NATIONAL INSTITUTE OF FOOD AND AGRICULTURE
Data Summit: Changing the Face, Place, and Space of Agriculture

TOPIC MODELING INPUT RECEIVED IN THE "IDEAS ENGINE"

PREAMBLE

An analysis was conducted of stakeholder input received through the Ideas Engine in conjunction with the National Institute of Food and Agriculture's (NIFA) Agriculture *Data Summit: Changing the Face, Place, and Space of Agriculture*. An empirical approach to stakeholder input analysis was used to provide an objective and comprehensive global view of stakeholder input to provide a sense of the areas that appear to be of most interest to stakeholders. Major categories reflect relevance across NIFA program areas and demonstrate relationships among the key topics that were identified within the categories. Key topics generated from the analysis were further examined for specificity.

STAKEHOLDER INPUT ANALYSIS

The analysis provided here represents a depiction of over 170 written ideas received from six groups of individuals with 40-75 participants each. These stakeholders comprise industry, academe, professional organizations, practitioners, and producers.

Identification of key concepts using topic modeling

As an alternative to attempting to read and synthesize individual documents, a method of topic modeling was employed that uses automated natural language processing (NLP). This approach produces a term map that visualizes the structure of text by showing the relationships among important terms in the text fields.

Results

Topic modeling of stakeholder comments resulted in the identification of over 280 key terms, which were loosely clustered in relation to one another. Based on the modeling conducted, stakeholder comments fell into 14 general clusters, which were further grouped by relevance to each other, resulting in 12 major meaningful clusters for analysis. These clusters represent the following topics, weighted by dominance and depicted by different colored dots in Figure 1.

1. Data infrastructure and management (red)
2. Applications and use of data, entities affected by data (green)
3. Creation, collection, provenance, and characteristics of data (royal blue)
4. Training, programs, student, and knowledge needs around data (meconium)
5. Federal agencies, principles, and protocols associated with data (purple)
6. University, team, community, and public/private aspects of data (yellow)
7. Data producers, engineers, scientists, and researchers of data (baby blue)
8. Big corporations/commercial entities (Facebook, Amazon, Google, Microsoft, Apple, Oracle) in data (golden brown)
9. Privacy, security, confidentiality, and quality data issues (greenish brown)

10. Biological and interoperable data systems (pink)
11. Bibliometrics, altmetrics, text and data mining (brown)
12. Data sharing, repositories, and analysis (blue green)

Comments covered a wide range of aspects concerning data. The most dominant cluster (24% of total items) related to data infrastructure and management. The second most dominant cluster (17%) centered around the use of data and how consumers, producers, the environment, and other entities are affected by data. These two top-ranking clusters significantly overlapped, suggesting the issues in each are related and/or dependent on one another. The third largest cluster (10%) was related to creation, collection, provenance, and characteristics of data. Remaining clusters were of similar significance (3-8%).

Within each cluster of topics, specific issues were raised by stakeholders. Having identified the overarching topics of importance to stakeholders, further examination and prioritization of the issues can be seen in the output of the Ideas Engine.

Figure 1. Topic model of stakeholder input on data depicting 12 primary color-coded clusters of topics.