OECD.AI data from partners: methodological note

OECD.AI
Policy Observatory

Last updated on: 8/06/2020

Introduction

The OECD AI Policy Observatory (OECD.AI) aims to help countries encourage, nurture and monitor the responsible development of trustworthy artificial intelligence (AI) systems for the benefit of society. As an inclusive online platform for public policy on AI, the Observatory provides a comprehensive database of AI policies from across the world. Building on the momentum of the OECD's Recommendation on Artificial Intelligence, the first intergovernmental standard on AI, the Observatory combines resources from across the OECD with those of partners from all stakeholder groups to facilitate dialogue and provide multidisciplinary, evidence-based policy analysis on AI.

The third pillar of OECD.AI, "Trends & data", aims to help policy makers develop, implement and improve policies for AI. To do so, OECD.AI showcases data and metrics to allow countries and stakeholders to compare policy responses, engage in international co-operation, monitor progress and develop best practices.

The "Data from partners" subsection provides data from partner institutions, with the objective of showing AI-related trends over time from as many high quality sources and vantage points as possible.

The present note aims at providing a full and transparent account of the sources and methodology used to construct the "Data from partners" visualisations for OECD.AI. It covers data provided by the following partners:

- Microsoft Academic Graph
- LinkedIn
- Event Registry

OECD.AI visualisations using data from these three data sources were developed by the AI Lab at the Slovenian Jožef Stefan Institute (JSI). It is important to highlight that data sources and methodologies in OECD.AI may differ from those previously used for OECD research, and therefore results may vary. Because methods for measuring AI are still evolving, there remain definitional and other issues that can influence empirical results. Efforts to develop clear definitions, taxonomies and classifications are underway, as are efforts to compile accurate and comparable indicators. Given the current evolving situation, showing trends in a timely manner based on transparent methodologies by partner institutions can be of value to policy makers.

1 Microsoft Academic Graph data

The Microsoft Academic Graph (MAG) is a heterogeneous graph containing scientific publication records, citation relationships between those publications, as well as authors, institutions, journals, conferences, and fields of study (Sinha et al., 2015; Wang et al., 2019).

MAG employs advances in machine learning, semantic inference and knowledge discovery to explore scholarly information. It is a semantic search engine (i.e. not keyword-based), which means that it employs natural language processing (NLP) to understand and remember the knowledge conveyed in each document.

Information is indexed by paper type, author, time of publication, field of study, publication outlet and institution. The graph is updated on a bi-weekly basis. In 2019, close to 1 million new papers were added every month (Wang, 2019).

Each paper is automatically categorised into a field of study following a machine-learned taxonomy. An algorithm detects the concepts present in each paper and identifies, or “learns”, the hierarchy of different fields of studies. This concept detection operation is performed bi-weekly every time the graph[1] is updated and is applied to every new article, which is tagged accordingly. The taxonomy itself is adjusted every 6 months. Because top level disciplines are important and visible, the top two levels of the taxonomy are manually reviewed against Wikipedia’s hierarchy of topic classifications[2] (Shen et al., 2018).

Determining Artificial Intelligence papers for OECD.AI

The visualisations provided in the OECD AI Policy Observatory (OECD.AI) use a subset of the MAG comprised of papers related to AI. A paper is considered to be about AI if it is tagged during the concept detection operation with a field of study that is categorised in either the “artificial intelligence” or the “machine learning” fields of study in the MAG taxonomy. Results from other fields of study, such as “natural language processing”, “speech recognition”, and “computer vision” are only included if they *also* belong to the “artificial intelligence” or the “machine learning” fields of study. As such, the results are likely to be conservative.

Collaboration between countries or institutions

OECD.AI displays research collaborations between different entities (either institutions or countries[3]). This is done by assigning each paper to the relevant institutions and countries on the

[1] In discrete mathematics, graphs are mathematical structures used to model pairwise relations between objects. Graphs are made up of vertices (also called nodes or points) which are connected by edges (also called links or lines). For more information, please see https://en.wikipedia.org/wiki/Graph_(discrete_mathematics).

[2] For more information, please see https://en.wikipedia.org/wiki/Category:Main_topic_classifications.

[3] Country names and codes in OECD.AI abide by the “OECD Guidelines regarding the use of the list of names of countries and territories.”

basis of the authors' institutional affiliations[4]. OECD and CSIC (2016) define collaboration as "co-authorship involving different institutions. International collaboration refers to publications co-authored among institutions in different countries…National collaboration concerns publications co-authored by different institutions within the reference country. No collaboration refers to publications not involving co-authorship across institutions. No collaboration includes singled-authored articles, as long as the individual has a single affiliation, as well as multiple-authored documents within a given institution."[5]

To avoid double counting, collaborations are considered to be binary: either an entity collaborates on a paper (value=1) or it does not (value=0). The shared paper counts as one toward the number of collaborations between two entities. The following rules apply:

- For between-country collaborations: papers written by authors from more than one institution in the same country only count as one collaboration for that country.
- For between-institution collaboration: papers written by more than one author from the same institution only count as one collaboration for that institution.

Matching institutions to their countries

The Global Research Identifier Database (GRID)[6] was used to match institutions in MAG to a country. Information about an institution's geographical coordinates, city, and country from GRID allowed for the geolocation of 72% of the institutions in MAG. The remaining institutions were matched manually to their respective countries.

An artificial entity called "international organisations" was created to reflect papers written by international organisations. This avoids counting these papers as originating in the country where the relevant international organisation is headquartered.

Following the same logic, papers from multinational enterprises were attributed to the country in which the company's headquarters are located, regardless of the country in which the actual research was conducted.

Type of papers

The MAG classifies papers into the following types depending on publication outlets: conference; book; book chapter; repository; patents; journal; and other. The "Repository" type refers to archival sites, including arXiv, bioRxiv, and SSRN. There may be several versions of "Repository" papers, including some that may be published in conventional journals. "Other" is a category comprising papers from journals or conferences of which the quality is unknown. This includes one-off workshops, new journals, or venues that no longer exist.

[4] Information about an author's institutional affiliations is available for about 51% of AI publications in MAG. Thus, collaboration statistics may be underestimated.

[5] Institutional measures of collaboration may overestimate actual collaboration in the case of countries where it is common practice to have a double affiliation (OECD and CSIC, 2016).

[6] The GRID is an openly accessible database of educational and research organizations worldwide, created and maintained by Digital Science & Research Solutions Ltd. It contains the institution's type, geo-coordinates, official website, and Wikipedia page. For more information, see https://www.grid.ac/.

Arguably, patents are conceptually different from the other categories in this listing.[7] Therefore, for simplicity, the category "Research publications" – the default setting for most MAG data visualisations on OECD.AI – comprises all paper types except patents. The drop-down menu "Publication type" allows selecting and viewing results for patents only.[8]

Counting of publications: quantity measure

In absolute terms, each publication counts as one unit towards an entity (a country or an institution). To avoid double-counting, a publication written by multiple authors from different institutions is split equally among each author. For example, if a publication has four authors from institutions in the US, one author from an institution in China and one author from a French institution, then 4/6 are attributed to the US, 1/6 to China and 1/6 to France. The same logic applies to institutional collaborations.

Counting of publications: quality measure

MAG assigns a rank to each publication to indicate its relevance.[9] It does so by using a dynamic eigencentrality measure that ranks a publication highly if that publication impacts highly ranked publications, is authored by highly ranked scholars from reputable institutions, or is published in a highly regarded venue and also considers the competitiveness of the field. The eigencentrality measure can be considered as the likelihood that a publication would be evaluated as being highly impactful if a survey were to be posed to the entire scholarly community. For this reason, MAG calls this measure the "saliency" of the publication.[10] Similarly, the saliency of an author, institution, field, and publication venue represents the sum of all saliencies of the respective publications.

To adjust for temporal bias – i.e. older publications having more citations than more recent ones because they have been in circulation longer), MAG considers that saliency is an autoregressive stochastic process. This means that the saliency of a publication decays over time if a publication does not receive continuing acknowledgments, or its authors, publication venue and fields do not maintain their saliency levels. Reinforcement learning is used to estimate the rate of the decay and to adapt the saliency to best predict future citation behaviours.[11]

[7] MAG includes patent applications as publications since they "fit well into the model of publication entity…because they all have authors, affiliations, topical contents, etc., and can receive citations" (Wang et al., 2019).

[8] Note that results for patents come with some considerable lag, as "publication [of a patent] generally only takes place 18 months after the first filing. As a result, patent data are publicly available for most countries across the world, often in long time series" (OECD, 2009).

[9] Since papers may be published in different formats and venues, in some cases different versions of the same paper exist in MAG. These papers are grouped under a unique identifier called "family ID", that borrows the value or "paper ID" of the main paper in the family. MAG ranks its publications by family ID.

[10] Saliency rankings differ from traditional citation counts in that the latter treat each citation as equal and perpetual, whereas the earlier imposes weighting on each citation based on the factors mentioned above. While citation counts could be altered with relative ease, to boost the saliency of an article one would have to persuade many well-established authors publishing at reputable venues to cite it frequently.

[11] By leveraging the scale of Microsoft's web crawler in Bing, MAG observes tens of millions of citations each week, which serve as feedback from the entire scholarly community on its saliency assessments. For more information about how publications are ranked in MAG, please see https://academic.microsoft.com/faq and Wang et al. (2019).

OECD.AI uses the saliency rankings from MAG as a measure of quality. To provide fairer intertemporal comparisons, publication ranks are normalised according to the publication year.

OECD AI Principles

Classification of scientific publications by AI Principle

A pool-based active learning algorithm[12] was developed and trained to classify publications under each of the OECD AI Principles ("AI Principles").[13] A subset of the AI publications in MAG was selected to train the active learning algorithm. This was accomplished by estimating a semantic similarity score between a publication's title and abstract, and information on each of the AI Principles from the following resources:

- OECD Recommendation of the Council on Artificial Intelligence (OECD, 2019).
- Practical implementation guidance for the OECD AI principles (OECD, forthcoming).
- List of keywords purposely created for each AI Principle, assigning a specific relevance level to each term (i.e., either high or standard relevance).

A similarity score was determined using the following methodology:

- A "count score" was calculated by counting the total number of high relevance and standard relevance keywords in a publication's keywords, abstract and title using the following formula:
    - *Count score* = (count of high relevance keywords) + 0.3*(count of standard relevance keywords)
- A "cosine similarity score" was calculated between the publications and the three abovementioned resources using the following formula:
    - *Cosine similarity score* = (cosine similarity between publications and the list of keywords for each AI Principle) + (cosine similarity between publications and each AI Principle's section from the practical implementation guidance)
- The similarity score is defined as the sum of the count and the cosine similarity scores:
    - *Similarity score* = (Count score) + (Cosine similarity score)

The 10 000 publications with the highest similarity score were included in the training dataset for each AI Principle. Training and refinement of the active learning classifier is expected to continue throughout 2020.

Selection of related recent scientific research by AI Principle

After using the active learning classifier to identify publications relevant to each of the AI Principles, publications are sorted based on their MAG saliency rank. The highest-ranking publications from the last six months are then selected to be shown in the OECD.AI platform.

[12] Active learning is a type of iterative supervised learning that interactively queries the user to obtain the desired labels for new data points. By choosing the data from which it learns, an active learning algorithm is designed to achieve greater accuracy with fewer training labels than in traditional supervised learning (Settles, 2010).

[13] Please see OECD (2019) for more information about the OECD AI Principles.

Policy areas

Classification of scientific publications by policy area

A list of the most relevant fields of study from the MAG taxonomy was created for each policy area[14]. An AI-related publication from MAG must contain at least one of the relevant MAG topics for a given policy area to be classified in that policy area.

Selection of related recent scientific research by policy area

For each policy area, relevant AI-related publications are sorted based on their MAG saliency rank. The highest-ranking publications in the last six months were then selected to be shown in the OECD.AI platform.

Additional metrics

Additional metrics are used to construct the y-axis of the "AI publications vs GDP per capita by country, region, in time" chart. These indicators include:

- **GDP**: GDP at purchaser's prices is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without deducting the depreciation of fabricated assets or the depletion and degradation of natural resources. Data are in current US dollars. Dollar figures for GDP are converted from domestic currencies using single year official exchange rates. For a few countries where the official exchange rate does not reflect the rate effectively applied to actual foreign exchange transactions, an alternative conversion factor is used. *Sources*: World Bank national accounts data and OECD National Accounts data files (data.worldbank.org/).
- **GDP per capita**: GDP per capita is gross domestic product divided by midyear population. GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in current U.S. dollars. *Sources*: World Bank national accounts data and OECD National Accounts data files (data.worldbank.org/).
- **Population**: Total population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. The values shown are midyear estimates. *Sources*: United Nations Population Division, World Population Prospects: 2019 Revision; Census reports and other statistical publications from national statistical offices; Eurostat: Demographic Statistics; United Nations Statistical Division, Population and Vital Statistics Report; U.S. Census Bureau: International Database; and Secretariat of the Pacific Community: Statistics and Demography Programme (data.worldbank.org/).
- **R&D expenditure (% of GDP)**: Gross domestic expenditures on research and development (R&D), expressed as a percent of GDP. They include both capital and

[14] Policy areas include agriculture, competition, corporate governance, development, digital economy, economy, education, employment, environment, finance and insurance, health, industry and entrepreneurship, innovation, investment, public governance, science and technology, social and welfare issues, tax, trade, and transport.

current expenditures in the four main sectors: Business enterprise, Government, Higher education and Private non-profit. R&D covers basic research, applied research, and experimental development. *Source*: UNESCO Institute for Statistics (uis.unesco.org).

For these metrics, data is interpolated in years where no data is available. If last year’s value is missing for an indicator, the value of the latest available year is used.

2 LinkedIn data

Skills

LinkedIn members self-report their skills on their LinkedIn profiles. Currently, more than 35,000 distinct, standardised skills are identified by LinkedIn. These have been coded and classified by taxonomists at LinkedIn into 249 skill groupings, which are the skill groups represented in the dataset.[15] The top skills that comprise the AI skill grouping are Machine Learning, Natural Language Processing, Data Structures, Artificial Intelligence, Computer Vision, Image Processing, Deep Learning, TensorFlow, Pandas (software) and OpenCV, among others.

Skill groupings are derived through a similarity index that measures skill composition at the industry level. Industries are classified according to the ISIC 4 industry classification (Zhu et al., 2018).

AI skills penetration (within country)

The aim of this indicator is to identify the penetration of AI skills in a country through the following methodology:

- Computing frequencies for all self-added skills by LinkedIn members in a given entity (occupation, industry, etc.) in 2015-2019.
- Re-weighting skill frequencies using a TF-IDF model[16] to get the top 50 most representative skills in that entity. These 50 skills compose the "skill genome" of that entity.
- Computing the share of skills that belong to the AI skill group out of the top skills in the selected entity.

For example, the top 50 skills for the occupation of "Engineer" are calculated based on the frequency with which they appear in LinkedIn members' profiles. If four of these skills that engineers possess are AI skills, then AI skills penetration is estimated to be 8% among engineers (i.e. 4/50).

To allow for skills penetration comparisons across countries, the skills genomes are calculated and a relevant benchmark is selected (e.g. OECD average). A ratio is then constructed between a country's and the benchmark's AI skills penetrations, controlling for occupations. For instance,

[15] LinkedIn continuously works on improving methodologies and taxonomies. LinkedIn and the OECD plan to work closely to update this information as the methodology and information evolve.

[16] In information retrieval, tf–idf or TFIDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus (Rajaraman and Ullman, 2011). The tf–idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general (Breitinger, Gipp, and Langer, 2015).

a country’s relative AI skills penetration of 1.5 indicates that, keeping occupations constant, AI skills are 1.5 times as frequent as in the benchmark.

AI skills migration (between countries)

Data on AI skills migration comes from the World Bank Group – LinkedIn “Digital Data for Development” partnership.[17]

LinkedIn migration rates are derived from the self-identified locations of LinkedIn member profiles. For example, when a LinkedIn member updates his or her location from Paris to London, this is counted as a migration.[18]

LinkedIn data provide insights to countries on the skills gained or lost due to migration trends. Skill migration is considered for skills at time t for country A as the country of interest and country B as the source of inflows and destination for outflows. Thus, net skill migration between country A and country B – for country A – is calculated as follows:

$$Net\ Skill\ migration_{a_s,b_s,t} = \frac{Net\ skill\ flows_{a_s,b_s,t}}{Member\ skill\ count_{a_s,t}}$$

Net flows are defined as total arrivals minus departures within the given time period. LinkedIn membership varies considerably between countries, which makes interpreting absolute movements of members from one country to another difficult. To compare migration flows between countries fairly, migration flows are normalised for the country of interest. For example, if country A is the country of interest, all absolute net flows into and out of country A, regardless of origin and destination countries, are normalised based on LinkedIn membership in country A at the end of each year and multiplied by 10 000. Hence, this metric indicates relative talent migration from all countries to and from country A.

To protect LinkedIn member privacy, a minimum threshold of 50 LinkedIn members meeting the criteria being queried is required for the results to be shown on OECD.AI (i.e. gross migration of at least 50 members in each skill group every year).

[17] For more information about this initiative and the methodologies used, please see https://linkedindata.worldbank.org/ and Zhu et al. (2018).

[18] LinkedIn migration rates for 2015 were compared with international migration flow data from the OECD, finding that LinkedIn covered roughly 21.4% of all migration flows in the OECD dataset. Coverage is best for migration of skilled workers between high-income countries.

3 Event Registry data

Classification of AI news by AI Principle and policy area

A list of keywords was created for each AI Principle and policy area by selecting the most relevant related concepts in Event Registry's search engine. To be classified under a certain AI Principle or policy area, a news article must a) be related to AI, and b) contain at least one of the concepts specified for that AI Principle or policy area in its title and/or abstract.

Selection of related online news by AI Principle and policy area

The most relevant news events from the past six months are shown in the "Related online news from Event Registry" lists in both AI Principle and policy areas dashboards. A news 'event' is defined as a collection of news articles covering the same news. It serves as a measure of news impact.

News sentiment analysis

A 'sentiment' score is assigned to each news article and event to determine whether an article or event is 'positive', 'neutral', or 'negative'. The sentiment score is calculated using the "VADER" (Valence Aware Dictionary and sEntiment Reasoner) open-source service, "a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media."[19]

VADER contains a lexicon of words for which a sentiment score has been computed. The sentiment score is a floating value between -1 (negative) and 1 (positive). Additionally, VADER contains a set of rules to determine how sentiment is emphasised with individual words (for instance, "very" emphasises the sentiment, while negations inverse it). The sentiment score of an article results from the analysis of the first five sentences of the article. In practice, news articles rarely have extreme values close to 1 and -1; thus, scores below -0.1 are classified as negative; between -0.1 and 0.1 as neutral; and above 0.1 as positive.

[19] For more information about VADER, please see https://github.com/cjhutto/vaderSentiment.

4 References

Sinha, A.; Shen, Z.; Song, Y.; Ma, H.; Eide, D.; Hsu, B.; and Wang, K. (2015), An Overview of Microsoft Academic Service (MAS) and Applications. In Proceedings of the 24th International Conference on World Wide Web (WWW ’15 Companion). ACM, New York, NY, USA, 243-246. DOI: http://dx.doi.org/10.1145/2740908.2742839.

Breitinger, C.; Gipp, B.; and Langer, S. (2015), Research-paper recommender systems: a literature survey, International Journal on Digital Libraries, 17 (4): 305–338, https://dx.doi.org/10.1007/s00799-015-0156-0. ISSN 1432-5012.

OECD (2019), Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

OECD (forthcoming), Revised outline for practical guidance for the Recommendation of the Council on Artificial Intelligence, https://one.oecd.org/document/DSTI/CDEP(2019)4/REV2/en.

OECD and SCImago Research Group (CSIC) (2016), Compendium of Bibliometric Science Indicators, OECD Publishing, Paris, http://oe.cd/scientometrics.

OECD (2009), OECD Patent Statistics Manual, OECD Publishing, Paris, https://doi.org/10.1787/9789264056442-en

Rajaraman, A. and Ullman, J. (2011), Mining of Massive Datasets, pp. 1–17, https://doi.org/10.1017%2FCBO9781139058452.002, ISBN 978-1-139-05845-2.

Settles, B. (2010), Active Learning Literature Survey, Computer Sciences Technical Report 1648, University of Wisconsin–Madison, http://burrsettles.com/pub/settles.activelearning.pdf.

Shen, Z.; Ma, H.; and Wang, K. (2018), A Web-scale system for scientific knowledge exploration, arXiv, https://arxiv.org/abs/1805.12216.

Wang, K.; Shen, Z.; Huang, C.; Wu, C.; Eide, D.; Dong, Y.; Qian, J.; Kanakia, A.; Chen, A.; and Rogahn, R. (2019), A Review of Microsoft Academic Services for Science of Science Studies, Frontiers in Big Data, https://doi.org/10.3389/fdata.2019.00045.

Zhu, T.; Fritzler, A.; and Orlowski, J. (2018). World Bank Group-LinkedIn Data Insights: Jobs, Skills and Migration Trends Methodology and Validation Results (English). Washington, D.C.: World Bank Group. http://documents.worldbank.org/curated/en/827991542143093021/World-Bank-Group-LinkedIn-Data-Insights-Jobs-Skills-and-Migration-Trends-Methodology-and-Validation-Results.