Michael Daw Balma - 2006520746

Review WarSampo Knowledge Graph: Finland in the Second World War as Linked Open Data

1. *Background*

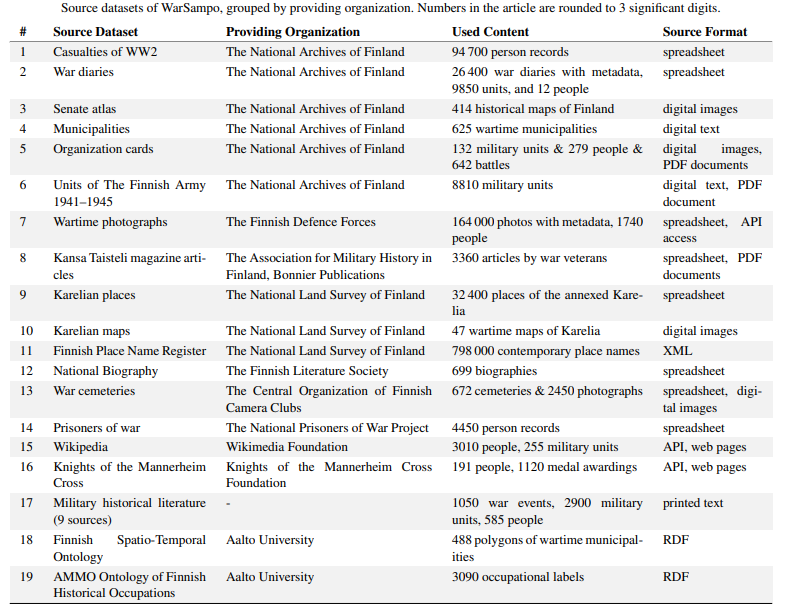
Informasi mengenai Perang Dunia ke-2 (WW2) tersebar dalam berbagai dokumen militer, pemerintahan maupun organisasi lainnya, sehingga informasi tersebut sulit untuk digunakan dalam pembuatan *global views* dari sejarah dan peristiwa perang. Adapun paper ini bertujuan untuk menunjukkan bagaimana *Linked Data* (LD) dapat menyelesaikan permasalahan tersebut. Data yang terdapat pada *source* akan di-*publish* sebagai *Linked Open Data* (LOD) pada SPARQL endpoint di dalam portal WarSampo yang telah dibuat sebelumnya, yang tercatat sudah mencapai 550.000 pengguna. Dataset description ini menjadi pelengkap pada publikasi lain yang berkaitan tentang WarSampo, yang lebih membahas secara detail mengenai KG serta proses dalam *maintain* data.

1. *Related Work*

Masalah umum yang dihadapi dalam LD untuk *digital humanities* dan *digital history* adalah proses penggabungan dan penggunaan berbagai macam *cultural heritage datasets* yang tersedia. Beberapa projek terkait menggunakan CIDOC *Conceptual Reference Model* (CRM) pada pemodelan datasetnya. Terdapat beberapa referensi pekerjaan sebelumnya, yang sudah dilakukan, yang mencakup *use case* dan *end-user perspective*, integrasi data menggunakan NEL (*Named Entity Linking*), *maintenance* dari *interlinked dataset*, serta pembuatan dan penggunaan bagian KG secara terpisah.

1. *Datasets*

Sumber data yang digunakan pada proyek ini berasal dari beberapa organisasi Finlandia. Beberapa dataset yang digunakan pada project WarSampo (serta pekerjaan yang terkait) juga digunakan. Dataset utama yang digunakan adalah Arsip Nasional, yang berisi mengenai detail dari setiap orang yang gugur dalam peperangan Finlandia pada WW2. Tujuan utama dari WarSampo adalah untuk menyatukan/menggabungkan kisah hidup pejuang dengan menggabungkan informasi melalui *data linking*. Oleh karena itu, unit militer beserta sejarahnya menjadi poin utama dalam dataset proyek ini, yang juga mengacu pada 1050 WW2 *events based on military historical literature*. Adapun dataset yang digunakan terdapat pada tabel berikut:

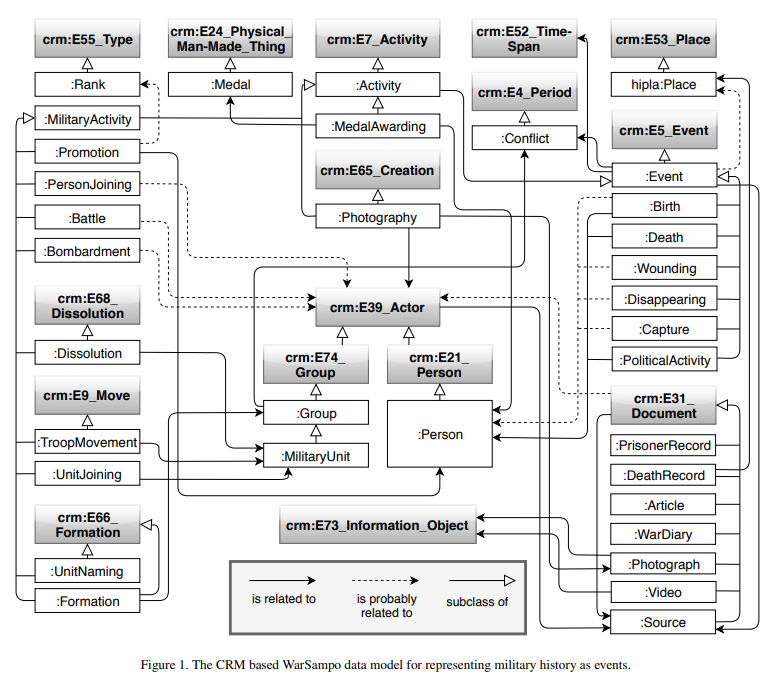


1. *Data Model*

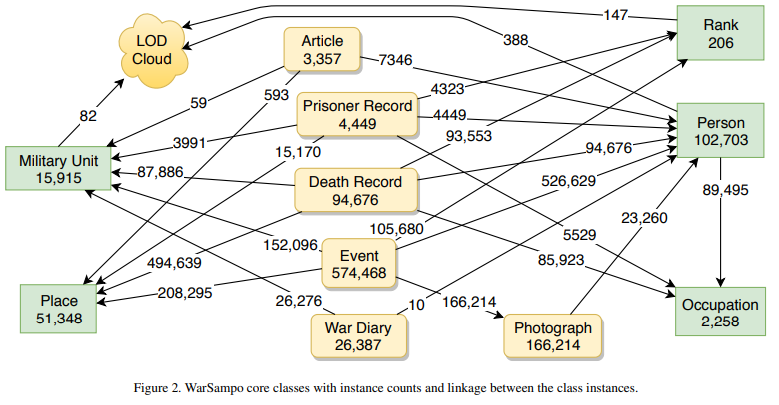
Dataset yang digunakan pada tabel diatas diubah dalam bentuk RDF dan disatukan dengan bentuk *event-based* data model. Proyek ini merepresentasikan perang sebagai sebuah ‘*events*’. Peneliti menggunakan CRM dengan tambahan konsep sejarah militer sebagai *conceptual framework*. Alasan penggunaan CRM sebagai model antara lain dikarenakan CRM cocok untuk menggabungkan sejarah peperangan, CRM merupakan ISO standard yang dokumentasi dan datanya tersedia dan dapat digunakan dengan mudah oleh orang/pihak lain, serta CRM dapat mendeskripsikan ‘*real world*’ lebih baik dibandingkan dokumen-dokumen yang tersedia, sehingga dapat digunakan untuk menggabungkan berbagai macam dataset yang ada untuk menyatukan representasi dari perang dan hal-hal yang terkait.

Penggunaan relasi dari *class* dan *subclass* CRM di-desain dengan membagi menjadi dua kategori utama berdasarkan *certainty*, yaitu relasi yang dapat diambil dari informasi dataset, serta relasi yang didapatkan melalui *entity linking methods*. Terdapat beberapa informasi yang tidak dapat dimodelkan menggunakan CRM, seperti status pernikahan, yang mana tidak termasuk dalam *grouping dan events*. Pada hal tersebut, informasi akan dituliskan dalam bentuk *simple SKOS vocabularies*.

Adapun terdapat bentuk CRM model yang digunakan sebagai berikut:



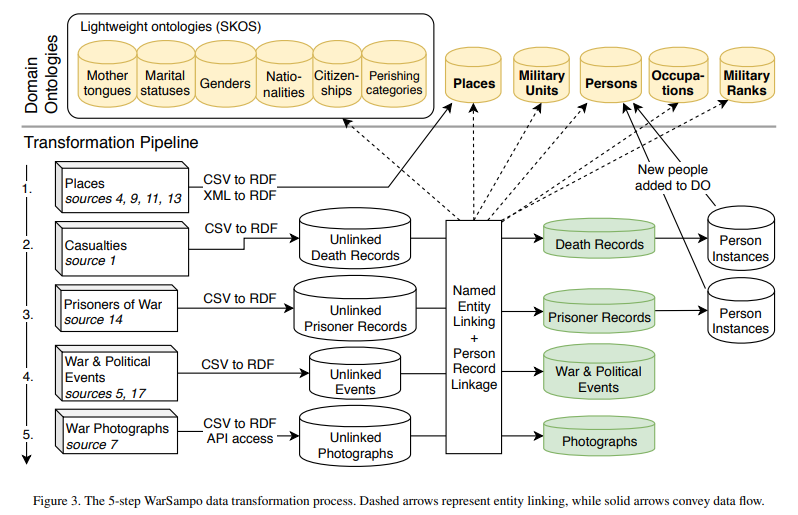
Berdasarkan *Core Class*, terdapat *instance counts* serta hubungannya terhadap *class* yang sudah didefinisikan, seperti pada gambar berikut:



1. *Data Model Population*

Pada awalnya, dibuat sebuah *shared* DO (Domain Ontologies) yang menggunakan proses secara manual, yang mana proses yang digunakan tidak dapat digunakan ulang. DO tersebut di-*maintain* secara langsung dalam bentuk RDF serta dibuat sebuah *repeatable data transformation pipeline*. Untuk menggabungkan *view* dari peristiwa perang, perlu dilakukan penyesuaian terhadap shared-DO. Penyesuaian tersebut berdasarkan pada *probabilistic NEL*, yang mana apabila terdapat error / *missing-link* yang kecil, maka tidak dianggap sebagai suatu masalah. Secara umum, proses populasi lebih berfokus kepada *recall* dibandingkan *precision*. Dengan demikian, proyek ini juga *open-feedback* terhadap user apabila user menemukan kesalahan pada suatu *link*. Pada beberapa kasus, seperti ambiguitas referensi terhadap seseorang, peneliti berusaha untuk memaksimalkan *recall* dan *precision*.

Proyek ini juga menggunakan *Transformation Pipeline* sebagai pipeline apabila dataset awal dilakukan pembaruan, maka KG yang terbentuk tidak akan menyebabkan *invalid data*, dikarenakan pipeline sudah melakukan *handle* dengan membuat ulang KG, serta melakukan perubahan *propagation*  yang disebabkan oleh MD/DO. Terdapat beberapa cara yang digunakan penulis apabila terdapat perubahan *propagation scenarios* saat pembaruan dataset dan DO. Berikut adalah alur yang digunakan dalam *data transformation pipeline*, serta *links* yang terbuat dengan menghubungkan ke DO, dengan catatan simbol kotak menggambarkan *structured data*, sedangkan silinder menggambarkan RDF data, serta warna kuning merepresentasikan DO sedangkan warna hijau merepresentasikan MD.



1. *Data Quality and Stability*

Data yang digunakan pada WarSampo KG berasal dari data yang dianggap *high quality*, dikarenakan kebanyakan data dataset berasal dari *national authorities*. Data tersebut tidak diubah maupun diganti, hanya dilakukan konversi menjadi RDF dan *linked* data. KG yang dibuat juga mendapatkan 5th *star level of the 5-star LD publishing principles*. Kualitas dari *vocabulary* yang digunakan mencapai *4th star level of the five stars of vocabulary use*, yang berarti metadata schema dari WarSampo dapat dianotasi menggunakan SCT, SKOS, dan OWL *vocabularies.* Sedangkan pada *entity linking*, NEL dari ‘*event description*’ pada DO pada ‘*people*’, *‘military* units’, *‘place’* memiliki F1 score 0,88. 1, dan 0.88 secara berurutan. Sedangkan NEL pada photograph metadata dari DO ‘*people*’, *‘military* units’, *‘place’* memiliki F1 score 0,8, 1, dan 0,77 secara berurutan.

*Linked Data* pada WarSampo dengan sumber eksternal juga dapat digunakan ulang. WarSampo terhubung dengan National Finnish Ontology Infrastructure, dengan total 6110 *links*.

KG yang terbentuk sudah cukup static, dengan beberapa perbaikan kecil yang diprediksi perlu dilakukan koreksi. DO yang baru dapat ditambahkan kedalam *ontology infrastructure* yang sudah ada, tanpa merubah DO yang sudah ada, dikarenakan DO berdiri sebagai komponen yang terpisah, yang mana di-*handle* juga secara terpisah oleh *transformation pipeline*.

1. Simpulan

Knowledge Graph yang terbentuk pada proyek ini sudah baik dalam membuktikan penggunaan *linked data* dalam menyelesaikan permasalahan mengenai informasi suatu peristiwa perang, yang tersebar dalam berbagai dataset. Hal ini dibuktikan pada bagian Data Quality dan Stability, yang sudah mencapai rating star yang cukup tinggi. KG ini juga dapat melakukan *queries* yang sebelumnya tidak dapat dilakukan karena alasan *relation* antar data. Walaupun demikian, terdapat beberapa kekurangan seperti penggunaan *event-based model* untuk melakukan *modeling* keseluruhan data, dimana terlalu kompleks, sehingga dapat berujung pada *slow queries*. Selain itu, penggunaan CRM menyebabkan kesusahan dalam melakukan edit data secara langsung, dikarenakan karakter CRM yang *verbose* serta *interlinking* antar dataset source yang tinggi. Penggunaan *transformation pipeline* juga belum menyelesaikan seluruh masalah yang ada, seperti contohnya, DO memerlukan update secara manual pada suatu kasus tertentu. Ambiguitas pada *entity linking* juga merupakan salah satu tantangan dalam menyelesaikan proyek tersebut.