Better Living In North Carolina: Challenges of Presenting Agricultural Statistics From The Past

James R. Stewart Jr. :: “Better Living” Digital Project Librarian :: North Carolina State University Libraries :: Special Collections Research Center :: Campus Box 7111 :: Raleigh, NC 27695-7111

Introduction

Agricultural legacy datasets are now available in the new LSTA digital project “Better Living In North Carolina: Bringing Science and Technology To The People”. “Better Living” is a partnership project between NC State University Libraries and the F. D. Bluford Library at NC A&T State University to digitize NC Cooperative Extension Service reports & materials.

The sought after statistical data in these reports proved to be perplexing to users. Explanations for color coded data were missing, throughout these reports, two sets of numbers were used for the sought after statistical data in these reports proved to be perplexing to users. Explanations for color coded data were missing, the number of data sets are so numerous. We planned for ways to make them more accessible, only to became aware of the complications in extracting this data digitally for modern researchers.

Challenges of Extracting Data From Our Digital Collection.

How can these figures be used if their meaning is lost? Throughout these reports, two sets of numbers were used for data (above). After extensive research the key to the color coded data was found to be the number of corresponding county agents. Would the data have much use if this was not discovered?

Data collections standards change over time. According to a NASS report, probability sampling and computer based measurements became more common after 1957. How much does that affect validity of pre-1957 data?

Which data sets should be prioritized? A single annual report can contain over 40 categories and over 200 questions. Ultimately, digital projects must decide on priority data sets based on patron use. Our data would be more valuable with multiple data entry points, but this would be costly time wise.

Effective (efficient) ways to extract data

Isolating text/writing with ICR and OCR technology? Several of the extension reports have handwritten stats and comments. These may require specialized handwriting recognition models.

Manual data extracting at the metadata collection level?

Recording the statistical figures, (questions and numbers) as part of the metadata process would be tedious and add significant time to the project.

Text Mining for Keywords

Keywords, names, and tags related to agricultural data and the NC Cooperative Extension Service can be added to an original “process lexicon ” for processing.

Value for today’s researchers

Figures and numbers from archival documents can confirm the accuracy of historic (qualitative) sources in archives. All agricultural annual statistical reports parallel an annual narrative or qualitative report (above). Extracted statistical figures may prove to be more accurate than, or validate, text found in narrative reports.

Archives + Digital Libraries + Data Science

Digital collections are potential clearinghouses of statistical data. All 50 United States have at least one library with a digital collection of Cooperative Extension documents.

Data Visualizations. Most visualizations of digital collections are created form descriptive metadata. Data visualizations of extracted keywords and figures from annual reports would be a unique reference tools for agricultural research.

Opportunities for computer & data scientists to explore digital collections. Outreach of the “Better Living” collection has sparked interest in computer science majors and visualization specialists. Manual or computational data extraction could be done as a graduate fellowship or a separate grant project.

Challenges of Presenting Agricultural Statistics From The Past

Introduction

Agricultural legacy datasets are now available in the new LSTA digital project “Better Living In North Carolina: Bringing Science and Technology To The People”. “Better Living” is a partnership project between NC State University Libraries and the F. D. Bluford Library at NC A&T State University to digitize NC Cooperative Extension Service reports & materials.

The sought after statistical data in these reports proved to be perplexing to users. Explanations for color coded data were missing, throughout these reports, two sets of numbers were used for the sought after statistical data in these reports proved to be perplexing to users. Explanations for color coded data were missing, the number of data sets are so numerous. We planned for ways to make them more accessible, only to became aware of the complications in extracting this data digitally for modern researchers.

Challenges of Extracting Data From Our Digital Collection.

How can these figures be used if their meaning is lost? Throughout these reports, two sets of numbers were used for data (above). After extensive research the key to the color coded data was found to be the number of corresponding county agents. Would the data have much use if this was not discovered?

Data collections standards change over time. According to a NASS report, probability sampling and computer based measurements became more common after 1957. How much does that affect validity of pre-1957 data?

Which data sets should be prioritized? A single annual report can contain over 40 categories and over 200 questions. Ultimately, digital projects must decide on priority data sets based on patron use. Our data would be more valuable with multiple data entry points, but this would be costly time wise.

Effective (efficient) ways to extract data

Isolating text/writing with ICR and OCR technology? Several of the extension reports have handwritten stats and comments. These may require specialized handwriting recognition models.

Manual data extracting at the metadata collection level?

Recording the statistical figures, (questions and numbers) as part of the metadata process would be tedious and add significant time to the project.

Text Mining for Keywords

Keywords, names, and tags related to agricultural data and the NC Cooperative Extension Service can be added to an original “process lexicon ” for processing.

Value for today’s researchers

Figures and numbers from archival documents can confirm the accuracy of historic (qualitative) sources in archives. All agricultural annual statistical reports parallel an annual narrative or qualitative report (above). Extracted statistical figures may prove to be more accurate than, or validate, text found in narrative reports.

Archives + Digital Libraries + Data Science

Digital collections are potential clearinghouses of statistical data. All 50 United States have at least one library with a digital collection of Cooperative Extension documents.

Data Visualizations. Most visualizations of digital collections are created form descriptive metadata. Data visualizations of extracted keywords and figures from annual reports would be a unique reference tools for agricultural research.

Opportunities for computer & data scientists to explore digital collections. Outreach of the “Better Living” collection has sparked interest in computer science majors and visualization specialists. Manual or computational data extraction could be done as a graduate fellowship or a separate grant project.

Sources

