DISCUSSION PAPER SERIES

DP15825

Human Biographical Record (HBR)

Arash Nekoei and Fabian Sinn

ECONOMIC HISTORY

ISSN 0265-8003

Human Biographical Record (HBR)

Arash Nekoei and Fabian Sinn

Discussion Paper DP15825
Published 18 February 2021
Submitted 17 February 2021

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

Human Biographical Record (HBR)

Abstract

We construct a new dataset of more than seven million notable individuals across recorded human history, the Human Biographical Record (HBR). With Wikidata as the backbone, HBR adds further information from various digital sources, including Wikipedia in all 292 languages. Machine learning and text analysis combine the sources and extract information on date and place of birth and death, gender, occupation, education, and family background. This paper discusses HBR's construction and its completeness, coverage, accuracy, and also its strength and weakness relative to prior datasets. HBR is the first part of a larger project, the human record project that we briefly introduce.

JEL Classification: N/A

Keywords: Bid data, Machine Learning, economic history

Arash Nekoei - arash.nekoei@iies.su.se
IIES and CEPR

Fabian Sinn - fabian.sinn@iies.su.se
IIES

Human Biographical Record (HBR)*

Arash Nekoei Fabian Sinn

February 17, 2021

Abstract

We construct a new dataset of more than seven million notable individuals across recorded human history, the Human Biographical Record (HBR). With Wikidata as the backbone, HBR adds further information from various digital sources, including Wikipedia in all 292 languages. Machine learning and text analysis combine the sources and extract information on date and place of birth and death, gender, occupation, education, and family background. This paper discusses HBR's construction and its completeness, coverage, accuracy, and also its strength and weakness relative to prior datasets. HBR is the first part of a larger project, the human record project that we briefly introduce.

This paper describes the Human Biographical Record (HBR), a new dataset of notable humans across human existence, created from various digital sources. The HBR is the first part of a larger project, the human record project. This project aims at capturing several aspects of human history. We are planning to make all datasets available at http://humanrecord.org/.

Section 1 describes our motivation to build a new dataset. Section 2 and 3 describes HBR's foundation and its construction. Section 4 describes variables available in HBR. Section 6 reports coverage and accuracy results. Section 7 compares HBR with prior datasets to provide users an overview about the advantages of each of them, and the last Section 8 offers some caveats of HBR and our ideas to improve it for the next version.

1 Motivation

Understanding the general patterns of human history has motivated many attempts to combine a large number of biographies. In this section, we briefly summarize past attempts and explain the two novel developments - crowd-sourcing and machine learning - that make our attempt possible and fruitful.

After a long tradition of collecting biographies of eminent people in ancient China, Greece, and Rome, starting from the 9th century, this practice flourished in the Middle East.[1] Two specific features of Arabic

*Arash Nekoei: Institute for International Economic Studies (IIES) at Stockholm University and CEPR, arash.nekoei@iies.su.se. We are grateful for help from the following colleagues at different stages of data construction for this project: Jonatan Ribert, Xueping Sun, Chih-Yu Tsou, and Yushi Wang.

[1]This includes, e.g., "Lives of outstanding generals" by Cornelius Nepos in 44 BCE and "Parallel Lives" by Plutarch published about 80 CE. Slightly earlier, Sima Qian (135– 86 BCE), the father of Chinese historiography, founded a tradition that presents history in a series of biographies.

biographical dictionaries relate to our purpose: their large number of entries and information about family relations.[2] In the post-medieval period, notable work has been done in Europe following this tradition.[3] Starting in the 19th century, "Who is who" books have become popular.[4] Most of these collections of biographies restrict their scope to a single region or profession. Because they are generally books that contain small text snippets with the individual's biography, it is challenging to utilize these in any quantitative analysis. Recently, a set of projects use large web-based encyclopedias and databases to construct datasets of eminent people.We review this literature in Section 7.

2 Foundation

The purpose of the Human Biographical Record (HBR) is to complement these datasets on several dimensions: the number of observations and the amount and accuracy of each entry's information. Wikidata is the backbone of HBR, where machine learning is used to gather additional information from other traditional and crowd sources, e.g., Britannica or Wikipedia.

Three recent developments are the foundations of HBR: crowd-sourcing in general, Wikidata in particular, and machine learning.

I. Crowd-sourcing enables people worldwide to collaborate and create the largest encyclopedia, Wikipedia, in 292 languages with more than 55 million articles as of today. However, there are three obstacles to overcome: First, free-form text sources do not follow a standard structure, making the information difficult to access. Second, sources can be more comprehensive in one version than in another. Third, sources can contain inaccurate or contradictory statements.

II. Wikidata is a crowd-sourced knowledge-base with a defined structure. Wikidata has more entries than Wikipedia of all 292 languages combined because it does not follow Wikipedia's criteria of including an entry. However, it is less comprehensive; it has less information for each entry.

III. Machine Learning allows us to combine the structure of Wikidata with information from Wikipedia and other digital sources. Using these tools, we can extract extra information from the text, enabling us to use the accumulated knowledge across all 292 languages of Wikipedia. We complement this with information extracted from the articles' network using the information in the cross-references, and we then optimally combine information from different sources, languages, text, and networks.

The dataset is based on work by contributors worldwide, covering all cultures, all countries at all times. The coverage across time is shown in Figure 1. An additional benefit of the data sources being available to edit on the internet is a continually growing dataset. Every day new individuals are added, old information is corrected, and further data gets supplemented. The digital form and connection to various other databases make the dataset easily extendable.

[2] Some examples of this tradition include "Book of the Governors and Judges of Egypt" by Al-Kindi (897–961), "Classes of Physicians and Philosophers" by Ibn Juljul (d. 1009), "Obituaries of Eminent Men and Notices of the Sons of the Epoch" by Ibn Khallikan (1211–1282). Quotation from Khalidi [1973]. See also Young [1990], Nawas [2006] and Douglas and Fourcade [1976].

[3] Young [1990]. A pioneering example is Giorgio Vasari's "The Lives of the Most Excellent Painters, Sculptors, and Architects", the birthplace of the term "Renaissance".

[4] It was pioneered by Adolphe Quetelet and Alphonse de Candolle, among others. See Cattell [1903] and references therein. The 19th century also witnessed the emergence of empirical studies of people appearing in such "who is who" books by biologists and psychologists. E.g. see Galton [1869]. Some of these studies of 'great men' had dubitable intentions.

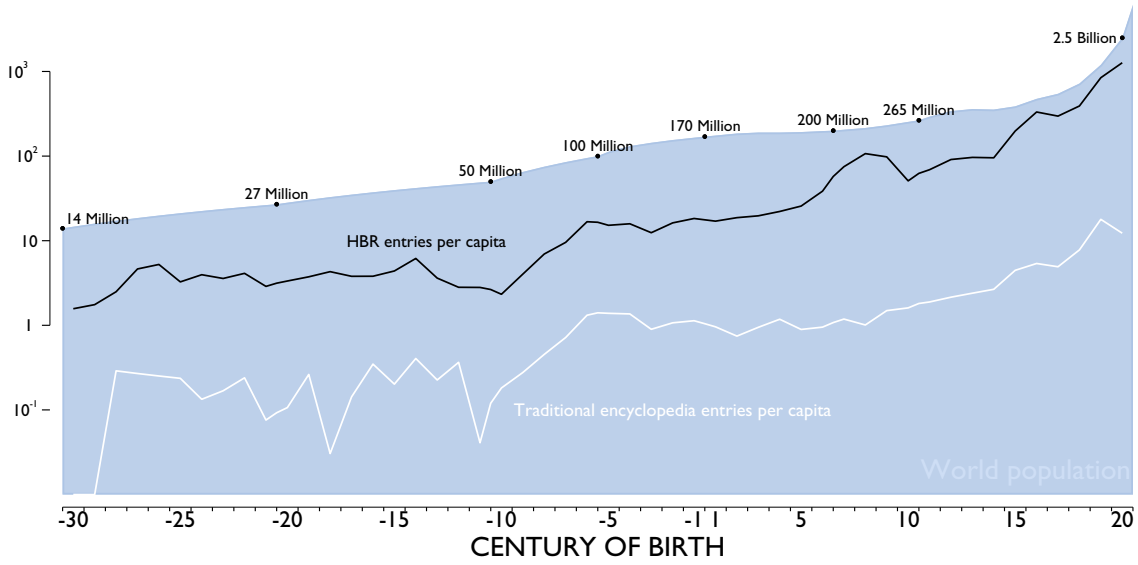

Figure 1: **HBR, TRADITIONAL ENCYCLOPEDIAS AND WORLD POPULATION** Number of entries in Human Biographical Record (HBR) vs. Traditional Encyclopedias per millions of world inhabitants.

3 Construction

3.1 Sources

Two datasets form the backbone of HBR. The first dataset is Wikipedia. Wikipedia is a free online encyclopedia created and edited by volunteers around the world.[5] Wikidata started as a structured representation of individual Wikipedia articles. Similar to Wikipedia, anyone can edit Wikidata. In contrast to Wikipedia, which allows a free-form text, information entered in Wikidata follows specific categories and must conform to particular rules.

The second backbone is Wikipedia in all 292 languages. It contains three components, the text itself, infoboxes, and categories. The information on different parts of Wikipedia appears in various forms of complexity, allowing us to extract data and identify the most frequently appearing pieces of information in case of internal contradictions across languages or other data sources.

We also use the network of all connections between Wikipedia pages across all languages to infer further information about the individuals in the data.

We further linked the data to the corresponding articles in four traditional Encyclopedias, namely, the Encyclopedia Britannica, the Great Russian Encyclopedia, the Gran Enciclopèdia Catalana, and the Encyclopædia Universalis. The links to the Encyclopedia Britannica were verified and are complete. Every article has a corresponding item in the HBR, and we recorded the corresponding encyclopedia page size. We also tried to match HBR with other encyclopedias, in particular a Chinese one. Unfortunately, none of Chinese encyclopedias is sufficiently digitized. Instead, HBR is matched to the Chinese Biographical Database (CBDB) [Harvard University et al., 2019]. We also use this link to complete the 470,000 biographies of Chinese indivduals that are in the CBDB.

[5] Wikipedia's accuracy was repeatedly tested and proven by comparing it to experts in the field or against encyclopedic articles like in Giles [2005]. Measuring Wikipedia's bias shows the bias exists only for living individuals, see, e.g. Kalla and Aronow [2015].

HBR entries are also linked to libraries' identifiers, like the Library of Congress id and other possible useful datasets like twitter handle, WorldCat, or ORCiD.

It is also matched with other parts of the human record project, like the Human War Record (HWR). HWR contains information about the location and time of each conflict throughout human history. HWR's novel aspect is that every conflict is matched with the entries from HBR who are involved, and records the ones who die in it [Nekoei et al., 2021].

The data creation date is the 26th of February 2020; however, because the data creation process takes a considerable amount of time, some parts were downloaded at a later point in time The last data was downloaded at the 20th of August 2020.

3.2 Methods/Tools

To gather information from the text, we use various tools commonly used in natural language processing. We extract data from the Wikipedia text using a cased multilanguage BERT neural network [Devlin et al., 2019] to assign country, continent, and occupation to individuals.

We extract information about an individual using pages linked to her and pages linking from her by utilizing gradient boosting on network features on the Wikipedia network (see Section 2.1). Using different connectedness measures, we train a gradient boosted forest to predict how likely it is to share a common characteristic like the century of being active. Intuitively individuals are more connected to their contemporaries for the century of being alive than to individuals who lived centuries apart. We then calibrate these predictions to achieve at least a 95% out of sample accuracy.

Combining the methods with a variety of data sources and languages allows us to solve ambiguities by picking the most frequently appearing pieces of information. If a method contains hyperparameters, we calibrate them by guaranteeing at least 95% accuracy in a holdout sample.

4 HBR Variables and their definition

The data contains one row for each individual. We identify individuals by using the crowd-sourced "human" identifier on Wikidata. Sometime the human status is not clear, therefore we also include dummy variables for being biblical, legendary, or possibly fictional. In a similar fashion we collect the gender variable mainly from Wikidata. Entries that do not have "male" or "female" as value are coded as missing. We recorded dates for each entry, first the time of birth and the time of death. All variables about time come in varying accuracy, from the exact year to millennia. We record for each individual a century that either corresponds to his year of birth or inferred from the year he flourished in.

HBR records the country at birth, death, and a country representing the places where the individual flourished (we refer to this country as main country). For example, Leonardo da Vinci was born in Italy, died in France, and was active in Italy. Countries are defined according to McEvedy et al. [1978] that provide historical population estimates for each country. The recorded location of individuals vary from villages to empires. While villages are contained in the area of a country, empires are often not. If the empire is the finest location specified for an entry, we use the country where the empire's capital is located. Based on this country definition, regions are defined as Asia, Americas, North Africa, Near East, and Western and Eastern

Europe. In this way, HBR entries are mapped to a country or region in a consistent way over a long span of time. Further, we record family relationships like spouse, parents, children, and siblings.

We also document the occupation by categorizing individuals into political, spiritual and intelligentsia and provide a more nuanced breakdown of the political, and intelligentsia categories.[6] Irrespective of their occupation, we also record the individual's religion and place of education.

We record several pieces of information about their Wikipedia characteristics. For example, the number of languages an individual appears in Wikipedia, the number of characters or words on each of these Wikipedia pages. Characters can be a letter from the Latin script, Chinese characters, etc. While we count words by tokenizing texts of various languages, the date when the page was first created, or how often Wikipedia users visited the page since 2016. HBR contains indicators for some external datasets that an entry is linked to, e.g., e CDBD, the Encyclopedia Britannica, or the Library of Congress Authority ID. The linking of Wikidata to both the Encyclopedia Britannica and the CBDB was checked manually to guarantee its accuracy and completeness.

5 Completeness

Completeness measures the share of non-missing values. Higher percentages imply that more individuals have the value of that variable. Table 1 describes the completeness for HBR and various sub-populations.

	HBR (1)	4-Encyclopedia (2)	Wikipedia (3)	Validation (4)
Gender	83.83%	99.93%	99.74%	99.93%
Century	72.82%	99.01%	97.3%	100%
Decade	58.84%	95.24%	87.39%	33.44%
Year of birth	60.43%	97.69%	89.68%	58.73%
Year of death	30.05%	84.47%	42.96%	52.61%
Country	77.59%	99.07%	99.63%	99.28%
Country of birth	32.38%	86.19%	57.15%	21.51%
Country of death	12.46%	70.16%	21.92%	16.66%
Continent	77.59%	99.07%	99.64%	99.35%
Occupation	75.9%	97.97%	99.33%	98.68%
Number of Observations	7,015,353	72,006	3,645,704	3,547,385

Table 1: **Completeness** Completeness measures the share of non-missing values in HBR in column 1. Column 2: 4-Encyclopedia is the union of the four encyclopedias, namely the Encyclopedia Britannica, the Great Russian Encyclopedia, the Gran Enciclopèdia Catalana and the Encyclopædia Universalis. Column 3: Wikipedia are all entries in HBR with a Wikipedia page in any of 292 languages. Column 4: Validation stratifies HBR entries with a Wikipedia page by century of birth similar to the validation sample of Section 5.
42.96% in column 3 is due to a large share of individuals still being alive.
100% in column 4 is by construction due to stratification by century. Stratification increases the share of ancient individuals with unknown exact Year of Birth and Death.

For HBR, column 1, the completeness decreases as the level of detail of the data rises. For example, from century to decade and decade to exact year, the completeness diminishes. The Year of Death can be missing for individuals who are still alive. The completeness is the highest for individuals that appear in the

[6] See Nekoei and Sinn [2020]

encyclopedias, column 2, as they are the most prominent individuals of the entire population. Individuals with a Wikipedia page, column 3, have the next highest completeness. This is because the presence of text makes it feasible to extract most values, see Section 3.2. The exception is the Year of Death, which is caused by manyt 20th-century individuals who are still alive.

For the Validation sample, column 4, the century is by construction complete as the sample is stratified by century. Stratification weighs each century equally, emphasizing individuals from the distant past, for whom the exact Year of Birth, Year of Death, or decade is often unknown.

6 Validation

This section compares HBR with the full information available in our sources. We do this across two dimensions, Coverage and Accuracy.

6.1 Method

We ask several individuals to read the Wikipedia pages in the most popular languages and collect the same information as the HBR. We then compare their results with the HBR in terms of both coverage and accuracy. Coverage compares how often the validators find values versus our methods. Accuracy measures the share of values in HBR, which are correct according to individuals' reading.

The sub-sample of HBR that we use for validation is drawn randomly from the population of individuals with a Wikipedia page and stratified across centuries in the following way. We select ten individuals from each century, starting from the 6th century BCE. To not overweight the BCE period, we only add 40 individuals from the period before the 6th century BCE. We also add ten individuals who do not have a century recorded in HBR. The size of the final validation sample is 310 individuals.

6.2 Coverage

Coverage compares how often the validators find values versus our methods. It is described in Table 2. Coverage, column 1, is the share of maximum information available captured by HBR. Namely, the sum of columns 2 and 3 divided by columns 2, 3, and 4. "Common", column 2, is the percentage of cases in which both HBR and the validators find the values. "Only HBR", column 2, is the percentage of cases that only HBR finds. "Only validators" are instances in which only the validators find the values, and the last column, "Neither", is the share of cases in which neither HBR nor validators can discover the information about the individual.

Identifying the gender for individuals is a simple task for both HBR and the validators. The time variables are easier to extract for the HBR because it uses more languages and the network of contemporaries, see Section 3.2. Finding a value for the main country is more easily done by the HBR because it uses all languages of Wikipedia. However, the validators frequently find additional information for the exact country of birth/death compared to the HBR. If there is uncertainty about whether a country is the place of birth/death, HBR falls back using the country as the main country. Hence, countries categorized by the validators as birth/death place are mostly categorized as the main country by the HBR.

	Coverage (1)	Common (2)	Only HBR (3)	Only validators (4)	Neither (5)
Gender	100%	99.68%	0.32%	0%	0%
Century	98.04%	70%	26.77%	1.94%	1.29%
Year of birth	97.25%	26.13%	8.06%	0.97%	64.84%
Year of death	95.77%	26.77%	17.1%	1.94%	54.19%
Main country	99.67%	77.1%	21.29%	0.32%	1.29%
Country of birth	75%	13.23%	9.03%	7.42%	70.32%
Country of death	73.97%	7.42%	10%	6.13%	76.45%

Table 2: **Coverage** Coverage compares how often the validators find values versus our methods; column 1 reports the share of maximum information avaialable captured by HBR. Namely, the sum of Column 2 and 3 divided by Column 2, 3 and 4. "Common", column 2, is the percentage of cases in which both HBR and the validators find the values. "Only HBR", column 2, is the percentage of cases that only HBR finds. "Only validators" are instances in which only the validators find the values and the last column, "Neither", is the share of cases in which neither HBR nor validators can find the values about the individual. The sum of columns 2-5 sum to one, by construction.

6.3 Accuracy

Accuracy measures the share of values in HBR, which are correct according to individuals' reading. To find a value, the majority of validators need to agree on a single value.

	Accuracy (1)	Instances (2)
Gender	100%	308
Century	99.54%	217
Year of birth	97.53%	81
Year of death	100%	83
Main country	96.23%	239
Country of birth	100%	41
Country of death	100%	23

Table 3: **Accuracy** Accuracy measures the share of values in HBR which are correct according to individuals' reading.

Table 3 describes the accuracy of the different variables. HBR's accuracy is above 95% for all variables. Meaning, if validators find a value for the variable, it corresponds to HBR's value in more than 95% of the cases. In most cases, errors are caused by Wikipedia articles in different languages providing different information, and the validators are only exposed to one language, whereas HBR combines the information across all articles.

7 Comparison with prior datasets

This section juxtaposes HBR with prior similar datasets to help users choose the best dataset to answer their research question. We focus on datasets capturing information about individuals. There are more general datasets, often refer to as knowledge-base, that collect information not only about individuals but also other entities. Yago is the leading example here(Mahdisoltani et al. [2013] and Rebele et al. [2017]). As of 2019, YAGO3 included 10 million entities and contained more than 120 million facts about these entities.

	HBR	CoST[7]	HoHT[8]	Longevity[9]	Pantheon[10]	HA[11]
Number of Entries	7,015,353	21,906	1,243,776	297,651	11,341	3,869
Main Source	Wikidata Wikipedia	Freebase	Freebase Wikipedia	IBN[12]	Wikipedia	various books
Variables						
Gender	Yes		Yes	Yes	Yes	Yes
Occupation	Yes	Yes	Yes	Yes	Yes	
Family relation	Yes					
Religion	Yes					
Education	Yes					
Location detail	Country	City	City	City	Country	Country
Accuracy tested	Partial	Partial	Partial		Yes	Yes

Table 4: **Comparison of HBR with prior biographical datasets**

8 Caveats and potential improvements

While digital crowd-sourced sources have advantages, they also provide difficulties.

One problem is the focus of Wikipedia editors on more recent individuals. This leads to older individuals of minor importance having worse data quality, both in quantity and quality, as errors take longer to be corrected. Because Wikipedia is inherently self-correcting and self-completing, users will update errors over time and add missing data. A second way to address this problem is to supplement the current content by adding more external data.

A further problem is malicious changes or vandalism. This problem affects few, albeit the most important, individuals in our dataset. Vandalism can be avoided in the future by taking several snapshots of the data, detecting vandalism, and taking an older version of the content if vandalism is detected. Our few experiment shows the promise of this approach.

References

J McKeen Cattell. A statistical study of eminent men. Popular Science Monthly, *1903.* 2

David De la Croix and Omar Licandro. The longevity of famous people from hammurabi to einstein. Journal of Economic Growth, *20(3):263–303, 2015.* 8

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), *pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.* 4

Fedwa M. Douglas and Genevieve Fourcade. The Treatment by Computer of Medieval Arabic biographical Data: an introduction and guide to the Onomasticon Arabicum. *Paris, 1976.* 2

Francis Galton. Hereditary genius: An inquiry into its laws and consequences, *volume 27. Macmillan, 1869.* 2

[7] Serafinelli and Tabellini [2018]
[8] Gergaud et al. [2017]
[9] De la Croix and Licandro [2015]
[10] Yu et al. [2016]
[11] Murray [2003]
[12] Index Biobibliographicus Notorum Hominum

Olivier Gergaud, Morgane Laouenan, and Etienne Wasmer. A brief history of human time. exploring a database of" notable people. 2017. 8

Jim Giles. Internet encyclopaedias go head to head. Nature, *438:900–901, 2005.* 3

Harvard University, Academia Sinica, and Peking University. China biographical database (CBDB), April 2019. URL https://projects.iq.harvard.edu/cbdb. 3

Joshua L Kalla and Peter M Aronow. Editorial bias in crowd-sourced political information. PloS one, *10(9):e0136327, 2015.* 3

Tarif Khalidi. Islamic biographical dictionaries: a preliminary assessment. The Muslim World, *63(1):53–65, 1973.* 2

Farzaneh Mahdisoltani, Joanna Biega, and Fabian M Suchanek. Yago3: A knowledge base from multilingual wikipedias. 2013. 7

Colin McEvedy, Richard Jones, et al. Atlas of world population history. *Penguin Books Ltd, Harmondsworth, Middlesex, England., 1978.* 4

Charles A Murray. Human accomplishment. *2003.* 8

John A. Nawas. Medieval Islamic Civilization: An Encyclopedia, *volume 1, chapter Biography and Biographical Works, page 110?112. New York: Routledge, 2006.* 2

Arash Nekoei and Fabian Sinn. Herstory: The rise of self-made women. 2020. 5

Arash Nekoei, Fabian Sinn, and Yushi Wang. Human c records (hcr). 2021. 4

Thomas Rebele, Arash Nekoei, and Fabian M Suchanek. Using yago for the humanities. In Workshop on Humanities in the Semantic Web-WHiSe II, *2017.* 7

Michel Serafinelli and Guido Tabellini. Creativity over time and space. 2018. 8

MJL Young. Religion, Learning and Science in the ?Abbasid Period, *chapter Arabic biographical writing. 1990.* 2

Amy Zhao Yu, Shahar Ronen, Kevin Hu, Tiffany Lu, and César A Hidalgo. Pantheon 1.0, a manually verified dataset of globally famous biographies. Scientific data, *3:150075, 2016.* 8