Bilaga 1: ETL-dokumentation av VOOKA-projektet

Innehåll

1. Allmänt

ETL-processen genomfördes med hjälp av Python-programmeringsspråket och alla programkoder som utvecklats finns öppet tillgängliga under MIT-licensen på projektets GitHub-sidor. I GitHub finns en förteckning över ETL-verktygets tekniska krav (requirements.txt), varav de viktigaste modulerna är GeoPandas och Shapely. Processens metodologi kombinerar både algoritm och ett heuristiskt förhållningssätt.

Det programmässiga arbetsflödet beskrivs i sin helhet nedan. De separata faserna beskrivs däremot närmare i egna avsnitt. Som allmän beskrivning bestod ETL-processen av sex huvudfaser (länkar till GitHub-sidorna):

1. Datainsamling och förbehandling
2. Sammanställning av uppgifter
3. Korrigeringar av uppgifter
4. Jämförelse av uppgifter
5. PDF-länkningskonvertering
6. Implementering av plandatamodellen

2. Datainsamling och förbehandling

2.1 Gränssnitt

I ETL-verktyget är det tekniskt möjligt att samla in uppgifter från tre olika källor:

1. från gränssnitt enligt OGC:s WFS-standard (kommunernas planmaterial),
2. från Esrin ArcGIS Feature Layers (kommunernas planmaterial) samt
3. OGC API Features -gränssnitt (LMV:s fastighetsuppgifter)

De två förstnämnda returnerar den inmatade informationen enligt URL i sin helhet som GeoPandas GeoDataframe. Skriptet OGC API Features returnerar däremot API:s innehåll i GeoJSON-format med en avgränsning av det landskap som ska granskas (bounding box), som har hårdkodats för genomförandet. I den fortsatta utvecklingen ska den hårdkodade geografiska avgränsningen ersättas med koordinater för det önskade området. I OGC API Features-skriptet har man också lagt till hjälpfunktioner med vilka LMV:s dataschema i geoJSON-format för fastighetsuppgifter kan normaliseras till tabulär information i Pandas DataFrame.

Vid utnyttjandet av ETL-verktyget är det också bra att beakta att fas 1.1 inte fungerar för alla lösgöringar av material från WFS-gränssnitten. Detta observerades i samband med genomförandet av VOOKA i Norra Savolax. Felet beror förmodligen på den version av Geopandas som används i ETL-verktyget och som inte kan hantera geometrierna i gränssnittet på rätt sätt. Vid behov går det dock att manuellt lösgöra material från WFS-gränssnitten. Versionsuppdateringen har antecknats i verktygets behov av vidareutveckling.

2.2 Inhämtande av uppgifter från url-länkar

Vid genomförandet i Norra Savolax observerades att flera kommuner har url-länkar på webbplatsen som leder direkt till de dokument som behövs i projekten (plankarta, planbestämmelser). För detta ändamål skapades ETL-fasen, med hjälp av vilken man kunde effektivisera nedladdningen av dokument. För nedladdning behövs en csv-tabell som innehåller url-adresser. I processen laddas filerna från URL-adresserna automatiskt till en önskad mapp. Om filen inte kan laddas ner från adressen i fråga ger programmet användaren ett felmeddelande.

OBS! Ladda ner filer endast från sådana adresser som är tillförlitliga. Den utvecklade koden utför automatisk nedladdning för alla URL-adresser som finns i CSV-filen. Det är användarens ansvar att säkerställa att url-länkarna är från en tillförlitlig källa.

2.3 Förbehandling

För en viss kommun hade WFS-gränssnittet genomförts i KommunGML-format, som inte kunde läsas in med traditionella metoder. Problemet löstes med ett separat XML-parser-skript, där planinformationen lösgjordes direkt från märkspråkets struktur.

En del av kommunmaterialet erhölls som CAD-ritningar som levererades separat i stället för som gränssnitt, där egenskapsuppgifterna var bundna till punktgeometrier i stället för de egentliga plangränserna. Egenskapsuppgifterna kombinerades med plangränserna med ett separat skript.

I LMV:s FDS-material anges den gamla kommunkoden för planerna för kommunsammanslagningsområden. Dessa uppdaterades så att de motsvarar den gällande kommunkoden för att möjliggöra en omfattande jämförelse med kommunmaterialet.

2.4 PDF-länkkonvertering

Det är möjligt att i ett enhetligt format länka PDF-planhandlingarna till indexet i geodataformat för planerna, om man vet till vilken planindexbeteckning varje handling är kopplad. De automatiseringar som presenteras i detta ETL-verktyg kräver validering, dvs. materialet som länkaren producerar ska för sin del kontrolleras manuellt. Dessutom kan inte behovet av manuell länkning helt elimineras med hjälp av automatiseringen.

I arbetet med PDF-länkningen sammanställdes en länkningstavla där varje rad innehöll information om planens indexbeteckning (FDS, kommunens material eller båda) samt planens dokumenttyp (t.ex. plankarta). Schemat på länktavlan (tabell 1) var följande:

Tabell 1. PDF-länkningstavlans egenskapsdatafält och deras förklaringar.

Egenskapsuppgifter	Förklaring
Kommunnummer	Kommunens officiella kommunkod.
Kommunens planbeteckning	Planbeteckning för kommunens planmaterial i geodataformat.
FDS-indexbeteckning	FDS-materialets indexbeteckning (planbeteckning_1).
Original filename	Det ursprungliga filnamnet på kommunens planbilaga.
New filename	Kolumn för nytt filnamn
Planslag	Koduppsättning för planslag på övre nivå.
Manuellt kontrollerad	Har materialet granskats manuellt (boolean)
Dokumenttyp	Koduppsättning för dokumenttyp, numerisk
Match equivalency %	Motsvarighetsprocent för automatisk länkning
Anmärkningar	Fält för registrering av observationer
Multipage	Finns det fler än en sida i dokumentet (boolean)
Status	Kodsystem som anger om planhandlingen är giltig.
Gällande	Är planhandlingen giltig eller inte (boolean).
Geometry origin	Geometrins källmaterial
Feltyp	Beskrivning av eventuellt observerat fel, numerisk
Beskrivning	Beskrivande text som genererats från fälten geometry origin och feltyp för plandatamodellen

Beskrivningar av klassificerade egenskapsuppgifter:

Koduppsättning för planslag:

- ak
- rak
- yk

Dokumenttyp -koduppsättning (verbal och numerisk):

- 1 = plankarta (inkl. märkningar och bestämmelser)
- 2 = plankarta (inkl. inte märkningar och bestämmelser)
- 3 = märkningar och bestämmelser (separat)
- 6 = annat

Status-koduppsättning:

- ok
- inte ok

Geometry origin-kodsystemet:

- kommun
- FDS
- digitaliserad i VOOKA
- ingen geometri

Feltyp-Kodsystemet:

- 0 = inget fel
- 1 = Fel/brist i handlingen
- 2 = Fel/brist i avgränsningen
- 3 = Fel/brist i handlingen och avgränsning
- 4 = Annat fel/brist
- 5 = Ingen indexbeteckning

2.4.1 Ändringar i länktavlan

I genomförandet i Norra Savolax lades data till och data togs bort från länktavlan. I projektet samlade man in plankartor och bestämmelser, därför lämnades beskrivningarna och programmen för deltagande och bedömning bort från PDF-länktabellen. I tabellen lades match equivalency-procenten till. Den berättar om den automatiska länkningens motsvarighet och hjälper vid manuell granskning. Om procenten är låg är det mycket rekommenderat att kontrollera om handlingen matchar indexet. Dessutom kan automatiken generera fel kanalbeteckning. Även detta kan härledas från en låg motsvarighetsprocent. Den automatiska länkningen är alltså inte helt tillförlitlig och kräver i viss mån manuell granskning och validering.

Dessutom lades följande nya fält för egenskapsuppgifter till i tabellen: felklassificering, geometry origin samt beskrivning. Kolumnen Felklassificering anger om det finns ett eventuellt fel i materialet (antingen index eller handlingen) som bör granskas. Felen klassificerades enligt typ av handling, avgränsning, båda dessa, andra fel samt fel som gällde avsaknad av indexbeteckning. I projektet användes felregistreringen för materialets länkning och validering. Det bör dock observeras att kommunerna är experter på planläggning inom sitt eget område, så det är i praktiken kommunerna som ansvarar för att granska felen.

Kolumnen Geometry origin anger varifrån indexmaterialets geometri härstammar. Detta är väsentligt för kommunerna så att de kan granska kvaliteten på sitt eget material. Kolumnen Beskrivning innehåller källan till geometrin samt felklassificeringen i samma kolumn i textform. Beskrivningsfältet finns både i geoPackage-formatets geodatamaterial och i den slutliga JSON-filen enligt plandatamodellen, där man dessutom har samlat FDS-koden som motsvarar planindexet.

2.5 Planhandlingarnas katalogstruktur

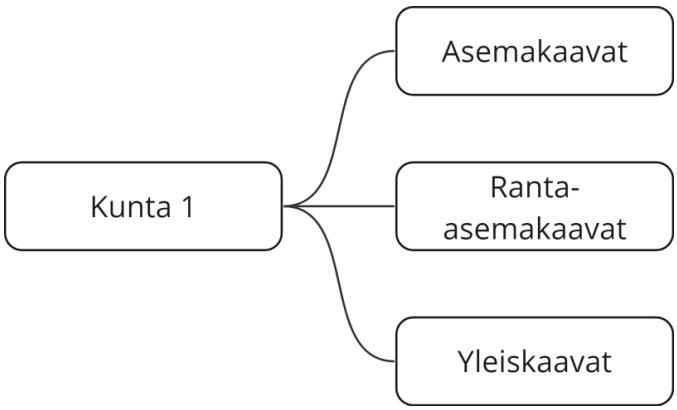
Planhandlingarna från kommunerna sparades i en standardiserad katalogstruktur i resurshanteringen. Katalogstrukturens stomme hade formen:

```
documents
|
└──kommunkod
|   |
|   |
|   └──ak
|   |   asiakirja.pdf
|   |   asiakirja2.pdf
|   |   ...
|   └──rak
|   |   asiakirja.pdf
|   |   asiakirja2.pdf
|   |   ...
|   └──yk
|     asiakirja.pdf
|     asiakirja2.pdf
|     ...
 …
```

ETL-verktyget utnyttjar katalogstrukturen i fråga för att namnge filer på nytt!

2.5.1 Automatisk mappsortering

I genomförandet i Norra Savolax utvecklades en automatisk mappsortering för såväl handlings- som indexmaterial. I projektet laddade kommunerna upp material till en Sharepoint-mapp, inom vilken materialet kunde finnas i flera olika undermappar. Då är det utmanande och tidskrävande att hitta rätt uppgifter i strukturen. Som lösning utvecklades en automatisk mappsorterare. Minimikravet för att koden ska fungera är att mapparna har sorterats kommunvis innan koden körs. Dessutom ska undermapparna inom den kommunspecifika mappen vara sorterade enligt planslag (bild 1). Därefter väljer koden önskade filformat och för dem under en mapp enligt planslag.



Figur 1. Minimikrav för den automatiska mappsorteraren för den ursprungliga mappstrukturen.

Koden identifierar PDF-filer som planhandlingar. Om de inlämnade planhandlingarna är i något annat filformat behandlar programmet dem inte. Koden identifierar SHP-, GPKG-, DWG- och DXF-material som index. Om indexmaterialet är något annat än de ovan nämnda formaten behandlar programmet dem inte.

Användaren ska slutligen manuellt kontrollera att de sorterade objekten har hamnat i rätt mapp för plantyper. Programmet kan inte heller identifiera om en fil är väsentlig för VOOKA-materialet, dvs. en plankarta eller en bestämmelse, eller någon annan fil som inte är relevant för projektet, till exempel en planbeskrivning. Efter den automatiska sorteringen ska användaren granska resultaten av sorteringen och eventuellt ta bort onödiga filer.

2.6 Länkning av handlingar som utsetts med planbeteckning till kommunens indexmaterial med planbeteckning

Ett av projektets huvudmål var att länka planhandlingarna till motsvarande planindex. Tidigare gjordes länkningen helt manuellt i pilotprojektet i Södra Savolax. Under pilotprojektet identifierades behovet av att effektivisera länkningen. I det inledande skedet av ETL-processen i Norra Savolax skapades skedet 1.4.6, med hjälp av vilket handlingar som utsetts med planbeteckningen kan länkas till kommunens planindexmaterial, om indexmaterialet innehåller planbeteckningar.

Programmet jämför filnamnen med kommunens planbeteckningar i indexmaterialet och länkar dem till varandra om motsvarigheten är tillräckligt bra. Samtidigt producerar programmet en motsvarighetsprocent.

Uppgifterna sammanställs i csv-tabellen i den form som länkningstabellen i PDF-länkningskonverteringen kräver. Eftersom programmet endast producerar vissa värden blir en del av datafälten tomma. I programmet har minimivärdet för motsvarighetsprocenten definierats som ganska lågt (35 %) för att länkningen ska vara så effektiv som möjligt trots eventuella fel. Därför ska användaren kontrollera de rader på länkningstavlan där motsvarighetsprocenten är låg.

I det här skedet utvecklades också en definition av dokumenttypen, som avgör filnamnets typ. Vid tidpunkten för publicering av detta dokument (2/2024) returnerar verktyget dokumenttypen som en bestämmelse(3) om filnamnet innehåller en ordlista som hänvisar till bestämmelsen. Om den aktuella ordlistan inte hittas ger verktyget en plankarta samt bestämmelser (1) som standardvärde för dokumenttypen. Ordlista med hjälp av vilken koden definierar dokumenttypen finns här: document_type.

OBS! Filtypen blir en bestämmelse om ordet i fråga förekommer i filnamnet. Detta orsakar fel i tabellresultaten och kräver att dokumenttypen i vissa fall granskas manuellt. Denna automatisering underlättar dock behandlingen av materialet, men kräver noggrannhet av användaren. Behovet av vidareutveckling i detta skede beskrivs i kapitel 8.

Med hjälp av automatisering kan man alltså underlätta länkningen, men inte helt eliminera behovet av manuell granskning.

För ETL-processen utvecklades också ett eget skede för sådana fall där filnamnet saknar planbeteckning och indexmaterial med planbeteckning saknas. Då kan man köra steg 1.4.7. Programmet skapar nödvändiga uppgifter i länktabellen, men länkar inte dokumentmaterial till indexmaterial.

3. Kombination och korrigering av data

3.1 Kombinerade plandata

Innehållet i planmaterialet från kommunerna varierade enormt. För att möjliggöra stabil jämförelse och validering skapas planmaterial i ETL-verktyget som kombinerats med både FDS- och kommunmaterial med ett enhetligt schema (tabell 2). Uppgifterna sparas som egna nivåer under den gemensamma master-geopackage. Nivåerna är följande:

- Detaljplaner_kommun
- Detaljplaner_fds
- Generalplaner_kommun
- Generalplaner_fds
- Kommunernas_registerföringsområden_fds (vid behov)

Detaljplanenivåerna omfattar både detaljplaner och stranddetaljplaner.

Planslagen har meddelats i FDS-materialet i enlighet med koduppsättningarna i Fastighetsdatasystemets fastighetsregister. De omvandlas i ETL-verktyget så att de motsvarar koderna i datasystemet för den byggda miljön[1]. I kommunernas material har plantyperna inte angetts i ett separat egenskapsfält. ETL-avgör dessa utifrån planförklaringarna i materialet. Om inga förklaringar har angetts i materialet anges som plantyper enligt antagandet i generalplanerna 23 = delgeneralplan, i stranddetaljplanerna 33 = stranddetaljplan och detaljplanerna 31 = detaljplan.

[1] https://koodistot.suomi.fi/codescheme;registryCode=rytj;schemeCode=RY_Kaavalaji

Tabell 2. Egenskapsuppgifter för kombinerade plandata och förklaringar till dem. Krysset beskriver för kommun- och FDS-materialets del om egenskapsuppgiften ingår i materialets schema.

Egenskapsuppgifter	Förklaring	Kommunmaterial	FDS-material
FID	Kod som identifierar raden	X	X
originalref	Utgångsmaterialets ursprungliga koordinatsystem som epsg-kod	X	X
den gamla kommunkoden	Den gamla kommunkoden för områden med kommunsammanslagningar		X
kommunkod	Gällande kommunkod enligt Statistikcentralen	X	X
kommunnamn	Kommunens namn	X	X
planbeteckning	Planbeteckning som kommunen gett	X	
kaavatunnus_1	Första delen av planbeteckningen för FDS-material		X
kaavatunnus_2	Slutdelen av planbeteckningen för FDS-material		X
planbeskrivning	Beskrivning av/namn på plan eller planindex	X	X
planslag	Kod för planslag enligt koduppsättningen i datasystemet för den byggda miljön	X	X
godkännandedatum	Datum för godkännande av planen	X	X
fastställelsedatum	Datum då planen fastställdes	X	X
ikraftträdandedatum	Datum då planen träder i kraft	X	X
objektregisterenheter	Förteckning över fastighetsbeteckningar i anslutning till planindex	X	X
kaavakartta_ja_maaraykset	Namn eller hyperlänk på plankartan och bestämmelsens PDF-dokument	X	X

plankarta	Namn eller hyperlänk på plankartans PDF-dokument	X	X
bestämmelser	Namnet på planbestämmelsernas PDF-dokument eller hyperlänk	X	X
Beskrivning	Beskrivning av källan till geometrin och eventuella feltyper	X	X

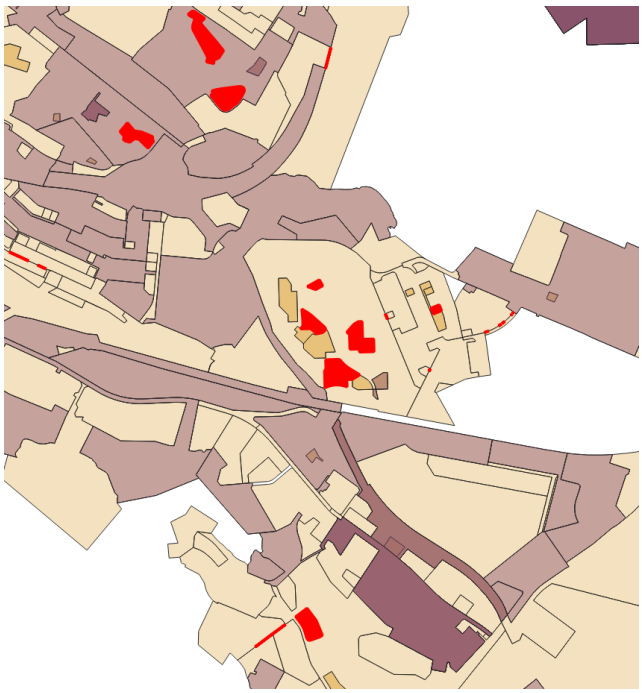
Kombinerade plandata har en central roll i alla skeden av ETL-verktyget, eftersom hela verksamhetslogiken grundar sig på ett schema i ett masterdataset.

3.2 Geometrikorrigeringar

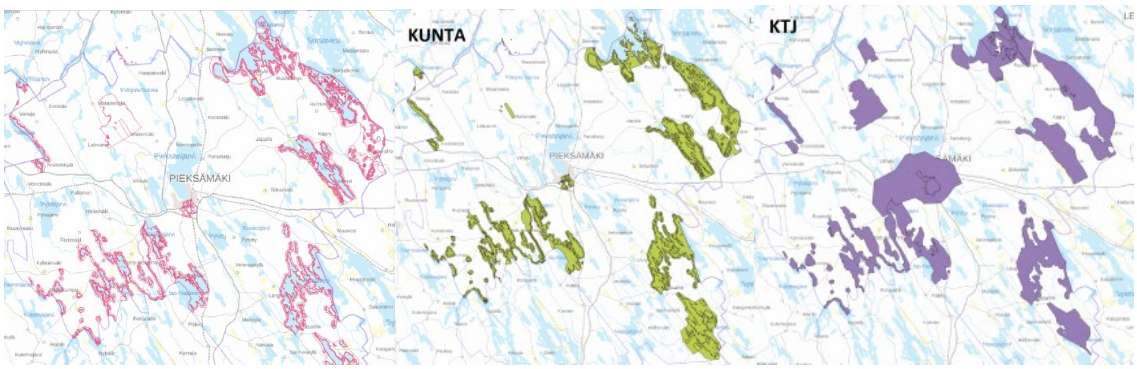
Vid geometrikorrigeringar stöder sig ETL-verktyget starkt på Pythons Shapely-bibliotek. För varje planindex körs explain validity-funktionen som returnerar uppgiften som textfält om geometrin i sig är ogiltig. Om formelindexet till exempel korsar sig själv eller innehåller en s.k. sliver-polygon, återställer funktionen textfältet "Ring Self-intersection" samt koordinaterna för den problematiska positionen. Om planindexets geometri är felaktig, korrigeras den med Shapelys make valid-funktion. I vissa ovan nämnda fall löser inte Shapely-funktionen geometriproblemet (särskilt CAD-material), varvid problemfallens geometrier validerades med geodataprogrammet QGIS.

Ofta ska den ursprungliga geometrin delas upp i flera objekt för att geometrin ska bli duglig. Om flera objekt ska med samma geometrityp ska skapas vid uppdelning återställs geometrin i flera delar (MultiPolygon). Om objekt ska skapas med olika geometrityper, återställs GeometryCollection. I fråga om planerna är endast områdesgeometrier tillåtna, varmed ETL-verktyget behandlar eventuella GeometryCollections separat och återställer dem till områden.

I en kommun producerade konverteringen av CAD-materialet till geodata överlappande "spökgeometrier" för planindexen. ETL-verktyget filtrerade bort dessa överlappningar utifrån jämförelsen av egenskapsuppgifter.

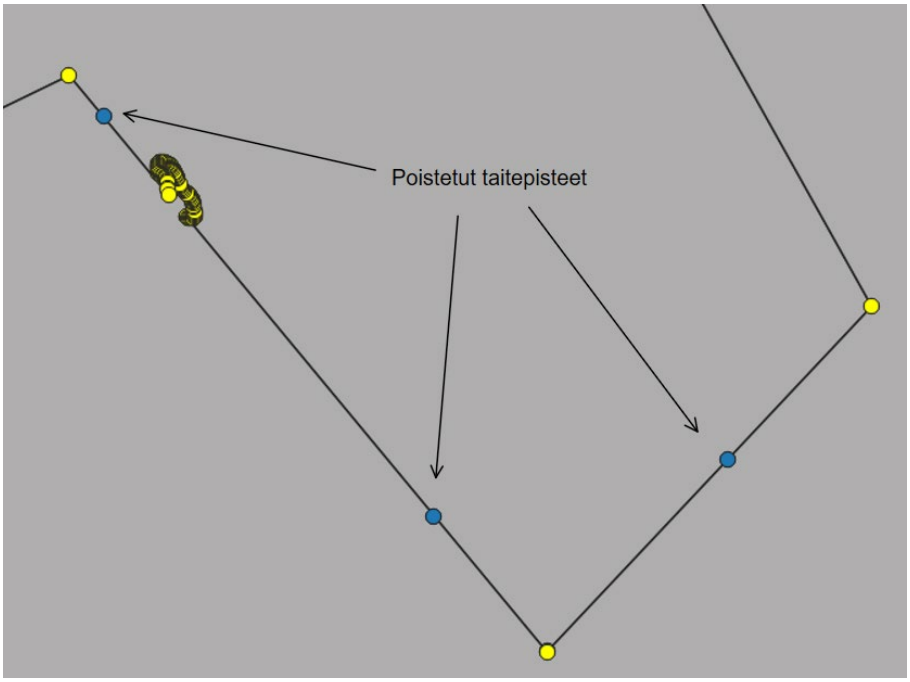


Figur 2. Kommunernas egna planindexgeometrier kan vara utmanande. I det här exemplet fanns det 852 polygoner med 352 geometrifel i kommunens detaljplanematerial.



Figur 3. Det är svårt att automatiskt omvandla sträckliknande avgränsningar till polygoner, eftersom man måste tolka avgränsningarnas riktighet. Till vänster originalmaterialet. I mitten omvandling till polygoner och till höger i FDS visas avgränsningar.

I genomförandet i Norra Savolax utvecklades ett skede i ETL-verktyget för att avlägsna extra vändpunkter mellan slutpunkterna på ett rakt linjesegment. Vissa system accepterar inte data där vändpunkter mellan intilliggande geometrier längs kanten av en rak linje (i princip onödiga) avviker från varandra. I ETL-verktyget körs funktionen removeUnnecessaryVertices för materialet, vilket tar bort dessa överflödiga vändpunkter samtidigt som materialets topologiska integritet bevaras (se bild 4).



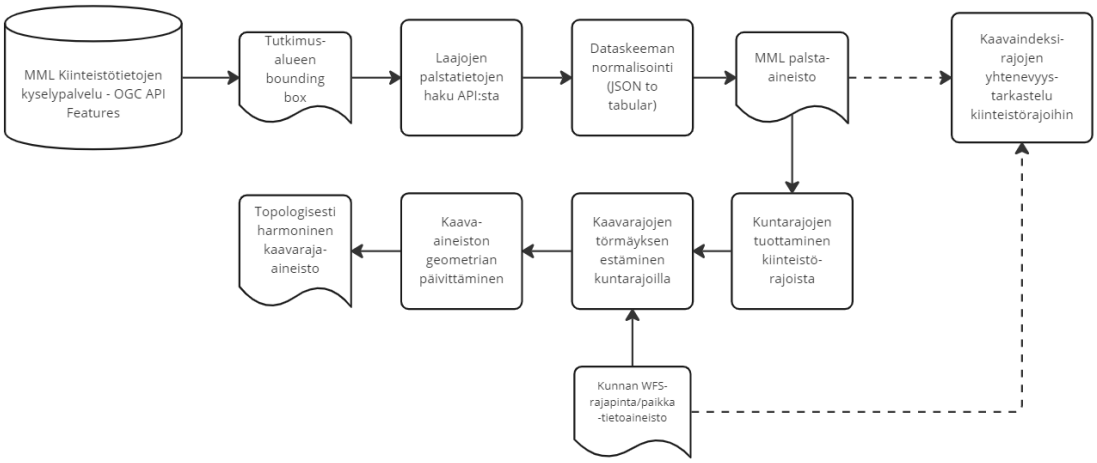
Figur 4. Exempel på korrigerat material. På bilden visas det ursprungliga och korrigerade materialet med Select by location-valet, där alla punkter som överlappar varandra har valts (i gult). Extra vändpunkter som tagits bort har visas i blått. Den topologiska integriteten bevaras också med bågformer som härstammar från CAD-material och som består av korta och raka linjer i GeoPackage-vektorformat.

3.3 Länkning av objektregisterbeteckningar

Lokaliseringsuppgifter för fastighetsbeteckningarna finns i LMV:s OGC API Features-gränssnittet för fastighetsuppgifter som en egen poängnivå. ETL-omvandlaren länkar till varje planindex information om fastighetsbeteckningar i anslutning till den huruvida lokaliseringspunkten för en enskild fastighetsbeteckning är inom planindexets yttre gränser. Detta innebär att många planer har flera fastighetsbeteckningar, varvid dessa listas efter varandra i planens egenskapsuppgifter.

3.4 Korrigering av kommungränskollisioner i generalplaner vid fastighetsgränser

I ETL-processens avsnitt allmänt, nämns att processens metodologi kombinerar både algoritm och ett heuristiskt förhållningssätt. Korrigeringen av kommungränskollisioner i generalplanerna vid fastighetsgränserna representerar det senare - genomförandet grundar sig på slutledning och bedömning av data. Bild 5 visar processens faser.



Figur 5. Processen för korrigering av kommungränskollision i generalplanerna. Det avsnitt som separerats med en streckad linje beskrivs närmare som en del av den geometrisk-topologiska jämförelsen.

Verksamhetslogiken för ETL-omvandlaren vid generalplanegränserna är följande:

1. För kommunen bildas en maskering av kommungränsen över fastighetsgränserna som fås från LMV:s tomtdata. Användaren kan om hen så önskar besluta om huruvida exklaverna ska beaktas eller inte (fastighetsgränser som ligger utanför den egentliga kommunavgränsningen).
2. Planindexen itereras individuellt och man kontrollerar om planindexet finns inom den bildade kommungränsmasken. Om indexet omfattas helt av maskeringen görs inget (det finns inget behov av gränsändringar).
3. Om planindexet inte omfattas av maskeringen helt, skär det fastighetsgränsen delvis eller är helt utanför maskeringen (det senare fallet förekom inte i kommunerna i Södra Savolax). Då finns det ett behov av att korrigera gränsen.
4. Man beräknar geometrin för de avsnitt där planindexet korsar maskeringen av kommungränsen (=intersection).
5. Man beräknar geometrin för de avsnitt där maskeringen av kommungränsen avviker från planindexet (=difference).
6. Man filtrerar difference-geometrier. De delar av diffrence-geometrin som ligger inom de yttre gränserna för det planindex som jämförs förkastas. Dessutom förkastas de difference-geometriandelar där arealen är exceptionellt stor. Varje kommun har ett tolerans-gränsvärde för arealen som finns listat i ETL-skriptets docstring och Jupyter Notebook.
7. Till sist beräknas kopplingen mellan den bildade intersect-geometrin och den filtererade difference-geometrin (= union), som bildar en ny geometri för det planindex som behandlas.

3.5 Korrigeringar av egenskapsdata

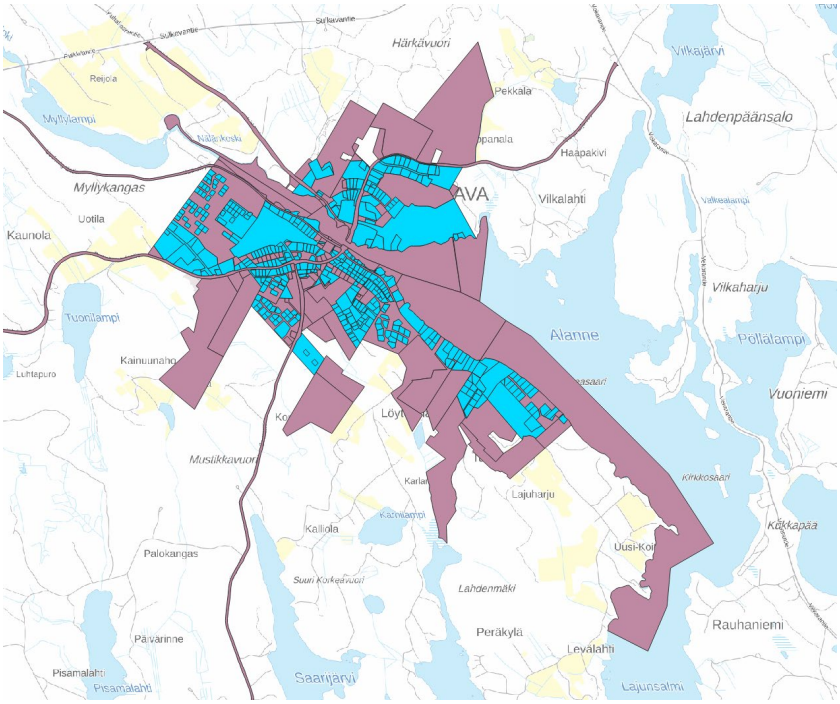
Datumen (t.ex. datum för godkännande av planen, datum för fastställande) hade angetts i mycket varierande former och delvis bristfälligt i planmaterialet från kommunerna. Olika kommuner använde som datumformat bl.a. DD.MM.YYYY, YYYMMD och YYYY. Dessutom hade uppgiften ofta angetts som textfält i stället för som datum-uppgiftstyp.

Som en del av ETL-processen skapades en funktion som översätter tabulär data i textform till date-datatyp i formen YYYY-MM-DD. Om utgångsuppgiften endast är ett år, gavs som datum den första dagen i januari året i fråga.

4. Geometrisk-topologisk jämförelse

4.1 Jämförelse av fastighetstomt och planindex

I jämförelsen av fastighetstomter och planindex lyfter man fram hur väl fastighetsgränserna samstämmer med gränserna för planindexen. Den topologiska samstämmigheten uttrycks som ett procenttal som beräknas som förhållandet mellan arealen för ett enskilt fastighetsskifte och för en korsningspunkt i planindexet (intersection) och fastighetstomtens areal (Bild 6).



Figur 6. Exempel på jämförelse av tomter och planindex. Med blått de tomter som motsvarar planindexgränserna med minst 90 % noggrannhet. Med rödaktigt de tomter vars motsvarighet till planindexgränserna är under 90 %.

4.2 Jämförelse av FDS- och kommunmaterial

Geometrisk-topologisk jämförelse görs i ETL-verktyget enligt OGC:s DE-9IM-standard (Dimensionally Extended 9-Intersection Model), med hjälp av vilken man identifierar enhetliga, överlappande och olika geometrier. Dessutom jämför ETL-omvandlaren plangränsernas form och arealer automatiskt.

I DE-9IM-modellen anges relationen mellan två geometrier med en 9-siffrig modell (Tabell 3). I VOOKA-projektet var den överlägset vanligaste relationsformen mellan FDS och kommunmaterialet "212101212".

Tabell 3. Exempel på DE-9IM relationsmodell.

DE-9IM	Insida (Interior)	Gräns (Boundary)	Utsida (Exterior)
Insida (Interior)	2	1	2
Gräns (Boundary)	1	0	1
Utsida (Exterior)	2	1	2

Modellen läses enligt följande:

- 2: Korsningspunkten för geometrierna är ett område (Polygon) dvs. tvådimensionell geometri
- 1: Korsningspunkten för geometrierna är en linje (LineString) dvs. endimensionell geometri
- 0: Korsningspunkterna för geometrierna är en punkt (Point), dvs. nolldimensionell geometri

Andra värden i DE-9IM-modellen kan vara:

- F: False - korsningspunkten förekommer inte
- T: True - vilket icke-false värde som helst är tillåtet (0,1 eller 2)
- *: allt är tillåtet

Om relationsmodellen mellan två geometrier motsvarar formen "T*F**FF*", är geometrierna helt lika. I VOOKA-projektet uppfylldes inte detta villkor för en enda plan som motsvarade FDS och kommunmaterial. Detta beror sannolikt på att digitaliseringen av planmaterialet har skett på olika sätt och DE-9IM-modellen tillåter inte ens en bråkdels avvikelse. **Därför tolkades motsvarande planer i VOOKA-projektet som geometrisk-topologiskt enhetliga om deras topologiska noggrannhet var minst 98 %.**

Topologisk noggrannhet uttrycks i ETL-omvandlaren med ett iou-värde. Förkortningen kommer från orden intersection of union, korsningspunkten för geometriernas beröringspunkt. Om de två planerna är geometriskt-topologiskt exakt likadana, är deras korsningspunkt och beröringspunktens arealer exakt desamma. Förhållandet mellan dessa är således 100 %. Om det finns geometrisk-topologiska skillnader i planerna är deras beröringspunkt av en annan storlek än korsningspunkten. Då skiljer sig förhållandet mellan beröringspunktens och korsningspunktens arealer också från 100 %.

Geometrisk-topologisk integritet uttrycks i ETL-verktyget totalt i fälten i Tabell 4.

Tabell 4. Datafält som uttrycker geometrisk-topologisk integritet och som produceras i ETL-verktyget.

Egenskapsuppgifter	Förklaring
area_ha	Planens areal i hektar.
refe_area_ha	Areal i hektar för en motsvarande plan.
de9im_pattern	Relation enligt DE-9IM-modellen med modellen med 9-siffror.
topo_equal	Boolean-information. Är planerna som jämförs topologiskt enhetliga eller inte. Ja, om iou-värdet är minst 98.
iou	Intersection of union-procent. Anges som förhållandet mellan berörings- och korsningspunkten i de planer som jämförs.
refe_planbeteckning	Planbeteckning för motsvarande plan.
a_delta_%	Den relativa felprocenten för arealerna i planerna som jämförs.
false_a_delta	Boolean-information. 0 = felprocenten är inte false positive, 1 = felprocenten är false positive.

Under VOOKA-projektet fanns det ingen identifierare mellan kommunernas data och FDS-data. På basis av egenskapsuppgifterna kunde man således inte veta vilken kommuns plan som motsvarade FDS-planen och tvärtom. I ETL-verktyget har problemet lösts så att motsvarigheten till planen i det andra datasetet är den plan vars iou-procent är störst.

En geometrisk-topologisk jämförelse kan genomföras med ett verktyg som producerats i VOOKA-pilotprojektet antingen kommun- eller FDS-materialbaserat. Det rekommenderas dock att FDS-materialet används som grund för jämförelsen eftersom materialet i princip är enhetligare i många kommuner i Finland!

4.2.1 Identifiering av felaktiga resultat i materialjämförelse

I DE-9IM-beräkningen är det möjligt att jämförelsens motsvarande formel kan vara en s.k. false positive. Detta innebär i praktiken att de relativa felprocenten för arealerna stiger till tusentals och att iou-procenten är minimal. Då ger beräkningen ett felaktigt resultat som verkar vara korrekt. I praktiken godkänns inte som motsvarande plan en plan enligt vilken arealernas relativa felprocent överstiger 100 %, varvid felprocenten kan konstateras vara false positiv. I genomförandet i Norra Savolax utvecklades en del för verktyget som identifierar dessa felaktiga uppgifter.

Detta genomfördes genom att lägga till ett nytt datafält false_a_delta som berättar om värdet är false positive eller inte. Datafältet får värdet 0 om felprocenten (a_delta_%) är under 100. Om felprocenten är över 100 får datafältet värdet 1.

OBS! Denna egenskap utvecklades endast för en FDS-materialbaserad jämförelse, eftersom jämförelsesättet i fråga konstaterades vara lönsammare på grund av materialets integritet.

5. Jämförelse av egenskapsuppgifter

Medan det är möjligt att göra en geometrisk-topologisk jämförelse även utifrån kommunmaterialet har jämförelsen av egenskapsuppgifterna i VOOKA-pilotprojektet endast genomförts på FDS-basis. Sammanfattningsvis uttrycks enhetligheten i FDS- och kommunmaterialets egenskapsuppgifter i ETL-verktyget i fälten i Tabell 5.

Tabell 5. Datafält som uttrycker egenskapsuppgifternas enhetlighet och som produceras i ETL-verktyget.

Egenskapsuppgifter	Förklaring
kl_equal	Boolean-information. Är uppgiften om planslag i FDS-materialet densamma som i kommunmaterialet.
hyv_equal	Anger om datumet för godkännande av FDS- och kommunmaterialet är detsamma.
voim_equal	Anger om FDS- och kommunmaterialets ikraftträdandedatum är detsamma.

Det första egenskapsfältet "kl_equal" anger som boolean-information om de jämförda materialens planslag är desamma. I dessa finns det tidvis stora variationer i VOOKA-pilotprojektets olika kommuner, eftersom:

1. FDS ofta i detaljplanerna berättar om planen är en riktgivande tomtindelning eller inte. I pilotprojektets kommunmaterial saknades denna information.
2. FDS innehåller ingen information om planer utan rättsverkningar. Denna information var också sällsynt men förekommer tidvis i VOOKA-pilotprojektets kommunmaterial.

Det andra och tredje egenskapsfältet "hyv_equal" och "voim_equal" anger att datumen för godkännande och ikraftträdande är enhetliga samt eventuella brister i dem. I fälten kan förekomma:

1. “Godkdatum/ikrafttdatum saknas i både”
   - I både FDS- och kommunmaterialet saknas det datum som ska granskas för planens del.
2. “Godkdatum/ikrafttdatum saknas i FDS”
   - I både FDS-materialet saknas det datum som ska granskas för planens del.
3. “Godkdatum/ikrafttdatum saknas i kommun”
   - I både kommunmaterialet saknas det datum som ska granskas för planens del.
4. 1 - Datumen som granskas är desamma
5. 0 - Datumen som granskas skiljer sig från varandra

Datumen för fastställandet ingår inte i jämförelsen, eftersom de saknas helt i FDS-materialet. Som allmän princip är de egenskapsuppgifter som ingår i jämförelsen de enda som kunde jämföras i VOOKA-pilotprojektet.

5.1 Utveckling av jämförelsen av egenskapsuppgifter

Jämförelsen av egenskapsuppgifter utvecklades i Norra Savolax så att den grundar sig på Pythons Difflib-modul. Skrivsätten för datumen kan variera betydligt och dessutom kan de innehålla onödiga tecken: till exempel mellanslag eller överflödiga nollor. Koden utvecklades så att datum kan anses vara lika med hjälp av ett visst tröskelvärde. För detta användes Difflibs SequenceMatcher-klass. Tidigare gav jämförelsen av datum inte särskilt goda resultat i koden. Orsaken till detta är till exempel att datumen för FDS-materialet i kommunjämförelsen har formen ÅÅÅÅ-MM-DD. I kommunens material kunde datumen vara av annan form, vilket innebär att när man enbart granskar jämförelsen av likhet kanske det inte beaktas att datumen kan vara lika, men bara i olika skriftliga former.

OBS! Denna egenskap utvecklades endast för en kommunbaserad FDS-jämförelse. Orsaken till detta är att jämförelsemetoden i fråga är mer lönsam på grund av materialets integritet.

Datumen för kommunernas och FDS planmaterial inom detta projekt motsvarade varandra i väldigt många kommuners fall. Verktyget som utvecklats för att ändra datumens egenskapsuppgifter måste dock preciseras på grund av det mycket varierande sättet att registrera datumen. För fortsatt användning av ETL-verktyget finns det ett oändligt antal olika sätt att skriva datumen nationellt sett, så verktyget bör utvecklas med många olika scenarier som tar hänsyn till alla skrivsätt. Nationellt sett finns det ett behov av att standardisera hur datumen skrivs för att göra den automatiska hanteringen av material mer effektiv.

6. PDF-länkkonvertering

6.1 Nya namn på filer med plandokument

Principen för att ge nya namn på filer med plandokument är att skapa enhetliga namn. ETL-verktyget skapar standardiserade namn på anslutningstabellen i det nya egenskapsfältet "new_filename" enligt följande principer:

1. Dokumentnamnet beaktar den officiella kommunkoden
2. Dokumentnamnet beaktar planslaget enligt kodsystemen i Datasystemet för den byggda miljön (Ryhti)
3. Dokumentnamnet beaktar dokumentettypen enligt Ryhtis kodsystem
4. Dokumentnamnet beaktar planbeteckningen
5. Två eller flera dokument får inte ha samma namn

Exempel på ett filnamn för ett plandokument som producerats av ETL-verktyget (mer information om ETL-verktyget i bilaga 1), "402-31-05-95-2.pdf", där:

- 402 är Leppävirtas officiella kommunkod
- 31 är detaljplan enligt RYTJ:s kodsystem för planslag
- 05 hänvisar i enlighet med kodsystemet Typ av handling i RYTJ till att bilagan omfattar både plankartan och bestämmelserna
- 95 är kommunens indexbeteckning
- 2 är ett glidande prefix (402-31-05-95-1.pdf finns redan, varmed två bilagor med samma namn har hittats för en gällande plan)

ETL-verktyget gör det möjligt att namnge dokumentfilen på nytt antingen utifrån FDS-koden eller kommunens egen indexbeteckning.

OBS! Praxis för att ge filer med plandokument nya namn är endast ett förslag till namngivning av bilagor i Ryhti! Den standardiserade namngivningspraxisen kan ändras så länge Ryhti-genomförandeprojektet pågår!

6.2 Länkning av dokumentfiler till geodatamaterial

ETL-verktyget länkar de filnamn som namngetts på nytt från länkningstabellen till geodatamaterialet på basis av plantyp, kommunkod samt planens indexkod (antingen FDS eller kommunens kod, beroende på vilkendera som bilagorna har namngetts på nytt). Planslag och kommunkod behövs eftersom:

1. Inom kommunen kan t.ex. stranddetaljplaner ha samma kommunens indexbeteckning som detaljplanerna
2. Olika kommuner kan ha exakt samma planbeteckning.

Efter körningen finns varje dokument i sitt eget fält i geodatamaterialet (plankarta).

6.3 PDF/A-konvertering

ETL-verktyget konverterar kommunernas PDF-dokumentfiler automatiskt med Ghostscript-konverteraren till PDF/A-arkivformat. Konverteringen grundar sig i tillämpliga delar på den Python-Ghostscript-konverterare som är öppet tillgänglig vid GitHub. En konverterare som anpassats för VOOKA-projektets behov finns som en egen version på projektets GitHub-sidor.

Om det erhållna PDF-formatet har skapats med en version som inte kan fås i PDF/A-format, sparas det som sådant. Om vissa av kommunens PDF-filer av tekniska skäl som beror på kommunens material inte konverteras till PDF/A-format, sparas dessa i normalt PDF-format.

7. Implementering av plandatamodellen

I ETL-verktyget implementeras plandatamodellen genom att omvandla planmaterialet i GeoPandas GeoDataFrame-format till GeoJSON-format. För ETL-omvandlare anges som materialkälla antingen "kommun" eller "FDS" samt planbeteckningarna för båda materialkällorna. Som en del av jämförelsen av materialet har man i punkt 4.2 skapat datafältet "refe_planbeteckning", som t.ex. i FDS-baserat material anger den kommunkod som motsvarar planen. Om motsvarande kod inte hittas (t.ex. om planobjektet endast finns i FDS:s material), lämnar ETL-omvandlaren uppgiften om kommunens indexbeteckning i GeoJSON tomt. Vid tillämpningen av plandatamodellen är det bra att beakta den kontinuerliga förändringen i plandatamodellen, så det lönar sig att kontrollera mängden inmatade uppgifter eller hur de namnges samt den hierarkiska strukturen för den information som sparas hos den instans som upprätthåller plandatamodellen innan uppgifterna matas in. Kommunernas material kan ha formen av antingen polygon eller multipolygon. I det slutliga JSON-formatet sparades dock alla geometrier som multipolygoner.

Med hjälp av ETL-verktyget är det dessutom möjligt att normalisera JSON-materialet enligt plandatamodellen tillbaka till geopackage-format så att materialet är lättare att läsa och granska för kommunerna i geodataprogram.

I plandatamodellen har det fastställts att det är obligatoriskt att anteckna beslutsfattarens art. I projektet samlade man inte in uppgifter om huruvida planen har godkänts av till exempel stadens nämnd eller en enskild tjänsteinnehavare. Informationen varierar från plan till plan, men sannolikt finns det endast några, om alls, beslut som fattas av enskilda tjänsteinnehavare. Därför definierades beslutsfattarens art i Norra Savolax i alla planer som 02 - Kollegialt beslutsorgan..

8. Behov av vidareutveckling

8.1 OGC API Features geografisk avgränsning

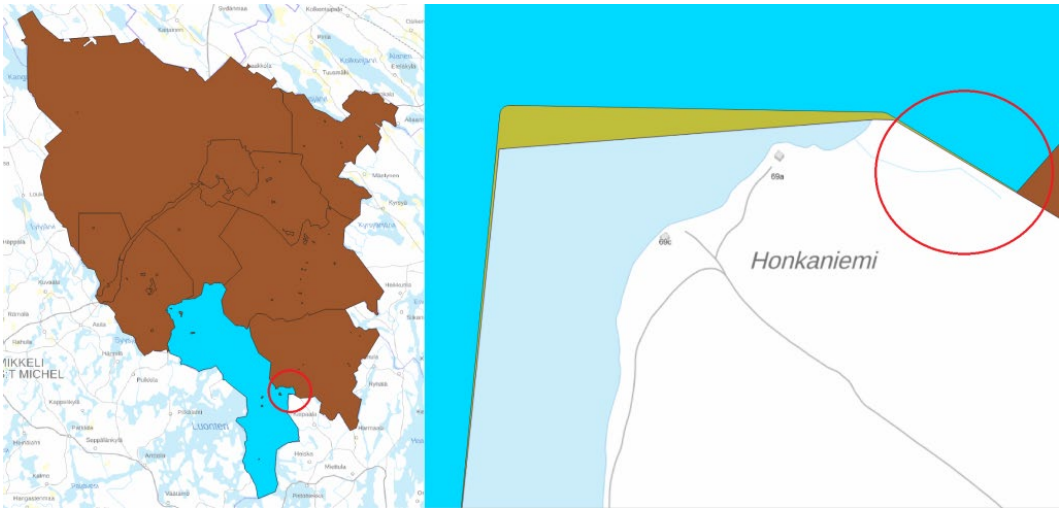
Både i pilotprojektet i Södra Savolax och i VOOKA-projektet i Norra Savolax har de geografiska gränserna för det landskap som granskas hårdkodats som en bounding box till OGC API Features getter-skriptet. I fortsättningen ska bounding box inom andra geografiska områden ersättas med önskade koordinater.

8.2 Versionuppdatering av geopandas

Under projektet i Norra Savolax i ETL-processens fas 1.1 observerades att ETL-verktyget inte fungerar för alla lösgöringar av material från WFS-gränssnitten. Felet beror förmodligen på den version av Geopandas som används i ETL-verktyget och som inte kan hantera geometrierna i gränssnittet på rätt sätt. De versioner av Geopandas och dess tillbehör som används i ETL-verktyget kan uppdateras i den fortsatta utvecklingen till nyare versioner, vilket kräver ändringar och testning i alla skeden av ETL-verktyget.

8.3 Kommungränskollisioner i generalplaner

Korrigeringslogiken för kommungränskollision på generalplanerna fungerar bra, men inte fullständigt. Under VOOKA-pilotprojektet identifierades en tydlig situation där logiken ännu inte hade gett ett svar: när planindexet förblir "öppet" längs gränsen för ett annat planindex, vilket innebär att differensgeometrin som bildas i beräkningen inte filtreras bort genom arealjämförelsen (se bild 7).



Figur 7. Exempel på en situation där korrigeringen av kommungränskollisionen i generalplansindexet inte fungerar fullständigt. Vid beräkningen av maskeringen av kommungränsen och planindexets difference blir indexet på gränsen till det andra planindexet "öppet" i förhållande till maskeringen, varvid skillnaderna inte filtreras bort med en jämförelse av arealen. Det blåa området (planindexet) förenas inte med det gröna (fastighetskommungränsen), även om det borde.

Filtreringen av difference-geometrier som baserar sig på arealens storlek grundar sig på försök eftersom det finns många skillnader i de spatiala precisionerna i kommunmaterialet. För VOOKA-projektets del har varje kommuns fungerande gränsvärde testats separat och listats som referens för vidareutveckling. Även om kommunernas toleransvärden är enhetliga kan man inte blint lita på dem i olika geografiska områden, utan varje kommun ska hitta ett fungerande gränsvärde separat genom testning. Som idé till fortsatt utveckling av detta behov kunde man också utbilda AI/ML-funktioner för att bedöma det lämpligaste toleransvärdet för varje kommun.

8.4 Objektregisterbeteckningar från tomtuppgifterna

Lokaliseringsuppgifterna för fastighetsbeteckningarna finns förutom på poängnivå även på områdesnivå från LMV:s OGC API Features-gränssnitt som en del av tomternas omfattande lokaliseringsuppgifter. Att utnyttja den regionala referensnivån för att länka objektregisterbeteckningarna till planindexen är ett noggrannare sätt än att utnyttja poängnivån. I genomförandet av VOOKA-projektet kan en del av fastighetsbeteckningarna inte länkas till planindex som i verkligheten tangerar fastighetsbeteckningen, eftersom uppgifterna om fastighetsbeteckningens läge är bundna till endast en punkt i stället för till influensområdet.

Dessutom vill man i plandatamodellen för datasystemet för den byggda miljön i fråga om objektregisterbeteckningarna veta om de helt och hållet ingår i planen samt arealen för området som ingår i planen. Denna information kan inte fås endast genom att utnyttja poängnivån.

Obs! Under genomförandet i Norra Savolax framkom det att dessa egenskapsuppgifter inte längre behövs för den plandatamodell som uppdaterats efter pilotprojektet, därför förblir utvecklingen av denna ett frivilligt skede.

8.5 Jämförelse av fastighetstomt och planindex

ETL-verktyget som utvecklats inom VOOKA-projektet identifierar inte om fastighetsgränserna har bildats innan planen trätt i kraft eller efter det. Således kan den inte indikera om syftet med planindexets gräns är att sammanfalla med fastighetsgränsen eller inte.

8.6 Jämförelse av egenskapsuppgifter

I VOOKA-pilotprojektet är det möjligt att jämföra egenskapsuppgifterna endast för tre datafält (planslag, godkännandedatum och ikraftträdandedatum) på FDS-materialbasis. I andra geografiska områden kan egenskapsuppgifterna vara mer omfattande, varvid även andra datafält kan tas med i jämförelsen. Dessutom är det möjligt att utvidga den FDS-baserade jämförelseprincipen så att den blir kommunbaserad.

8.7 Uppföljning av mappsortering av material

Under projektet i Norra Savolax upptäckte man att kommunerna ofta uppdaterar materialet i delar i Sharepoint-tjänsten, vilket gör det utmanande att följa upp materialinsamlingen. För att följa upp materialmottagningen vore det nödvändigt att uppdatera mappsorteringen med ett verktyg som identifierar duplikat i det nyuppladdade materialet. Detta skulle möjliggöra uppföljning av mängden mottaget material.

8.8 Skapande och uppdatering av tabellen för PDF-länkkonvertering

För närvarande skapas länkningstabellen så att det i processen skapas en separat csv-fil från mappen vid varje körning, från vilken informationen överförs till en tabell i molnet. Under projektet diskuterades möjligheten att direkt uppdatera tabellen i molnet, men det största hindret för detta var datasäkerhetsfrågor. Programmet kan inte ges rättigheter att ändra filer som finns i ett företags moln.

Det skulle dock vara möjligt att skapa en lösning där informationen uppdateras lokalt på datorn i en csv-tabell. För detta ändamål vore det bra att utveckla programkoden så att man med hjälp av den kan uppdatera filen utan att de redan skapade värden försvinner. Dessutom bör en mer omfattande behandling av fel utvecklas för detta ändamål.

Denna fas skapades under projektet i Norra Savolax, så det finns sannolikt ett behov av att vidareutveckla fasens programkod och dess verksamhetslogik för att den ska fungera även för andra landskap.

8.9 Skapande av tabell för PDF-länkkonvertering - dokumenttyp

För att skapa en länktavla utvecklades ett verktyg som avgör filnamnets typ. För närvarande returnerar verktyget dokumenttypen som en bestämmelse, om filnamnet innehåller en ordlista som hänvisar till bestämmelsen. Om den aktuella ordlistan inte hittas ger verktyget en plankarta samt bestämmelser som standardvärde för dokumenttypen. Verktyget är för närvarande inte perfekt varför dokumenttypen måste manuellt verifieras för att säkerställa dess korrekthet. Verktyget ger för närvarande delvis felaktiga resultat, så för vidareutveckling skulle det vara nödvändigt att skapa ett säkrare sätt att identifiera dokumenttypen. I praktiken innebär detta att det är nödvändigt att använda till exempel Python Difflib-modulen eller någon annan motsvarande metod.

8.10 Identifiering av inre plangränser med hjälp av maskininlärningsmetoden

I fas 3.2.1 av ETL-processen begränsades kommunernas plangränser till de faktiska kommungränserna. Testningen av de toleransvärden som används vid automatiseringen av processen genomfördes genom försök, men eftersom antalet planer eller delar av planen utanför kommunens yttre gränser som förekommer i Norra Savolax material var mycket få, följde planernas yttre gränser mycket väl kommunens gällande yttre gränser. Samma logik kunde användas för exempelvis att harmonisera detaljplanernas gränser med gränserna för intilliggande tomter. Vid automatiseringen av denna process borde man dock utnyttja maskininlärningsmetoder som kan läsa gränserna för inre planer och söka de närmaste eller lämpligaste punkterna eller linjerna som kommunens plangränser jämställs med. I maskininlärningsmetoden kunde man också överväga att använda jämförelsematerial såsom Lantmäteriets fastighetsgränser, om de är konsistenta med till exempel detaljplanernas yttre gränser.