

**МОНГОЛ УЛСЫН ИХ СУРГУУЛЬ**

**ХЭРЭГЛЭЭНИЙ ШИНЖЛЭХ УХААН ИНЖЕНЕРЧЛЭЛИЙН СУРГУУЛЬ**

**МЭДЭЭЛЭЛ КОМПЬЮТЕРИЙН УХААНЫ ТЭНХИМ**

**СИСТЕМ ИНТЕГРЭЙШН, АРХИТЕКТУР**

Даалгавар №8 OpenRefine ашиглах

Шалгасан: ..................................................................................../ З.Цолмон /

Гүйцэтгэсэн: ................................................................................/У.Болормаа/

2023.04.21

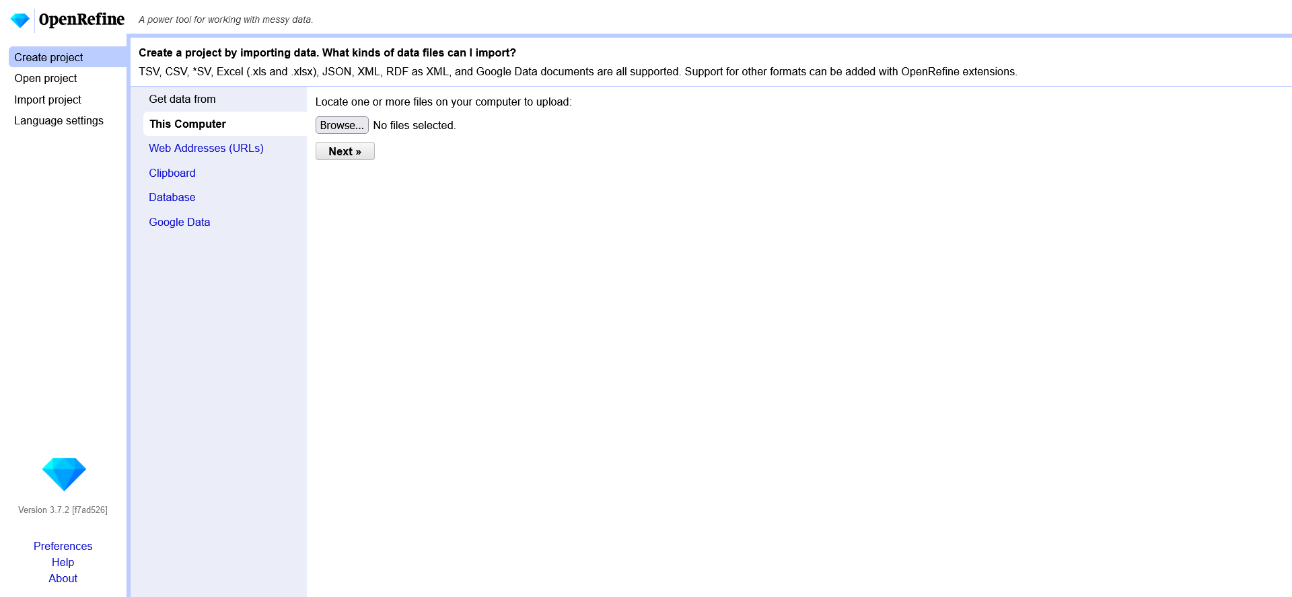
Агуулга

1. OpenRefine
2. Data validation
3. Muliple representation
4. Duplicated records detection

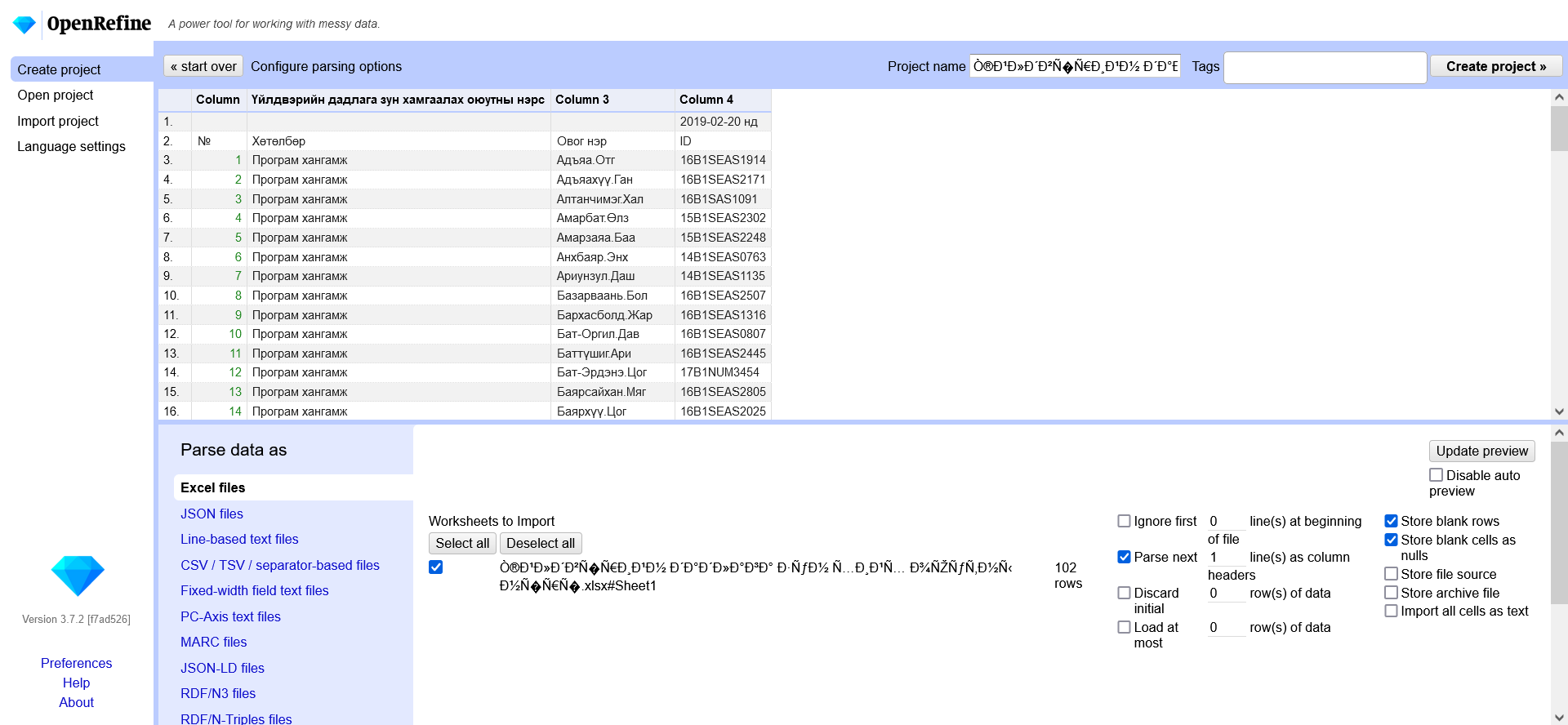
1. **OpenRefine** өмнө нь Google Refine гэж нэрлэдэг байсан ба энэ нь хэрэглэгчдэд том өгөгдлийн багцыг судлах, цэвэрлэх, удирдахад туслах зорилготой үнэгүй, нээлттэй эх сурвалжтай өгөгдөл цэвэрлэх, өөрчлөх хэрэгсэл юм. **OpenRefine** нь хэрэглэгчдэд шүүлтүүр, хуваах, нэгдэх, хувиргах зэрэг янз бүрийн үйлдлүүдийг гүйцэтгэх замаар текст болон тоон өгөгдөл зэрэг эмх замбараагүй өгөгдлийг хялбархан цэвэрлэж, цэгцлэх боломжийг олгодог. Үүнийг хэргэлснээр бид өөр өөр эх сурвалж дахь өгөгдлийг нэгтгэх, тааруулах, өгөгдлийн алдаа, зөрчлийг илрүүлэх, засах боломжтой байдаг. OpenRefine нь хэрэглэгчдэд ээлтэй интерфэйс, өргөн хүрээний өгөгдөл хувиргах функцүүдийн ачаар том, нарийн төвөгтэй өгөгдлийн багцтай ажилладаг өгөгдлийн шинжээч, өгөгдөл судлаач, судлаачдын түгээмэл хэрэгсэл гэж хэлж болно.

OpenRefine нь CSV, TSV, Excel, JSON, XML, RDF гэх мэт өргөн хүрээний өгөгдлийн форматтай ажиллахад зориулагдсан. Энэ нь мөн олон төрлийн програмчлалын хэлийг ашиглан давтагдах даалгавруудыг автоматжуулах, өөрчлөн өгөгдлийн хувиргалт хийх боломжийг хэрэглэгчдэд олгодог хүчирхэг скрипт бичих чадварыг өгдөг.. Энэ нь ялангуяа вэбээс хасагдсан өгөгдөл, дижитал түүхэн баримт бичиг, олон эх сурвалжаас авсан өгөгдөл зэрэг эмх замбараагүй, нарийн төвөгтэй өгөгдөлтэй ажиллахад тохиромжтой.

**Openrefine.exe нээх үед:**

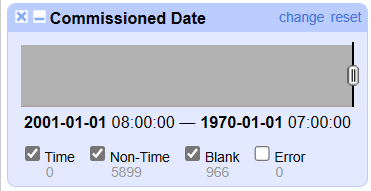


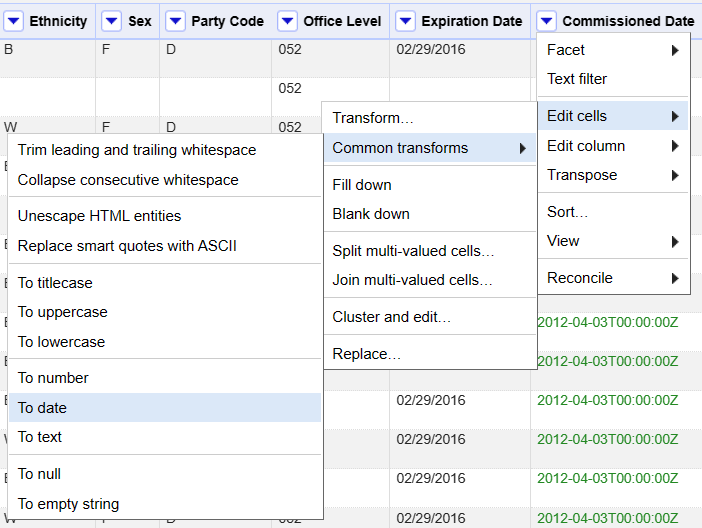
Файлаа оруулах:

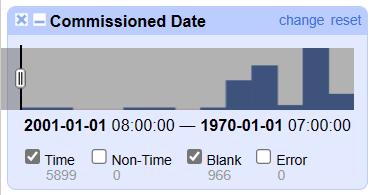


**2. Data validation**

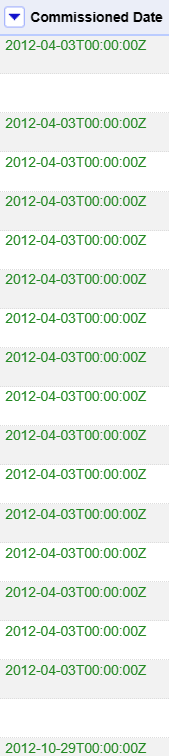
Өгөгдлүүд нэг форматад байгаа эсхийг шалгаж зөв болго. - Огнооны өгөгдлүүд дээр ажилла

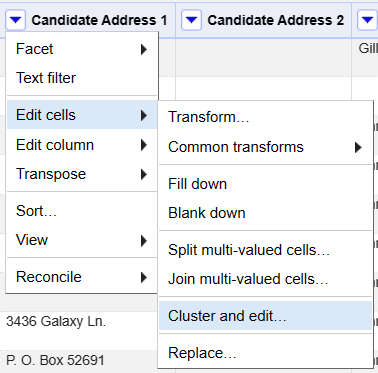


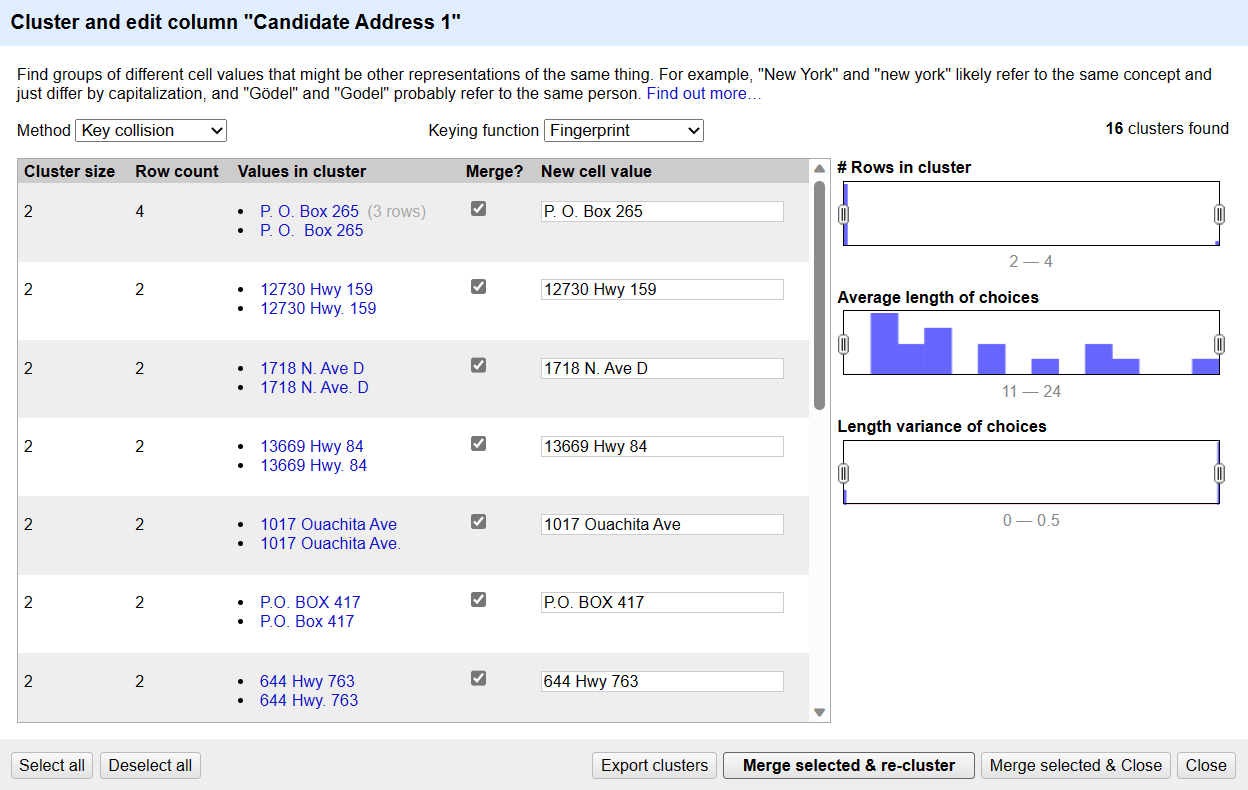




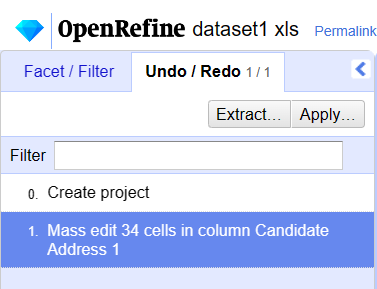
Зассан дата:









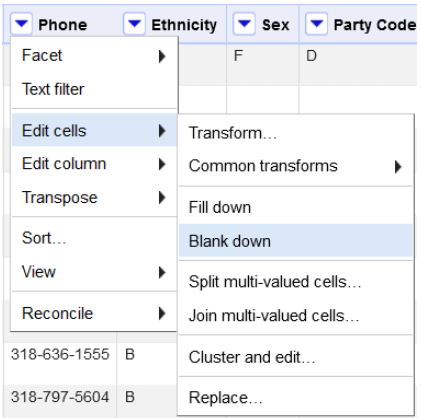


1. **Duplicated records detection**

Давхардсан утгатай өгөгдлүүдийг олж устгах

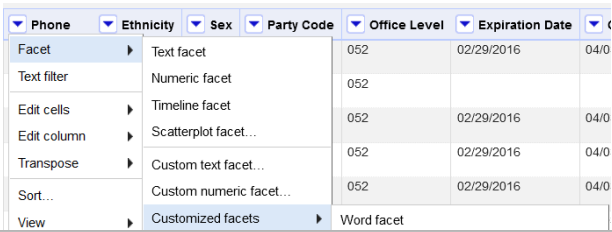
“Edit cells” сонгоод дараа нь “Blank down.” сонгоно.

"Blank down" нь хоёр мөр ижил утгатай эсэхийг илрүүлэх болно. Хэрвээ ижил байвал хоёр дахь мөрийг "хоосон" болгож, нүдний утгуудыг арилгах болно.



"Blank down" ашигласны дараа “Facet by blank” ашиглаж тухайн баганын хоосон нүдний утгууд бүхий мөрүүдийг тодорхойлох юм.

Фасет цонхноос "true" сонголтыг сонгоно.



Үүний дараа “Remove all matching rows.” ашигласнаао давхардсан бүх мөрийг устгах болно

