**Review WarSampo Knowledge Graph**

Judah Ariesaka Magaini – 2006463042

**Latar Belakang**

WarSampo *Knowledge Graph* (KG) dirancang untuk memfasilitasi penggunaan data Perang Dunia II (WW2) dalam analisis dan aplikasi berbasis komputasi. Sebelumnya, data sejarah WW2 tersedia dalam berbagai format yang tersebar di berbagai sumber seperti buku, artikel, dan lainnya. Ketidakseragaman data sejarah ini membuatnya sulit untuk dipelajari dan diproses untuk penelitian humaniora digital. Dengan pembuatan WarSampo sebagai salah satu *Linked Data* (LD), masalah ini dapat diatasi. Tercatat bahwa sebanyak lebih dari 550 ribu pengguna telah memanfaatkan LD ini melalui portal yang dibangun di atas LD yang telah dibuat. Selain itu, tercatat juga terdapat beberapa aplikasi yang telah memanfaatkan data dari WarSampo KG.

***Related Works***

Pemanfaatan LD untuk mengintegrasikan data yang heterogen bukanlah konsep baru, terutama dalam bidang humaniora dan sejarah digital. Telah ada beberapa proyek sebelumnya yang serupa dengan WarSampo KG, yang menggunakan CIDOC *Conceptual Reference Model* (CRM). Bahkan, beberapa proyek tersebut telah dipublikasikan dan berkaitan dengan data sejarah Perang Dunia Pertama (WW1) dan WW2.

***Data Model***

Dalam proses pembangunan WarSampo KG, digunakan berbagai dataset dari sumber yang berbeda-beda, masing-masing dengan format dan isi yang beragam. Dataset-dataset ini kemudian digabungkan dan ditransformasi menjadi sebuah KG yang harmonis. Selama proses pembangunan, didefinisikan sebuah *data model* yang menggunakan kerangka konsep CRM, dilengkapi dengan beberapa ekstensi. *Data model* ini dapat divisualisasikan dalam gambar berikut.

A diagram of a program

Description automatically generated

Penggunaan CRM dalam WarSampo didasarkan pada pemahaman bahwa perang merupakan rangkaian dari berbagai kejadian (*event*), sehingga CRM cocok digunakan untuk membangun model berbasis *event*. Selain itu, CRM juga merupakan standar ISO, yang memudahkan proses pembuatan dan dokumentasi dengan menggunakan *tools* yang sudah ada, sehingga menghindari ‘*reinventing the wheel*’. CRM lebih berfokus pada deskripsi kejadian di dunia nyata daripada dokumen itu sendiri, sehingga dapat digunakan untuk mengintegrasikan data sejarah dari sumber yang heterogen. Selain CRM, digunakan skema atau *vocabulary models* seperti RDFS untuk merepresentasikan relasi antar kelas (*is related to*, *is probably related to*, dan *subclass*), SKOS untuk merepresentasikan relasi informasi yang tidak dapat direpresentasikan oleh CRM seperti status pernikahan, kewarganegaraan, dan lain-lain. SKOS juga digunakan untuk penamaan literal dengan properti *skos:prefLabel* dan *skos:altLabel*. DCT juga digunakan untuk merepresentasikan relasi sumber (*dct:source*) dan deskripsi tekstual (*dct:description*).

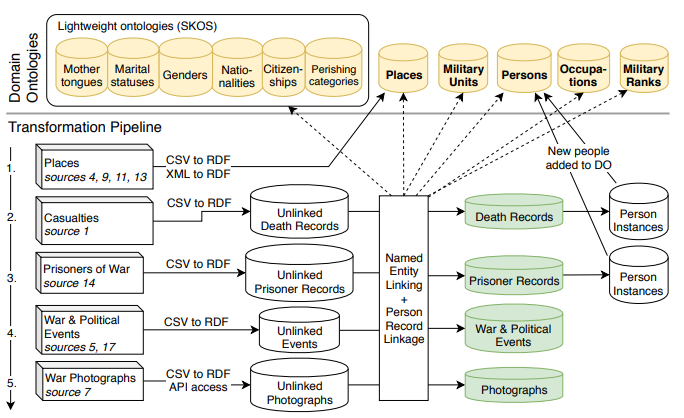
Secara umum, WarSampo KG memiliki kelas utama (*core classes*) dan metadata yang dapat divisualisasikan sebagai berikut. *Core classes* dalam *domain ontology* (DO) ditunjukkan dengan kotak warna hijau, sedangkan metadata (MDF) ditunjukkan dengan kotak warna kuning.

A diagram of a data flow

Description automatically generated with medium confidence

**Populasi Data**

Setelah mendefinisikan *data model* dan *core classes* beserta metadatanya, WarSampo KG dipopulasikan dengan menggunakan *transformation pipeline* yang *repeatable* dan efisien. Pipeline ini mengubah dataset sumber yang heterogen menjadi RDF dan menghubungkannya dengan *domain ontologies* (DO) yang dibuat sebelumnya. Hanya sebagian kecil data awal yang dipopulasikan secara manual, seperti 5610 orang, unit militer, pangkat, dan medali. Untuk sebagian besar data, penghubungan entitas dari berbagai sumber dilakukan dengan metode *probabilistic named entity linking* (NEL), yang menghasilkan sejumlah kecil *link* yang salah atau hilang. Gambaran umum dari *transformation pipeline* ditunjukkan pada gambar berikut.



Semua dataset akan diubah menjadi format CSV yang lebih terstruktur jika belum dalam format tersebut. CSV tersebut kemudian ditransformasikan menjadi format RDF, menghasilkan entitas yang belum terhubung (*unlinked*). Dengan menggunakan *probabilistic* NEL dan *record linkage* (berbasis *machine learning*), entitas-entitas tersebut secara otomatis dihubungkan dengan entitas lain yang telah ada dalam DO. Hasil dari proses populasi WarSampo ini menghasilkan sekitar 14 juta *triple* yang terdiri dari beberapa DO dan MDS.

**Evaluasi**

Dari segi kualitas data, WarSampo KG menunjukkan kualitas yang baik karena berasal dari berbagai sumber data yang kredibel dan tidak melakukan modifikasi pada isi data, hanya struktur datanya. WarSampo KG telah mencapai *5th star level* pada *LD publishing principles*. Namun, masih ditemukan beberapa kekurangan, seperti kesalahan semantik saat melakukan *validation tests* dengan SPARQL *queries*. Salah satu contohnya adalah adanya catatan individu yang menerima perawatan setelah mereka dinyatakan meninggal. Untuk kualitas *vocabulary use*, WarSampo KG telah mencapai *4th star level* karena telah menggunakan *vocabulary* lain seperti DCT, SKOS, dan OWL. Sedangkan untuk kualitas *entity linking*, WarSampo KG telah menunjukkan kualitas yang baik dengan hasil NEL yang menghubungkan deskripsi kejadian (event), foto, dan artikel majalah dengan orang, unit militer, dan tempat dengan skor F1 yang cukup baik, meski masih ada ruang untuk peningkatan pada beberapa hubungan. Kualitas *person record linkage* juga cukup baik berdasarkan skor presisi hasil *random sampling* dari keseluruhan *person record linkage*. Kualitas WarSampo KG juga didukung oleh penggunaannya sebagai sumber eksternal oleh pihak ketiga seperti KOKO dan YSA, yang merupakan infrastruktur ontologi milik pemerintah Finlandia, serta DBpedia, Wikidata, dan lain-lain untuk dihubungkan ke ontologi mereka.

**Masalah dan *Future Works***

Seperti yang telah disebutkan sebelumnya, WarSampo KG telah digunakan oleh berbagai aplikasi dan dihubungkan dengan ontologi lain yang kredibel, sehingga manfaat yang diberikan oleh WarSampo KG tidak perlu diragukan lagi. Namun, masih ada beberapa aspek yang perlu diperhatikan dan ditingkatkan untuk meningkatkan kualitas WarSampo KG. Salah satunya adalah permasalahan model berbasis *event* yang dapat meningkatkan kompleksitas dan memperlambat *query*. Selain itu, penggunaan CRM dalam *data model* dapat menyulitkan pemeliharaan data secara langsung, meski hal ini dapat diatasi dengan *data transformation pipeline* untuk sebagian besar kasus. Permasalahan umum lainnya, seperti disambiguasi entitas yang tidak memiliki informasi yang jelas, juga masih ada dalam WarSampo KG. Selain itu, masih ada kekurangan dataset tentang perang yang tersedia untuk mencakup semua peristiwa dan individu yang terlibat.