Schema, ontologies, archives and next generation IR problems

Dr. Robert Warren[1]

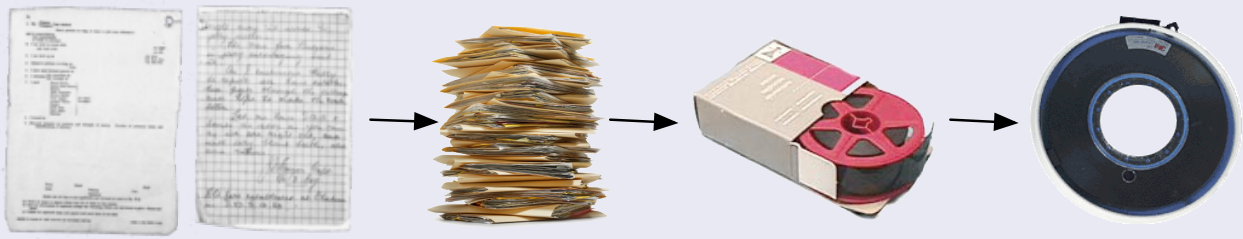
[1]rwarren@math.carleton.ca
Math and Statistics
Carleton University, Canada (now)
DRS School of Computer Science
University of Waterloo, Canada (previously)

Webis, Bauhaus Universität, Weimar

Muninn Project (Great War)

- Support for changes in organizations, people and relationships.
- Taxonomy, multilingual, multicultural.
- Record instances and classes of objects that don't exist anymore.
- Current bias is towards Canada and the British Empire.
- Less ontological engineering than design-by-exception.
- Talking to and integrating to library systems is misery.

The problem with archives



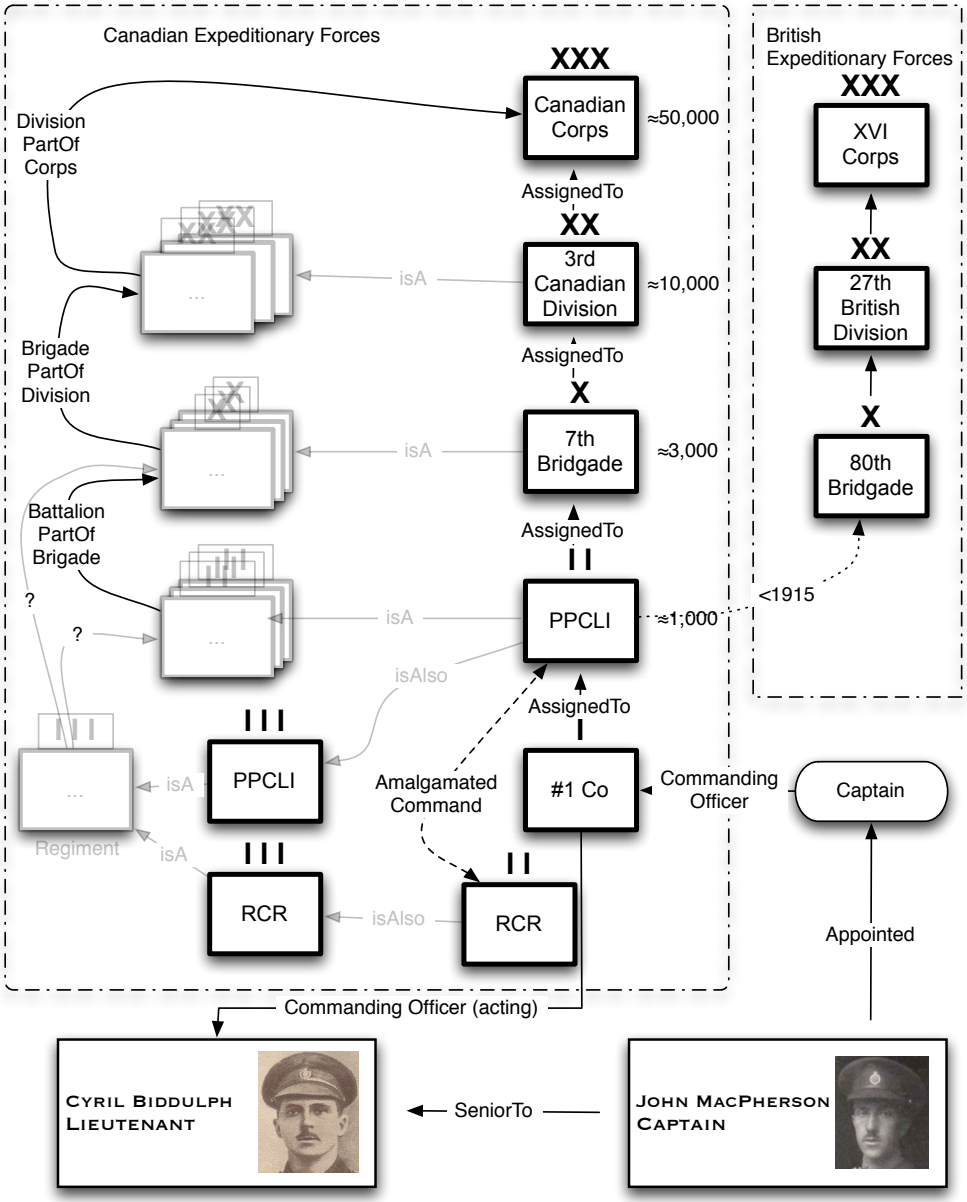
- “We’d love to give you the data but we can’t get it out of the computer system more than one page at a time.”
- “We’d love to give you the data but it’s on 300 unlabeled tapes.”
- “We’d love to give you the data but we don’t know where it is.”
- “We’d love to give you the data. Please fill out this licensing agreement.”
- “We’d love to give you the data but we can’t afford to store it so we’ll burn it.”

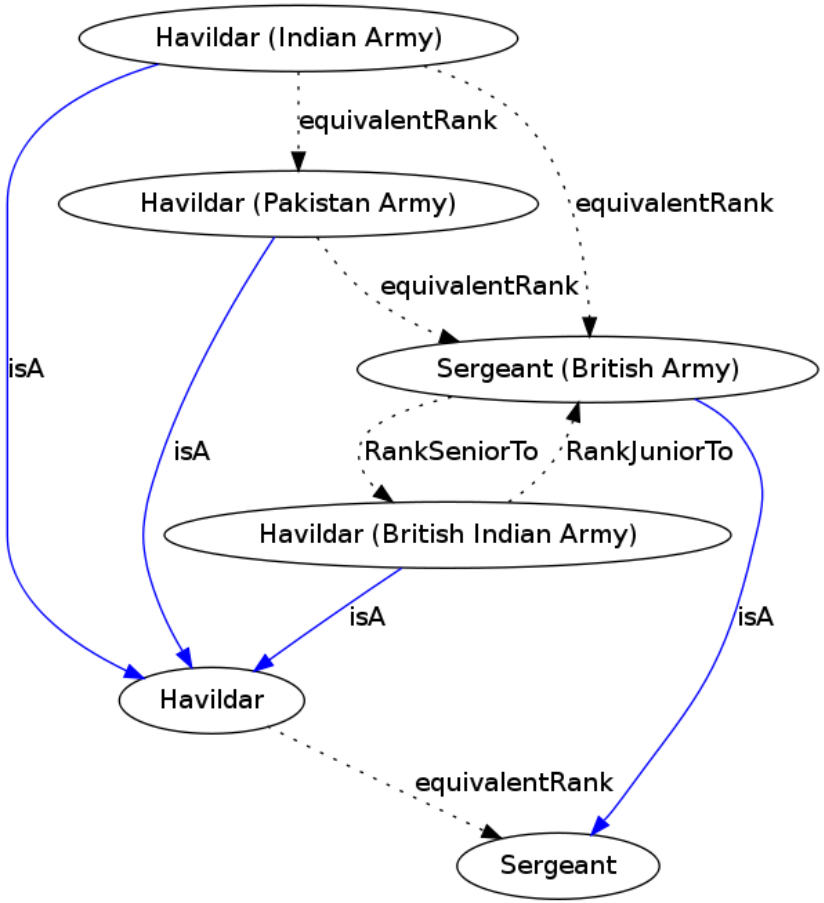
Previously in the search world...

- Faceted search didn't work out
- *Everything* is a free form query with *magic*(tm).

...This is what is going to happen

- The last problem was the data bloat, we now have the meta-data bloat.
- Meta-data is now an ontology, a schema, documents already annotated with word net, cancorp, etc... some of this has already happened with word processors.
- Querying this is somewhere between classical databases and information retrieval.
- *Data management is a nightmare. It's about to get worse.*


Canadian Expeditionary Forces
British Expeditionary Forces
XXX
Canadian Corps
≈50,000
Division PartOf Corps
AssignedTo
XX
3rd Canadian Division
≈10,000
isA
Brigade PartOf Division
AssignedTo
X
7th Bridgade
≈3,000
isA
Battalion PartOf Brigade
AssignedTo
II
PPCLI
≈1;000
isA
?
?
isAlso
AssignedTo
III
PPCLI
isA
Regiment
isA
III
RCR
isAlso
II
RCR
Amalgamated Command
#1 Co
Commanding Officer
Captain
Appointed
Commanding Officer (acting)
XXX
XVI Corps
XX
27th British Division
X
80th Bridgade
<1915
Cyril Biddulph
Lieutenant
SeniorTo
John MacPherson
Captain


Havildar (Indian Army)
equivalentRank
Havildar (Pakistan Army)
equivalentRank
equivalentRank
isA
Sergeant (British Army)
isA
RankSeniorTo
RankJuniorTo
Havildar (British Indian Army)
isA
isA
Havildar
equivalentRank
Sergeant

Exploratory IR in unknown domains

- Ongoing work with Shelley Hulan, English Literature.
- Large collections of mixed documents.
- Retrieval needs aren't traditional - "soft" requirements.
- "Operationalizing" those requirements is painful.
- Ongoing problems with complex information retrieval problems.

Prototypes using passage voice

Home Blog Search | People Events Organizations Locations Documents Forms | About

Muninn Text Search

red flares sos very lights Submit

Voice: Imperative ○ Subjunctive ○ Passive ○ All ○

Results: 15

Results for query: red - **344** flares - **275** sos - **30** very - **1575** lights - **95**

#	Document	Result	I S P
1	2143703	Gun Schemes. 24 SOS 25 SIGNALS 26 CONTACT PLANES The SOS Signal will be made by a rapid succession of either RED Rockets or VERY LIGHTS (1½ or 1") Three White Very Lights will be fired	0 0 16
2	2143791	hereby cancelled and the following substituted: (i) SOS The Corps SOS Signal is RED, either Rockets or Very Lights as many as possible being fired in rapid succession. These	0 2 12
3	2147118	Flares. 2 carried by each Platoon Commander. 2 carried by each Section Commander. Rifle Wire Cutters. 2 per Section. SOS Rockets. Per Company H.Q. – 2. Per Platoon H.Q. – 1. Red Rockets. Per Company H.Q. – 1. Per Platoon H.Q. – 1 Red Very Lights.	3 6 2
4	2143787	Battalion Report Centre. 15 LIGHT SIGNALS (1) SOS The Corps SOS Signal is RED, either rockets or very lights, as many as possible being fired in rapid succession. In	3 2 10
5	2144248	Each Company will be equipped with a quantity of ground flares, SOS Rockets and Verey Lights White, 1 inch. 13 COMMUNICATION Communication will be maintained by	0 2 12

Prototypes using passage voice

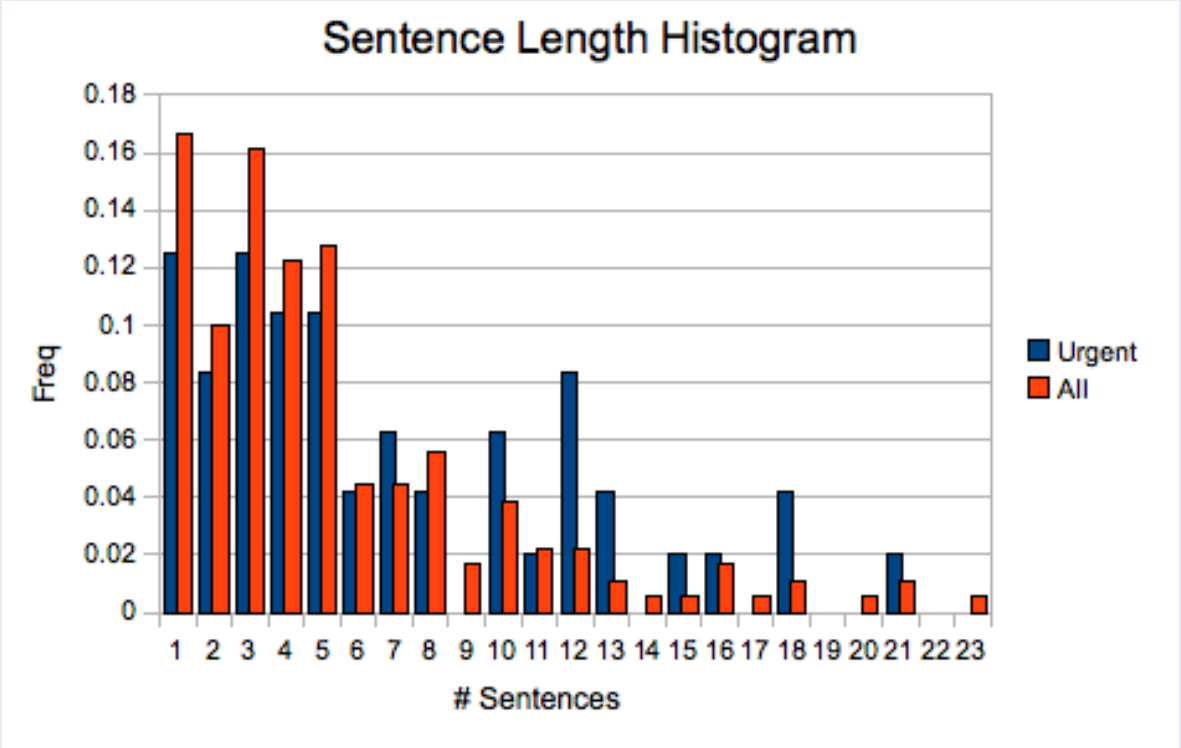
Home Blog Search | People Events Organizations Locations Documents Forms | About

---

Voice: Imperative Subjunctive Passive

2143458 No 13 PARTY. A Stretcher Bearer Party with 4 Stretchers and First Aid Appliances will be in waiting inside our parapet to evacuate casualties. 7 SUPPORTS (a) Artillery. 2 Barrages will be required. (1) Enemy front Line from 0. 7. a. 9. 2. to 0. 7. b. 3. 3. and 0. 7. b. 9. 6. to 0. 8. a. 1. 8. leaving Trench from 0. 7. b. 3. 3. to 0. 7. b. 8. 6. clear for attacking Parties. Enemy Support Line from 0. 7. b. 3. 3. to 0. 7. b. 9. 6. Fire will commence on the code Word GO from the Telephone Officer in the Sap. (2) Enemy Front Line from 0. 7. b. 3. 3. to 0. 7. b. 9. 6. Fire will commence at the Code Word COME from the Telephone Officer in the Sap. The first code word will be given when the 3 Attacking Parties are inside the enemy Trench and nd the 2 code word when the whole raiding Party has reached Sap 50 Yards from enemy Wire. Artillery will cease fire on receiving code word HAPPY from the Officer at the Telephone in Trench 0. 4. (b) MACHINE GUNS] (1) Battalion Machine Gun Officer will arrange for short burst of fire over head while the wire cutting is going on. He will also arrange that all available Guns sweep enemy parapet from 0. 7. b. 4. 5. to 0. 7. c. b. 0. and 0. 7. d. 6. 5. to 0. 8. a. 1. 6. from the time the attacking Parties enter Trench until they have returned to our own line. The Signal to commence fire will be the commencement of the firs Artillery Barrage. They will cease fire when the Artillery cease fire. (c) Trench Mortars and Stokes Guns While the Artillery is firing Stokes Guns and Trench Mortars will fire on point 0. 7. a. 9. 0. Volume of fire to be decided on by the Officers concerned. Rifle Fire (d) While the wire cutting is going on sentries in the Front Line should fire about once every ten minutes. There will be no rifle fire while the Artillery is firing. Retirement When the KLAXON HORN sounds, the FLANK PARTIES will work back as quickly as possible to the point of exit. When they get into touch with the CENTRE PARTY the LATTER will at ONCE leave the Trench and retire through the Sap to Trench 0. 4. When the FLANK PARTIES come in contact with each other toe first man of each party will leave the trench and using the supply of reserve bombs, bomb outwards while the remainder are leaving the trench and retire into the Sap. The RIGHT FLANKING party will leave the trench FIRST. When all are clear of the Trench the two men who are bombing to the Flank will throw the 4 smoke bombs, and will retire with the Officer in charge of the LEFT PARTY. No 4 PARTY will retire when the Reserve bombs have been handed over to the 2 bombers of the LANKING PARTIES. No 5 PARTY will retire as the LAST OFFICER and MAN reaches them. When the last Man has passed the Supporting Party in the Sap the Lewis Gun with the Supporting Party will sweep the parapet on both sides of the Hedge until the Retiring Party are well on their way to our Lines. The Supporting Party will THEN pass the Telephone Party and return to our trench. The Code Word for the second Barrage will be given by the Officer at the Telephone and the Telephone wire will be then CUT and the Officer and Telephonist return to our Trench.

Messages and Signals - a data-mining perspective



SNA driven Information Retrieval

- Information Retrieval on large sets of legal documents.
- Used a novel social networking method to modify document rankings.
- Above median for 2/3 topics and top score for one topic.
- Problem: Some normalization problems with senders outside of the network.

$$P_{\text{doc}} = P_{\text{bm25}} \dot{A} VG(\forall P_{\text{doc}}(\text{sender})) \quad (1)$$

Current Era

"Kill them all, let god sort them all" (Local madman)

Crusades Era

"Massacrez-les, car le seigneur connaît les siens." (Arnaud Amalric, French madman)[a]

[a] http://fr.wikipedia.org/wiki/Arnaud_Amaury

Research problem

Track both the idiom *and* the underlying themes across all **8** centuries.

RDF (Resource Description Framework) / Linked Open Data

XML like, but with cross references between files.

```
1 <org:University>Bauhaus</org:University>
```

OWL (Web Ontology Language)

```
1 <owl:Class rdf:ID="University">
2 <owl:subClassOf rdf:resource="#Educational_Organization
    " />
3 <rdf:Type rdf:resource="http://xxxx.de/de_universities"
    />
4 </owl:Class>
```

Library Catalogs (MARC)

Warren, Baby Boy, 1919-1977

Differing views of the same data

```
1 <foaf:firstName>Robert</foaf:firstName>
2 <foaf:lastName>Warren</foaf:lastName>
```

GNL view

```
1 <gnd:preferredNameForThePerson>Rob Warren</
    gnd:preferredNameForThePerson>
2 <gnd:forename>Robert</gnd:forename>
3 <gnd:surname>Warren</gnd:surname>
4 <gnd:locQualifier>Academic</gnd:locQualifier>
```

Library View

```
1 <foaf:Author>Robert H. Warren (Academic, 1973-)</
    foaf:Author>
```

The human being is the IR implementation

- MARC / Library Catalogs mimic library cards.
- In 1967, about 600k books published worldwhile. In 2011, 600k in the UK alone (+2M others). Expect 14M books published in the US in 2012.
- None of this will scale! (An the LOC knows it.)
- The *thing* and the name of the *thing* are not the same *thing*!

Both *MARC* and the *FOND* reference change!

RG 4353.664 550-670, Robert Warren

We have to make this work together:

- Taxonomies (Canada is a realm, a constitutional monarchy, a confederation, a chunk of land and a Dominion. Newfoundland used to be a Dominion but is now a Province and part of Canada.)
- Good quality, cross referenced information systems are coming and the data deluge will be replaced with the meta-meta-data deluge.
- Cultural translation versus word translation (see wikipedia)
- We urgently need a super-class to the “bag-o-words” model.

Conclusion

Questions?