6.23.17

Center *for* Research Libraries
GLOBAL RESOURCES NETWORK

Defining the “Critical Corpus”

A Status Report on the CRL Analysis

Critical Corpus

Introduction

Critical Corpus

Mellon Funded Planning Project

To plan and measure the strategic print preservation efforts of North American libraries for Social Sciences and Humanities

Examined

- 19 collections
- 2.5 million records

Identified

~462,850 unique SSH titles

- ~40,720 (9%) in PAPR

Critical Corpus

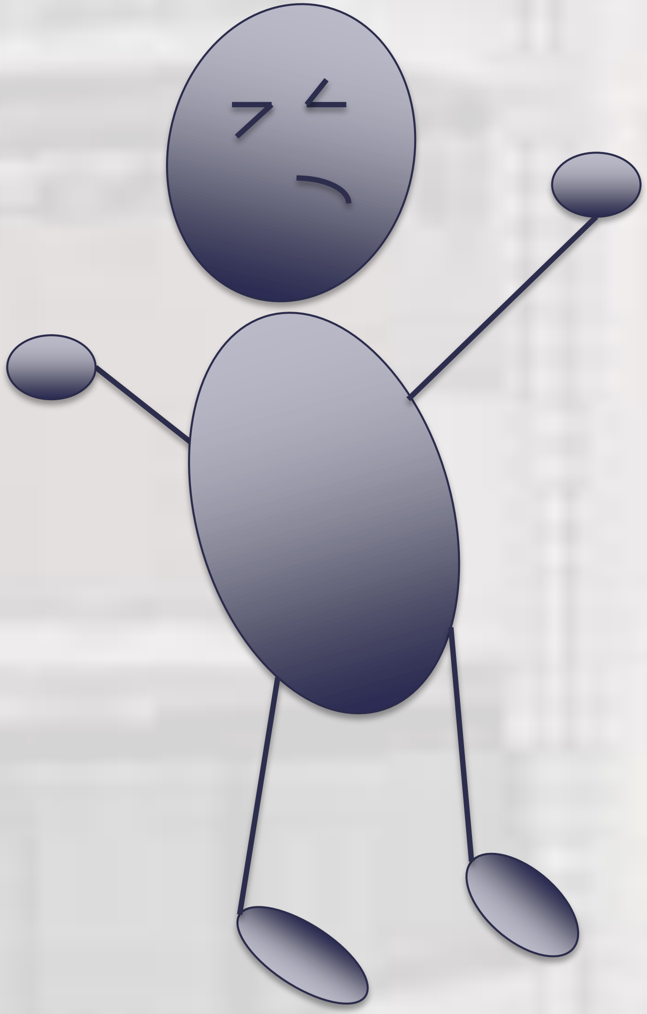
Objectives

- To define the costs and requirements for preserving the “universe” of Humanities and Social Sciences serials.

- To develop and cost out a methodology and strategy to identify the “critical corpus” of journal literature published in print form and important to academic research in the humanities and social sciences.

- To develop a significantly large list of titles to lay the groundwork for review and curation of a final list.

Major Problems with SSH Records



- 31,209 title records (6.7%) lacked a numeric identifier. The range for missing identifiers were .5%-17.3%
- ~129,000 records (28%) lacked classification in the final analysis.
- 49,009 records (11%) lacked country, language, or date information
- Five lacked titles

Critical Corpus

Highlights

Center *for* Research Libraries
GLOBAL RESOURCES NETWORK

Starting Point

Requested full MARC records from Columbia, Cornell, NYPL & Princeton

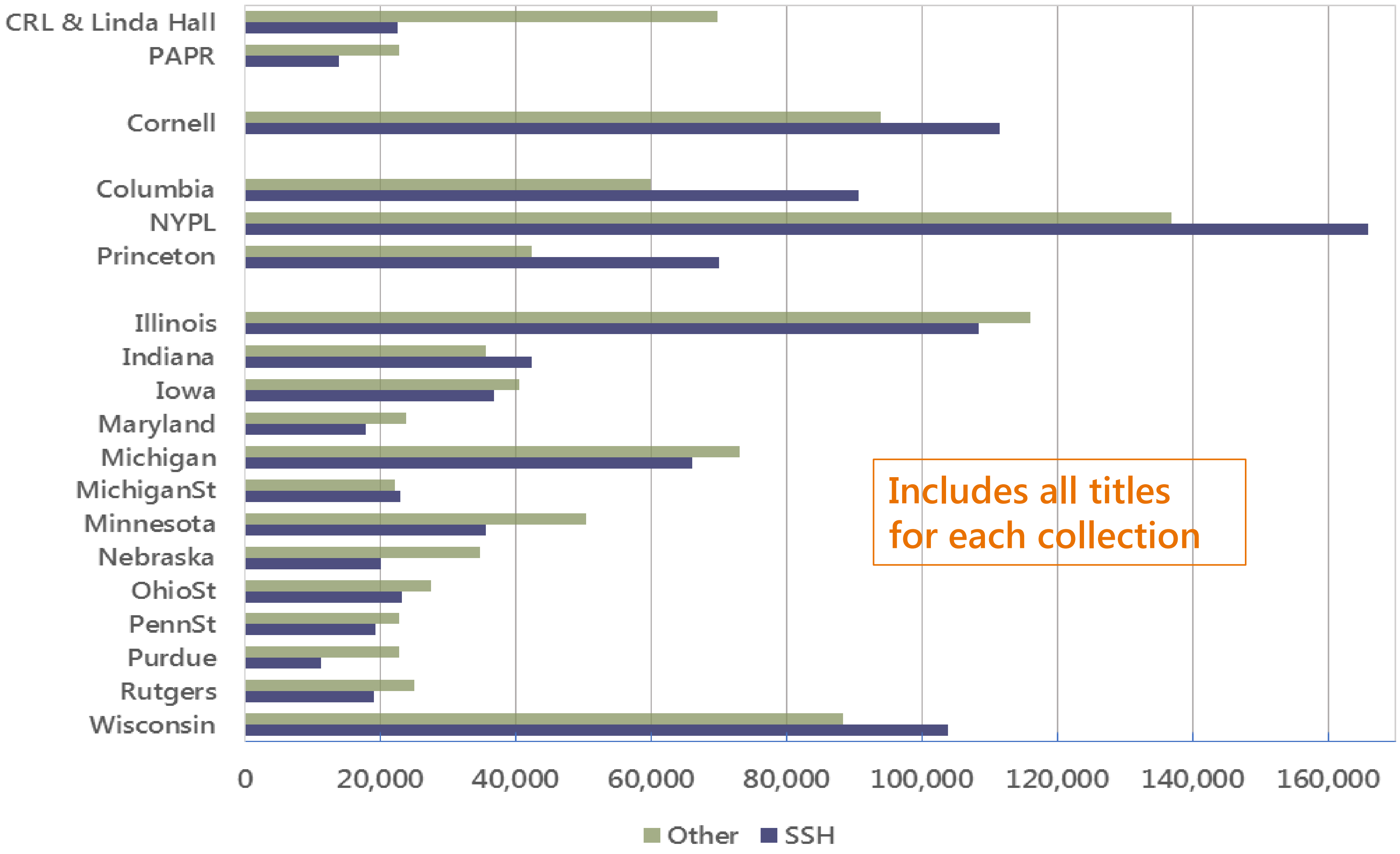
Added previously processed records from 13 BTAA libraries

Downloaded records from PAPR

CORNELL UNIVERSITY FOUNDED A.D. 1865
B1G ACADEMIC ALLIANCE
PAPR

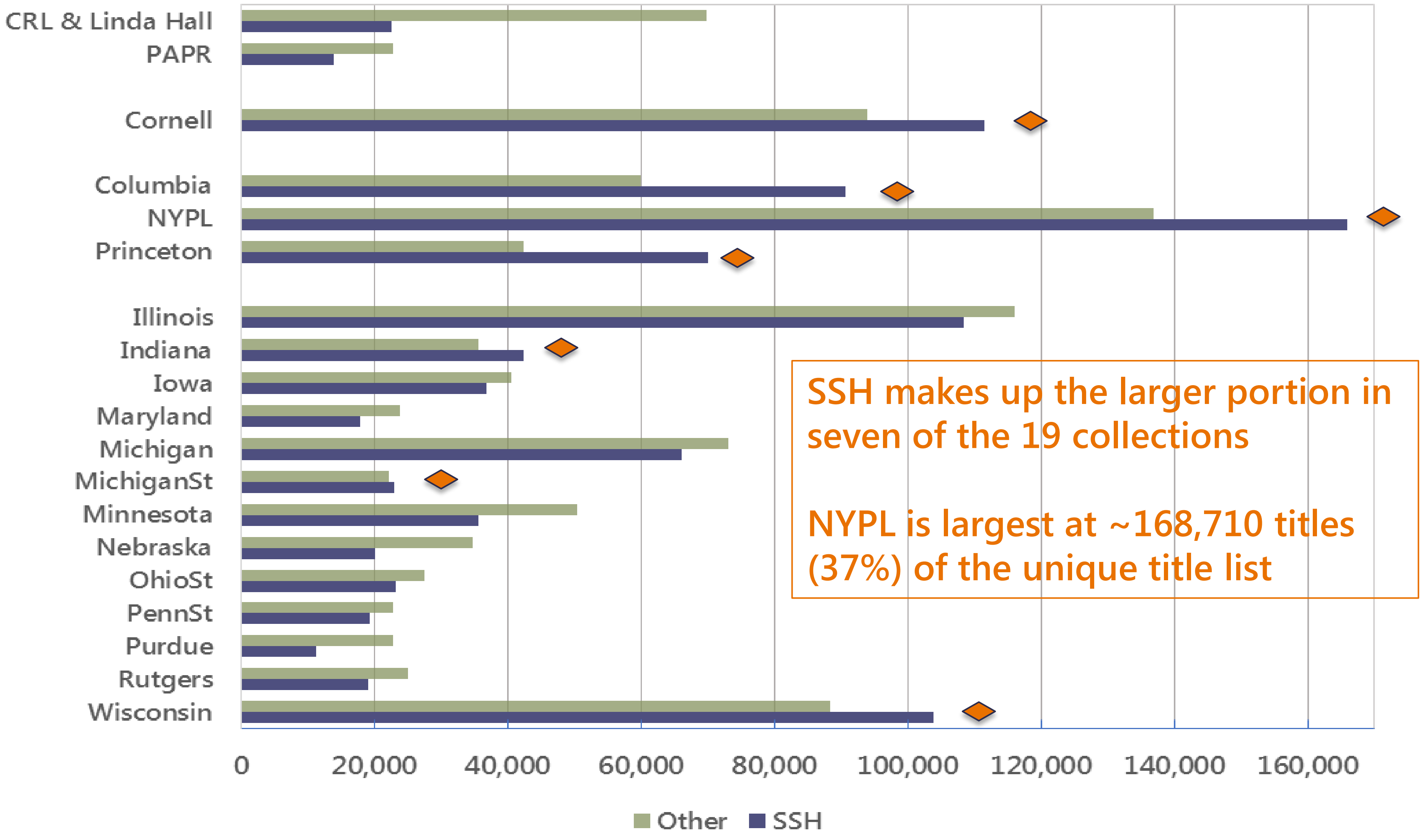
2.5 M Records

Breakdown of Social Sciences & Humanities vs. Other LC Classes in Each of the Collections



Includes all titles for each collection

Breakdown of Social Sciences & Humanities vs. Other LC Classes in Each of the Collections



SSH makes up the larger portion in seven of the 19 collections

NYPL is largest at ~168,710 titles (37%) of the unique title list

Highlights of Breakdown of Each Collection by SSH LC Class

Social Sciences (H) was strongest in all collections

Industries. Land use. Labor. (HD) made up 28.3% of H

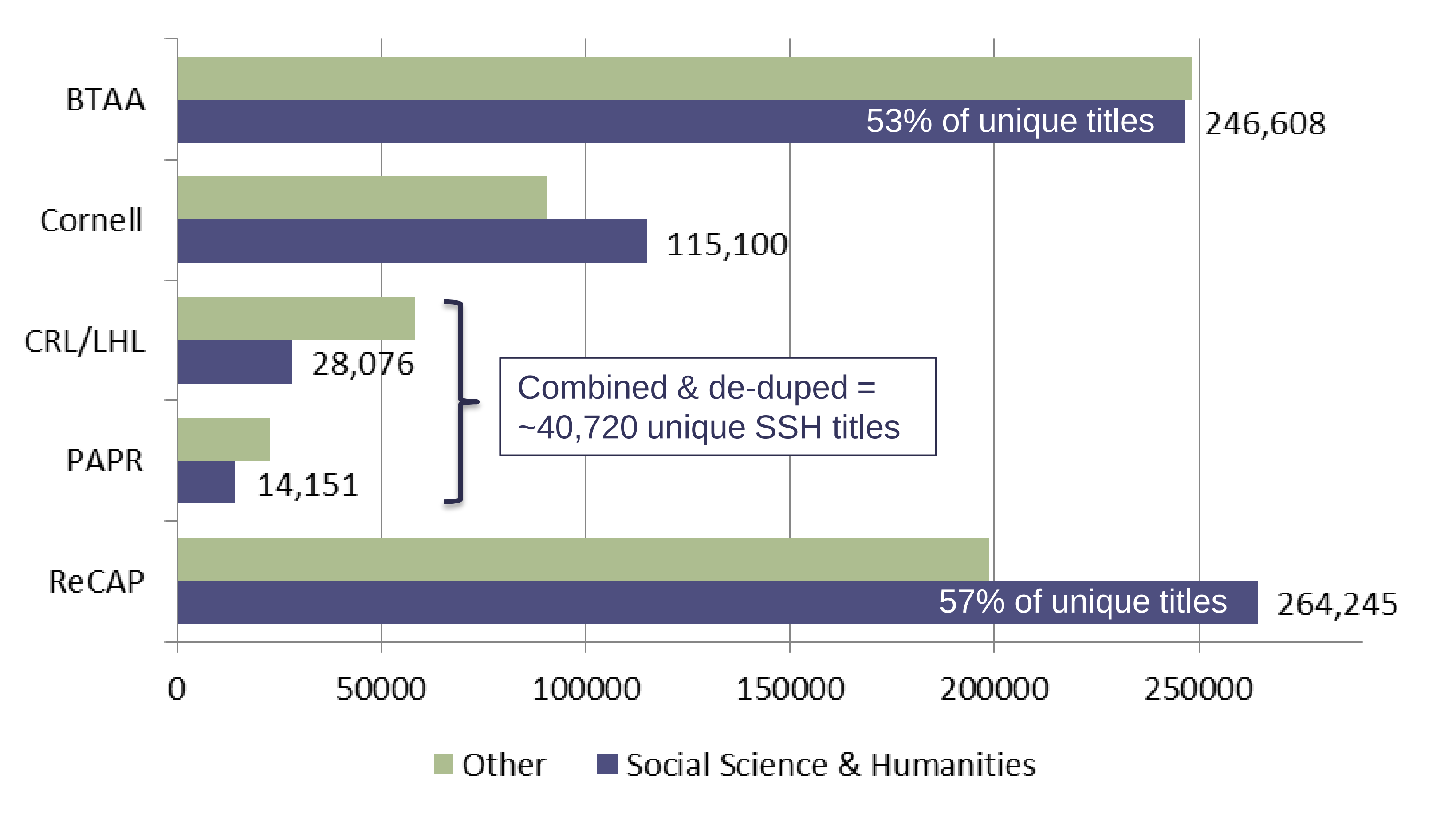
Second strongest classes

European History (D)	Language & Literature (P)
Columbia	Illinois
Cornell	Indiana
Michigan	Iowa
MichiganSt	Maryland
NYPL	OhioSt
Princeton	PennSt
Wisconsin	Purdue
	Rutgers

Corpus collection strengths & weaknesses within individual collections.

LC Class	Min	Collection	Max	Collection	Median	Range
B	2.4%	PennSt	12.9%	Columbia	5.0%	10.5
C	0.6%	CRL/LHL	1.9%	Wisconsin	1.1%	1.3
D	6.5%	Purdue	16.8%	Michigan	11.6%	10.3
E	0.7%	CRL/LHL	2.6%	PAPR	1.8%	1.9
F	1.2%	CRL/LHL	6.1%	PAPR	3.1%	4.9
G	2.7%	Columbia	7.2%	CRL/LHL	5.2%	4.5
H	34.6%	OhioSt	56.8%	MichiganSt	46.2%	22.2
J	4.2%	CRL/LHL	7.4%	MichiganSt	5.6%	3.2
M	0.7%	CRL/LHL	4.9%	Maryland	2.0%	4.2
N	2.1%	MichiganSt	11.1%	Maryland	5.0%	8.9
P	8.4%	MichiganSt	24.9%	OhioSt	13.1%	16.6

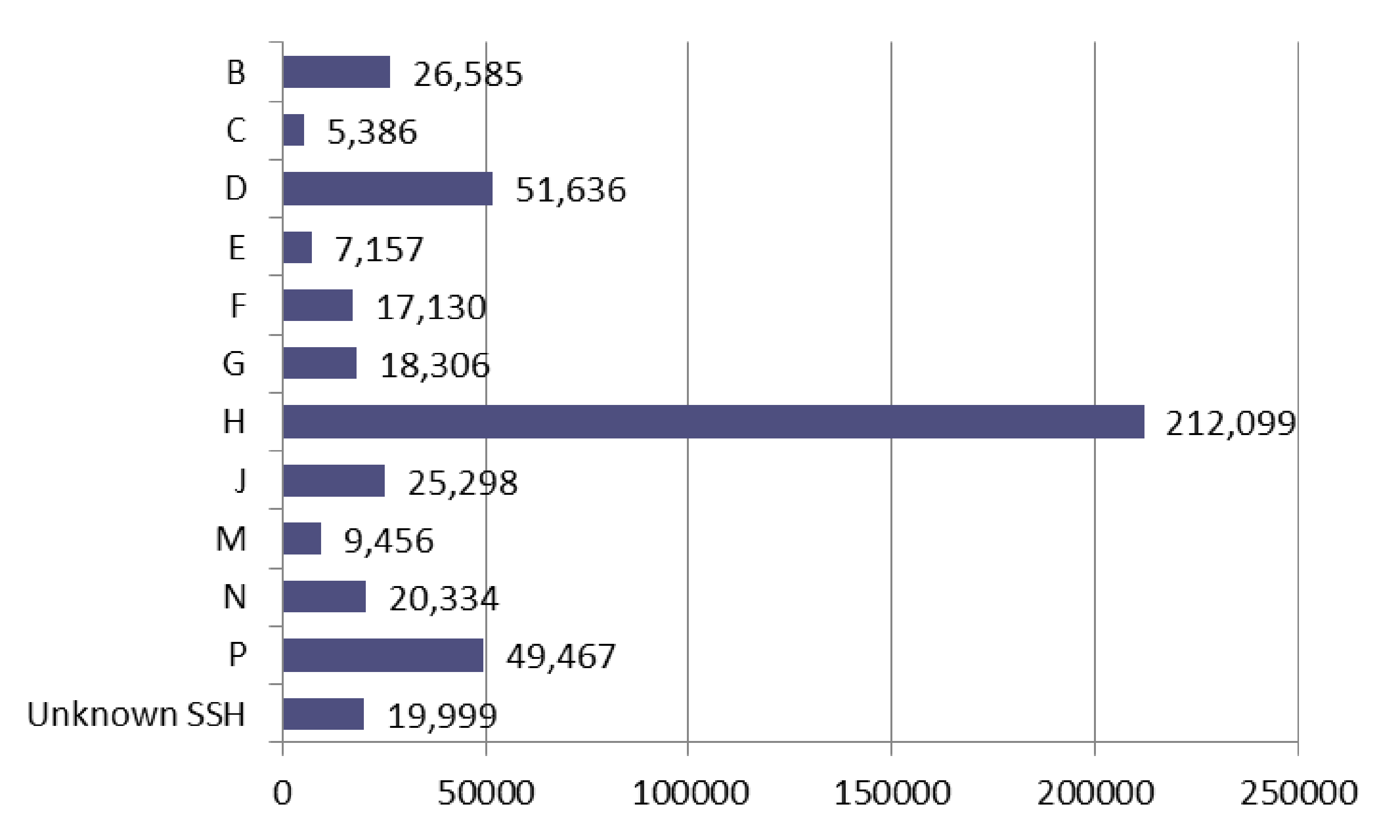
Aggregated Corpus: Breakdown of Social Sciences & Humanities vs. Other



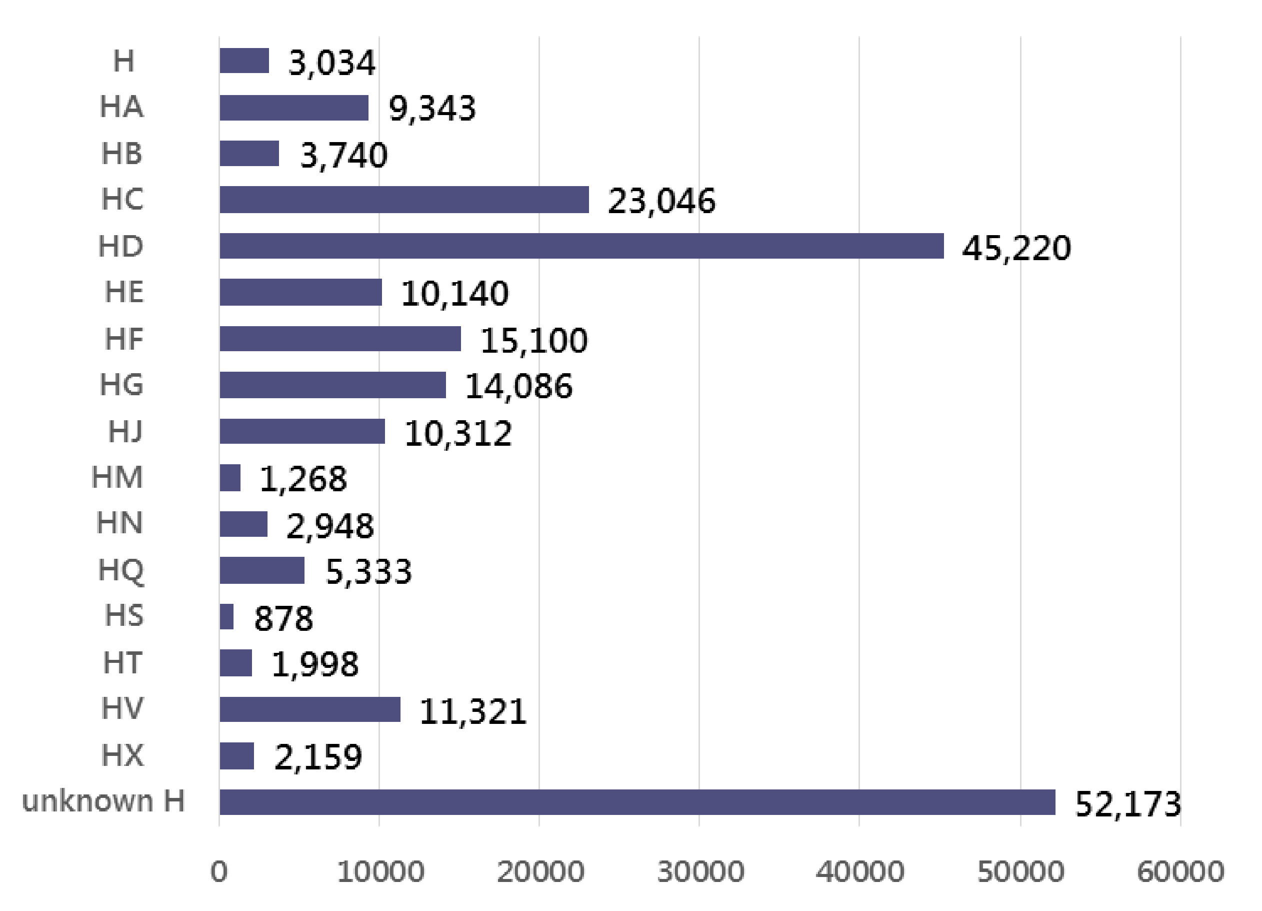
Combined collections were de-duped within groups

Aggregated Corpus

After de-duping across partnerships,
462,850 unique SSH titles were identified

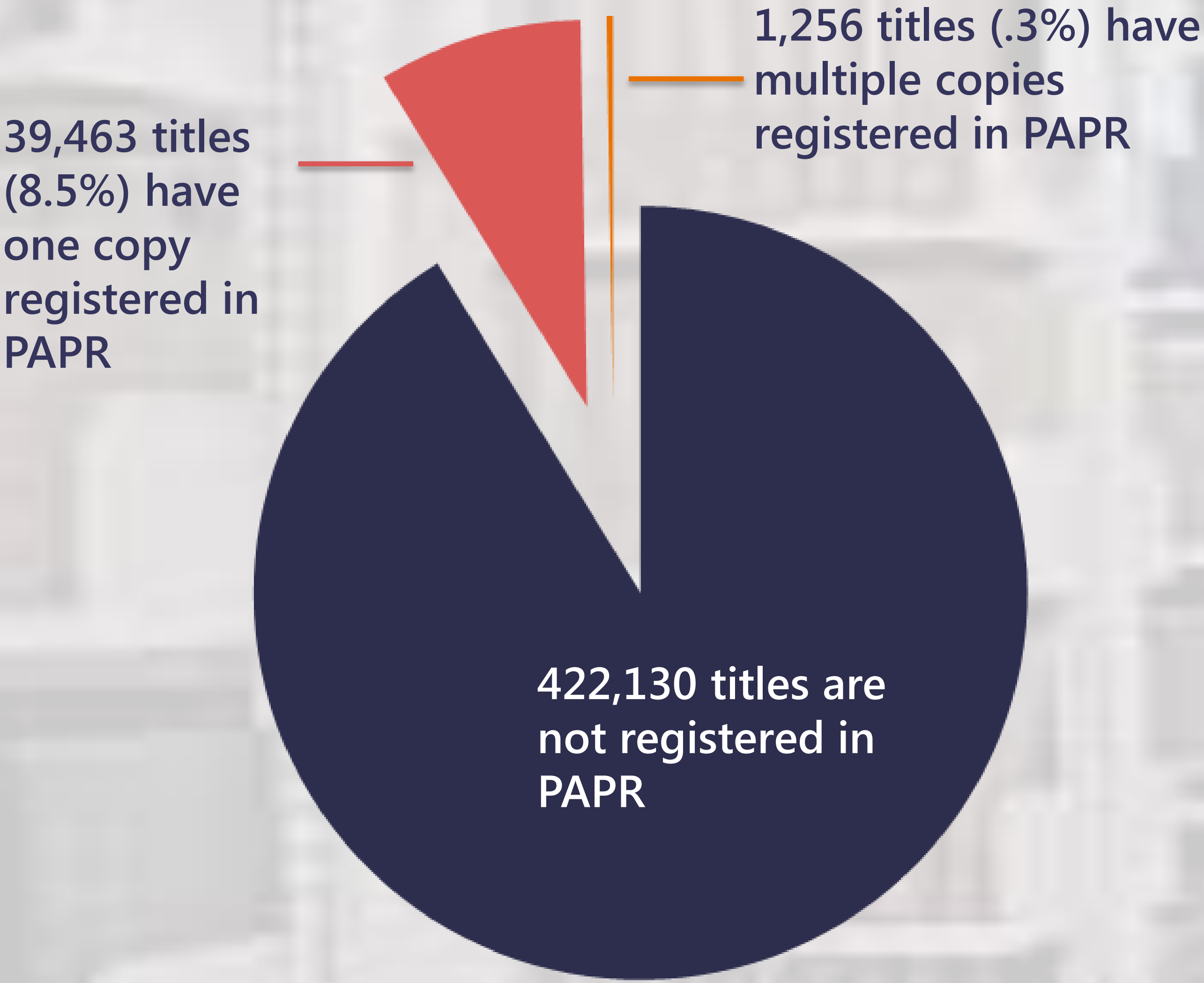


Aggregated Corpus
Distribution of H subclasses

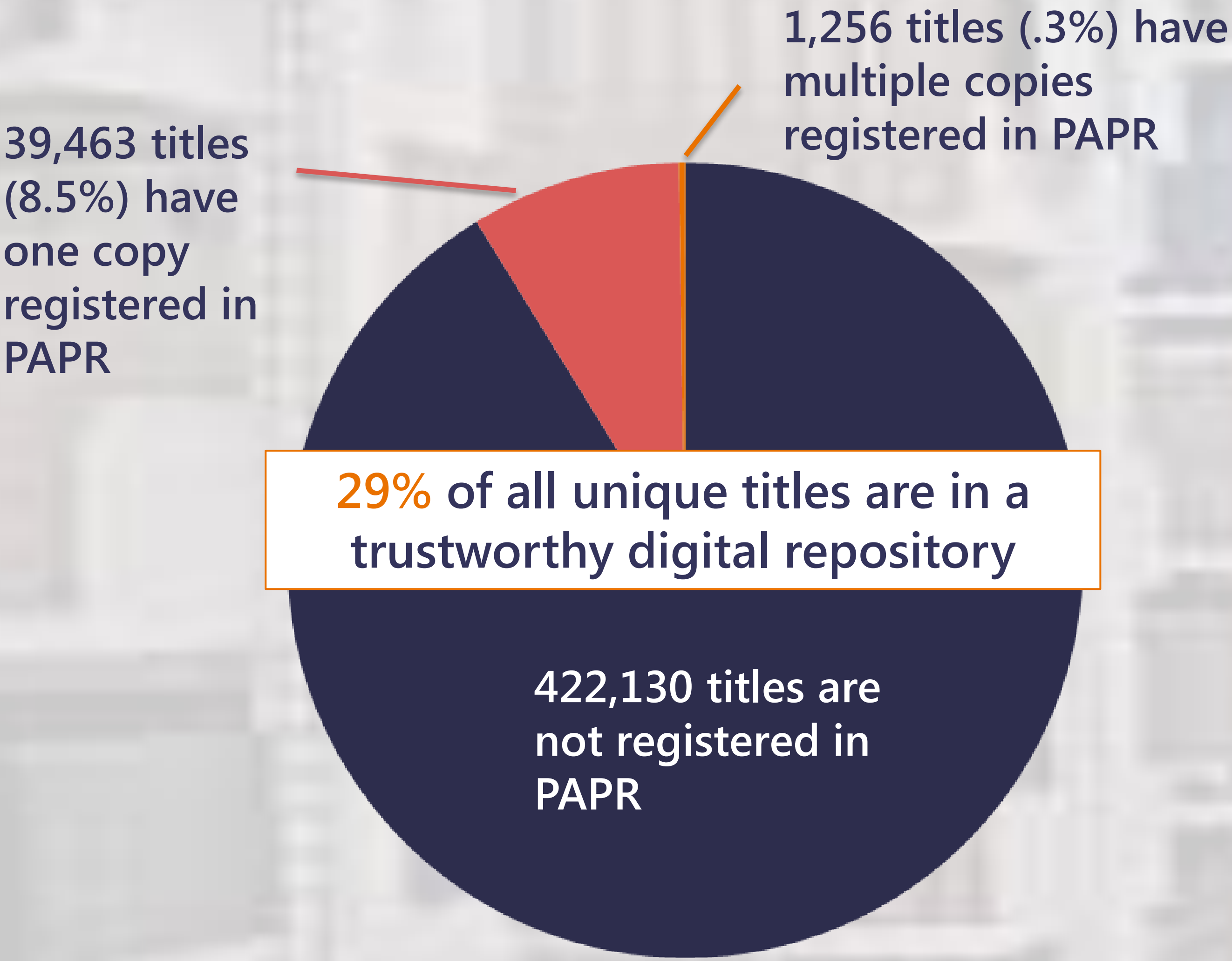


Over 50,000 titles (25%) could be assigned a classification letter but not subclass with an automated routine.

Archive Status of Unique Titles

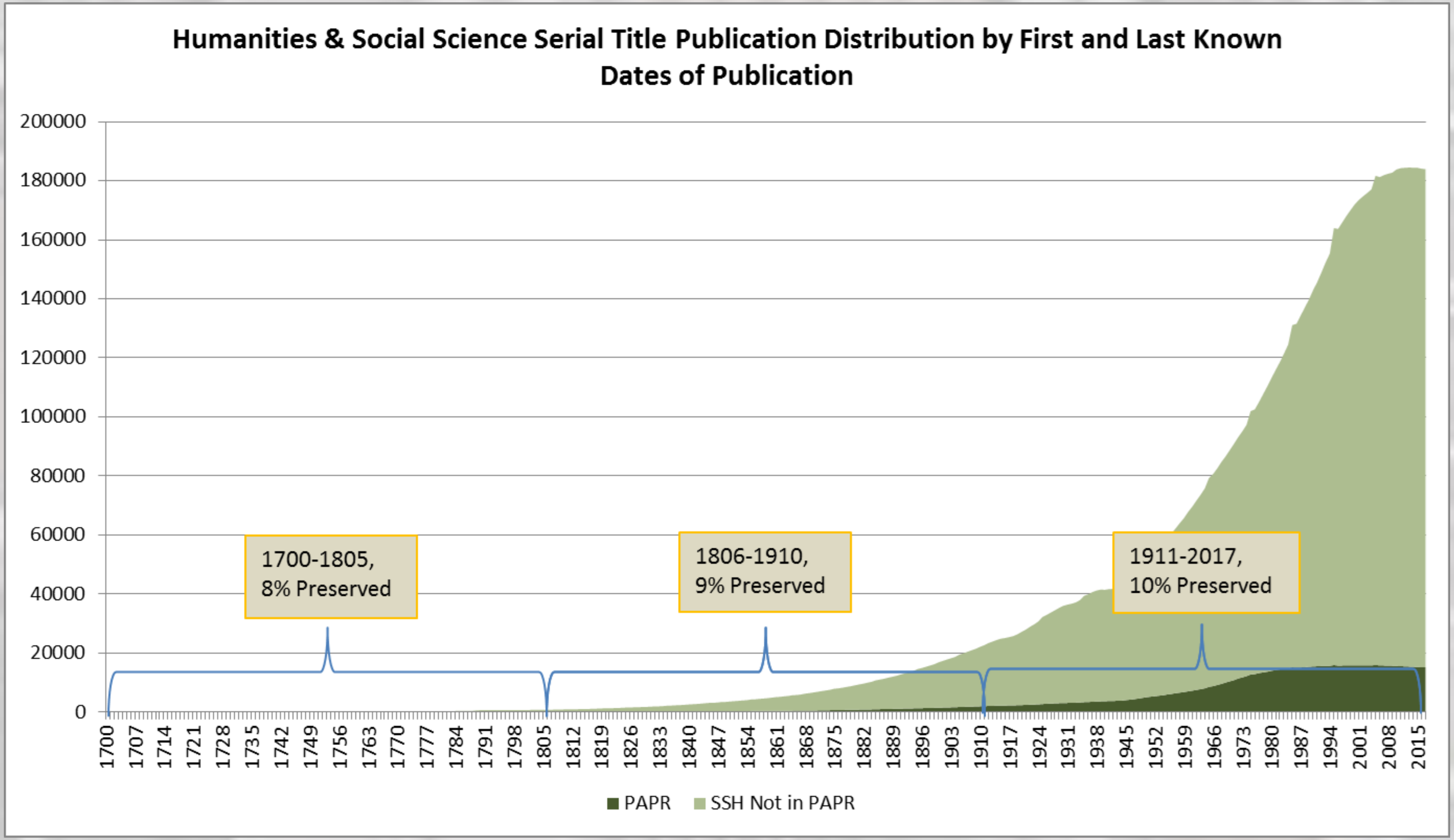


Archive Status of Unique Titles



29% of all unique titles are in a trustworthy digital repository

Aggregation of Holdings



Critical Corpus

Next Steps

Critical Corpus

Next Steps

- Add titles from additional libraries
- Post the list of critical corpus titles
- Obtain feedback from researchers
- Determine best approach to encompass the entire corpus in archiving plans
- Identify and engage the partners, who can best help us meet the new goal

Questions

Amy Wood
awood@crl.edu