BiographyNED
Project meeting: 28-October-2010
Verslag: Piek Vossen

- Serge zal een update maken van zijn document over de cases naar aanleiding van input van Els en Susan. Een belangrijk aspect zal zijn eventuele kritiek op de selectie van personen in het portaal. In bijna alle cases is genealogische informatie belangrijk. Case 5 (Gouverneurs) lijkt het beste geschikt voor een pilot. Het gaat om een overzichtelijke set individuen die handmatig in kaart gebracht kunnen worden en die we later kunnen uitbreiden, ook met genealogische gegevens.
- Antske heeft een inventarisatie gemaakt van de bronnen: welke velden en waarden komen hoe vaak voor. Data sparseness is de regel. De vraag is of alle entries ook 'belangrijk' zijn. Veel personen zijn geimporteerd maar niet geannoteerd. Wellicht kan Jelle aangeven waarom bepaalde velden niet voldoende gevuld zijn. Antske gaat een aantal vragen verzamelen die we voor de volgende bijeenkomst naar Jelle sturen en in de Skype meeting kunnen doornemen.
- Identiteit van personen gaat een belangrijke functie worden. Antske heeft een aantal gevallen gevonden waarbij verschillende records naar dezelfde persoon zouden kunnen verwijzen en andersom. Dit gaat nog belangrijker worden als we 'mentions' van personen in de teksten zelf gaan proberen te herleiden tot unieke ids en als we andere bronnen gaan koppelen. We verwachten dat we een functie kunnen ontwikkelen die voor iedere persoon kan bepalen in wat voor mate die identiek is aan ieder andere persoon. Daarbij gaan bepaalde features (naam, geboorte- en sterftedatum, geboorte- en sterfteplaats) harder zijn dan andere features. Verder kunnen we de betrouwbaarheid van de bronnen meenemen bij de weging: verschillen in informatie van 'goede bronnen' zal zwaarder tellen. Een identificatiefunctie kan harde en zachte eigenschappen meenemen. Zachte eigenschappen (bijv. alle tekstuele data rond een persoon als een zak van woorden) zullen een grotere rol spelen bij niet belangrijke personen en bronnen buiten het portaal (bijv. Wikipedia). Niet beroemde mensen kunnen toch belangrijk zijn voor bijv. netwerk analyses: twee beroemde personen kunnen met elkaar verbonden worden a.h.v. een onbekend iemand.
- We hebben een onderscheid gemaakt in drie verschillende lagen aan data sets:

    1. Records in het portaal
    2. Andere bronnen bijv. Wikipedia, PDC (in zoverre niet in portaal)
    3. Een formele database in RDF die we in het project definiëren

    In het project gaan we niet het Biografisch portaal of Wikipedia wijzigen maar extraheren we gestructureerde en formeel interpreteerbare gegevens uit de bronnen 1) en 2) die we stoppen in 3). Met formeel interpreteerbaar bedoelen we dat we kunnen redeneren met de data in 3) en daarop escience tools kunnen ontwikkelen. De extractie van 1) en 2) naar 3) moet automatisch (herhaalbaar) en leiden tot gelinkte data. Iedere veld binnen ieder record in 1) en 2) dat gebruikt wordt om een veld en/of record te maken in 3) blijft gelinkt als de bron voor de target. Indien de identiteitsfunctie een te hoge score geeft voor twee records in 1) of tussen een record 1) en een record in 2) dan wordt er een record in 3) gemaakt dat wordt gelinkt aan meerdere records in de bronnen. Uiteindelijk krijgen we many-to-many mappings tussen de databases. Dit laat iedereen vrij om vele gedistribueerde data te wijzigen binnen eigen redacties terwijl we voortduren die data kunnen linken aan de BiographyNED database in RDF. Verder kan iedereen die dat wil gebruik maken van de data in de BiographyNED RDF database om zijn/haar eigen database aan te passsen.
- Els gaat ???? uitnodigen om een lezing te geven over ????
- Else geeft aan Antske door wat de ranking is voor de betrouwbaarheid van de bronnnen.