Investigation of the Update and Reach of ORCID-Registered Researchers and Publications for THOR (Technical and Human Infrastructure for Open Research)

Ahmad Al Marzook, Siqi Hong, Yu-Chen Huang, Olivia Wikle

Abstract—THOR is a project analyzing the persistent identifier (PID) type Open Researcher and Contributor ID (ORCID). Because ORCID is relatively new, few visualizations exist to help THOR team members understand and act on ORCID data. The objective of this project is to create visualizations that reveal in which disciplines and geographic areas ORCID has become most popular, and between which countries collaborations have become most frequent, over the years 2012-2016. Tableau was used to create a bar chart displaying top *Web of Science* categories of ORCID publications as well as line graphs and geospatial representations of countries with the top number of ORCID-registered Person and Publication IDs. Sci2 was used to create a geospatial network visualization of countries with the most collaborations. These visualizations provide greater insight into this subset of data on ORCID-users, and will allow THOR to determine where to focus future outreach efforts in the interest of promoting increased use of ORCID identifiers.

Index Terms—THOR, ORCID, *Web of Science*, visualization, bar chart, line graph, geospatial, network

INTRODUCTION

The Technical and Human Infrastructure for Open Research (THOR) project is a European Union-funded enterprise that aims to improve the interoperability and sustainability of persistent identifiers (PIDs) among researchers. The THOR group has access to the basic metadata connected with PIDs, and is monitoring the landscape of PIDs for trends. One popular PID type currently monitored by THOR is the Open Researcher and Contributor ID (ORCID).

Initially implemented in 2012, ORCID identifiers are especially important in the academic research environment because the attribution of sources is integral to the trustworthiness of new research, and academic careers depend upon the accuracy and amount of published research a scholar produces.[1] Crediting others by name can lead to inadvertent negative consequences: a name may be misspelled, may change over time, or may not include appropriate accents and diacritics.[2] Researchers using ORCID identifiers each "own" a specific identifying number that they can attach to the research they produce, and each scholar's ORCID record shows connections between identifiers and source items such as publications, grants, or dissertations (an example of an ORCID profile can be found here).[3] In this format, using ORCID identifiers becomes more effective than simply using researchers' names; if implemented in a correct and detailed fashion, a unique identifier makes it easier for researchers to discover exactly what information or data their colleague has contributed to a project.[4] As Laurel L. Haak states, one of ORCID's main goals is to develop standardized "authoritative attribution processes."[5] ORCID developers hope that such a standard will in turn promote increased sharing of research by scholars who, thanks to ORCID, can trust that they are guaranteed to get credit for it.[6]

Because ORCID identifiers are still a recent development and not yet wide-spread, few visualizations have been made using data gathered from ORCID. While it is relatively easy to find academic literature that discusses the merits of ORCID IDs and strategies for increasing their use in academic institutions, a search for scholarly or informal studies that incorporate visualizations into their analysis of ORCID yields few results. An exception is the blog *Data Science Lab*, which constructed a brief visual analysis of ORCID data in 2013, about one year after ORCID's inception.[7] The *Data Science Lab* visualizations are comprised of bar charts and a pie chart. One of the blog's bar charts (featured here in Figure 1) shows that in the first year of ORCID's existence (Oct 2012 - Oct 2013), the number of ORCID profiles increased steadily, though the number of profiles created day to day in each month varied.[8]

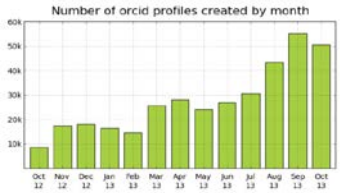

Figure 1: ORCID Profiles Created by Month, as visualized by *Data Science Lab*[9]

A pie chart produced by the blog (Figure 2) shows that the majority of profiles created during ORCID's first year were left incomplete: only one-sixth of the circle contains profiles with one or more publications, while the rest of the circle is comprised of profiles with no publications.

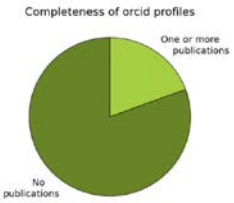

Figure 2: Completeness of ORCID Profiles, as visualized by *Data Science Lab*[10]

Our goal for this project was to assist a team made up of members from THOR and ORCID by visualizing ORCID data. After completing our background research, we entered into a consultation with our client team, which was made up of members from THOR and ORCID. By the time of our meeting in March of 2017, the client had created a few visualizations of ORCID data, consisting mainly of bar graphs and a geo-spatial visualization created using VosViewer. The client provided us with pre-processed datasets and asked that we try to diversify the types of visualizations that we produce with this data, noting their interest in having us use the advanced tools we have learned in this class to create an interactive geo-spatial visualization that might display some aspects of ORCID in finer detail.

Based on our background knowledge of the types of visualizations of ORCID data already produced and our understanding of the data variables given to us, we resolved to create visualizations with the potential to allow the client team to discover in which disciplines and geographic areas ORCID has become most popular and between which countries collaborations have become most frequent over the years 2012-2016. This information is valuable in that it will allow our client team determine where to focus future outreach and engagement efforts in order to promote increased use of

- *Ahmad Al Marzook, Indiana University. E-mail: aalmarzo@iu.edu.*
- *Siqi Hong, Indiana University. E-mail: siqihong@iu.edu.*
- *Yu-Chen Huang, Indiana University.. E-mail: yuchhuan@indiana.edu.*
- *Olivia Wikle, Indiana University. E-mail: owikle@iu.edu.*

Manuscript revised 24 April 2017.
For information on obtaining reprints of this article, please send *e-mail to: owikle@umail.iu.edu.*

ORCID identifiers. Additionally, these visualizations will spark further research questions, as they have the potential to alert the client team to particular trends that will require future qualitative analysis. Ultimately, THOR will report on the relevant findings that result from our visualization, thereby helping ORCID to adjust its system accordingly and increase its relevancy to academic research.

1 Data Sets Used

Two data files were used for the following visualizations. The first file, which we utilized for most of our visualizations, was a csv file. This pre-processed dataset, entitled "Matched ORCID data for MOOC," was created by our client, who informed us that the following procedures were undertaken to produce it: From a raw JSON and XML dataset containing a large amount of data on records in the ORCID registry, all ORCID Person IDs that were associated with registered ORCID Publication IDs were extracted. This group of Person IDs and associated Publication IDs was further reduced to those publications were also registered in the database *Web of Science*, an online citation indexing service maintained by Clarivate Analytics that includes citations for specialized subject categories within an academic discipline. If a publication was found to exist both in ORCID and in *Web of Science*, the publication and its relevant Person ID were kept in the dataset, and the publication's *Web of Science* categories were added to the dataset. Those ORCID publications that were not cross-listed in *Web of Science* were removed from the dataset. This means that the "Matched ORCID data for MOOC" csv file contains data on a subset of ORCID IDs that are in the ORCID registry: it contains information on the subject categories of publications that are registered in both ORCID and *Web of Science*, along with the relevant Person IDs of the researchers who created each publication.

"Matched ORCID data for MOOC" contains data on 185 countries and 251 *Web of Science*-defined subject categories. The data was collected over the years 2012-2016, and is separated into 16 temporal quarters. The quarters are labeled Q1-Q16, and are specified as follows:

Q1: 2012 Oct-Dec
Q2: 2013 Jan-Mar
Q3: 2013 Apr-Jun
Q4: 2013 Jul-Sep
Q5: 2013 Oct-Dec
Q6: 2014 Jan-Mar
Q7: 2014 Apr-Jun
Q8: 2014 Jul-Sep
Q9: 2014 Oct-Dec
Q10: 2015 Jan-Mar
Q11: 2015 Apr-Jun
Q12: 2015 Jul-Sep
Q13: 2015 Oct-Dec
Q14: 2016 Jan-Mar
Q15: 2016 Apr-Jun
Q16: 2016 Jul-Sep

The dataset includes information on how many publications were registered by each researcher to his or her ORCID account each quarter (in other words, how many new ORCID Publication IDs were added each quarter), as well as how many ORCID IDs have been created in each country as of each quarter. Each quarter is cumulative, adding the number of newly-registered IDs to the number of the preceding quarter (for instance, if Q1 contains five IDs, and three new IDs were added in Q2, then the value of Q2 is eight IDs). The column labels in our dataset, from left to right, are defined as follows:

Column 1 = "**wosid**": Identification number of the *Web of Science* category
Column 2 = "**woscat**": *Web of Science* category
Column 3 = "**cid**": Country name identification number
Column 4 = "**country**": Country name
Columns 5-20 = "**persorcid1-16**": Cumulative total number of Person ORCID IDs with works matched in *Web of Science* for each quarter, based on the creation date of their ORCID ID
Columns 21-36="**pubsorcid1-16**": Cumulative total number of Publication ORCID IDs that also match to subject categories in *Web of Science* for each quarter, based on the creation date of the ORCID Publication ID within the ORCID registry

	A	B	C	D	E	F	G
1	wosid	woscat	cid	country	persorcid1	persorcid2	persorcid3
2	72	evolutionary biology	2	Albania	1	1	1
3	88	entomology	2	Albania	1	1	1
4	2	automation & control systems	3	Algeria	1	2	2
5	42	chemistry - physical	3	Algeria	1	2	4
6	75	energy & fuels	3	Algeria	1	1	4
7	76	engineering - multidisciplinary	3	Algeria	1	3	4
8	79	engineering - chemical	3	Algeria	1	2	3
9	86	engineering - electrical & electronic	3	Algeria	1	3	6
10	119	instruments & instrumentation	3	Algeria	1	2	2
11	137	materials science - ceramics	3	Algeria	1	1	1

Figure 3: Screenshot of "Matched ORCID data for MOOC" csv file

This dataset indicates that, as of Q16 (Jul-Sep 2016), the cumulative total of Person ORCID IDs across all subject categories equaled 2,952,719. It is important to understand that this number does not represent the total number of Person ORCID IDs created in the ORCID registry as of Jul-Sep 2016. Rather, the number 2,952,719 represents the total number of researchers who, from Q1 (Oct-Dec 2012) to Q16 (Jul-Sep 2016), had created ORCID Person IDs for themselves and had at least one publication registered not only under an ORCID Publication ID, but also under a *Web of Science* (WoS) ID that matched to a WoS subject category. In other words, more ORCID IDs had been created as of Q16 than are in this dataset, but not all fit the criteria of being associated with registered ORCID publications that were cross-referenced in WoS and assigned WoS subject categories.

The second file that we utilized was an XML file provided to our client by Clarivate Analytics, the company that oversees *Web of Science* and provides data analytics to customers. Clarivate compiled the data in this file by matching ORCID publications that are registered both in ORCID and in the *Web of Science* database with information about the countries of residence of the authors who collaborated on those publications. This dataset therefore only includes data for those ORCID publications which were registered in both ORCID and *Web of Science* as of Jul-Sep 2016, and does not contain data for every publication ever registered through ORCID. The file is labeled "PubCountry" (indicating publication country), and includes records showing connections between ORCID Person ID, ORCID Publication ID, *Web of Science* (WoS) Publication ID, and a list of countries that collaborated on each individual publication. For clarification, the data layout is represented in Figure 4 below. From left to right,

Column 1 = ORCID Person ID
Column 2 = ORCID Publication ID
Column 3 = WoS Publication ID
Column 4 = Countries where the researchers who collaborated on the publication reside

```
0000-0001-5000-0736|11557873|000328099000017|SPAIN;PORTUGAL
0000-0001-5000-0736|11557875|000329552000008|SPAIN;IRAN;PORTUGAL
0000-0001-5000-0736|11557878|000337748800003|SPAIN;FRANCE;PORTUGAL
0000-0001-5000-0736|11557880|000337748800016|GERMANY;PORTUGAL
0000-0001-5000-0736|11557881|000313064100010|SPAIN;PORTUGAL
0000-0001-5000-0736|11557882|000309573600009|SPAIN;NORWAY;GERMANY;PORTUGAL;CANADA
0000-0001-5000-0736|11557883|000309573600008|GERMANY;PORTUGAL
0000-0001-5000-0736|11557884|000307802900002|SPAIN;PORTUGAL;AUSTRALIA;CANADA
0000-0001-5000-0736|11557885|000299979500010|GERMANY;PORTUGAL
```

Figure 4: Screenshot of "PubCountry" XML file

Figure 4 shows a number of different publications on which the researcher with the ORCID ID 0000-0001-5000-0736 collaborated. From the countries listed for each publication record in Column 4, one can make the inference that this researcher resides in Portugal, and has collaborated on *Web of Science* publications with researchers from Germany, Spain, Australia, Canada, France, and Iran. Multiple ORCID Person IDs (Column 1) may be connected to one WoS Publication ID (Column 3).

2. Visualization 1: *Web of Science* Categories

Using the "Matched ORCID data for MOOC" dataset, we created a pivot table employing the columns "country" and "pubsorcid16." We sorted the pivot table from the highest value in "pubsorcid16" to the lowest. The results of this sorting allowed us to extract the five countries with the highest number of ORCID-registered publications. The five countries were (in order of highest number of ORCID publications to lowest) Italy, Spain, USA, UK, and Australia.

Wishing to create a visualization showing the top WoS subject categories in which each of these five countries had the most registered publications, we generated a pivot table using the columns "country," "woscat," and "pubsorcid16." The pivot table was arranged so that "country" and "woscat" were rows, and "pubsorcid16" was a value, showing the total number of publications in each category and for each country. The resulting pivot table looked like this:

Row Labels	Sum of pubsorcid16
Afghanistan	4
food science & technology	1
physics - multidisciplinary	2
surgery	1
Albania	74
automation & control systems	1
biophysics	1
biotechnology & applied microbiology	2
chemistry - multidisciplinary	1
clinical neurology	2
computer science - information systems	1
computer science - theory & methods	1
ecology	1
engineering - civil	1
engineering - electrical & electronic	10
engineering - manufacturing	1
engineering - mechanical	2
engineering - multidisciplinary	1
entomology	2
ethics	1
evolutionary biology	1

Figure 5: Pivot Table Created from "Matched ORCID data for MOOC" dataset

Each of the five top countries were selected, and the WoS categories for each country were sorted from highest to lowest according to the sum of each category's publication count as of Q16 ("pubsorcid16"). The top five subject categories for the top five countries were combined into one Excel file, which was then imported into Tableau, where the Dimensions "country" and "woscat" were added to Columns, while the Measure "pubsorcid16" was added to Rows. The following visualization resulted, arranged from lowest publication count on the left to highest publication count on the right, both between countries and between subject categories.

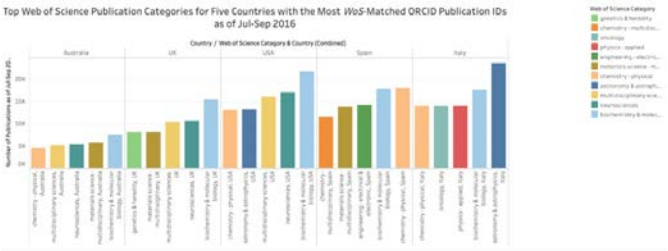

Figure 6: Top *Web of Science* Publication Categories for Five Countries with the Most *WoS*-Matched ORCID Publication IDs as of Jul-Sep 2016

[Visualization URL]

It is interesting to note that the only subject category that is common across all five countries is "biochemistry & molecular biology," which is the highest-ranking category in Australia, the UK, and the USA, and second-highest in Spain and Italy. Other categories which are common but are not universally ranked across the top five countries are "chemistry-physical" and "materials science - multidisciplinary." Spain seems to have the most unique combination of top categories: "engineering - electrical & electronic" and "chemistry - multidisciplinary" are found in Spain's top five categories but remain absent from the top five categories of Australia, the UK, the USA, and Italy.

3.1 Visualization 2: Person ORCID IDs

Line graphs were implemented to show the change in number of ORCID Person IDs and ORCID Publication IDs in our dataset over time. In order to do this, we first had to decide on which time periods we would use. We wanted to show the count of Person and Publication IDs during at least one point each year, over the years 2012-2016, and we set about discerning which quarters in our data would best accommodate this kind of representation. Because our data begins with the quarter Oct-Dec 2012, and ends with Jul-Sep 2016, in order to keep our temporal intervals equal between each year, we had to omit either 2012 or 2016 from our visualization. We decided to omit 2012, and used the following quarters for each year from 2013 to 2016:

2013: Q4 (Jul-Sep 2013)
2014: Q8 (Jul-Sep 2014)
2015: Q12 (Jul-Sep 2015)
2016: Q16 (Jul-Sep 2016)

To create a visualization showing the number of ORCID Person IDs with matched publications in WoS added to the ORCID registry each year over the years 2013-2016, pivot tables were created from the "Matched ORCID data for MOOC" dataset. The first pivot table was comprised of the columns "country" and "persorcid4," and was arranged from the highest "persorcid4" value (the number of ORCID Person IDs as of Q4 in our dataset) to lowest.

Some researchers in our dataset do not have a country associated with them; their country is instead represented in the dataset as "Unknown." The Unknown category is by far the largest throughout quarters 1-16. In Q16, the country category Unknown was associated with 1,920,755 ORCID Person IDs in our dataset, while the next country category, USA, was associated with 121,215 ORCID Person IDs. This is a huge difference, especially considering that the total number of all ORCID Person IDs with publications matched to WoS subject categories as of Q16 was 2,952,719; over half of the ORCID Person IDs in our dataset do not have a country attached to them. In the analysis that follows, whenever running visualizations concerning countries using the "Matched ORCID data for MOOC" file, we have used only on the 1,031,964 ORCID Person IDs in our dataset that are associated with a country. After the Unknown country category was removed from the sorted data, the ten countries with the highest numbers of ORCID Person IDs as of Q4 were extracted and put into a separate Excel file. This process was repeated using the columns "persorcid8," "persorcid12," and "persorcid16" in place of "persorcid4." The data collected from these pivot tables was used to create a new Excel file made up of counts for persorcid 4, 8, 12, and 16 for each of the top twenty countries. This aggregated file was uploaded to Tableau. The Dimension "year" (2013-2016) was placed into Columns and the Measure "SUM(persorcid)" was placed into Rows, producing the following visualization:

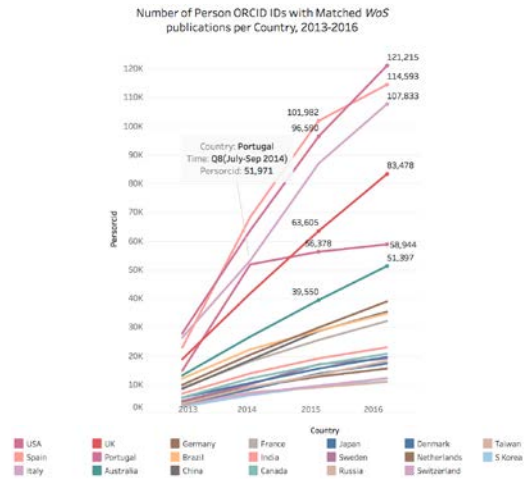

Figure 7: Person ORCID IDs with Matched *WoS* Publications per Country, 2013-2016

[Visualization URL]

3.2 Visualization 2: Publication ORCID IDs

The process described above was undertaken again to produce a line graph showing the number of ORCID Publication IDs with matched records in WoS added to the ORCID registry over the years 2013-2016, except that the columns persorcid 4, 8, 12, and 16 were replaced with pubsorcid 4, 8, 12, and 16.

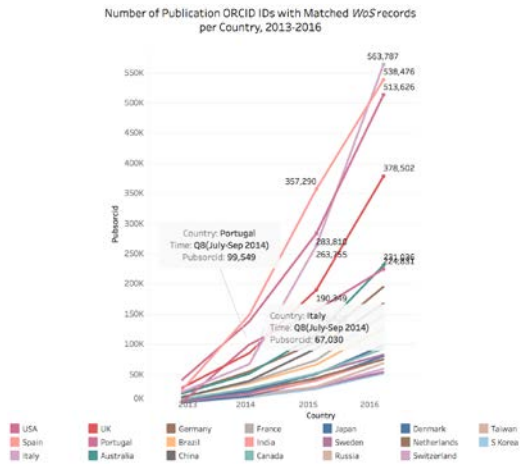

Figure 8: Publication ORCID IDs with Matched *WoS* Records per Country, 2013-2016

[Visualization URL]

In the line graph displaying the countries with the most Person IDs within the "Matched ORCID data for MOOC" dataset, there is a steady growth of Person ORCID IDs with matched WoS publications in Spain, the USA, Italy, and the UK from 2013-2016. In 2014, Portugal's initial growth rate of Person ORCID IDs with matched WoS publications dropped to a lower growth rate. The USA, Spain, and Italy experienced a slackened growth rate in Person ORCID IDs with matched WoS publications beginning in 2015, while the growth rate of Person IDs in the UK has remained more constant over the years 2013-2016.

The top four countries in the line graph displaying the countries with the most Publication IDs within the "Matched ORCID data for MOOC" dataset are the same as those with the most Person IDs: Italy, Spain, USA, and the UK. With a total of 378,502 publications added to the ORCID registry and WoS in 2016, the UK hovers in the space between the top three countries and the rest of the countries. Portugal does not experience a severe drop-off in publication additions over these four years, as it did for person additions in 2014. It is interesting to note that the top six countries for both the number of Person and Publication IDs are the same across all four years. However, the ranking between these six countries varies; for instance, Portugal has more publications added in 2015 than Australia, but Australia has more publications added in 2016 than Portugal. To reiterate our explanation of the data used above, it is essential to remember that this information is only representative of our dataset, which is a subset of all the scholars and publications currently registered with ORCID.

4. Visualization 3: Geospatial View of Person & Publication ORCID IDs

The Excel files created to produce the line graphs above were also used to produce geospatial visualizations via Tableau. To summarize the pre-processing procedure that was described above, these files utilize data from quarters 4, 8, 12, and 16 to represent the sum of ORCID Person IDs and ORCID Publication IDs in the "Matched ORCID data for MOOC" dataset during Jul-Sep of 2013, 2014, 2015, and 2016. To produce these geospatial visualizations, we did not extract the top twenty countries with the highest number of ORCID Person and Publication IDs. Rather, we represented the growth of Person and Publication IDs for all countries in our dataset over the years 2013-2016. Two Excel files were uploaded to Tableau: one file with the sum of Person ID count for each country in our dataset and one file with the sum of Publication ID count for each country in our dataset. In Tableau, a geospatial layout was selected, and the Dimension "Year" was placed in Columns, along with the Measure "Longitude." The Measure "Latitude" was placed in Rows. The proportional circle size was adjusted to represent the number of Person IDs and Publication IDs, and the circle colors were adjusted to represent different countries. Tableau automatically recognized the longitude and latitude for most of the countries in the dataset, but for those that it did not recognize, longitude and latitude were manually added.

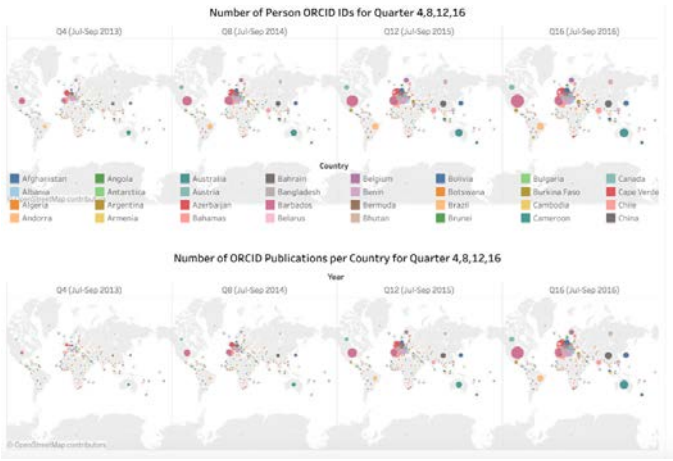

Figure 9: ORCID Person and Publication IDs, 2013-2016

[Visualization URL]

Though this geospatial visualization displays similar information to the line graphs above, its interactive nature allows the user to understand the growth of Person IDs and Publication IDs on a global scale, lending itself well to visual comparison between global areas over four discrete time periods. One can easily view, for example, the large growth in and Person and Publication IDs in Western Europe and North America over the years 2013-2016.

5. Visualization 4: Country Collaborations

To produce our final visualization, we used the "PubCountry" XML file provided by Clarivate, described in part 1 of this paper. To discover which countries had researchers that collaborated with one another the most, we imported the XML file into Excel, creating a pivot table with Column 4 (collaborating country names) as a Row and the Distinct Count of Column 3 (WoS Publication ID) as a Value. This newly-created table displayed the number of collaborations that have taken place between each combination of countries (the screenshot of this file in Figure 10 shows Afghanistan and England have a total of 11 collaborations between them):

Row Labels	Distinct Count of Column3
(blank)	40161
AFGHANISTAN	7
AFGHANISTAN;AUSTRALIA	4
AFGHANISTAN;AZERBAIJAN;TURKEY	1
AFGHANISTAN;BARBADOS;BRAZIL;INDONESIA;SOUTH KI	1
AFGHANISTAN;BELGIUM	1
AFGHANISTAN;BELGIUM;NIGER;USA;PAKISTAN;BURUNI	1
AFGHANISTAN;BELGIUM;USA;DEM REP CONGO;CANAD	2
AFGHANISTAN;BELGIUM;USA;DEM REP CONGO;PAKIST	1
AFGHANISTAN;BELGIUM;USA;PAKISTAN;CONGO;CENT A	1
AFGHANISTAN;BELGIUM;ZAIRE;SYRIA;USA;HAITI;PAKIST	1
AFGHANISTAN;BRAZIL;ENGLAND;DENMARK;TANZANIA;	1
AFGHANISTAN;BRAZIL;PAPUA N GUINEA;INDONESIA;SO	1
AFGHANISTAN;BRAZIL;PAPUA N GUINEA;INDONESIA;SO	1
AFGHANISTAN;BRAZIL;PAPUA N GUINEA;SOUTH KOREA;	1
AFGHANISTAN;BRAZIL;SOUTH KOREA;SLOVAKIA;AUSTRI	1
AFGHANISTAN;BRAZIL;ZIMBABWE;AUSTRIA;SINGAPORE	1
AFGHANISTAN;CANADA	1
AFGHANISTAN;CHILE;IRAN	1
AFGHANISTAN;CHILE;MEXICO;USA;PERU;JAPAN	1
AFGHANISTAN;ENGLAND	11
AFGHANISTAN;ENGLAND;FRANCE;NETHERLANDS	1
AFGHANISTAN;ENGLAND;GERMANY	1

Figure 10: Screenshot of "PubCountry" XML File Imported into Excel

We sorted the data to find the top 20 country combinations with the highest number of collaborations, and added these 20 countries and their respective numbers of collaboration to a new Excel file. We imported this file into Sci2, and used Bing geocoder to find the Latitude and Longitude of each of these countries.[11] Using this data, we manually created an NWB file with 12 nodes and 20 edges

(representing each of the top 20 country collaborations). This NWB file was imported into Sci2, and a geospatial network layout with base map was created by selecting World Map, then selecting the correct latitude and longitude columns. The resulting .graphml network file was imported into Gephi, then Gephi was used to generate a PNG file of the network. We imported the base map that was generated by Sci2 and the network that was generated by Gephi into Adobe Photoshop, where we were able to combine the base map and network. The combined images were then saved as a PNG image file, which was imported into InkScape to create appropriate legends for the visualization. Using the Excel file containing the top 20 collaborations, we created a bar graph to add to our Sci2 network visualization, in order that viewers might have more information about the number of collaborations between each country.

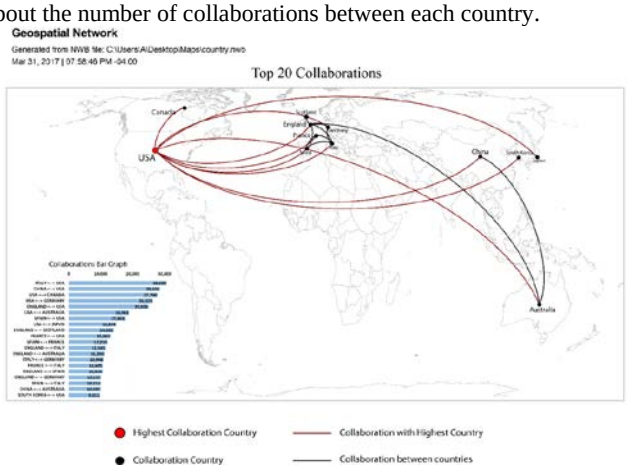

Figure 11: Top 20 Country Collaborations, as of Q16 (Jul-Sep 2016)

[Visualization URL]

This visualization shows that, according to the dataset "PubCountry" provided by Clarivate, which contains data on only those Publication IDs that are registered in both ORCID and *Web of Science*, the USA is emerges as the country with the most collaborations with other countries. Those collaborations in which the USA is not involved have been represented in the networks with black edges. Those in which the USA is involved are represented with red edges. The top five collaborations range from 25,025 to 30,689. Interestingly, there is a large margin between these top five collaborations and the next highest collaboration (between the USA and Australia, at 18,961).

CONCLUSION

One challenge for us in this project was in trying to find a way to visualize data regarding collaborations between researchers and countries. The scale of the "PubCountry" XML file from Clarivate was massive, containing over 7 million records. Such a large dataset prevented us from being able to visualize collaborations between individual Person ORCID IDs in the timeframe available for this project. Instead, we settled on creating Visualization No. 4 to show the countries with researchers that collaborated with one another the most, a visualization that required less data preprocessing and thus was more conducive to our time frame.

Overall, the variety of visualizations presented here work together to show many characteristics of a subset of data from the ORCID registry: specifically, data of ORCID-registered researchers who had ORCID-registered publications that were also registered in the *Web of Science* database, and were assigned *Web of Science* subject categories. The connection to WoS categories proves useful in interpreting and better understanding this community of ORCID-users. From this dataset we see, for instance, that when ORCID-registered researchers register publications in both ORCID and WoS, those publications are most likely to fall under scientific categories, such as biochemistry and molecular biology (Figure 6). Wos includes humanities and social science disciplines as well as STEM disciplines, yet non-STEM subject categories are not often assigned to the publications in this subset of ORCID data, and certainly do not fall within the top subject categories in our dataset.

As our line graphs and geospatial visualizations show, ORCID Person and Publication ID registration has been consistently highest in Western countries over the years 2013-2016 within this subset of data. The subtle shifts in the order of these high-ranking countries over the years may provide insight when considered in light of the economic and political context in which they occurred. For instance, it may be interesting to consider what happened in Portugal after Jul-Sep 2014 that made the rate of ORCID-registered Person IDs fall (Figure 7).

Our Sci2-generated geospatial network also provides valuable information about the subset of ORCID users in our dataset. THOR could possibly use this dataset to gain a better understanding of in which countries international collaborations are taking place in relation to those researchers with publications registered in both ORCID and WoS. Perhaps the knowledge that, within this subset of ORCID-users, collaborations with the USA are most common will spur outreach efforts by THOR to increase the number of ORCID-registered publications produced by collaborations between countries other than the USA.

In the future, we propose that a necessary further step in the analysis of ORCID IDs should include a move toward understanding more about the ORCID researchers themselves. We were not given data that describes qualities such as gender or ethnicity of ORCID researchers, but if this data could be obtained, visualizations could potentially be used to gain more insight into trends behind research and collaboration.

ACKNOWLEDGMENTS

The authors wish to thank collaborators Robin Desler, of THOR; Tom Demeranville, of ORCID; and Ade Deane-Pratt, of ORCID.

REFERENCES

[1] L. L. Haak. Persistent identifiers can improve provenance and attribution and encourage sharing of research results. *Information Services & Use*, 34:93-96, 2014.
[2] A. Meadows. Everything you ever wanted to know about ORCID… but were afraid to ask. *College & Research Libraries News*, 23-30, 2016.
[3] Ibid.
[4] L. L. Haak. Persistent identifiers can improve provenance and attribution and encourage sharing of research results. *Information Services & Use*, 34:93-96, 2014.
[5] Ibid.
[6] Ibid.
[7] A peek at the ORCID registry public data file. (2013). *The Data Science Lab*. https://datasciencelab.wordpress.com/2013/12/05/a-peek-at-the-orcid-registry-public-data-file/. Retrieved March 13, 2017.
[8] Ibid.
[9] Ibid.
[10] Ibid.
[11] Sci2 Team. Science of Science (Sci2) Tool. [Internet]. Indiana University and SciTech Strategies; 2009. Available: http://sci2.cns.iu.edu.