**USA National Foundation for Science**
**Research Collaboration Network meeting of 2018-09-18 (9 am-3 pm)**

Chairs:

- Stephen Weldon (IsisCB and University of Oklahoma) - **SW**
- Gavan McCarthy (ESRC, University of Melbourne) - **GM**

Others Present:

- Alison Pearn (Darwin Correspondence Project and Epsilon, Cambridge University) - **AP**
- Anne Barrett (Imperial College) - **AB**
- Louisiane Ferlier (Royal Society) - **LF**
- Elisabeth Smith (Darwin Correspondence Project and Epsilon, Cambridge University) - **ES**
- Mike Hawkins (Darwin Correspondence Project and Epsilon, Cambridge University) - **MH**
- Ana Alfonso-Goldfarb (CESIMA, Pontifical Catholic University of São Paulo) - **AA**
- Marcia Ferraz (CESIMA, Pontifical Catholic University of São Paulo) - **MF**
- Kathleen Vogel (University of Maryland) - **KV**
- Julia Damerow (via Zoom) (ASU, A Place Called Up) - **JD**

The meeting was convened at Imperial College. Many thanks to Anne Barrett for arranging this venue.

1. Epsilon discussion -- We began by discussing what Epsilon was and how it could be integrated with other projects of RCN members
    a. **MH, AP, ES**: discussed protocols for collaboration
    b. Key points to note: right now there are eight sets of letters/projects. The focus has been with legacy projects that began as print.
    c. They have guidelines for people delivering to Epsilon. Those guidelines require TEI submissions with a relatively minimal structure. There are only a few key fields that are absolutely required.
    d. We talked about linking to the IsisCB authorities, Royal Society authorities (many of which have ODNB record links), Encyclopedia of Australian Science authorities, etc.
2. Tools and protocols
    a. **GM** and **LF** can easily export records as EAC-CPF. This is a global standard for sharing all kinds of archival authority records and gives anyone providing this format a wide range of options for sharing.
    b. Where possible all projects should have the capacity to export and or import data using OAI-PMH.
    c. **SW** acknowledged that IsisCB must build EAC and MODS export for all its data.
3. We need to produce a service that will function as Black Goat was expected to do. It needs to be an **authority matching/disambiguating system**

a. **JD** will build this tool. She needs sample data and clear instructions on what the tool must do and how it is to operate. She thinks it will be about 12 months to create.
b. **GM** will be responsible for helping direct people to understanding EAC formats for data.
    i. The following groups currently have authority data that they can share: Darwin Correspondence Project, Royal Society, IsisCB, Mueller Correspondence Project.
    ii. Types of authorities that we will potentially want to share. The list below includes authorities that we will eventually want to be able to match using the tool:
        1. *Darwin*: People, but this includes many items that are not biological human beings. For them a person is any of the following: a human person; a group of people (the doctors); corporations (societies and associations); publications (*Nature*). They also will want to share Repository file identifiers (holders of items) and bibliographical items. Their bibliographical citations have not been have not been turned into machine readable citations.
        2. *Von Mueller*: People; Bibliographical citations; Mueller eponomy (plant names named after plant); Institutions; Places; Events; Subjects
        3. *IsisCB*: Bibliographic citations; Authorities; Subjects; Places; Classifications;
        4. *Royal Society*: People and bibliographical citations.
    iii. Other Notes:
        1. *Royal Society* uses the DOI as a unique identifier for bibliographical citations as does *IsisCB*
        2. Many projects have bibliographic data that is in a simple text string. It will be necessary to find ways to build authorities out of this data.
        3. **GM** suggested looking to HUNI (an Australian project that the Canadians have just provided major funding for) for a list of authority types.
c. What should the tool do?
    i. It is to be a disambiguation tool
    ii. It is to establish identity relationships between items and mark those relationships with confidence values based on the data that it can match from various records.
    iii. The service should be able to ingest data from all projects in bulk uploads. It cannot rely on the projects having data accessible via an API.
    iv. The tool should be able to read and parse EAC-CPF data for authorities and MODS data bibliographic records. It should also be able to work with

data in a prescribed CSV format for any projects that have data that simply cannot be converted to EAC-CPF or MODS.

v. It should have an API that will allow projects to extract data from it. This should be for both small batches as well as large bulk exports.

vi. The data will be viewed and cleaned by individuals. It would be good, but not necessary, to have the ability to edit on the service. However, we believe that most projects will have their own workflows for cleaning and correcting information. The tool should, thus be able to reingest previously ingested data from a project so that it can work with corrected and updated records from any project.

vii. The service should be able to provide visualizations of the data on service.

viii. All data should be clearly linked to its source and to its type.

ix. There should be a public face to the system, though only registered projects should be able to upload and modify data.

x. Updates: The tool will need to keep track of existing and past data. It will need to indicate when items have been deleted. All decisions about aligning data need to be logged--be they machine made decisions or human made decisions (machine decisions should carry a date; human-made decisions (whether done on the system, if that is a feature, or done through reingesting data through the bulk update function) must be linked to the person/project making the decision and the date. Subsequent uploads will be used to create a new ranking, and all people who have downloaded certain records will need to be alerted that the ranking has changed.

4. Miscellaneous notes:
    a. RCN reports need to be shared will all RCN participant. There was some concern that without such oversight the NSF might be given incorrect information about some of the projects and the relationship of their participation with RCN. This is especially important in different ways for projects when they are looking for funding. **So we request Manfred Laubichler to circulate all reports before they are finalized to the entire list of RCN affiliates**.
    b. There was some general discussion about the lack of resources for digital humanities in key areas. **GM** noted that we need institutions that train people in XML. There is an extreme dearth of support in this area. **ES** noted that the general lack of programmers familiar with the kinds of material that we are working on creates a serious shortage of help for all projects. **AP** concurred. We need to have ways to encourage training in the technical skills that are central to the work that we are doing.
    c. **GM** noted that Trove has user tools that we should look at
5. The steering committee for this project will be **SW, GM, LF, and ES**. Proposed name of this RCN working group: “Authorities integration & data aggregation working group.”
6. The next meeting

a. Held at Utrecht on 2019 July 23-27? (during the HSS annual conference)
b. **SW** proposed that we have an entire day devoted to developing tools that will work with and manage different classification systems. **LF** suggested that we look at the Royal Society user designated items from Phil Trans.

Specific tasks:

- **MH**: He will work to create XSLT stylesheets that convert between TEI and EAC. He will do the same for conversion between TEI and MODS.
- **SW** will have **JD** make all IsisCB data accessible at EAC and as MODS.
- **MH**: Will create Bitbucket repositories (for both code and data) for his work and with IsisCB-Darwin Correspondence links. **MH** will create and give **GM** admin rights
- **SW and AA** will discuss CESIMA collaboration of their bibliographic data and thinking about ways of sharing bibliographic, person, and subject authorities.
- **December 1**: Sample data by all parties should be given to **MH** to put in a Bitbucket repository for **JD** to access
- **LF** to get specifications for software development done within the month.
- **SW** to put pre-1974 data into bitbucket for **MH**: The key problem will be in converting bibliographic strings into fielded data. This conversion technology will be especially helpful in working with other projects -- since this bibliographic string data exists in most of the projects in this group.
- **SW** and **GM** to tell **JD** what we want to say on the website about this working group of the RCN committee.