About the data

RDF generation for Belfast Group Data

Rebecca Sutton Koeser
rebecca.s.koeser@princeton.edu
https://orcid.org/0000-0002-8762-8057

June 2015
https://belfastgroup.digitalscholarship.emory.edu/network/about/

This document describes the steps that are done by the "prep_dataset" script, which harvests and builds the RDF dataset for the website, which is used in part as the basis for the network graphs and chord diagrams. Prior to running the script, significant work was required to 1) to tag names in the EAD and TEI and 2) expose the tagged information as RDF so it could be harvested, but this work is documented elsewhere.

1. **Harvest and collect RDF**
    - **1.1.** Harvest RDF from specific collections we are interested in on the Emory Finding Aids website; includes logic to harvest "related" web pages using related links, which allows for automatically harvesting all parts of a multipart finding aid (e.g. a large collection with multiple series/subseries and index) and also allows us to automatically harvest data from collections that are labeled as related collections in the EAD via the "Related Materials in this Repository" section (see the Michael Longley papers for an example)
    - **1.2.** Harvest RDF for the TEI Group Sheets included on Belfast Group Poetry | *Networks*, for inclusion in the list of Group sheets and so data about names and places mentioned in the poetry can be included in the dataset
    - **1.3.** Generate RDF from two sets of local "fixture" sources that are not available as RDF elsewhere:
        - **1.3.1.** a list of the Group Sheets based on the collection at Queen's University Belfast (converted to HTML and cleaned up to make some corrections and improvements for consistency with our other data)
        - **1.3.2.** four HTML documents; one brief biographies pulled from the original Belfast Group website (annotated in RDFa and tagged with VIAF identifiers), to provide profile information for persons associated with the group that was otherwise not being included in our dataset; one with the Edna Longley biography (from the Michael Longley finding aid, but not harvestable); one with the Hannah Hobsbaum biography (from *International Who's Who in Poetry 2015*); and one with information about the one Group sheet known to be in private hands (Pakenham)

2. **Identify and "smush" Group Sheets**

The Group Sheets in the RDF data do not have unique identifiers (other than the ones available as digital editions), so for convenience the script includes work to identify Group sheet content and to de-duplicate the Group sheets where multiple copies are present in different collections. This work consists of going through the data and identifying manuscripts associated with the Belfast Group, tagging them as a Belfast Group sheet (using a custom RDF type), and then "smushing" them in the RDF, generating a new, distinct identifier so that a copy of the same Group sheet in different archival collections or in a TEI digital edition can all be connected to each other.

3. **Annotate the graph with related data**

Because we are using identifiers from other RDF data sources—specifically VIAF, DBpedia, and GeoNames.org—preparing the data includes a step to harvest a minimum set of useful data from those external systems for use in the website (e.g., names for people in VIAF, descriptions and Wikipedia links for people in DBpedia, and geographical coordinates for places in GeoNames.org)

4. **Make inferences about the data and add them to the dataset**

Some information that is necessary or useful for displaying parts of the site (e.g., the list of Group sheets) or the network graphs is present in implicit ways within the RDF dataset at this point, but not present in a way that can be easily queried for use. This step in the process makes a few specific inferences about the data and then adds that information back to the data in the form of new RDF triples that can then be used by the data or for generating network graphs.

- **4.1. Time period:** infer whether a Group sheet belongs to the first or second period of the Group, and add coverage information to the RDF graph for that Group sheet:
  - **4.1.1.** If there is a date associated with any copy of that Group sheet, use that date to determine which time period it belongs in
  - **4.1.2.** If there is a TEI copy of the Group sheet, use the coverage information from the TEI
  - **4.1.3.** Otherwise, assume that if a Group sheet is in the Hobsbaum collection at Queen's, it belongs to the first time period (since Hobsbaum was only involved in the first period)
- **4.2. Ownership:** infer ownership of a Group sheet based on the archival collection(s) where it can be found (e.g., if a Group sheet is listed in the Longley papers, add a triple stating that Longley owned that Group sheet)
- **4.3. Associate authors and owners of Group sheets with the Belfast Group:** In order to easily identify people associated with the group when generating profile pages and network graphs for the site, goes through each Group sheet and adds a triple (if not already present) to indicate that anyone who owns or authored a Group sheet should be considered affiliated with the group.
- **4.4. Mentions of people/places in the digitized Group Sheets:** For convenience, and to allow explicit connections to be included in the network graphs, go through all poems

in the RDF data (only harvested from the TEI) and add a direct relation between the author and the entities mentioned in the text.

5. **Generate Network Graphs**

For ease of processing, network analysis, and display on certain portions of the website, the RDF data is converted into network graph format using the Python NetworkX library.

5.1. full network: all triples in the RDF (with the exception of RDF sequences) are converted into a network where each subject and object is a node and each predicate is an edge between them, weighted based on some sense of the strength of that connection (see current weights used in the code)

5.2. specialized network based on the Group sheets by time period: nodes are the Group in period one (1963–1966) and period two (1966–1972), and the authors and owners of Group sheets; edges are added between authors and owners of Group sheets that belong to the Group in one period or the other, and between co-authors, co-owners, and owners and authors of the same Group sheet, and edges are weighted based on the number of Group sheets.

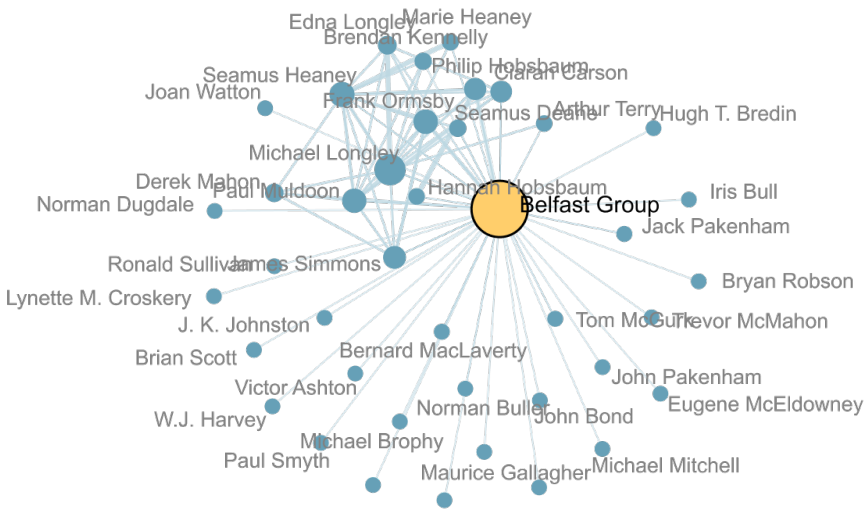

Figure 1. Network graph of people associated with the Belfast Group.
(view interactive version)

The "People Associated with the Belfast Group" graph (Figure 1) is generated from the full network graph (described above in 5.1): an egograph centered on the Belfast Group, filtered to only include persons and organizations and restricted to include nodes that are only one or two degrees away from the Belfast Group (and in the case of the two-degree graph, nodes are further filtered by a minimum degree of five because otherwise there is too much data to be presented or made sense of).

Similar logic is used to generate the egographs on individual profile pages, except that when filtering

nodes places are also included in addition to persons and organizations.

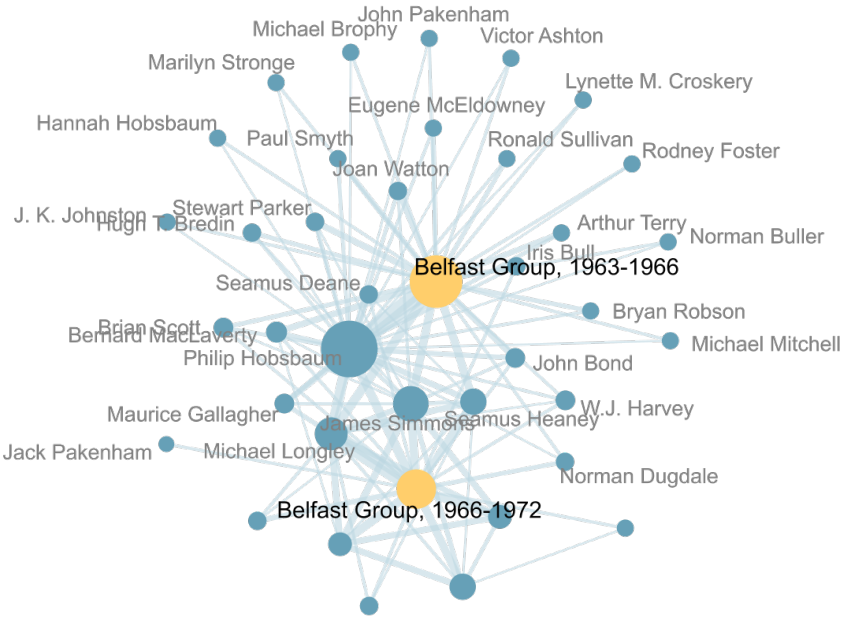

Figure 2. Network graph of Belfast Group authors by time period.
(view interactive version)

The "Belfast Group Authors by Period" graph (Figure 2) is generated from the specialized network described above in 5.2.