openRGD

RIKSTÄCKANDE GENEALOGISK DATABAS

Sammanslagen släktforskning

Sveriges historiska befolkning med släktrelationer

2015-04-30

delprojekt openRGD

Resultat

RIKSTÄCKANDE GENEALOGISK DATABAS (RGD)
är ett samarbetsprojekt mellan DIS (www.dis.se) och
EIT vid Lunds universitet (www.eit.lth.se)

Övergripande målsättning med en Rikstäckande Genealogisk Databas (RGD) är att skapa en databas över Sveriges historiska befolkning med släktrelationer och utan dubbletter – baserat på släktforskares samlade forskningsresultat.

RGD:s SYFTEN

- Stödja medlemmars släktforskning genom feedback
- Förbättra tillförlitlighetsnivån i släktrelationer och sakuppgifter genom sammanläggning av flera forskares uppgifter
- Att många forskare bidrar och granskar innehållet för att höja kvaliteten
- Utgöra en grund för en databas över Sveriges historiska befolkning med unika identitetsbegrepp
- Kunna länkas med annan forskning avseende svenskarnas liv och verksamhet

openRGD DELPROJEKT

openRGD är ett delprojektet inom RGD som avser att väsentligt bredda och förenkla möjligheterna för allmänheten att använda de verktyg som tas fram i huvudprojektet RGD genom att:

- Att utveckla och utvärdera en algoritm för fuzzy matchning (identifikation av gemensamma individer och familjer) mellan två släktträd. Fokus för algoritmen är noggrannhet och effektivitet.
- Utveckla fria Web-tjänster för access till RGD's verktyg.

openRGD är ett samarbete mellan Föreningen för datorhjälp i släktforskningen (DIS) och Lunds universitet (Institutionen för Elektro- och Informationsteknik, EIT) och har stötts med ett finansiellt bidrag från Interfonden .SE under tiden 2014-06-01 – 2015-04-30. Utöver personella resurser bidrar EIT med sin kunskap inom informationsteknik vad gäller algoritmer och databasstruktur.

Funktionerna i openRGD-Web kan användas för att höja tillförlitligheten i relationer och sakuppgifter i den egna forskningen dels genom vissa formella kontroller av den ingående GEDCOM-filen och dels vid jämförelse av samma fil mot en annan GEDCOM-fil. Det är ambitionen att släktforskaren skall kunna höja kvaliteten i den egna forskningen. openRGD's Web-tjänster syftar till att med "självbetjäning" kunna göra ovanstående kvalitetsgranskning (indatavalidering) samt att i två GEDCOM-filer kunna identifiera matchande individer men även identifiera avvikelser i form av släktrelationer eller sakuppgifter för dessa individer.

Eftersom databasen ej skall få innehålla dubblettposter på individer ställs mycket höga krav på identifiering av potentiellt matchande individer. RGD jämför hela familjebilder (ej enstaka individer) i en indatafil med motsvarande familjebilder i databasen för att i största möjlighet säkerställa en korrekt identifiering av gemensamma individer. Mer än 99.8 % av alla gemensamma individer upptäcks.

FORTSÄTTNING / VIDAREUTVECKLING

Resultatet av arbetet som omfattades av Internetfondens stöd är mycket värdefullt för det fortsatta arbetet med RGD som grovt kan delas in i tre delar:

1. att granska kvaliteten i ett bidrag och föreslå åtgärder/förbättringar
2. fuzzy matchning och uppdatering av en RGD-databas med ”sammanslagna” data
3. hur ska vi arbeta med RGD?

För den första delen kommer vi att arbeta vidare med att realisera och underhålla databaser med personnamn, källor och orter. Vid en första anblick kan det verka trivialt men rymmer en del komplexitet, t.ex. i form av att bygga en ortsstruktur som starkt påminner om ”Ingår i / Består av” för produkter och artiklar. Strukturen ska även inkludera en tidsdimension. Vidare behöver vi hitta en bra balans mellan de automatiserade och de manuella inslagen och skapa former för feedback till bidragsgivaren.

För den andra delen bygger vi vidare på erfarenheterna och ser hur det går att trimma matchningen ytterligare. Sammanslagningen av personer, familjer och relationer kräver också en bra balans mellan de automatiserade och de manuella inslagen samt en värdering av olika källor. Vi ser framför oss att våra medlemmar kommer att hålla ett öga på innehållet så att felaktigheter och förbättringar rapporteras regelbundet. En variant på temat ”ständiga förbättringar”, kanske inte lika öppet för egna åtgärder som Wikipedia men ändå samma grundtanke.

För den tredje delen kommer vi att ta fram processer, rutiner, rollbeskrivningar och liknande som krävs.

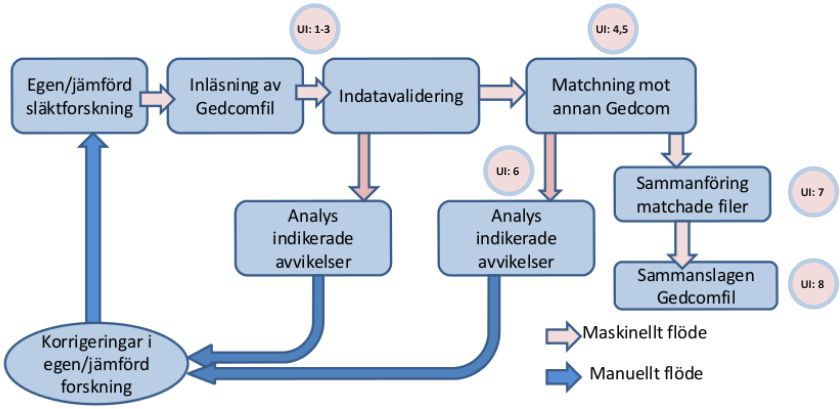
Tidplan: eftersom vi i stort sett bygger på enbart ideella krafter så måste det få ta den tid som krävs och i den takt vi finner rätt person för rätt uppgift.

ÖPPET TEST-SYSTEM - openRGD

Resultatet av Internetfond's projektet är ett antal öppna Web-tjänster som kan användas fritt av alla. Antingen med eget konto eller inloggad som guest. Enda skillnaden är att databaser sparas mellan sessioner för användare med konto.

Implementeringen av openRGD bygger på tekniker och verktyg som Machine Learning, Information Retrieval, Python, MongoDB, Lucene, och SVM.

Koden är fritt tillgänglig och kan laddas ner från GitHub.*



Figur 1: Funktionsöversikt/dataflöde openRGD.

* Mer information, URL:er, m.m. hittar du på sista sidan.

Indatavalidering / egenkontroll av GEDCOM-filer (UI:1-3)

- Namnkontroll, listar formella fel och möjliga felregistreringar av kön genom kontroll av alla namn i GEDCOM-filen mot DIS namndatabas (som är under utveckling). Programmet listar namn som saknas med personens angivna kön men som finns med motsatt kön. Dessa personer kan vara registrerade med fel kön.
- Ortkontroll, listar angivna platser som inte är svensk församling eller land. Programmet läser den angivna orten för händelserna född, död och vigd. Därefter jämför programmet detta mot generella tabeller för svenska församlingar och länder. De orter som då inte får träff listas tillsammans med ett urval av möjliga alternativa församlingar. Även länder kontrolleras mot en tabell med svensk stavning enligt ISO-standard.
- Dubblettkontroll, listar möjliga dubblettkandidater. Kontrollen avser att finna lika eller snarlika personer, som möjligen kan vara dubbletter. Programmet jämför alla personer i GEDCOM filen med varandra. Försöker att hitta och värdera likheter i individernas uppgifter. De personer, som har liknande uppgifter listas parvis.

Slutligen skapas en databas i openRGD som används för fortsatt bearbetning.

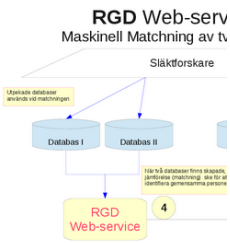
Matchning av två GEDCOM filer (UI:4, 5)

4. Maskinell Matchning - Läs mera

Matchning av två databaser innebär att programmet försöker att identifiera personer, som finns i båda databaserna.

Databas Välj databas I att matchas mot Välj databas II

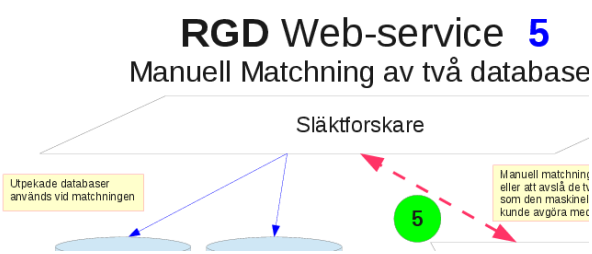
Matcha!



5. Manuell Matchning - Läs mera

Den maskinella matchningen kan inte med tillräcklig säkerhet matcha alla personer mot varandra, utan lämnar tveksamma matchningar till manuell bedömning.

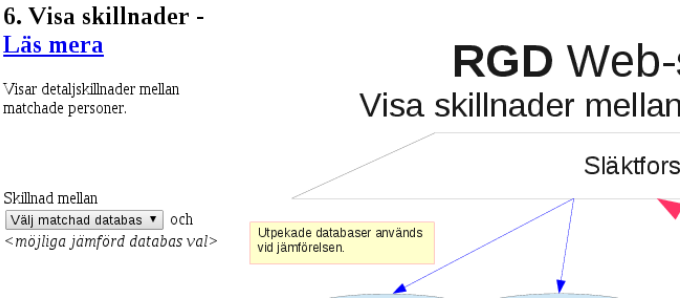
Detaljerad dokumentation (visas i nytt fönster): Matchinfo.pdf



Släktforskare, med viss del av forskningen gemensam, kan analysera likheter/skillnader genom att maskinellt matcha en databas mot någon av de andra databaser man har laddat upp. Här har man då möjlighet att familjevis hitta likheter och skillnader mellan de två maskinellt matchade databaserna.

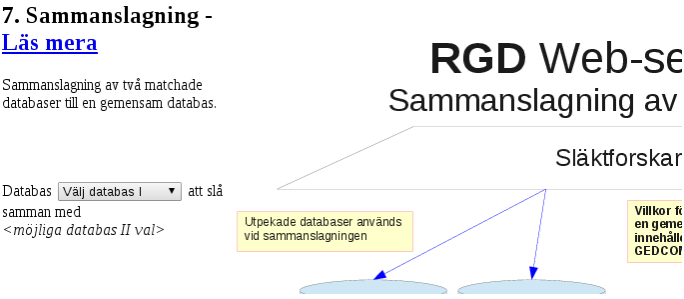
Den maskinella matchningen kan inte med tillräcklig säkerhet matcha alla personer mot varandra, utan lämnar tveksamma matchningar till manuell bedömning.

Avvikelser i matchad data (UI:6)



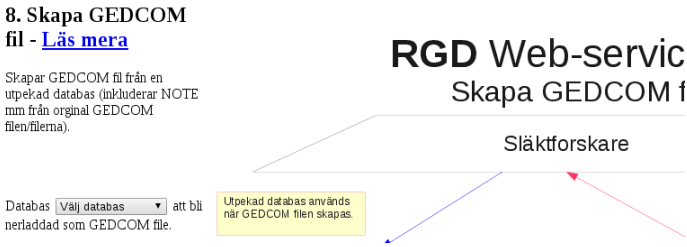
Forskare som hittat gemensamma anor kan hitta detaljskillnader hos matchade personer. Matchade databaser innehåller troligen skillnader i data för enskilda matchade personer. Skillnader på så sätt att data finns i båda, men där uppgifterna avviker från varandra. Personer där differenser finns listas och skillnaderna synliggörs.

Sammanslagning av matchad släktforskningsdata (UI:7)



Efter att manuell matchning (UI: 5) har genomförts, kan två databaser slås samman enligt resultatet från matchningen, personer som identifierats i båda databaserna läggs samman till en person. Resultatet blir en ny databas med all information från de två sammanslagna fast utan dubbletter. Sammanslagning kan bl.a. användas om man fått GEDCOM data från annan släktforskare och vill inkluderas den i sin egen.

Skapa GEDCOM-fil av sammanslagen släktforskningsdata (UI:8)



Den databas, som skapats och bearbetats från de ursprungliga GEDCOM filerna, kan användas för att generera en ny GEDCOM fil.

USE CASES

Kontrollera sina egna data

- Ladda upp GEDCOM-fil (UI:1-3)
- Gå igenom resultat-listorna
  - RGDN.txt - Namnfel eller namn som saknas i namndatabasen, men finns med avvikande kön
  - RGDO.txt - Ortnamn / Platser som ej kunnat identifieras som församlingar i GEDCOM filen
  - RGDD.txt - Individer med lika eller snarlika uppgifter, som bör kontrolleras avseende dubblett eller felregistrering.
  - Checklista - Andra fel och varningar
- Eventuellt kör alternativ dublettkontroll (UI:2A)

  Välj databas I som matchar den GEDCOM-fil som just laddades upp

USE CASES

Jämföra sin egen forskning (A) med en kollegas forskning (B)

- Ladda upp GEDCOM-fil A via (UI:1-3)
- Ladda upp GEDCOM-fil B via (UI:1-3)
- Kör maskinell matchning (UI:4)

  Välj den minsta av A och B som databas I och den andra som databas II
- Starta manuell matchning (UI:5)

  Välj databas I och II på samma sätt som ovan

  Jämför data för matchade individer
- Alternativt visa skillnaderna (UI:6)

  Välj databas I och II på samma sätt som ovan

Slå samman två personers forskning (A, B) till en GEDCOM-fil

- Ladda upp GEDCOM-fil A via (UI:1-3)
- Ladda upp GEDCOM-fil B via (UI:1-3)
- Kör maskinell matchning (UI:4)

  Välj den minsta av A och B som databas I och den andra som databas II
- Starta manuell matchning (UI:5)

  Välj databas I och II på samma sätt som ovan

  Gå igenom alla 'MultiMatch' respektive alla med status 'Manuell' och avgör vad som är korrekt.
- Slå samman databaserna (UI:7)

  Välj databas I och II på samma sätt som ovan
- Skapa en GEDCOM-fil av den sammanslagna databasen (UI:8)

  Välj databas som databas II ovan

Trimma din släktforskning

- Se artikel i Diskulogen med Släktforskarnytt nr 108, 2015-04, Rolf Carlsson: sid 36 -- 37 "Släkttrim Trimma din släktforskning", och sid 38 "Borås Släktforskare testar Släkttrim".

  http://www.dis.se/sv/publikationer/diskulogen/senaste-numren.html

KONTAKT

Christer Gustavsson (DIS)
christer.gustavsson@dis.se

Anders Ardö (EIT, Lunds universitet)
anders.ardo@eit.lth.se
www.eit.lth.se/staff/Anders.Ardo

INFORMATION

DIS
www.dis.se/sv/projekt/genealogisk-databas.html

Internetfonden
www.internetfonden.se/rikstackande-genealogisk-databas

Testsystem
http://rgd.eit.lth.se:8085/ (fram till 2015-06-01)
http://rgd.dis.se/

Programkod
https://github.com/andersardo/gedMerge.git

Diskussioner på DisForum
http://forum.dis.se/vb/forumdisplay.php?121-Rikst%C3%A4ckande-Genealogisk-Databas-(RGD)

DIS

LUNDS UNIVERSITET
Lunds Tekniska Högskola

FINANSIELLT STÖD FRÅN
.se | internetfonden