https://doi.org/10.14428/rete.v10i0/hc

Measuring Human Capital: from WorldCat Identities to VIAF

Matthew Curtis David de la Croix

IRES/LIDAM, UCLouvain

This note is a summary of how we can measure scholars' human capital after the retirement of the WorldCat Identities project on March 23, 2023. WorldCat Identities was previously providing us measures of scholars' output and recognition.

1 The WorldCat Identities Measure

Construction of the relational database of scholars and literati active in medieval and early modern European academia started in 2017. We harvested data from secondary sources on the history of universities and academies. For each scholar in the database, we searched manually for links in Wikipedia and in WorldCat. Using the content of Wikipedia and WorldCat, we computed a heuristic human capital index for each person in the database using a principal component analysis. More precisely, the individual human capital index q_i of an individual i is given by:

$$\begin{aligned} q_i \;=\; & -1.76 + 0.43\ln(\text{no. characters of the longest Wikipedia page}) \\ & +0.40\ln(\text{no. Wikipedia pages in different languages}) + 0.47\ln(\text{no. works in WorldCat}) \\ & +0.46\ln(\text{no. publication languages in WorldCat}) + 0.47\ln(\text{no. library holdings in WorldCat}) \end{aligned}$$

The weights (0.43, 0.40, etc) are obtained from the first principal component of the five indicators (De la Croix et al. 2020; De la Croix 2021).

The original description of the WorldCat Identities project, from which our measures were taken, reads as follows:

> WorldCat Identities has a summary page for every name in WorldCat (currently some 30 million names) including named persons, organizations and fictitious characters. The pages include information derived from WorldCat and other sources (VIAF, FAST) plus with unique data derived or created through a variety of special processing activities (e.g., WorldCat Identities provides statistical data about how widely held a work is). A typical WorldCat Identities page will include a list of most widely held-by-libraries works by and about the identity, a list of variant forms of name the identity has been known by, a FAST tag cloud of places, topics, etc. closely related to works by and about the person, links to co-authors, and more. Titles listed are linked to WorldCat.org, and in many popular WorldCat Identities pages, links to the corresponding Wikipedia (English language) article are provided. (From https://www.oclc.org/research/areas/data-science/identities.html).

We illustrate how data was collected from WorldCat identities with an example taken from De la Croix and Goñi (2020): Honoré Bicais (Figure 1). He is listed as a University of Aix professor in Belin (1905); see the corresponding RETE (De la Croix and Fabre 2021). WorldCat considers different spellings of the family name (Bicais, Bicaise, Bicays, and the latinized Bicaisius and Bicaissius), which

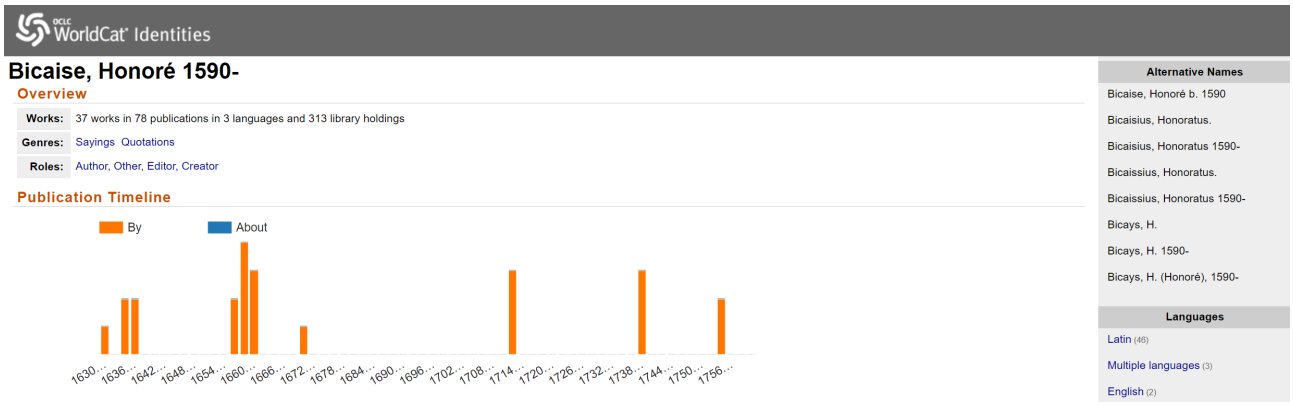

Figure 1: WorldCat Identities page of Honoré Bicais

ensures that the matching of authors to publications is accurate. Honoré Bicais was a relatively prolific scholar: there are 37 distinct works by or about him in the catalogues, involving 3 different languages. Libraries in the world hold 313 copies of these works (library holdings).

Combining the number of works, number of languages, and number of library holdings in a single measure of human capital allows us to take into account both the quantity and quality of publications. Indeed, high quality publications are more likely to be translated in other languages and to be held widely by libraries.

The human capital indexes have been used in all the RETE prior to this one (that is 66 notes) and in all the scientific publications exploiting these data (De la Croix and Fabre 2019; De la Croix et al. 2020; De la Croix and Goñi 2020; De la Croix and Morault 2022; De la Croix and Vitale 2023; Blasutto and De la Croix 2021; Baudin and De la Croix 2023; Curtis and De la Croix 2023; Zanardello 2023). Building a human capital index also means we can produce rankings of persons. We will come back to these rankings in Section 4.

2 The Retirement of the Identity Project

In early 2023, a cryptic message appeared on all WorldCat Identities pages: "The WorldCat Identities web application will be retired and shut down in the coming months and the data is no longer being updated." Asking for clarification, we received the following answer: "Support are not aware of the details surrounding WorldCat Identities, but we have been informed that this project was experimental for OCLC and it is currently on hold while the team responsible for it decide which direction would be best to take it down. As such, updates to Identities pages are not currently going through."[1] On March 23, 2023, all the information contained in WorldCat Identities pages was withdrawn.

There are probably millions of Wikipedia pages referring to a WorldCat Identities page. Now all those links are redirected to the new WorldCat Entities pages. These entities pages do not have any information on publications. They are similar to what can be found on Wikidata, hence of no direct interest for us.

3 The VIAF Measure

"The VIAF (Virtual International Authority File) combines multiple name authority files into a single OCLC-hosted name authority service. The goal of the service is to lower the cost and increase the utility of library authority files by matching and linking widely-used authority files and making that information available on the Web." (From https://viaf.org/). The address of the VIAF pages can be found from most of our WorldCat links, so we did not lose the effort put into finding these links.

1. "OCLC is a global library organization that provides shared technology services, original research, and community programs for its membership and the library community at large."

Figure 2 shows the VIAF page for Honoré Bicais.

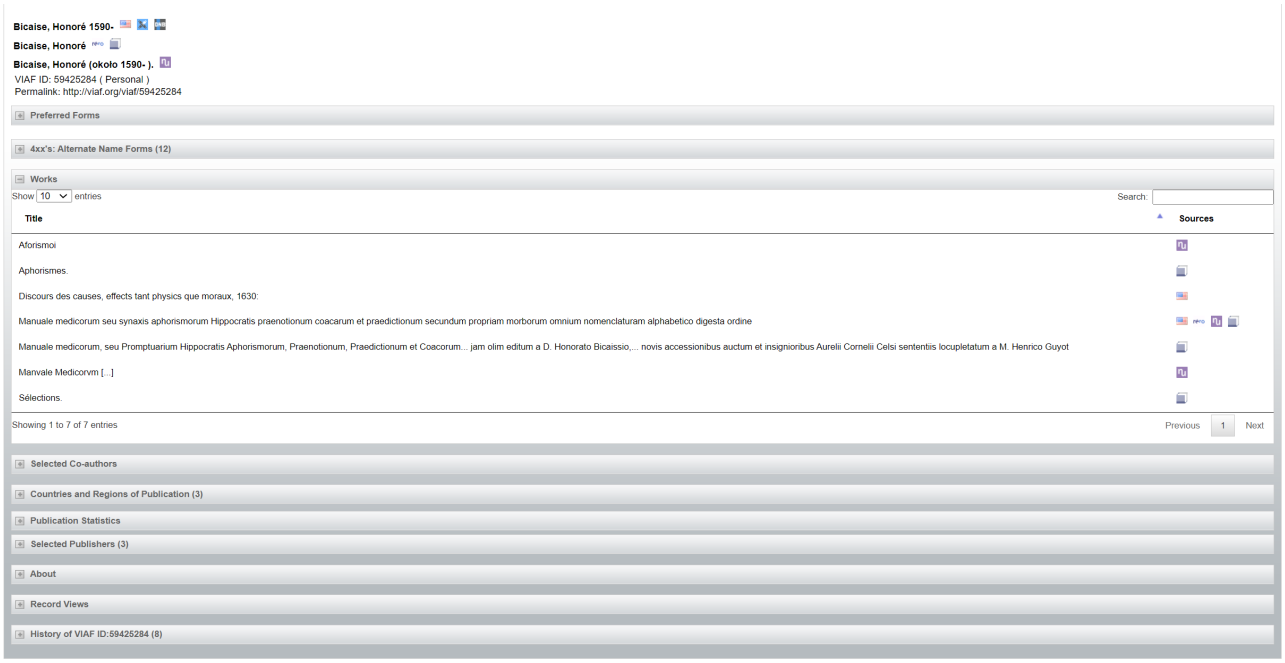

Figure 2: VIAF page of Honoré Bicais

From this VIAF page, we collect four numbers, reflecting four dimensions of the work by the author. In doing so we approximate both the quantity and the quality of the output of each scholar. The four numbers are: the number of titles in the category "Works," the number of alternate names, the number of countries of publications, and the number of publishers. The last three measures are proxies for the recognition of the works of the author.

4 Comparisons

To establish whether we can replace our WorldCat based human capital index by one built on VIAF data, we provide here a comparison between the two measures in a sample of 20,542 scholars of whom we have a VIAF page. We first look at whether we could predict WorldCat outcomes (number of works, number of languages, number of library holdings, and number of publications) with VIAF data on number of alternative names, number of countries of publications, number of publishers, and number of works. Results are shown in Table 1.

The VIAF variables appear both significant and powerful predictors of WorldCat variables. The R^2 are rather high, which is the sign that the information contained in VIAF is correlated with that in WorldCat. In particular, we find that the number of titles in VIAF is a strong predictor of number of works in WorldCat (which is not surprising) and that the number of countries and publishers are strong predictors of the number of library holdings (and hence are measures of notoriety).

The next step is to conduct the principal component analysis, combining this library based information with Wikipedia information. Table 2 shows the loadings of the first principal component under the two sets of included variables. The usual heuristic used to identify relevant principal components is to keep those having eigenvalues above one. In our case, the first principal component alone satisfies this condition. We use it as an index of a scholar's human capital. Compared to the equation shown in Section 1, the weights are slightly different because the sample is not identical; here it is restricted to scholars with a VIAF page. Note also that we have assumed that having no Wikipedia page is similar to having one page with a length of 60 characters and that having no WorldCat page is similar to having a page with one work in one language held by one library.

Having now two alternative measures of human capital, we can compare them. The Pearson linear correlation coefficient is equal to 0.92, while the Spearman's rank correlation coefficient is

	Dependent variable (WorldCat):			
	ln(works)	ln(languages)	ln(holdings)	ln(1 + pubs.)
VIAF variables	(1)	(2)	(3)	(4)
ln(1 + alt. names)	0.386***	0.125***	0.389***	0.329***
	(0.008)	(0.004)	(0.011)	(0.008)
ln(1 + countries)	0.222***	0.320***	0.730***	0.419***
	(0.018)	(0.008)	(0.025)	(0.019)
ln(1 + publishers)	0.318***	0.007	0.602***	0.510***
	(0.016)	(0.007)	(0.022)	(0.017)
ln(1 + titles)	0.578***	0.053***	0.359***	0.423***
	(0.013)	(0.006)	(0.019)	(0.014)
Constant	0.714***	0.101***	1.659***	1.218***
	(0.016)	(0.007)	(0.022)	(0.017)
Observations	20,542	20,542	20,542	20,542
R^2	0.711	0.499	0.654	0.706
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 1: OLS regression of WorldCat measures on VIAF measures

	(1) WorldCat	(2) VIAF
No. characters of Wikipedia page	0.419	0.358
No. languages Wikipedia	0.424	0.367
No. works in WorldCat	0.476	-
No. languages in WorldCat	0.430	-
No. library holdings in WorldCat	0.476	-
No. of alternative names in VIAF	-	0.413
No. of countries in VIAF	-	0.438
No. of publishers in VIAF	-	0.425
No. of titles in VIAF	-	0.440
No. Eigenvalues > 1	1	1
% variance explained by 1st PC	69.6%	70.6%

Table 2: First principal component of the human capital of scholars

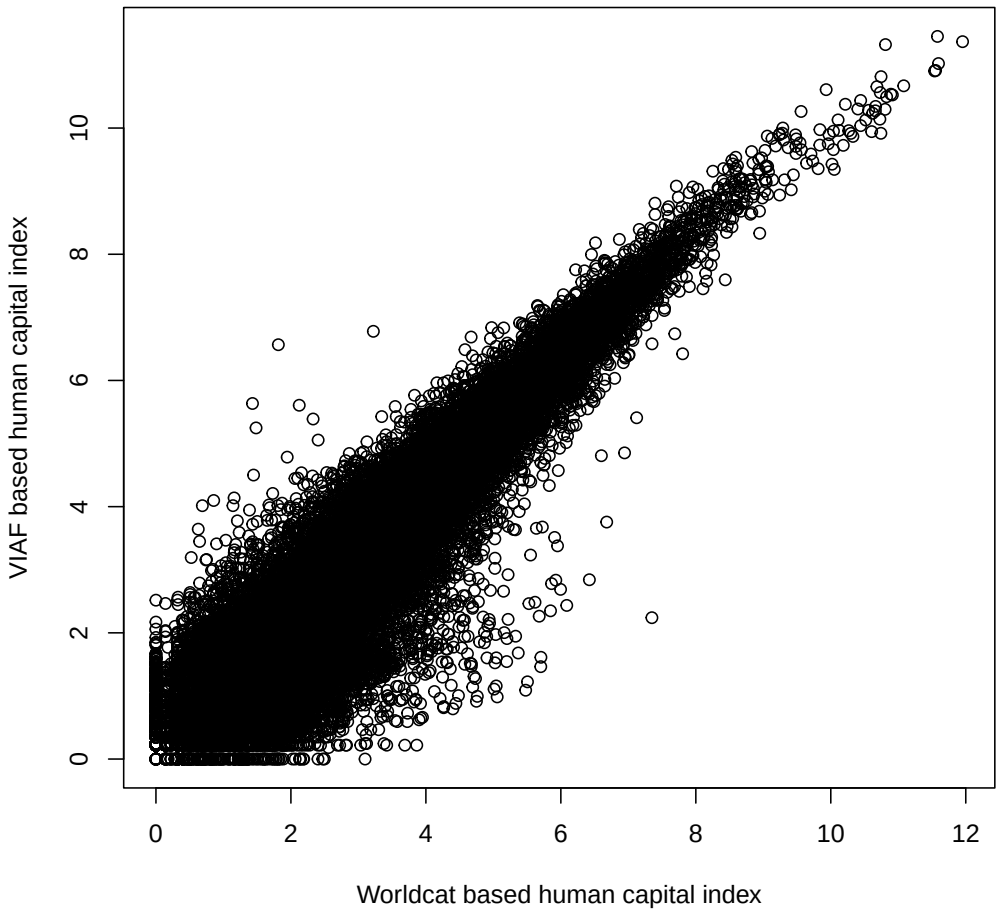

Figure 3: Scatter plot of the two human capital indexes

WorldCat based		VIAF based	
Scholar	Index	Scholar	Index
Martin Luther	11.95	François-Marie Arouet de Voltaire	11.45
Jean-Jacques Rousseau	11.60	Martin Luther	11.37
François-Marie Arouet de Voltaire	11.58	Thomas Aquinas	11.32
Immanuel Kant	11.55	Jean-Jacques Rousseau	11.02
JC Friedrich von Schiller	11.54	Immanuel Kant	10.91
Niccolo Machiavelli	11.08	JC Friedrich von Schiller	10.90
Jonathan Swift	10.91	Desiderius Erasmus	10.81
René Descartes	10.89	Niccolo Machiavelli	10.67
Giovanni Boccaccio	10.83	Gottfried Wilhelm von Leibniz	10.66
Thomas Aquinas	10.81	Carl Linnaeus	10.61
Benjamin Franklin	10.80	Jean Calvin	10.56

Table 3: Ranking of top scholars, based on two different human capital indexes

0.91. Both are very high, which reassures us that VIAF based measure can be used as a replacement for the WorldCat base measure. The x-y plot of the two measures in Figure 3 illustrates our results.

Finally, we can look at the top of the human capital distribution to see if switching from WorldCat to VIAF implies drastic changes in the ranking. Table 3 shows the top 0.5‰of the distribution. Many names appear in both columns. The VIAF based ranking puts Voltaire first, instead of Luther, but only by a hair's breadth.

5 Final Thoughts

Although the retirement of the WorldCat Identities project was bad news for those interested in measuring human capital from publications data — a bit as if Wikipedia suddenly disappeared — we found a viable alternative using statistics drawn from the VIAF platform.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 883033 "Did elite human capital trigger the rise of the West? Insights from a new database of European scholars."

Homepage: https://perso.uclouvain.be/david.delacroix/uthc.html
Twitter: https://twitter.com/UTHCerc
Database: https://shiny-lidam.sipr.ucl.ac.be/scholars/

First version April 18, 2023

References

Baudin, Thomas, and David De la Croix. 2023. The emergence of the child quantity-quality tradeoff – insights from early modern academics. Unpublished.

Belin, Ferdinand. 1905. *Histoire de l'ancienne université de Provence, ou histoire de la fameuse univesité d'Aix: d'après les manuscrits et les documents originaux.* Paris: A. Picard et fils.

Blasutto, Fabio, and David De la Croix. 2021. Catholic censorship and the demise of knowledge production in early modern italy. CEPR Discussion Paper 16409, conditionally accepted at *Economic Journal.*

Curtis, Matthew, and David De la Croix. 2023. Science, Theology, Medicine, Law. Academic disciplines and the rise of the West. Unpublished.

De la Croix, David. 2021. Scholars and literati in European Academia before 1800. *Repertorium Eruditorum Totius Europae* 5:35–41.

De la Croix, David, Frédéric Docquier, Alice Fabre, and Robert Stelter. 2020. The Academic Market and the Rise of Universities in Medieval and Early Modern Europe (1000-1800). CEPR Discussion Paper No. DP14509. https://cepr.org/publications/dp14509.

De la Croix, David, and Alice Fabre. 2019. A la découverte des professeurs de l'ancienne université d'aix, de ses origines á 1793. *Annales du midi* 131:379–402.

———. 2021. Scholars and Literati at the Royal Bourbon College in Aix-en-Provence (1603–1763). *Repertorium Eruditorum Totius Europae* 3:43–50. https://doi.org/10.14428/rete.v3i0/bourbon.

De la Croix, David, and Marc Goñi. 2020. Nepotism vs. intergenerational transmission of human capital in academia (1088–1800). CEPR Discussion Paper No. 15159. https://doi.org/10.2139/ssrn.3807314.

De la Croix, David, and Pauline Morault. 2022. Winners and losers from the Protestant reformation: an analysis of the network of European universities. THEMA Working Papers 2022-11, Université de Cergy-Pontoise. https://ideas.repec.org/p/ctl/louvir/2020029.html.

De la Croix, David, and Mara Vitale. 2023. Women in European academia before 1800 – religion, marriage, and human capital. *European Journal of Economic History,* https://doi.org/10.1093/ereh/heac023.

Zanardello, Chiara. 2023. Market forces in italian academia today (and yesterday). *Scientometrics* 128 (1): 651–698. https://doi.org/10.1007/s11192-022-04579-0.