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, plus 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.

15

POULTRY, DAIRY CATTLE, BEEF CATTLE, SHEEP, SWINE, AND HORSES

Report Only This Year's Extension Activities that are Supported by Records

Item	(a) Poultry	(b) Dairy cattle	(c) Beef cattle	(d) Sheep	(e) Swine	(f) Horses and mules	
	20	8			1		
131. Number of method demonstration meetings held	132	37			4		131
	16	7					
132. Number of adult result demonstrations completed or carried into the next year	2410	1119					132
	10	4					
133. Number of animals involved in these completed adult result demonstrations	58218	374					133
	3	2					
134. Total profit or saving on adult result demonstrations completed	4277	1436					134
	3	1			1		
(1) Boys	48	6			2		

What do these red numbers mean?

(Above) Image capture from Combined Annual Report of Extension Workers 1946

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

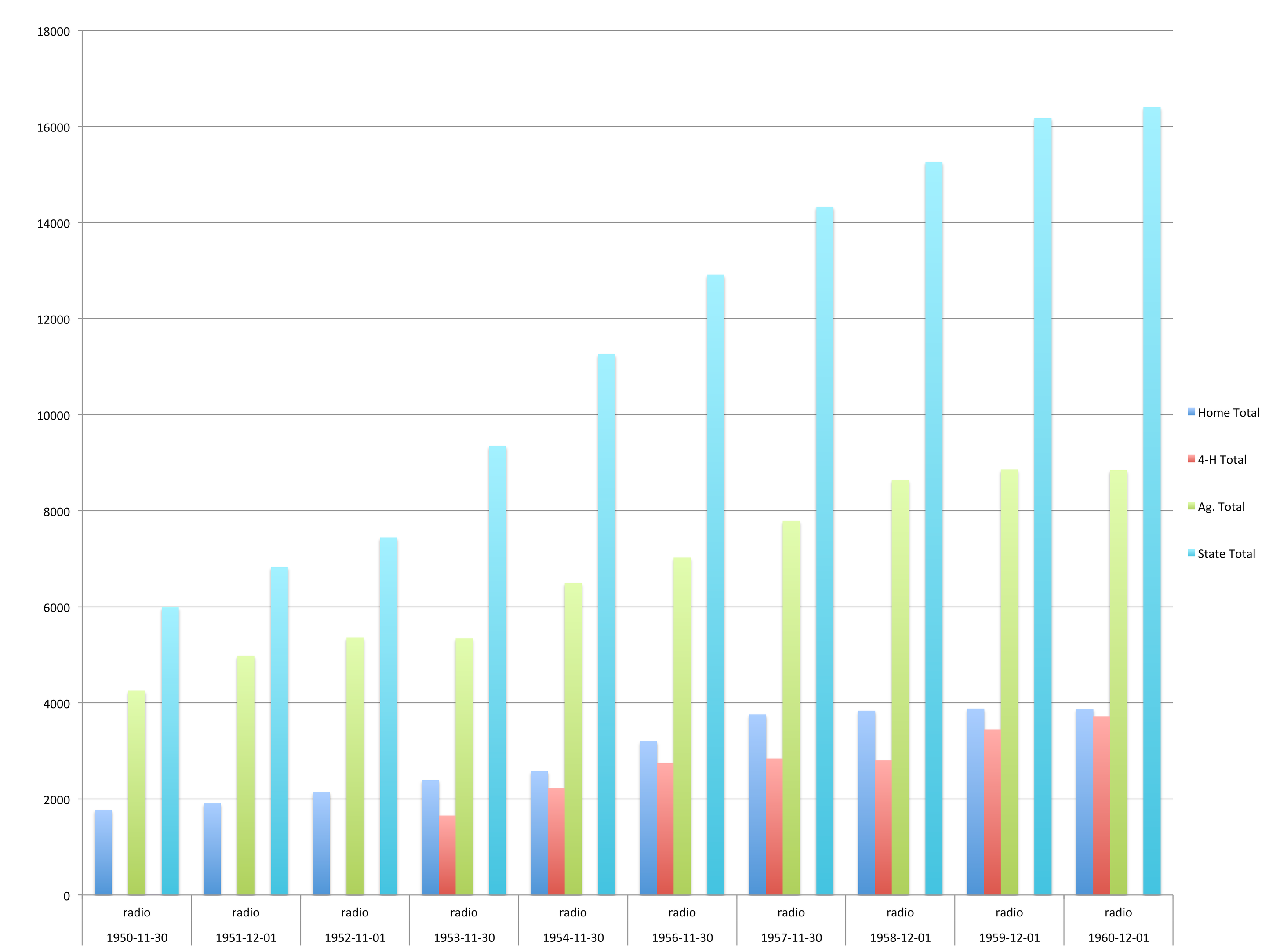
Isolating text/writing with ICR and OCR technology? Several of the extension reports have handwritten stats and comments. These may require specialized handwriting recognition models.

a. Used recipes for new ways to prepare and use: tomatoes 66 12419, green and yellow vegetables 78 20537, cabbage 55 8318, white potatoes 56 7859, sweet potatoes 66 10973, apples 58 18215, chicken 60 10209, eggs 57 9951, meat 56 13239, fresh milk 59 17007, dry milk 43 5027.

Image capture from an early county home demonstration report. Some reports have handwritten data.

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.



NC Cooperative Extension Service: Combined County Use of Radio – Report Years 1950 – 1960. Compiled from NC Cooperative Extension Annual Reports. Extracted data from this digital collection potentially fulfills many research questions related to agricultural growth in North Carolina

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

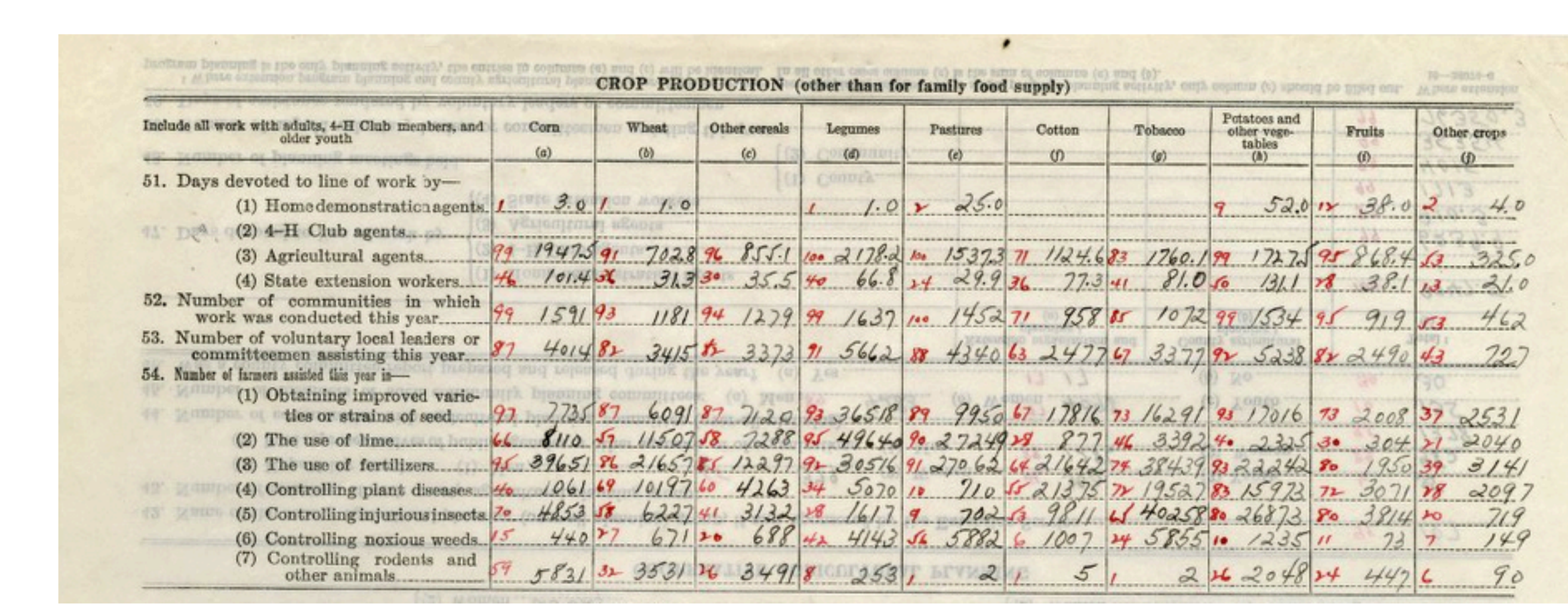
Digital collections are potential clearinghouses of statistical data. All 50 United States have at least one library with a digital collection of Cooperative Extension documents.

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.

Figures from the *Combined Annual Report of County Extension Workers 1944* (below) and a page from the *Cooperative Extension Annual Statistical Report of 1944*, (right)

Can keywords and numbers be extracted from digital scans like these to create visualizations of agriculture demonstrations during World War II.



Sources

Allen, Rich. 2007. *Safeguarding America's agricultural statistics a century of successful and secure procedures, 1905-2005*. [Washington, D.C.]: U.S. Dept. of Agriculture, National Agricultural Statistics Service. http://purl.access.gpo.gov/GPO/LPS116937.

Hybrid grammar language model for handwritten historical documents recognition. IbPRIA (Conference), João Miguel Sanches, Luisa Micó, and Jaime S. Cardoso. 2013. *Pattern recognition and image analysis: 6th Iberian Conference, IbPRIA 2013, Funchal, Madeira, Portugal, June 5-7, 2013.*

Li, J., H. J. Wang, and X. Bai. 2015. "An intelligent approach to data extraction and task identification for process mining". *INFORMATION SYSTEMS FRONTIERS.* 17 (6): 1195-1208.

NCSU LIBRARIES

FBluford Library NORTH CAROLINA A&T STATE UNIVERSITY