On a methodology to compute the topics and the topography of DESA publications

White paper

Summary: This is an exploration into the characteristics of DESA publications to help illuminate discussions on emerging topics. It uses a statistical technique to extract information on how the publications relate to each other through statistically generated “topics” and compares the results with an analysis of the WDR and the HDR. We find a high degree of concentration of all publications on a single “topic” that forms the core of the development vocabulary. The less central topics, however, are very informative to understanding the areas of focus of the texts. We also present the results visually using a network diagram.

A characterization of DESA’s publications requires a decision on a system of classification of the texts such as keywords, themes, methodologies used, etc. The set of categories to be applied is fundamental to the results and, as such, requires careful consideration for the purpose of the analysis. There is great value in doing this carefully and methodically, which can result in a powerful tool to help better understand a corpus of work. However, large complex documents such as the major DESA publications are difficult to categorize in just one dimension. Publications have many techniques, approaches, and intersecting topics. The need to better understand the interactions across topics and fields of study will continue to worsen this trend. Clearly, the growing need to analyze vast numbers of texts for relationships requires more sophisticated tools of analysis.

Topic models are a class of statistical techniques that looks for patterns in the use of words in a corpus of text. It goes beyond word frequency analysis and identifies lists of words that occur in statistically meaningful ways. The model’s premise is that any text is authored from a basket of possible words, called “topics”. Using this tool to analyze DESA’s publications should provide interesting insight into what are the topics that form the corpus of work, and what is the degree of specialization and centrality of each publication and topic relative to all others.

**Methodology:**

The selection of the number of topics to be computed can be done either by trial-and-error or by estimating a hierarchical LDA (HLDA) model that will output a best fit of topics. A trial and error method is more time consuming and relies on the observation of the computed topics for relevance and a sense of whether there are too many narrow topics. It can also be bound by the analytical needs of the exercise where a large number of topics may detract from the need to examine broad themes.

An initial consideration is the number of topics that would be of interest to the interest reader to make use of in understanding the themes covered by the publications. Less than ten topics would be too aggregated to provide sufficient information beyond broad topics. Too many topics has exactly the opposite problem as 50 or more topics would provide too much detail and many of the topics might look quite similar to the reader. Nonetheless, there is value in comparing the results of a reasonable number of topics with these extreme cases to see if any insight can be gained.

For this exercise we estimated models with 5, 10, 15, 30, 50, and 100 topics and analyzed the results. As the objective of this study is to inform internal deliberations on themes and tendencies of major publications, we think that the viable choices lie between 10 and 30 topics.

In addition, we also compute an HLDA model to try to gain some more precision into the number of topics that best fit the corpus of texts. The HLDA allows us to more precisely decide on the number of topics to be computed and to hopefully avoid overly specific topics.

**Results:**

The results include a list of the statistically generated “bucket of words”–topics– as well as the degree of association between each topic and publication. We compute the topics for three sets of texts: DESA, WESS, and main texts from DESA plus the WDR and the HDR (“major”).

|  |  |  |  |
| --- | --- | --- | --- |
|  | DESA | WESS | MAJOR |
| # of texts | 233 | 16 | 62 |
| Date range | 1999-2015 | 1999-2014 | 1998-2015 |
| Number of topics |  |  |  |

*DESA*

The DESA dataset includes 233 texts dating back to 2000. These include major flagships (WESS, RWSS), UNTT texts, SG reports, CDP notes, ECOSOC AMR, policy briefs, DESA working papers, and other important texts.

The HLDA analysis, detailed in the appendix, reveals that at the lowest levels of the topic tree we find that the topics quite detailed but still meaningful and insightful to the analysis. At the same time we observe [one] marginal and overly specific topic, which indicates that computing more topics would only result in further narrow results. With this computation we can be confident that a [xx]-topic model would sufficiently capture the themes that are meaningful to this study. This matches the trial-and-error intuition and observations.

The 15 topics identified by the model are ranked in the table below according to their average importance to all of the publications. The texts have a much stronger connection to a specific topic (topic 4, “Economic and social development”) than to any other topic. Nearly all of the documents have topic 4 as their most important topic by a large margin. As mentioned above, regardless of how many topics were calculated, the general concentration pattern was nearly identical. A single topic dominates the relationships with the texts in the corpus, while the other topics represent the degree of specialization of the texts.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 15 topic model of DESA publications, ranked by average importance of that topic to 233 DESA texts | | | | |
| Label | Topic | Mean weight | Median weight | Standard Deviation |
| Economic and social development | Topic 4 | 0.7145046 | 0.7248531 | 0.1063206 |
| Finance and credit | Topic 0 | 0.0506689 | 0.0035587 | 0.1074001 |
| Inequality | Topic 8 | 0.0375066 | 0.0035587 | 0.0940288 |
| Environment | Topic 14 | 0.0360824 | 0.0019704 | 0.0761569 |
| Consultations | Topic 10 | 0.0271975 | 0.0020924 | 0.0743457 |
| Global partnership | Topic 7 | 0.0271384 | 0.0021444 | 0.064495 |
| Trade | Topic 11 | 0.0237492 | 0.0027315 | 0.0583403 |
| Health | Topic 3 | 0.0184279 | 0.0015991 | 0.0508962 |
| Microfinance | Topic 12 | 0.0177164 | 0.0020793 | 0.0560552 |
| Achieving MDGs | Topic 13 | 0.0145957 | 0.0019162 | 0.0438863 |
| Law and corruption | Topic 9 | 0.0119003 | 0.0025381 | 0.0520761 |
| Governance | Topic 2 | 0.0114137 | 0.0020672 | 0.0343886 |
| Conflict | Topic 1 | 0.0104899 | 0.0018773 | 0.0352338 |
| Migration | Topic 6 | 0.0091512 | 0.0017455 | 0.040091 |
| Civil society | Topic 5 | 0.0068717 | 0.0016 | 0.0456011 |

The words that comprise topic 4 are used extensively in all of DESA’s texts analyzed. A quick review of the words and terms that make up that topic helps us understand why this topic is so influential.

|  |  |  |
| --- | --- | --- |
| Topic | Label | 40 leading words in the topic in order of importance |
| 4 | Economic and social development | developing; economic; social; development; world; international; global; united; poverty; public; policy; growth; income; national; high; important; poor; financial; developed; increase; economic\_growth; poverty\_reduction; young\_people; private\_sector; older\_persons; labour\_market; trade\_liberalization; social\_protection; south\_africa; east\_asia; rural\_areas; recent\_years; interest\_rates; saharan\_africa; developing\_world; human\_rights; human\_development; economic\_development; income\_inequality; natural\_resources |

“Economic”, “social”, “development”, and others are at the core of DESA’s mission and focus. As such, it makes sense that DESA’s texts have a common vocabulary that features prominently. The fact that topic 4 includes terms such as “poverty”, “income”, and “growth” speaks to the focus on economic tools and the importance of poverty in all aspects of development.[[1]](#footnote-1) The full list of the 15 estimated topics is below as a word cloud of the top 40 words by order of importance in each topic.

As topic 4 reflects DESA’s *raison d’être*, the other 14 topics can shed light on specializations in DESA publications, or themes that are occasional. Issues of finance, income distribution, the environment, and global partnerships and coordination are at the top. Some issues are also closely associated and indicate that perhaps fewer buckets would be optimal: Law and corruption; governance; civil society are three topics that seem similar to each other.

*WESS*

[wess description and results]

The dataset

The HLDA model

The xx-topic model results

*Major*

[description and results]

The dataset

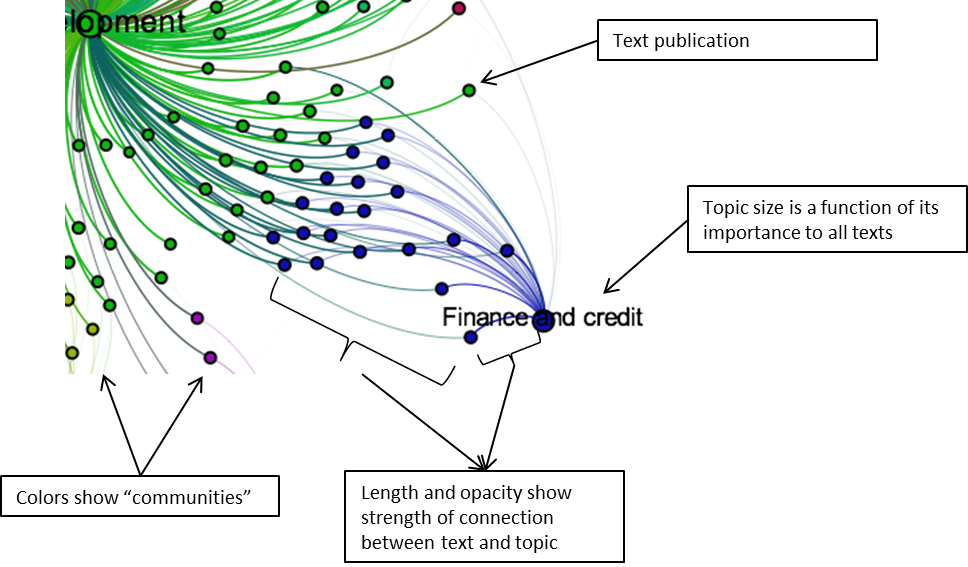
The HLDA model

The 20-topic model results

**Visualizing the results**

The connections between each publication and each topic can be visually represented by the network diagram shown below. The distance between the publication and a topic is a function of the strength of the connection, or how important is the topic to the publication. A topic that is important to many publications will be more centered while a topic that is more specialized and has fewer strong links will be at the periphery.

How to read the network diagrams:



The importance of topic 4 to all DESA publications is reflected in the star-shape of the diagram.[[2]](#footnote-2) The most important topics in DESA publications are those with more and stronger connections to each publication. From the diagram above we can quickly see that “inequality”, “finance and credit”, “environment”, “global partnership”, and “trade” tend to be the ones with more weight to more publications. “Migration” and issues of governance, civil society, and rule of law are further away reflecting their higher specialization in fewer publications. Interestingly, “health” and “achieving MDGs” (a topic that is related to specific interventions to achieve specific MDG goals) are also in the periphery. These topics in the periphery may well represent areas of opportunity for more specialized and focused policy research and advice.

|  |
| --- |
| Network diagram of 233 DESA publications according to importance of the 15 topics |
| C:\Users\Marcelo.Lafleur\Desktop\SunT15B1_filtered.png |
| Note: Links weaker than 10% are not shown. |

[WESS diagram]

|  |
| --- |
| Network diagram of 16 WESS texts according to importance of the 15 topics |
|  |
| Note: Links weaker than 10% are not shown. |

[MAJOR diagram]

|  |
| --- |
| Network diagram of 62 major texts according to importance of the 15 topics |
|  |
| Note: Links weaker than 10% are not shown. |

**Conclusion:**

To our knowledge, this is the first time DESA’s publications have been analyzed in this way. The topology of the network was surprising at first, but is reasonable given the specific mission of DESA and its publications. Preliminary results from a similar analysis of the WDR and the HDR publications show that such a structure is common.

This exercise served not only to better understand the corpus itself, but also to illustrate the value of this technique to DESA’s internal processes. The use of topic modelling techniques to analyze vast quantities of text is commonly used in the humanities. This technique can also serve to characterize DESA’s output and to identify any patterns than are of interest to discussions on emerging themes. The visual representation of the quantified relationships can also serve as an intuitive illustration of gaps or opportunities for research, or to motivate a more careful analysis of the results.

**Appendix:**

*Major*

HLDA results for major publications:

5 levels were requested. The tree of topics is presented below. 22 total topics were estimated. Of these, two look excessively specific and of little relevance to the corpus. Based on this result we select 20 topics to be computed.

151261/62 risk jobs social people labor employment workers financial poor insurance

1231150/62 development developing growth economic income public international world policy services

41568/4 young youth people education school programs health skills work secondary

54616/4 women gender men female womens girls mortality work differences access

27443/4 migration migrants international united labour workers states immigration population refugees

43824/7 climate change emissions carbon human energy adaptation mitigation people world

36625/7 climate carbon change energy emissions water mitigation adaptation costs emission

75706/7 energy development technology food technologies climate sustainable developing global emissions

23/3 unaddressedas significantspaceforchangefollowing movementand gangmember gutierrez backsliding monitoringis boer ment firmswhich

21054/3 private health education sector electricity agricultural privatization technology market public

39748/3 violence security international conflict institutions political national justice support fragile

82515/44 agricultural rural agriculture food poverty land farmers research market areas

211445/44 human development people rights health global political education hdi inequality

96236/10 financial trade growth capital development developing banks international debt markets

69198/10 poverty social labour inequality employment economic income poor growth workers

103633/20 water environmental sanitation people poor human urban development access consumption

54315/2 economic areas migration migrants cities urban transport people land density

21612/1 older age persons health population ageing care pension labour developed

19133/1 financing global development health mechanisms innovative funds tax fund international

42127/4 services health poor public service education providers people school water

167/4 dampens interviewed dwellers retrospect everyoneand profilewhich livedto guises rightsstemming regionoften

72617/4 firms information knowledge investment competition institutions market climate governments private

|  |  |  |  |
| --- | --- | --- | --- |
| DESA publications: 40 leading words in each topic weighted by their importance in each topic  (colors are not meaningful) | | | |
| Topic 0 - Finance and credit | Topic 1 - Conflict | Topic 2 - Governance | Topic 3 - Health |
| C:\Users\Marcelo.Lafleur\Desktop\T15b1 Topic clouds\T0.png | C:\Users\Marcelo.Lafleur\Desktop\T15b1 Topic clouds\T1.png | C:\Users\Marcelo.Lafleur\Desktop\T15b1 Topic clouds\T2.png | C:\Users\Marcelo.Lafleur\Desktop\T15b1 Topic clouds\T3.png |
| Topic 4 - Economic & social development | Topic 5 - Civil society | Topic 6 - Migration | Topic 7 - Global partnership |
| C:\Users\Marcelo.Lafleur\Desktop\T15b1 Topic clouds\T4.png |  |  |  |
| Topic 8 - Inequality | Topic 9 - Law and corruption | Topic 10 - Consultations | Topic 11 - Trade |
| C:\Users\Marcelo.Lafleur\Desktop\T15b1 Topic clouds\T8.png |  |  |  |
| Topic 12 - Microfinance | Topic 13 - Achieving MDGs | Topic 14 - Environment |  |
|  |  | C:\Users\Marcelo.Lafleur\Desktop\T15b1 Topic clouds\T14.png |  |

|  |  |  |  |
| --- | --- | --- | --- |
| WESS: 40 leading words in each topic weighted by their importance in each topic  (colors are not meaningful) | | | |
| Topic 0 - Finance and credit | Topic 1 - Conflict | Topic 2 - Governance | Topic 3 - Health |
| C:\Users\Marcelo.Lafleur\Desktop\T15b1 Topic clouds\T0.png | C:\Users\Marcelo.Lafleur\Desktop\T15b1 Topic clouds\T1.png | C:\Users\Marcelo.Lafleur\Desktop\T15b1 Topic clouds\T2.png | C:\Users\Marcelo.Lafleur\Desktop\T15b1 Topic clouds\T3.png |
| Topic 4 - Economic & social development | Topic 5 - Civil society | Topic 6 - Migration | Topic 7 - Global partnership |
| C:\Users\Marcelo.Lafleur\Desktop\T15b1 Topic clouds\T4.png |  |  |  |
| Topic 8 - Inequality | Topic 9 - Law and corruption | Topic 10 - Consultations | Topic 11 - Trade |
| C:\Users\Marcelo.Lafleur\Desktop\T15b1 Topic clouds\T8.png |  |  |  |
| Topic 12 - Microfinance | Topic 13 - Achieving MDGs | Topic 14 - Environment |  |
|  |  | C:\Users\Marcelo.Lafleur\Desktop\T15b1 Topic clouds\T14.png |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Major publications: 40 leading words in each topic weighted by their importance in each topic  (colors are not meaningful) | | | |
| Topic 0 - Finance and credit | Topic 1 - Conflict | Topic 2 - Governance | Topic 3 - Health |
| C:\Users\Marcelo.Lafleur\Desktop\T15b1 Topic clouds\T0.png | C:\Users\Marcelo.Lafleur\Desktop\T15b1 Topic clouds\T1.png | C:\Users\Marcelo.Lafleur\Desktop\T15b1 Topic clouds\T2.png | C:\Users\Marcelo.Lafleur\Desktop\T15b1 Topic clouds\T3.png |
| Topic 4 - Economic & social development | Topic 5 - Civil society | Topic 6 - Migration | Topic 7 - Global partnership |
| C:\Users\Marcelo.Lafleur\Desktop\T15b1 Topic clouds\T4.png |  |  |  |
| Topic 8 - Inequality | Topic 9 - Law and corruption | Topic 10 - Consultations | Topic 11 - Trade |
| C:\Users\Marcelo.Lafleur\Desktop\T15b1 Topic clouds\T8.png |  |  |  |
| Topic 12 - Microfinance | Topic 13 - Achieving MDGs | Topic 14 - Environment |  |
|  |  | C:\Users\Marcelo.Lafleur\Desktop\T15b1 Topic clouds\T14.png |  |

1. An analysis of the World Bank’s World Development Report and of UNDP’s Human Development Report mirrors this pattern: one key topic is strongly represented in all publications, while specialization is indicated by the other “satellite” topics. [↑](#footnote-ref-1)
2. The colors represent estimated modularity, or “communities” of publications and topics into sub-groups. The size of each label is a weighted function of the degree of association between each topic with each text. [↑](#footnote-ref-2)