Keyword Ontology - An Outline Rationale

University of Dundee (Research Partner)

During the development of the ArtsAPI project, we established that one way of making sense of the data generated via email is to filter it in some way to show different aspects of the network and it's connections. We established a number of scenarios that proposed the potential of filtering the data in different ways. These included:

1. Filtering networks by 'sector', to provide evidence of the interdisciplinary working and knowledge exchange of an organisation.
2. Filtering networks by 'country', to provide evidence of the international impact/reach of an organisation.
3. Filtering networks by 'activity' to provide different views of who is involved in key aspects of organisational activity across the network.

The first two challenges can be met by capturing user input data that links network nodes to specific countries and sectors. The third however, is more challenging. For example, it is difficult to establish sets of roles and responsibilities or job titles that are uniformly reflected across the arts sector. Many people have similar responsibilities but have different titles and many people have multiple roles within arts organisations. This is extremely difficult to formalise in any meaningful way let alone provide a structure by which to describe data that could be handled by a computer. Similarly, it is equally, if not more difficult, to identify skills and skillsets that individuals might have as each individual has a unique set of experiences that make them who they are. It soon became clear that an alternative approach was necessary. Enter Keyword searching. Keyword search offers us the opportunity to mine email for particular keywords that we think are valuable in determining the kind of 'activities' that arts organisations are involved in. By establishing keywords that are aligned to particular activities it becomes possible to infer that people that are using those key words in their email are involved in the related activity. The difficulty in achieving this is twofold. Firstly, it is difficult to establish what the activities are that arts organisations are involved in without doing some prior work and secondly it is difficult to establish the best key words to search for that are associated to those activities.

Happily for us, we had already done the prior work necessary in terms of establishing some of the activities that arts organisations are engaged in. This data came out of our initial workshop with the arts organisations where we asked them to describe the Artefacts, Articles and Activities that they engage

with during their day-to-day work. This piece of work established that things like budget setting, management meetings, conferences, writing funding bids etc were all part of the daily routine of arts organisations.

Analysis of this data provided our initial structural ontology of the Arts Organisation space where we were able to apply a hierarchical structure of categories to the data, in order to make it useful. But this only took us so far. The second part of the problem was in identifying the keywords associated with those categories. For this we had to analyse example email between FutureEverything and University of Dundee. In working by hand with raw email data we carefully considered how people actually talk about their work and which words frequently appeared in email trails about certain topics. For example it became clear that in threads related to funding applications words like ‘draft’ ‘amend’ ‘written’ input’ and ‘edit’ became very common. These, we proposed, could be grouped together under a subheading of ‘Writing’ which in turn has been grouped under ‘Development’. Interestingly though the word ‘funding’ itself falls under a different category which is associated with ‘Management’. By working in this way we have been able to link our analysis of raw email data to the overall ontology to create a new ‘keyword ontology’ that allows us to mine email to look for evidence of discussion around particular activities.

If this works, we should be able to identify the particular ‘actors’ in the network that can be identified with particular ‘activities’ and show the relative strength of their contribution amongst a network of other people also involved in that activity. Thus is should be possible to build a network map of an activity such as Operations that shows the people involved in discussions that feature words such as ‘planning, timetabling, submission, deadline, meeting etc. likewise it should be possible to differentiate a network map of an activity such as Networking that shows the people involved in discussions that feature words such as ‘chat, coffee, party, dinner, lunch etc. Some people may be involved in both activities and the network maps should be able to show this. Of course, it’s not as straightforward as would first appear. There are a whole host of issues to resolve such as false positives generated by keywords and the limitations of the chosen keywords themselves in establishing the context of an email. However, we feel that these are issues that can be worked on to improve the outcomes if this initial strategy proves to be successful.