da hier die Koordinaten nicht als String angegeben wurden und es somit keine Anführungszeichen zum entfernen gibt. Trotzdem werden hier die Koordinaten als eine Spalte angezeigt. Zu unserem Glück befindet sich aber zwischen lat und lon immer `" / "`, weswegen sich das Problem einfach und schnell mit Notepad++ beheben lässt.

3. CSV-Datei in Notepad++ öffnen
4. Ersetzfunktion öffnen (Strg + H)
5. Im oberen Feld `" / "` eingeben Warnung! Man muss aus dem Dokument irgendein Vorkommen von `" / "` **mit Abständen, ohne Anführungszeichen** kopieren und dann in die obere Zeile des Replace-Fensters einfügen, und **nicht** vom Arbeitsblatt kopieren.
6. Im unteren Feld `,` (Komma) eingeben

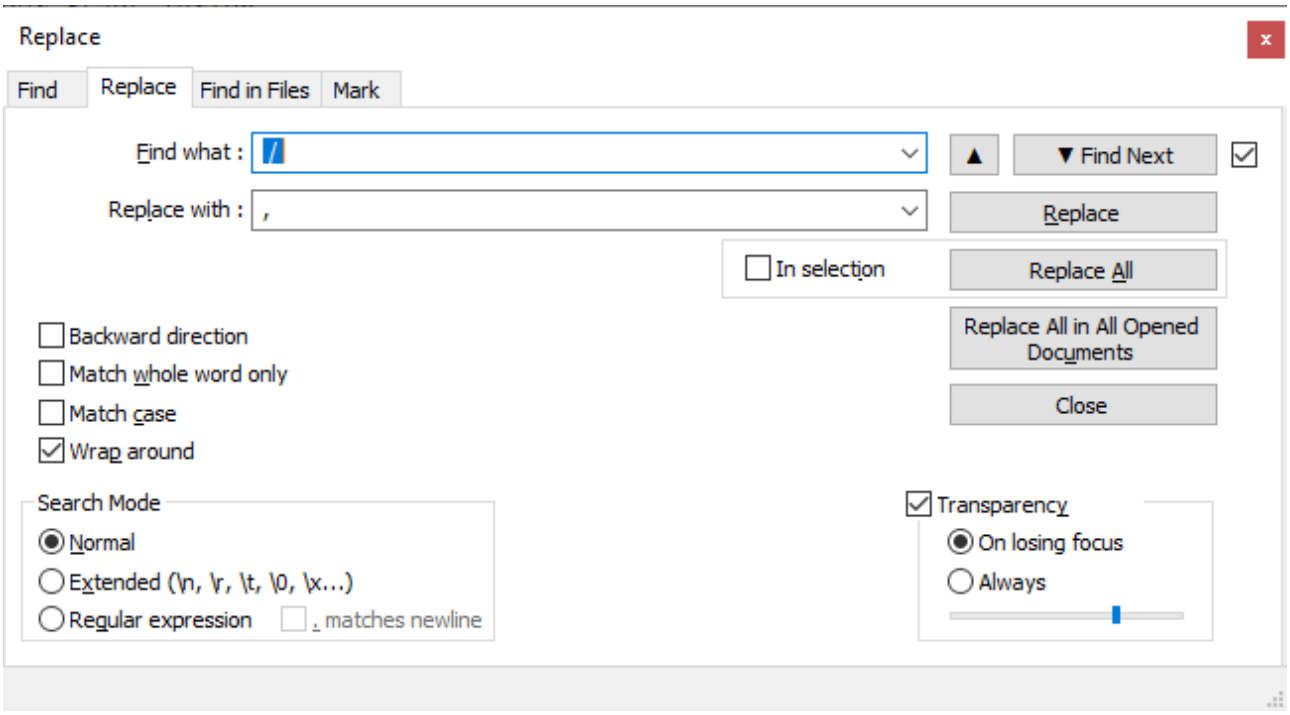

Abbildung 2. So sollte das Fenster danach aussehen

7. Auf "Replace All" (dt. "Alle ersetzen") drücken

8. Spalte hinzufügen Nun haben wir eine Spalte mehr, als in der ersten Zeile vorgesehen, also wählen wir einfach `Geokoordinate` aus und tauschen es mit `lat,lon` aus
9. Auf leere Koordinaten achten

   Da es möglich ist, dass einzelne Elemente keine Koordinaten besitzen, kann es sein, dass diese Elemente nun eine Zeile weniger haben (beim Ersetzen haben wir ja pro Zeile mit Koordinaten ein Komma, ergo eine weitere Zeile, hinzugefügt). Ob dies der Fall ist, können wir untersuchen, indem wir unsere Daten durch einen Lint untersuchen lassen, dieser gibt dann Unstimmigkeiten aus und in welcher Zeile sich diese befinden. ToolkitBay bietet hier einen nützlichen Lint an. In unserem Beispiel ist der "Record 58" fehlerhaft, was bedeutet, dass wir in der 59. Zeile (da ja die 1. Zeile nur die Feldnamen enthält, und keinen "Record") ein Komma hinzufügen müssen, an der Stelle, an der die Koordinaten wären.

10. Die CSV Datei in ein GeoJSON konvertieren mit convertcsv.com Beachte, dass du die Latitude und Longitude Felder angeben musst (in "Step 3: Choose output options"). Die Latitude befindet sich in Feld 6, die Longitude in Feld 7.

Nun kann die GeoJSON-Datei direkt in QGIS importiert werden.

Beispiel 3: HTML-Tabelle im Web

Wir müssen uns aber nicht nur auf Wikitabellen beschränken, sondern können auch rein anhand des HTML-Tags Tabellen in eine CSV-Datei konvertieren, und somit können wir dies nun bei nahezu jeder erdenklichen Tabelle im Netz anwenden. Zudem ist es auch sehr einfach und schnell. Als Beispiel befindet sich eine Tabelle mit Bahnbildern am Ende des Arbeitsblattes; das Ziel ist es, diese Tabelle in ein GeoJSON zu konvertieren, welches wir in GQIS importieren können.

Beispieldateien:

Siehe Anhang

```
Banhnfoto,Name,Kurzbescheibung,lat,lon
,Zug im Abendrot in Ilanz,Ziemlich selbst beschreibend,-,-
,Lock und Wagon,Eine Lock und ein Wagon,46.8428393,9.48085999972222
,Blick zurück,Ja schaut aus der Lock nach hinten und so,-,-
,Zug bei Landquart,Zügig voran,46.9395796,9.55894379972222
,Zwei Wagons,Zwei Wagons in einer Berggegend,-,-
,Fahrt durch das Dorf,Zwei Wagons fahren durch irgend ein Dorf im Bündnerland,-,-
,Gewarteter Zug,Ein Zug wird gewartet,46.9642584997222,9.55621529972222
,Halt bei Thusis,Ein Zug hält in Thusis,46.6986813997222,9.44036089972222
,Zug hält,"Zug hält bei einer Haltestelle, Endstation Arosa",-,-
,Zwei Locks,Zwei Locks stehen sich gegenüber an einer Haltestelle,-,-
```

Aufgabe 3

1. Die HTML-Tabelle in eine CSV-Datei konvertieren mithilfe dieser Seite (**Anmerkung:** Gib bei "Enter URL" die URL des Arbeitsblattes an. Da aber die HTML-Table-Tags nur erkannt werden, wenn man

das HTML-File angibt, soll man auch die URL dessen angeben und nicht die des PFDs. Ausserdem nicht zu vergessen ist, die richtige Tabelle auswählen (Step 3: Generate output), da manchmal das Programm fälschlicherweise Tabellen meint zu finden, die es gar nicht gibt. In unserem Beispiel handelt es sich hier um Tabelle 5)

2. Die CSV Datei in ein GeoJSON konvertieren mit diesem Link Da in diesem Beispiel die Koordinaten nicht als String abgespeichert wurden, bedarf es keiner Bearbeitung, weswegen man die Daten direkt in das GeoJSON Datenformat konvertieren kann. Beachte, dass du die Latitude und Longitude Felder angeben musst (in "Step 3: Choose output options"). Die Latitude befindet sich in Feld 4, die Longitude in Feld 5.
3. Überprüfen mit geojson.io

Nun kann die GeoJSON-Datei direkt in QGIS importiert werden.

3. Der Umgang mit CSV

CSV ist ein de-facto-Standard. Es gibt einen Standard datz - und es gibt Excel, das davon abweicht (v.a. Encoding und Delimiter).

Komma oder Strichpunkt als Delimiter. Eher Komma verwenden.

Strings in Hochkomma - besonders wenn möglicherweise Komma oder Strichpunkte vorkommen.

Umlaute ⇒ UTF8 verwenden.

Feldnamen bilden die erste Zeile in einer CSV-Datei und legen fest, was für Werte in die respektive Spalte gehören. Da Spezialzeichen in CSV-Dateien mit Feldbegrenzerzeichen gekennzeichnet werden, diese aber, wenn man Daten herunterlädt, nicht automatisch hinzugefügt werden, ist es besser, sie gleich zu entfernen.

Schritt-für-Schritt Anleitung um Feldnamen mit Notepad++ zu säubern

1. Gewünschte Datei mit Notepad++ öffnen
2. Oberste Zeile markieren
3. (Strg + H) um in den Ersetzmodus zu gelangen
4. Im Search Mode links unten den Knopf "Regular expression" anhaken
5. Rechts den Knopf "In Selection" anhaken
6. In die oberste Eingabe die Regex `[^,A-Za-z0-9_-]` einfügen
7. Die untere Eingabe leer lassen
8. Auf "Replace All" klicken (rechts vom Knopf "In Selection")
9. Man beachte, dass der Zeilenumbruch auch gelöscht wird, diesen muss man manuell wieder

hinzufügen.

Für Interessierte an Ersetzen in Notepad++ und für Interessierte an RegEx: regex 101 und RegExr

Noch Fragen? Wende dich an die QGIS-Community.

4. Anhang

Liste zum Beispiel 3:

Banhnfoto	Name	Kurzbescheibung	lat	lon
	Zug im Abendrot in Ilanz	Ziemlich selbst beschreibend	46.463126	9.122776
	Lock und Wagon	Eine Lock und ein Wagon	46.842839	9.480859
	Blick zurück	Ja schaut aus der Lock nach hinten und so	46.769662	10.111101
	Zug bei Landquart	Zügig voran	46.939579	9.558943
	Zwei Wagons	Zwei Wagons in einer Berggegend	46.792075	9.821148
	Fahrt durch das Dorf	Zwei Wagons fahren durch irgend ein Dorf im Bündnerland	46.967543	9.554941
	Gewarteter Zug	Ein Zug wird gewartet	46.964258	9.556215
	Halt bei Thusis	Ein Zug hält in Thusis	46.698681	9.440360
	Zug hält	Zug hält bei einer Haltestelle, Endstation Arosa	46.79786	9.70235

Banhnfoto	Name	Kurzbescheibung	lat	lon
	Zwei Locks	Zwei Locks stehen sich gegenüber an einer Haltestelle	46.705088	8.855182
	Zug im Schnee	Ein Zug steht im Schnee (Recht typisch dort oben)	46.792075	9.821148
	Zug hält am Bahnhof in der Nacht	Es ist Nacht, wir sind am Bahnhof und ein Zug hält	46.85362	9.52901
	Bündner Aussicht	Wir sehen aus dem Zug und bekommen eine typische bündner Aussicht zu sehen	46.82115	9.86087
	Gelagerter Zug	Ein abgesteller Zug von der Vorderseite	46.939579	9.558943
	Gehaltener Zug von der Seite aus	Ein Zug, welcher gerade hält, sehen wir von seiner rechten Flanke	46.77532	9.20995
	Panoramawagen	Aufnahme eines Zuges mit Panoramawagons	46.50766	9.84958
	Güterwagen	Ein Güterwagen in der graubündischen Landschaft, der Baumstämme transportiert	46.649673	9.723689
	Gelbe Lock	Eine gelbe Lock mit Fokus auf Wald und Berg	46.89206	9.84731
	Rangierlock	Eine Rangierlock im Abendrot	46.8775	9.53323
	Zug im Schnee (weit)	Ein Zug im bildnerischen Schnee	46.84431	9.86526

Banhnfoto	Name	Kurzbescheibung	lat	lon
	Halt bei Ilanz	Ein Zug hält bei Ilanz	46.775251	9.206711

Noch Fragen? Wende dich an die QGIS-Community!