Overview

It will likely be useful to obtain additional publications for each award in order to have additional textual information related to the award. This will be useful for community detection via topic modelling approaches, as well as to provide additional features or enrich existing features for other purposes.

Methodology

The most immediately apparent way to do this is to scrape the references from the NSF award listings. The NSF offers a search service at nsf.gov. This can also be accessed programatically with POST requests:

```
curl -d "AWD_ID=0963183&HistoricalAwards=false"
http://www.nsf.gov/awardsearch/showAward
```

The result of this POST request is an HTML page with a listing of information regarding the award. Most interesting for our purposes is a section at the bottom titled: "PUBLICATIONS PRODUCED AS A RESULT OF THIS RESEARCH". This section contains citations for papers that were published as a result of the research funded by this particular award. This list of citations is not present in the XML data for each award, so it will be necessary to scrape the HTML pages to obtain the citations, then parse them into their relevant fields.

Data Storage

Publication

```
*********************************
Field              Type
*********************************
id (PK)            int
title              string
journal            string
volume             string (int?)
pages              string
year               int
uri                string
award_id (FK)      int
*********************************
```

Author

```
*********************************
Field              Type
*********************************
person_id (FK)     int
pub_id (FK)        int
```

```
********************************
```

Person

```
********************************
Field               Type
********************************
id (PK)             int
fname               string
lname               string
middle_init         char
email (unique)      string
********************************
```

Example

For example, the following publication listing is from the POST request above:

---

1. Yuwen Sun, Lucas F Wanner, Mani B Srivastava. "Low-cost Estimation of Sub-system Power," Proceedings of the Third International Green Computing Conference (IGCC'12), 2012.
2. Shafiee, A., Brandenberg, S.J., and Stewart, J.P.. "Laboratory investigation of the cyclic and post-cyclic properties of Sherman Island peat," GeoCongress, San Diego, 2012.
3. Reinert, T., Brandenberg, S.J., Stewart, J.P., and Lemke, J.. "Remote monitoring of consolidation of peaty organic soil beneath a model levee," GeoCongress, San Diego, 2012.
4. Zainul M Charbiwala, Paul D Martin, Mani B Srivastava. "CapMux: A Scalable Analog Front End for Low Power Compressed Sensing," Proceedings of the Third International Green Computing Conference (IGCC'12), 2012.
5. Reinert, E.T., Brandenberg, S.J., Stewart, J.P., and Moss, R.E.S.. "Dynamic field test of a model levee founded on peaty organic soil using an eccentric mass shaker," 15th World Conference on Earthquake Engineering, Lisbon, Portugul, 2012.

The first of these would be parsed into the following record:

```
Publication
********************************************************************************
*************
Field               Value
********************************************************************************
*************
id (PK)             0
title               "Low-cost Estimation of Sub-system Power"
journal             "Proceedings of the Third International Green Computing
Conference (IGCC'12)"
volume              NULL
pages               NULL/ALL
year                2012
uri                 NULL
```

```
award_id (FK)       0963183
*********************************************************************************
*************
```

This record would then be matched up with each person, and if the person does not already exist in the db, he/she would be added, then matched as an author.

Notes on Extensibility

One thing which is evident from the example above is the discrepancy between a "conference" publication and a "journal" publication. In order to produce maximum value from the data, it seems like it may be worthwhile to create another table called something like "Vehicle" or "PubType" or something like that. This table would include the name, abbrevation, and perhaps some metadata about the conference/journal, as well as a type indicating whether it is a conference or a journal (enum). Each entry would be given an integer ID and this could be referred to by a FK in the Publication table.

This would not only be useful to more clearly distinguish between a conference publication and a journal publication, but also to create the opportunity to attach keywords such as prominent topics to the conference/journal. This might aid in community detection efforts.