## Scoping out options for approaches to large sets of textual data (Phase Two)

Given the large amount of textual data and the general consensus that NVivo is not adequate for that scale, the procedures used for qualitative analysis need to be re-evaluated for Phase 2.

Possibilities discussed previously include using natural-language processing (machine-learning) or an existing R-package. As these types of approaches are outside RAT areas of expertise we need to consider hiring someone with the relevant experience for whichever approach we take (in addition to learning ourselves where possible). Ideally, we’d have someone with the relevant expertise to help decide which approach to take (I presume we’ll need to clarify what structure RAT will take, and how it integrates within the broader project, for phase 2 anyway).

In the meantime, this document is to keep track of the options we’ve scoped out.

We’ve budgeted 0.6 of a person at same level as Eden for RAT - David to help with hiring decisions

## Progress Details:

### 20201008 – Meeting with David

1. RAT is going to be moving our documentation over to Github. To set it up in ways that work well with how you're all working in the Data Management & Analysis team:
   1. Make sure the Github project is a private repo . (need to invite collaborators). Use own Github account can transfer ownership to the organization if/when David sets it up.
   2. Use R Studio as the GUI
   3. Write documentation using R-Markdown (or plain text).
   4. Invite to project: Martin & David (& research assistants as needed).
2. We'll be abandoning NVivo in phase 2 and are looking for a better tool that works with large amounts of textual data. We've identified some options:
   1. David will look at RQDA and see if anything more recently maintained that is equivalent. <https://rqda.r-forge.r-project.org/> - although I noticed it was last updated 2016, <https://github.com/Ronggui/RQDA> and may not be supported (e.g., <https://github.com/Ronggui/RQDA/issues/38>) so need to check that out. )
   2. NLTK – natural language processing toolkit (Originally in python, can be used in R. Can call python from R. <http://www.nltk.org/> )
3. Given points 1 & 2, RAT is going to be hiring someone to help us with managing large sets of textual data for the purpose of qualitative analysis and mixed methods research questions and we'd like your advice and input on that position description (which Martin has started drafting) and the hiring decision given they'll need experience with:
   1. R experience.
   2. NLK experience (in python or R).
4. Given points 1-3 , Eden is going to be learning to use Git better and start learning R – David to send links to resources and provide help as needed..

### 20201008 – Meetings with Fiona & Martin

Martin’s team-lead role formalized – so, going forward he will act as the liason with all other teams and if I need to know something about the project as a whole he’s to communicate to the rest of RAT.

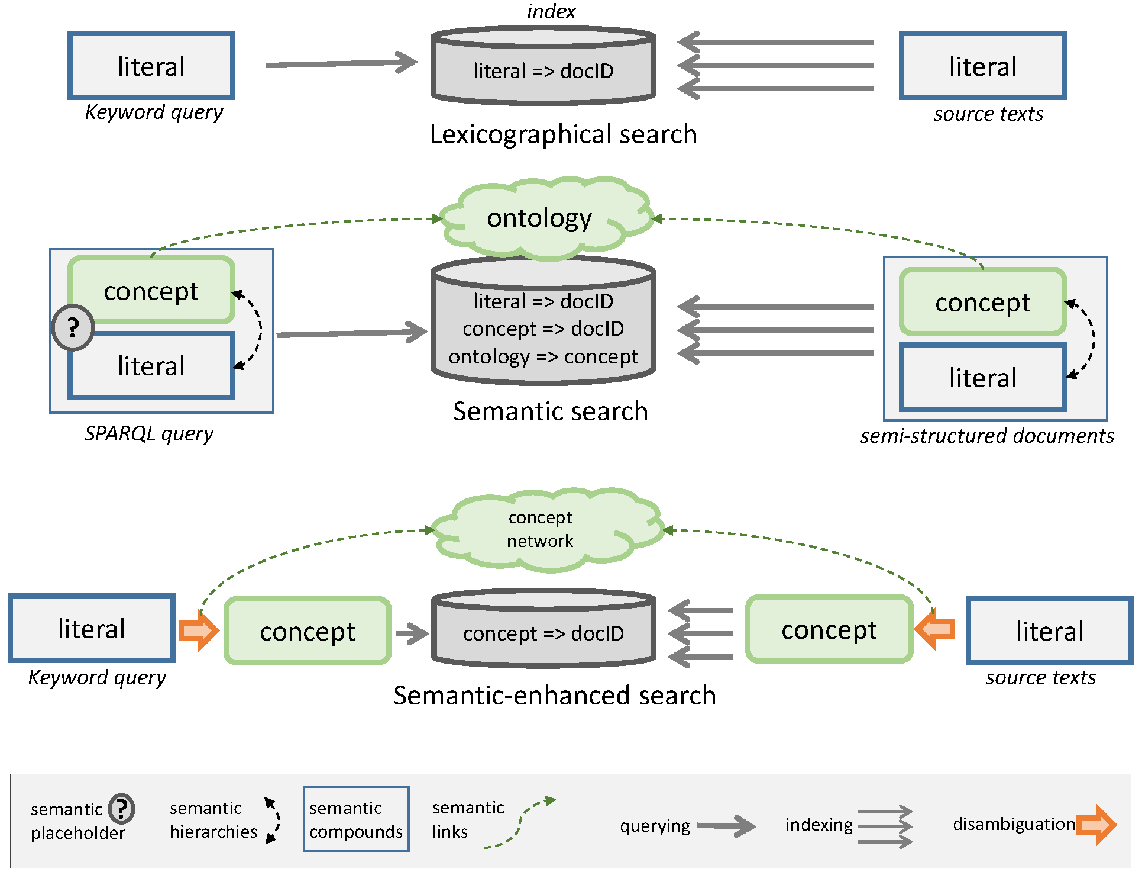
Eden to learn Git and R – and to set up the documentation for RAT on Github so that any shared data can be better managed with the Data Management & Analysis Team.

The budget also includes 0.6 of a person at same level as Eden for RAT - David to help with hiring decisions so that we can get the relevant experience in R and/or NLTK that we need to manage and do the analysis on the large scale of textual data that we’ll have

***20201007***

Meeting with Craig Taube-Schock, a computer scientist from the University of Waikato who helped develop Capisco (a semantically-enhanced search and discovery system for large-scale text corpora) – see <https://doi.org/10.1145/2833219.2833223>

Notes:

* Capisco is a semantically-enhanced search process designed to handle large collections of pre-existing documents – could we use non-document based textual data sets (e.g., short textual units extracted from text-boxes in an online-discussion platform). - Capisco uses a knowledge-based engine for topic extraction – and you can send any arbitrary text (requires text format) to this engine for topic extraction (if documents aren’t in text, you can automate text-extraction and then fed that into Capisco). (see details below)
* Limitations of ‘semantic approaches to text-based documents’ that use machine-learning include the difficulty of getting the training data sets and that as a humanities researcher it’s difficult me to improve or influence the algorithm and it’s outcome. (In contrast, Capisco allows users to manually extend/refine the concepts-in-context (CiC) network).
* In the into to Capisco paper (2015), there is a mention that, while the automatic disambiguation using the CiC network done during the indexing phase could be employed for semantic query expansion, the reliance on long query sentences presents a hurdle.
* Craig specializes in enhanced semantic-search processes precisely because he could not bet machine-learning approaches to function reliably for classifying context-dependent content in text. He was supervised by Ian H. Witten (a ML specialist who has recently retired) and continues to work with ML people. Based on this experience, his advice is to not use ML for qualitative-analysis of any kind because none of the ML approaches available adequately account for context. Even with supervised learning, you need to feed in features for classification of context and this requires multiple instances of every possible context-based use of word. The ML algorithms can learn (i.e., extrapolate) but they aren’t able to self-reflect, so the very human problem of ‘You don’t know what you don’t know’ become a problem of garbage-in/garbage-out – i.e., misclassification. The best we get with ML is to ensure that the context is extracted from a really specific context – i.e., you can extract semantic context because you already know the context that context will appear within.
* If wanting to explore the ML options anyway, advice is to get specifics about what we are expecting ML to do for us and then track down and fact-check the assumptions embedded in those expectations. **Craig is going to talk to two ML experts he knows about my questions and put me in touch with them if we want that.**
* More on Capisco. “In comparison to the ontologies and dictionaries used in classic semantic search and other semantic approaches (Sections 4.2 and 4.3), the CiC network does not merely contain concepts but also their explicit contexts of use. Furthermore, only two types of (interlinked) relationships are modelled (term in context and term for concept). Though this may appear to pose a restriction in terms of expressiveness, this very structure allows scholars and technical novices to extend and manipulate the network”
* Compared to other search strategies:
* 

### 20201006

Meeting with Ross about what phase 2 and the feasibility and process of incorporating machine-learning as a technique into the qualitative data analysis process (using the existing manually-coded data as a training set for example).

A few ideas discussed - regardless of which, we want to get onto this sooner rather than later (but especially if there is a PhD project involved).

Ideas:

* Hire a machine-learning expert with experience in semantic-content analysis who is willing to set up and execute the machine-learning system for analyzing out qualitative data as part of RAT
* Get a PhD student on-board who is willing to spend their first 18 months setting up and executing the machine-learning system for analyzing repliCATS qualitative data (principally supervised by a machine-learning expert, and co-supervised by Eden and/or Martin sitting within RAT).
* Look into context-sensitive semantic-search tools such as Capisco that could be used without machine-learning expertise -although it’d be good to still hire another person for RAT with relevant experience in this type of semantic-search practice (ideally one who was interested in qualitative methodology).

### 20201006

Suggestions on slack:

* <https://rqda.r-forge.r-project.org/>

### Ideas List:

* Use an R package developed for qualitative analysis of large data sets,
  + such as: CAQDAS/ RQDA <https://rqda.r-forge.r-project.org/> - although I noticed it was last updated 2016, <https://github.com/Ronggui/RQDA> and may not be supported (e.g., <https://github.com/Ronggui/RQDA/issues/38>) so need to check that out.
  + Upsides: R is something already used in the repliCATS team.
  + Downsides:
* Hire a NLK expert with experience in semantic-content analysis who is willing to set up and execute the machine-learning system for analyzing out qualitative data as part of RAT
  + Upsides:
  + Downsides: (see noted from 20201007)
* Use a context-sensitive semantic-search tool and adapt for qualitative analysis, such as:
  + Capisco (Hinze, A., Taube-Schock, C., Bainbridge, D., Cunningham, S. J., & Downie, J. S. (2015). Introducing Capisco: a semantically-enhanced search and discovery system for large-scale text corpora. *ACM SIGWEB Newsletter*, *Autumn 2015*, 4:1–4:14.<https://doi.org/10.1145/2833219.2833223>).
  + Upsides/ Downsides: (see noted from 20201007)
* Get a PhD student on-board who is willing to spend their first 18 months setting up and executing the machine-learning system for analyzing repliCATS qualitative data (principally supervised by a machine-learning expert, and co-supervised by Eden and/or Martin sitting within RAT).
  + Upsides:
  + Downsides: (see notes from 20201007)