RICardo World Trade Web, 1834-1938

Béatrice Dedinger, Sciences Po, Center for History (CHSP), Paris, France
Paul Girard, Sciences Po, médialab, Paris, France

1. Introduction

RICardo is a project initiated at the beginning of the 21th century with the aim of constructing a database of bilateral trade covering all of the world's countries over the nineteenth and mid-twentieth centuries. Nearly fifteen years have passed before the first results of our work could be published. The final RICardo project includes a database (~300,000 data; 12.2017 version) and a website to explore the data (http://ricardo.medialab.sciences-po.fr/#/). Contrary to the usual procedure, we did not try to publish a research paper before making our work available to the public. Given the time we have spent on this project, it was most important for us to publish our work as both a dataset and an exploration tool. We can now turn to the analytical part of the undertaking.

The ultimate goal of our current research is to exploit the entire database to give an overview of the geographical distribution of world trade over the period 1834-1938. The question is simple: who trades with whom? Surprisingly, this question has rarely been raised, and for the first time with a database of this magnitude. What we currently know about the geography of trade in the nineteenth century is rather confused, if not contradictory. Bairoch was the first scholar to have built a bilateral trade database. According to his study, the geographical structure of European foreign trade during the nineteenth century – as described by trade shares – reflects "a preponderance of inter-European trade and trade between developed regions".[1] In 1993, Anderson & Norheim published many papers that tap the trade-intensity and trade-openness indicators to demonstrate the preponderance of inter-continental over intra-continental trade for a hundred or so years from 1830.[2] In their vast inquiry on the history of world trade, Findlay & O'Rourke describe the international economy of the nineteenth century as fundamentally different from what had gone before, this break translating in an intercontinental commodity price convergence that points to a globalizing world.[3] More recently, Fouquin & Hugot have built a large trade dataset (1827-2014) to estimate the distance elasticity of trade over the last two centuries. They conclude that the First Globalization was fuelled by a relative intensification of short-haul trade, in other words, by a process of regional-biased internationalization.[4] New approaches have been emerging over the last decade, which apply network analysis to the study of international trade. A group of physicists, using methods of network geometry, has created the World Trade Atlas 1870-2013 that points to a hyperbolic world.[6] This leads to a greater importance of distance due to trade networks becoming more

[1] Bairoch (1974).
[2] Anderson & Norheim (1993a); (1993b); (1993c).
[3] Findlay & O'Rourke (2007).
[4] Fouquin & Hugot (2016).
[6] García-Pérez, Boguñá, Allard & Serrano (2016).

hierarchical (small economies are moving away) and to a tendency towards trade regionalization. In the same vein, De Benedictis et al.,[7] considering the period 1950-2000, estimate that the world is still far from being fully connected. In the end, we do not really know what the geographic distribution of world trade looked like in the aftermath of the Napoleonic wars, nor how it has evolved until the next great war.

It must be stressed that the methodology used to measure or estimate the pattern of trade globalization is of key importance. The conclusions of these selected works differ according to the way the process of integration is being portrayed. The network approach seems better suited to gain knowledge about the structuring process of international trade over the nineteenth and mid-twentieth centuries. Our ambition is to describe this early globalization process from the data we built without aiming at producing a model. When we started considering the dataset that could be relevant to answer our question with the network method, we faced a number of issues inherited from the choices we made in the elaboration of the RICardo base. Indeed, the RICardo database is not "perfectly cleaned". We have decided to keep some of the imperfections that are unavoidable in historical bilateral trade statistics.[8] Of course, a number of filtering choices have been implemented to make the database manageable, but we stayed as close to the sources as possible because we think that imperfections themselves are a valuable information, telling a lot about the way the data were built. The result is a great complexity of the database. This complexity must be investigated before being able to move forward into aggregating the data to support quantitative analysis.[9] That is why our presentation will focus on this preliminary and critical stage of our research. We would like to share with you the complexity issues we are facing and the process of simplification we have devised to handle them.[10] This will come after a brief presentation of the RICardo dataset.

2. The RICardo database in brief

The first data were collected in 2004. By chance, we found an archive source at the *Bibliothèque nationale de France* that provides a quantitative overview of countries' bilateral trade as early as 1829. The discovery of the *Extraits d'Avis Divers* was the starting point of the RICardo project. From then until the end of 2017, 400,000 data have been collected from a variety of sources, a process of homogenization and conversion of data has been elaborated, and a website has been created.[11] We will focus here on some major characteristics of the RICardo database.

---

[7] De Benedictis L. & L. Tajoli (2011).
[8] See Dedinger & Girard (2017).
[9] Drucker (2011).
[10] Note that this process of simplification reduces the precision of our data but it is necessary to amplify the signals we are trying to analyse. Besides, we keep track on how we did it so that we can go back and challenge our quantitative findings with the closer-to-the-source data. Cf. Latour (1993).
[11] For more details about the realisation of the project, see Dedinger & Girard (2017).

- *RICardo compared*

The RICardo database includes around 300,000 data points in the last version (December 2017). Each point is a bilateral/total trade flow of exports or imports. This is currently the most exhaustive trade database dedicated to historical bilateral trade statistics. The experiment closest to RICardo is the Correlates of War (COW) trade data elaborated by Barbieri, Keshk, and Pollins.[12] However, there are two important differences between the two databases. First, the Barbieri database focuses on 1870-2014, it gathers more than 1,700,000 bilateral trade flows, and 90 % of all observations are for the post-1938 period. In RICardo, the time span is quite different. The purpose is to know better the pre-1870 period by tapping new archives that were largely unexplored at the time we began the project. Second, COW records around 60 countries over the period 1870-1938. As we will see in this presentation, the world depicted by RICardo is much more sophisticated. That is why we chose the term 'entity', better suited than 'country', for labelling the many kinds of territorial entities. In its last version, there are 1,713 entities in the RICardo entities list.

- *Sources*

Three types of sources are identified in RICardo. The primary sources are the original sources, the customs returns published by the national statistical offices. The secondary sources are compilations of primary sources. For the pre-1850/60 period, they include the French series '*Annales du commerce extérieur*'. For the post-1860 period, the two British Statistical Abstracts series (*Statistical Abstract for the principal and other foreign countries*, *Statistical Abstract for the several colonial and other possessions of the United Kingdom*) are the main secondary sources. The third type of source, named 'estimation', concerns total trade data. It includes works that re-estimate historical total trade series or publish historical total trade series (such as national historical abstracts). Total trade series provided by Mitchell's *International Historical Statistics* are also included in this category.

There are two remarks regarding the sources. First, when we began the project, our guideline has been to prioritize secondary sources. In fact, considering human and financial constraints, the ambition of collecting only primary sources was not realistic. The main flaw of this strategy appeared some years later when we drew an overview of the database. There was a significant decrease in the number of total observations over the interwar period, mainly due to the use of the League of Nations's *Mémorandum sur le commerce international et sur les balances des paiements*. So we decided to return to the digitization of trade archives and the collection of trade data. This is still a work in progress. Second, the heterogeneity of sources can accentuate the heterogeneity in the name and type of entities recorded in the RICardo database. Since this problem of heterogeneity is the main focus of our presentation, it will be developed in more details thereafter.

[12] Barbieri & Keshk (2016); Barbieri, Keshk and Pollins (2009).

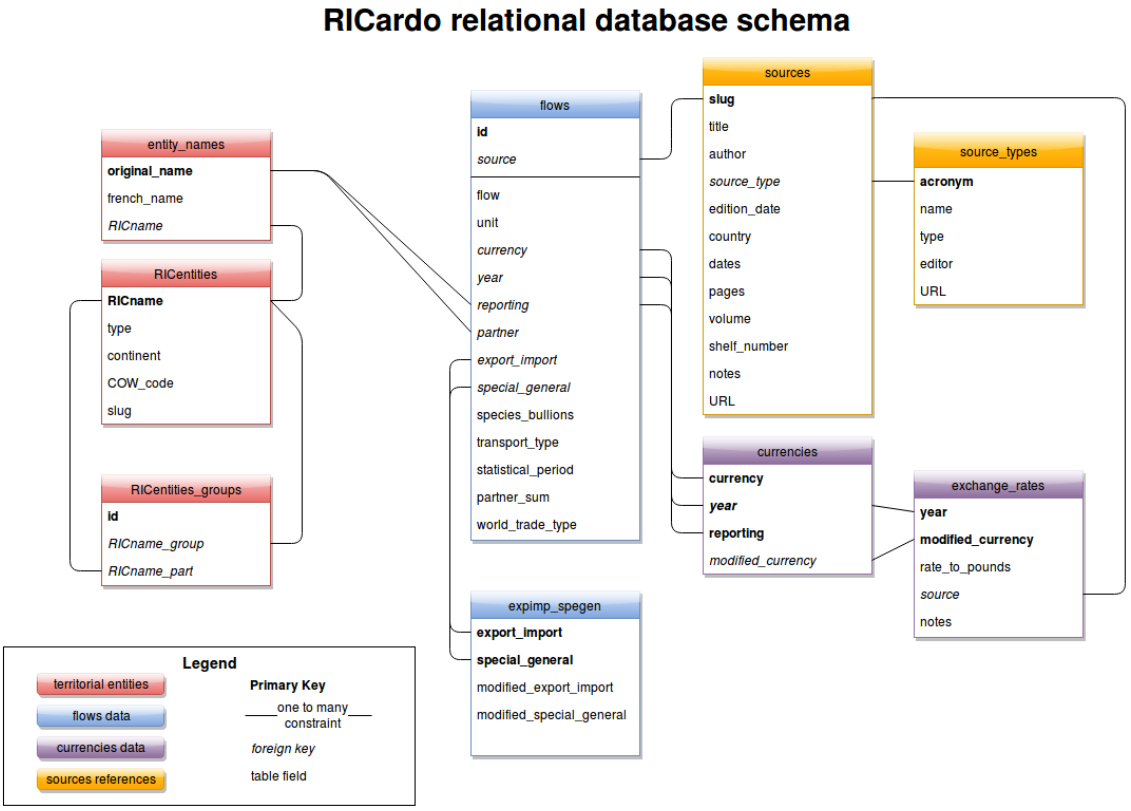
- *Data homogenization*

The RICardo database assembles tens of thousands of data collected from a variety of sources over more than a hundred years. Data are trade flows expressed in a host of currencies. Each trade flow implies two actors. Originally, the total number of entities in the database amounted to more than 4,000. It was a real mess (see an example in the table 1). We have created a relational database linking the core 'flows' table (original data) to the 'entities', 'currencies', and 'sources' tables where each type of information is homogenized, converted, and referenced (table 2). The standardization process we have developed resulted in a sharp reduction in the number of entities, to less than 2,000 (2017 version). Note that for the conversion of trade data into a common currency unit, the pound sterling, we have built a new exchange rate database.

Table 1. Extract of the RICentity table

original_name	RICname
africa - east coast	East Africa
africa (east coast)	East Africa
africa east coast	East Africa
africa- oriental coast	East Africa
africa, east coast	East Africa
afrique orientale	East Africa
côte orient. afrique	East Africa
côte orientale afrique	East Africa
côtes orientales d'afrique	East Africa
east africa	East Africa
east coast of africa	East Africa
eastern coast of africa	East Africa
foreign east africa	East Africa
katch	East Africa
afrika (oostkust)	East Africa

Table 2



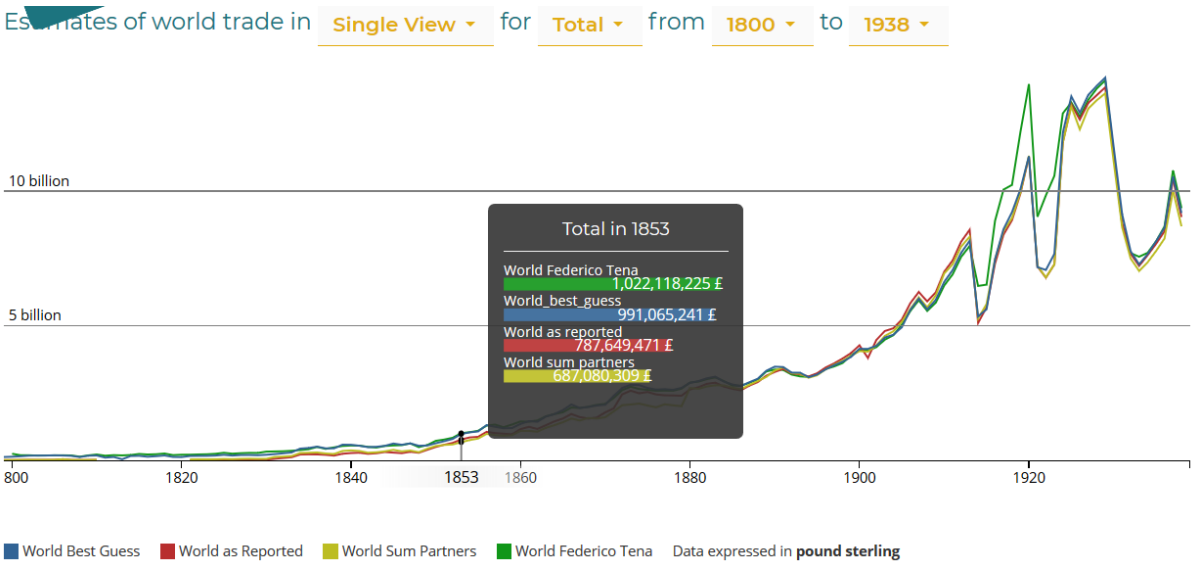
- *A bilateral and total trade database*

The RICardo database combines two sets of data, total trade data and bilateral trade data. The user has access to comprehensive series of historical total trade data, that is, total exports and imports of world's countries over c.1800-1938. In the RICardo vocabulary, total exports/imports of a reporting entity are assimilated to bilateral trade flows with the entity partner 'world'. Two types of partner 'world' co-exist in the base. 'World as reported' refers to data extracted from primary or secondary sources and to the line 'Total' in the trade tables of these sources. 'World estimated' coincides with total trade data extracted from sources classified as 'estimation' that provide only countries' total trade and that we have tapped to complement countries' total trade series. Note that the differentiation between 'reported' and 'estimated' does not exist for bilateral trade flows. In the current version of RICardo, there are 22,082 total (exports and imports) trade flows (reported + estimated) and 272,731 bilateral (exports and imports) trade flows from 1800 to 1938. The problem of complexity is a matter only for bilateral trade data.

The RICardo website (*World* view) provides a comparison between different world trade estimations. In addition to the '*World as reported*' series, three other series are available. '*World Sum Partners*' is total trade as obtained from the addition of all partner entities (except World) trade flows of the database. '*World best guess*' selects the "best" available data for a reporting/year on a priority basis (first 'World estimated', second 'World as reported', third 'World sum partners'). '*World Federico & Tena*' traces their new estimates of world trade that

are the most elaborated estimation of world trade to this day.[13] As can be seen in the graph below, the RICardo and Federico & Tena trade series are very similar (except over 1914-1923). Besides, the Federico & Tena database assembles an average number of polities much smaller than the number of entities of the current version of the RICardo base. We will use it to gauge the degree of quality of our simplified list of entities (see point 5).

Figure 1. World trade series 1800-1938



Source: RICardo website

3. Complexity of the RICardo trade matrix

Studying the geographical structure of world trade over the nineteenth and mid-twentieth centuries would require a perfect bilateral trade matrix, that is, a matrix providing bilateral exports/imports of each of the world's countries with each of the world's countries for each year from ca. 1800 to 1938. Unfortunately, the RICardo bilateral trade matrix is far from this ideal-type model due to three main shortcomings: the heterogeneity of entities, the variability in time, the problem of missing data.

- *Heterogeneity of entities*

In historical trade statistics documents, it is quite common to find trading entities that are not 'countries'.[14] To help structure this heterogeneity, we have created five types of entities:

[13] https://www.uc3m.es/ss/Satellite/UC3MInstitucional/es/TextoMixta/1371246237481/Federico-Tena_World_Trade_Historical_Database

[14] The term 'country' has to be understood in the sense of 'sovereign state" as defined in Correlate of Wars' *State System Membership*. Sovereign states are defined by COW as: 1. Before 1920 all political entities with a

city/part of, colonial area, country, geographical area, group (of entities). Through the homogenization process, each entity is given a RICname identified by a number of three variables: type, continent, and COW code (see table 3).[15] Only RICardo entities of the type 'country' are given a COW code. In other words, each entity referenced in the COW list is identified as a 'country'. For those entities which do not show up in the COW database – i.e. 'city/part of', 'colonial area', 'geographical area', and 'group' – new English entity names have been created without a numerical code. Furthermore, new entity names and new codes have been created for three 'countries' that are not included in the COW database: Kingdom of Sardinia (325S), Prussia (255P) and Germany (Zollverein) (255Z).[16]

Table 3. Extract of the RICentity list

RICname	type	continent	COW_code
A Coruna	city/part_of	Europe	
Aden	country	Asia	681
Aden & Algeria & Egypt/United Arab Republic & Iran (Persia) & Morocco & Asian Turkey	group	World	
Aden & Arabia	geographical_area	Asia	
Aden & India	group	Asia	
Antigua and Barbuda	country	America	58
Asian Turkey	city/part_of	Asia	
British Africa	colonial_area	Africa	
British Antilles	colonial_area	America	
British Asia	colonial_area	Asia	
British Colonies	colonial_area	World	

population of at least 500,000 people having entertained diplomatic relations (in the person of at least a chargé d'affaires) with Britain and France, and 2. After 1920, all country members of the League of Nations (later, of the United Nations) or alternatively all entities with a population of at least 500,000 and diplomatic representation with at least two "great powers" (including according to COW: Germany, China, the US, France, Italy, Japan, the UK and Russia-USSR).

[15] The COW project, initiated in 1963 by American political scientists, has collected quantitative information about armed conflicts in the post-Napoleonic period and resulted into the constitution of several databases, two of which are concerned with the definition and inventory of state entities. Thus, the *State System Membership List* contains the list of all entities which have enjoyed the internationally recognized status of sovereign state as of 1815. The *Colonial/Dependency Contiguity Data* variable identifies every contiguity situation (land or river boundaries, or bodies of water) of political entities of the international system (sovereign states, colonies and dependencies) and leads to drawing up a subsidiary list of colonies and dependencies belonging to sovereign states. A ciphered code is attributed to each of these entities with dates of changes of political status. All this information (including entity name and code, political status, and relevant time periods) is a 50 odd-page document entitled *Entities.pdf*; that served as a basis to define the name and code of each 'country' type *RICname*.

[16] Created in 1720, Kingdom of Sardinia (assimilated to Italy/325 in the COW list) consisted, before the unification of Italian states in 1861, of Savoy, Piedmont (Turin), Aosta, Nice and the island of Sardinia. Prussia is assimilated to Germany/255 in the COW list. But before the foundation of the German empire (1871), several German entities appear in the RICardo list of partner names – Prussia, German states, Germany and German Zollverein – which do not correspond to similar territories. Thus, new country names have been created for Prussia and the German Zollverein (which may be considered as an economic union). We have translated 'German states' and 'Germany' into 'Germany' although 'Germany' did not actually exist before 1871.

In the current version of the database (total + bilateral), there are 1,713 RICentities, divided into 573 groups, 425 countries, 395 city/part of, 172 geographical areas, and 148 colonial areas. If only the bilateral dataset is considered, we find 1,428 entities of which 81 % are of the 'group', 'country' and city/part of type. Table 4 separates these entities into 'reporting that are also partner entities' and 'entities that are only partners'. It highlights the fact that the problem of heterogeneity mainly stems from the entities that are only partners. To answer the question "who trades with whom", it is clear that we have to reduce the heterogeneity of entities. This implies that we convert as much 'non-country' entities as possible into 'country' reporting. In this presentation, we propose solutions to disaggregate over-country entities ('group', 'colonial area') and aggregate sub-country entities ('city/part of'). We will not address the two remaining issues brought by 'geographical areas' and countries that are 'only partner'. For the later, the issue is not heterogeneity but a problem of trade structure that is precisely what we would like to address.

Table 4. Type of entities in bilateral trade data

Type of entity	*Entities in bilateral data*		*Reporting also Partner*	*Only partner*
	Nb	%	Nb	Nb
Group	518	36	7	511
Country	366	26	187	179
City/part of	266	19	34	232
Geographical area	147	10	0	147
Colonial area	131	9	1	130
Total	1428	100	229	1199

The problem of heterogeneity seems quite limited when we consider the number and value of flows (table 5). Around, respectively, 80 % and 90 % of the number and of the value of bilateral trade flows refer to a partner 'country'. But as we focus on the network structure of bilateral trade, those flows might well be of a high influence and need to be converted. Besides, the problem of heterogeneity does not depend on the type of sources. The proportion of each type of entity is very similar in primary and secondary sources.

Table 5. Number and value of flows by source and by partner type

(*Number of flows*)

Type of source	*Country*	*Group*	*Colonial area*	*Geographical area*	*City/part of*	*All partners*
Primary	79	7	6	4	4	55
Secondary	84	6	5	3	2	45
Total	81	7	5	4	3	

(Value of flows)

Type of source	*Country*	*Group*	*City/part of*	*Colonial area*	*Geographical area*	*All partners*
Primary	85	9	5	1	1	68
Secondary	91	6	1	1	1	32
Total	87	8	3	1	1	

- *Variability in time*

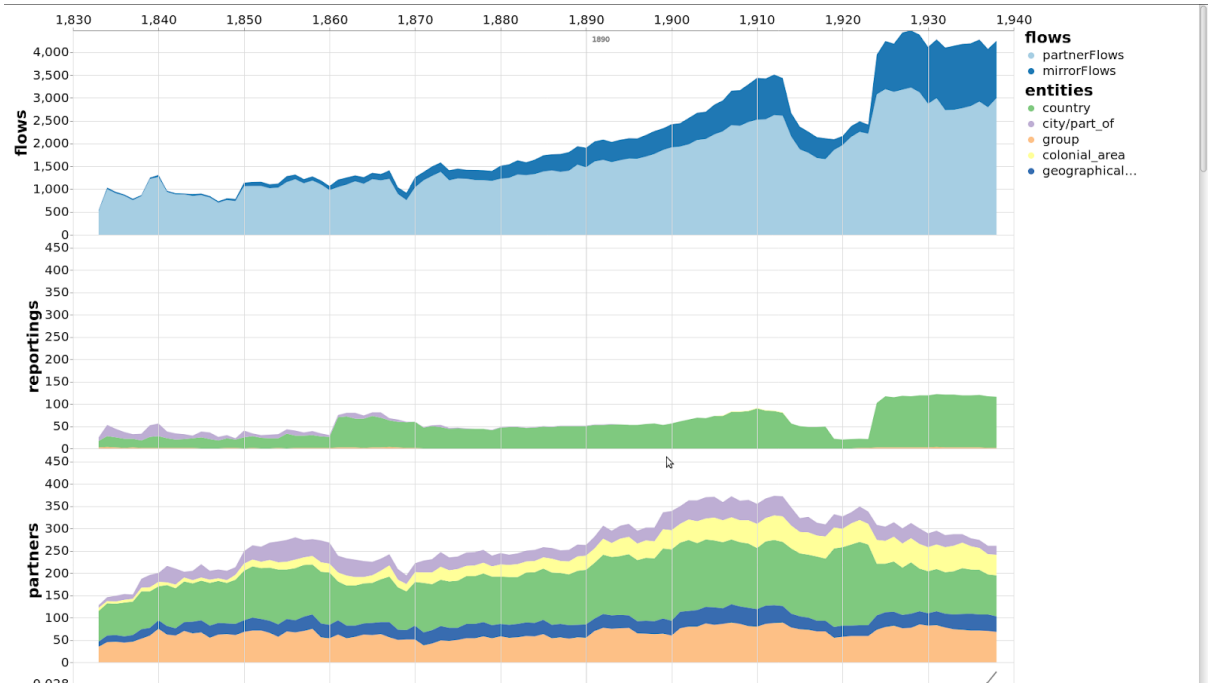
How does this problem of heterogeneity evolve over 1800-1938? Two remarks surface from the graphs below. First, it confirms that the heterogeneity problem concentrates on the partner entities. On the reporting side, 'country' and 'city/part of' are the only types throughout the period.[17] What kind of city/part of could publish trade statistics? Actually, these trade flows emanate from harbors that are the registration areas for trade statistics at the entry/exit of merchandises into/out of a country. For example, reporting cities are Bahia, Rio de Janeiro, Lisbon, Porto, Bilbao, Santander, or Smyrna (Izmir). As can be seen on the graph, the 'city/part of' type disappears from the reporting list around the 1860s (see below). The heterogeneity problem does not disappear on the partner side. The good news is that it remains relatively constant over the whole period, making the search for solutions more practicable.

Second, the graphs point to another flaw of the database: the number of trade flows fluctuates over the period. Again, the ideal-type model should exhibit a constant (and high) number of annual flows from 1800 to 1938. A number of reasons account for the variability in the number of observations. An obvious one is that the work is not yet finished. This applies especially to the 1914-1923 period where the number of flows collapses. Another important reason is the change in the sources used to collect the data. This explains the breaks in 1833/34 and 1839-1840, which coincide with the use of new sources: the *Extraits d'Avis Divers* in 1834, the *Bulletin du commerce extérieur* in 1839. It causes a jump in the number of reported reporting entities that translates into a jump in the number of registered trade flows. Inversely, the adoption of the British statistical abstracts as major secondary sources from 1861 has a great impact on the number of reporting entities – multiplied by 2.5 in 1861/1860 – but a smoothed one on the number of trade flows. Another impact of the British statistical abstracts is the disappearing of the reporting type 'city/part of' that explains the slight decrease in the number of reporting entities over the 1860s-1870s. Historical reality must also be taken into account. Over time, changes may occur in the political definition of a country, countries "disappear", "new" countries appear. Examples are the creation of federations – such as Italy (1860), Canada (1867), Germany (1871) – that reduces the number of world's countries. Or inversely, the disintegration of empires – including colonial empires, but this is not relevant for the RICardo period – that raises the number of world's countries (for instance, disintegration of Austria-Hungary after 1918). Another variable explaining fluctuations in flows' number is the number of partner entities recorded by each reporting entity. The metadata view (see the RICardo website) provides some clarification on this point by allowing a categorization of reporting entities according to the number of partners. For example, from the mid-1880s the number of reporting countries that record more than 50 partners exceeds the number of reporting countries that record less than 10 partners. Eventually, a major reason is the availability of data. Trade records of a large number of the world's countries are not available before the last quarter of the 19th century, either because they were not published, or because

[17] The type 'group' also appears but these are very scarce cases.

we did not find them. This may explain the steady increase in the number of flows' curve from the 1880s until WWI.

Figure 2. Number of flows, number of reporting entities and partner entities by type, 1830-1938



- *Missing data*

The problem is well known to the trade economic historian and the trade statistician.[18] The first question is whether missing data is tantamount to a zero value. The answer is "probably no", when one state does not report a trading relationship and the other does. A second-best option thus seems to substitute a missing bilateral import value with the exporting state's report for the same flow (mirror flow) to help complement the database. This is the option taken by the Barbieri team[19] and the IMF. A different approach has been adopted in RICardo that, unlike Barbieri, focuses on the pre-WWII period where it can be assumed that sources of divergence in mirror flows were more numerous. In RICardo, we recommend that the user takes advantage of the comprehensive character of the base to compare data and ultimately select the more reliable figures. Therefore, in the database supplied to the public, there has been no attempt at reallocating or estimating data. We may wonder if the use of mirror flows could help to offset the variability in the number of observations. The graph above shows that mirror flows help to complement information but this does not alter the general course of the curve.

[18] See the RICardo committee 2017 report for some comments on this question: https://f-origin.hypotheses.org/wp-content/blogs.dir/4050/files/2017/11/RIC_Workshop_2017_Report.pdf
[19] Barbieri & Keshk (2016), 4-8.

To sum up, the shortcomings of the RICardo database are for a great part due to shortcomings in historical trade statistics themselves. This mainly concerns the problem of variability for which no practical solution seems available apart from keeping completing the base. However, this flaw should not hinder the pursuit of our research project. Conversely, we cannot claim to study the geography of trade globalization in a historical perspective if we do not find practical solutions to alleviate the problem of heterogeneity of entities.

4. How to reduce the complexity of the RICardo matrix

One practicable way to come closest to the ideal-type matrix is by converting entities that are not 'countries' into the 'country' type. That is what we are currently working on. We have focused on the types 'group', 'city/part_of', and 'colonial area', leaving aside the type 'geographical area'. Remember that this type counts for 4 % of the number of bilateral flows and 1 % of their value (last version of the database).

- *Converting 'group'*

A 'group' is a partner entity that is composed of a number of entities. Groups were created to reflect sources where a reporting reports trade with a list of countries (or entities).

For a reporting R which trades with group G in year Y, we want to disaggregate the trade of R with each entity En of G. There are three cases:

- If all entities En are reporting in year Y and declare bilateral trade with R:

we can use the bilateral flows to disaggregate the 'group's flows according to the trade shares of En with R. To do so, we calculate trade ratios of each En in the group trade with R [ratio(E) = trade(E with R)/ sum(trade(En with R)]. Then we apply those ratios to the trade of R with G in year Y to recreate bilateral trade flows of R with each En using the figure reported by R. We call this case 'totally disaggregated with ratio'.

- If not all but only some entities En are reporting in year Y and declare bilateral trade with R:

we cannot calculate the ratio of original flows. In such cases, available En mirror flows are subtracted from R's trade with G and respective En are removed from the group G. If there is only one remaining entity composing G then trade of R with G becomes a trade flow to a single entity. We call this case 'totally disaggregated'. If not, the group G has been 'only partially disaggregated', the 'group' entity remains.

There are 18,123 bilateral flows whose partner entity is a 'group', of which 7,021 have been disaggregated: 1,162 were 'totally disaggregated with ratio'; 4,777 were 'totally disaggregated'; 1,080 were 'only partially disaggregated'.

Here is an example of the case 'totally disaggregated with ratio':
In 1893, China reports export trade with the 'group' "Australia & New Zealand" = 1.048.464 Chinese Haikwan Tael
In 1893, Australia reports imports from China = 344.518 £
In 1893, New Zealand reports imports from China = 11.469 £

- ⇨ The sum of Australia and New Zealand imports from China = 355.987 £
- ⇨ Australia's ratio = 96.8 % / New Zealand's ratio's = 3.2 %.

China's exports to "Australia & New Zealand" can be replaced by two new bilateral flows, created by applying the mirror ratios to the original flow:
China's exports to Australia = 96.8 % of 1.048.464 = 1.014.913 Chinese Haikwan Tael
China's exports to New Zealand = 3.2 % of 1.048.464 = 33.551 Chinese Haikwan Tael

Here is an example of the case 'totally disaggregated':
In 1885, China reports exports to "Australia & New Zealand" = 1.880.104 Chinese Haikwan Tael
In 1885, New Zealand reports imports from China = 129.154 £
There are no data for Australia.
New Zealand's imports are converted into the Chinese currency by using the RICardo exchange rate database: 129.154 * 3.77952755906 = 34.172 Chinese Haikwan Tael.
This amount is then subtracted from the original flow to get an estimated bilateral export flow of China to Australia: 1880104 - 34172 = 1.845.932 Chinese Haikwan Tael.
In this case, China's exports to "Australia & New Zealand" is converted to China's exports to Australia, only leaving the New Zealand part to the bilateral trade figure.

- *Converting 'City/Part of'*

In the bilateral database, 266 entities are identified as 'city/part of', of which 34 on the reporting side and 232 on the partner side. These 'city/part of' include port cities, islands and not clearly delineated areas of a country. Each 'city/part of' has been linked to the country it belongs to, i.e. to an entity classified in the COW table. Thanks to this proceeding, we can aggregate all 'city/part of' into the corresponding 'country' by summing them, delete 'city/part of' bilateral flows and create new 'country' bilateral flows. These new bilateral flows have then been incorporated into the database to supplement missing data with a flag stating that they were produced by an aggregation procedure. The two tables below illustrate the process. For example, it allowed us to reconstruct bilateral trade between Spain and two partners, the United Kingdom and France, for the years 1839 and 1840. In this case, the database has data of Spanish trade with the United Kingdom in 1840 from a primary source that will be favored instead of the reconstructed flows. Therefore, all aggregated 'city/part of' partner flows will be kept if they provide an estimation not available in the database and will be ignored otherwise.

Table 6. RICname of the type 'city/part of'

RICname	**type**	**continent**	**part_of_country**
A Coruna	city/part_of	Europe	Spain
Cadix	city/part_of	Europe	Spain
Cartagena (Spain)	city/part_of	Europe	Spain
Gijón	city/part_of	Europe	Spain
San Sebastián	city/part_of	Europe	Spain
Santander	city/part_of	Europe	Spain
Russia (Baltic Sea)	city/part_of	Europe	Russia/USSR
Russia (Baltic Sea) & Russia (North Sea)	city/part_of	Europe	Russia/USSR
Russia (Black Sea)	city/part_of	Europe	Russia/USSR
Russia (European Ports)	city/part_of	Europe	Russia/USSR
Russia (North and Pacific Ports)	city/part_of	Europe	Russia/USSR
Russia (North and South Ports)	city/part_of	Europe	Russia/USSR
Russia (North Ports)	city/part_of	Europe	Russia/USSR
Russia (Pacific Ports)	city/part_of	Europe	Russia/USSR

Russia (South Ports)	city/part_of	Europe	Russia/USSR
Russia (White Sea)	city/part_of	Europe	Russia/USSR
Russia and Siberia (Kiakhta)	city/part_of	Europe	Russia/USSR
Russia and Siberia (land)	city/part_of	Europe	Russia/USSR
Russia/USSR (North)	city/part_of	Europe	Russia/USSR
Russian Poland	city/part_of	Europe	Russia/USSR
Russian Turkestan	city/part_of	Europe	Russia/USSR
United States of America (Atlantic Coast)	city/part_of	America	United States of America
United States of America (Pacific Coast)	city/part_of	America	United States of America

Table 7. Examples of 'city/part of' (reporting/partner) aggregated into 'country' flows

year	reporting	partner	exp/imp	flow (£)	nb	city/part_of aggregated (reporting/partner)
1839	Spain	United Kingdom	Exp	1460113	9	Cadix,A Coruna,Málaga,Bilbao,Santander,Gijón, San Sebastián,Balearic Islands,Cartagena (Spain)
1839	Spain	United Kingdom	Imp	472520	9	Cadix,A Coruna,Málaga,Bilbao,Santander,Gijón, San Sebastián,Balearic Islands,Cartagena (Spain)
1839	Spain	France	Exp	504136	8	Cadix,A Coruna,Málaga,Bilbao,Santander, San Sebastián,Balearic Islands,Cartagena (Spain)
1839	Spain	France	Imp	561203	9	Cadix,A Coruna,Málaga,Bilbao,Santander,Gijón, San Sebastián,Balearic Islands,Cartagena (Spain)
1840	Spain	United Kingdom	Exp	1465429	7	Cadix,Málaga,Cartagena (Spain),San Sebastián, Bilbao,Santander,Gijón
1840	Spain	United Kingdom	Imp	474375	8	Cadix,Málaga,Cartagena (Spain),San Sebastián, Bilbao,Santander,Gijón,A Coruna
1840	Spain	France	Exp	531112	9	Cadix,Málaga,Cartagena (Spain),San Sebastián, Bilbao,Santander,Gijón,A Coruna,Balearic Islands
1840	Spain	France	Imp	559819	9	Cadix,Málaga,Cartagena (Spain),San Sebastián, Bilbao,Santander,Gijón,A Coruna,Balearic Islands

Thanks to this proceeding, 18,284 bilateral flows have been deleted and 12,571 bilateral flows have been created (31% of reduction) of which 1,388 were duplicates of existing flows (example of Spain-UK trade in 1840). In other words, it helps us to reduce the heterogeneity of entities and to complement the database with new estimated flows.

	Nb of bilateral flows/original		Nb of bilateral flows/aggregated	Reduction in the nb of flows
'City/part of' reporting	9.818	Converted into reporting 'country'	6.242	3.576 (- 36 %)
'City/part of' partner	8.466	Converted into partner 'country'	6.329	2137 (- 25 %)
Total	18.284		12.571	5713 (- 31 %)

- *Converting 'colonial area'*

With 'colonial area', the purpose is to be able to convert the trade of a reporting entity with a colonial area by disaggregating this area by the countries composing it. Colonial areas are declared in the sources without specifying the countries composing the area. We thus have to complement our data with the COW dataset to retrieve the area composition.

Let us consider a reporting entity R which does report trade with a colonial area Co in year Y. In cases in which member countries of Co (known thanks to COW dataset) report trade with entity R in year Y, we want to extract Cn mirror flows to R trade with Co. To disaggregate the

colonial area, value of trade which Cn reports with R in Y is subtracted from the total amount of R's trade with Co. The remaining balance can be positive, negative or null. A positive amount can possibly match a trade flow of R with a remaining member country of Co. It might happen that the subtraction leads to a negative flow. This anomaly might reflect a variability in trade figures, in which case we could simply delete the negative flow, or be the result of a wrong definition of the colonial entities represented either by us or by the statistical body. Such cases will be carefully and manually reviewed before stating about them.

The conversion process is not yet stable enough to include preliminary results in this paper. We hope to be able to share our first results in our presentation.

5. Conclusion: assessing the quality of the converted dataset

The work we are presenting you is in the making. Our wish is to get your comments on the choices we have made and hopefully new ideas on the matter. We are currently working on the 'colonial area' disaggregation process. Once this process is completed, we intend to describe the gain that has been made in terms of heterogeneity and submit the new database to three quality tests:

- Total trade: we will recalculate the new 'world sum of partners' figures to compare it with Federico & Tena series. This test will assess the quality of our new bilateral trade in terms of total value by year;
- Number of countries: the (dis)aggregation of entities we have described is meant for reducing the heterogeneity of the RICardo bilateral database. In order to gauge the trade entity discrepancy through time, we will compare the yearly number of countries in the new RICardo dataset with that of the Federico & Tena series.
- Quality level: to evaluate the quality of the simplified RICardo bilateral dataset, an indicator will be created by mixing the results of the first two tests described above with a source quality index. After our work on heterogeneity, the bilateral dataset will be composed of flows collected from primary and secondary sources and of newly estimated flows. For each year, we will therefore be able to differentiate the number of flows according to its source and to calculate a quality index on both completeness and trust-level. Compiling those quality indices will allow us to assess the years for which the RICardo dataset is the most or less reliable. Our goal is to define the years for which the quality is strong enough to support our quantitative analysis.

References

Anderson K. & H. Norheim (1993a), "From imperial to regional trade preferences: its effect on Europe's intra- and extra-regional trade", *Weltwirtschaftliches Archiv*, 129, p. 78-102.

Anderson K. & H. Norheim (1993b), "History, geography and regional economic integration", *in* K. Anderson and R. Blackhurst, eds., *Regional Integration and the Global Trading System*, New York 1993, pp. 19-51.

Anderson K. & H. Norheim (1993c), "Is World Trade Becoming More Regionalized?", *Review of International Economics*, 1, 91-109.

Bairoch P. (1974), "Geographical Structure and Trade Balance of European Foreign Trade from 1800 to 1970", *The Journal of European Economic History*, 3 (3), pp. 557-608.

Barbieri K. & O. M. G. Keshk (2016), *Correlates of War Project Trade Data Set Codebook, Version 4.0*. Online: http://correlatesofwar.org.

Barbieri K., O. M. G. Keshk & B. Pollins (2009), "TRADING DATA: Evaluating our Assumptions and Coding Rules", *Conflict Management and Peace Science*, 26(5): 471-491.

Chase-Dunn C., Y. Kawano, and B. D. Brewer (2000), « Trade Globalization since 1795: Waves of Integration in the World-System », *American Sociological Review*, 65(1), pp. 77-95.

De Benedictis Luca and Lucia Tajoli (2011), "The World Trade Network", *The World Economy*, 34 (8), pp. 1417-1454.

Dedinger B. & P. Girard (2017), « Exploring Trade Globalization in the Long Run: the RICardo Project », *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 50:1, 30-48.

Drucker J. (2011), « Humanities Approaches to Graphical Display », *Digital Humanities Quarterly* 5 (1). http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html.

Fagiolo G., J. Reyes and S. Schiavo (2010), "The evolution of the world trade web: a weighted-network analysis", *Journal of Evolutionary Economics*, 20 (4), pp. 479-514.

Federico, G. and Tena Junguito, A. (2016). World trade, 1800-1938: a new data-set. EHES Working Papers in Economic History, n. 93.

Findlay R. and K. H. O'Rourke (2007), *Power and Plenty. Trade, War, and the World Economy in the Second Millennium*, Princeton & Oxford, Princeton University Press.

Fouquin Michel and Jules Hugot (2016), *Back to the Future: International Trade Costs and the Two Globalizations*, mimeo.

García-Pérez, G., Boguñá M., Allard A., and M. A. Serrano (2016), “The hidden hyperbolic geometry of international trade: World Trade Atlas 1870-2013*”*, *Scientific Reports* 6, 33441.

Latour, B. (1993), « Le topofil de Boa-Vista. La référence scientifique: montage photophilosophique », *Raisons pratiques* 4: 187–216.