Cory Snavely
Library IT Core Services manager
University of Michigan

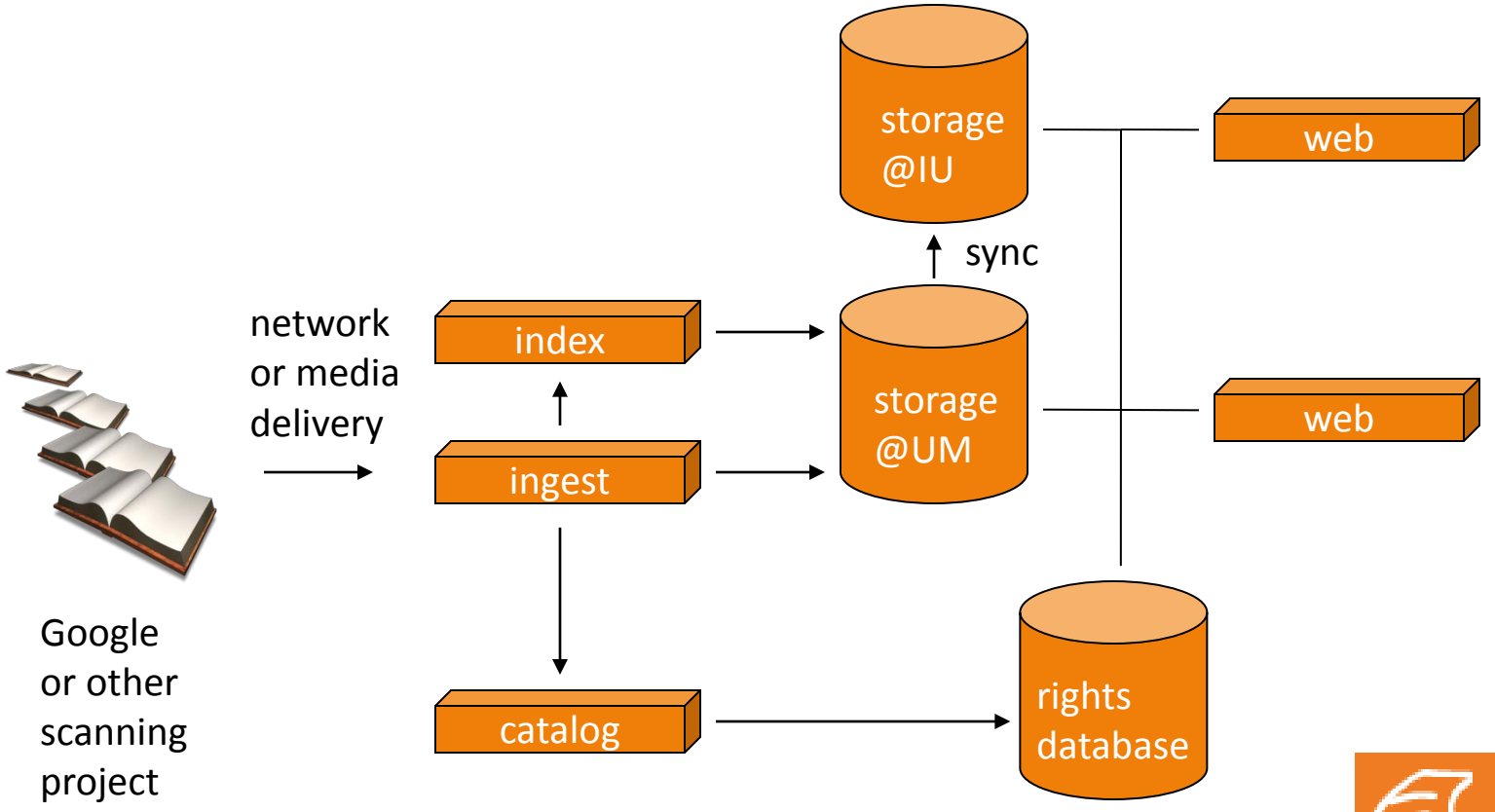
September 2010

HathiTrust project profile

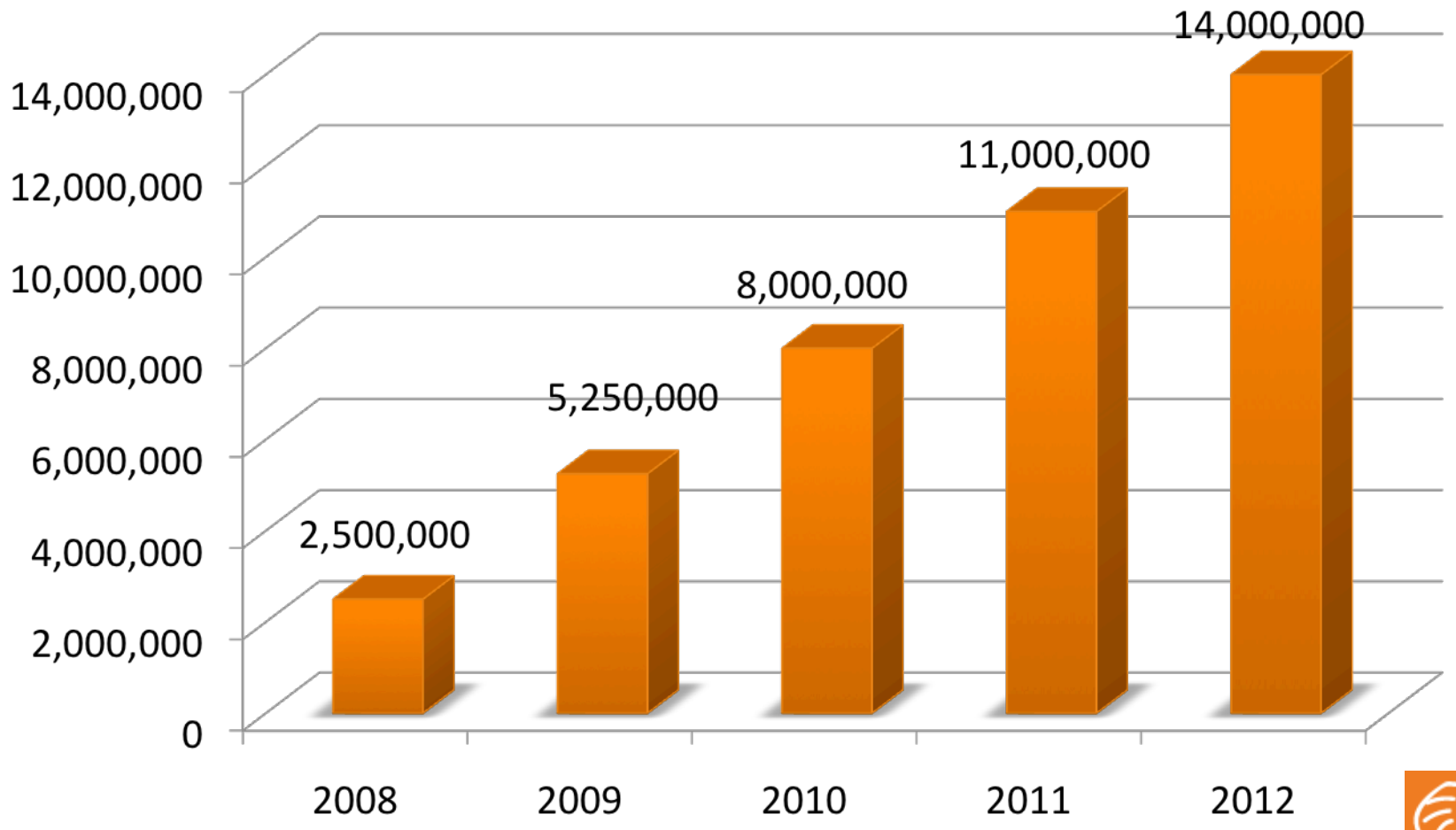
- Launched October 2008
- 29 member institutions and growing
- primarily Google-scanned materials but also other sources
- 6.7 million volumes, 350 pages average
- 250 terabytes in two US instances

www.hathitrust.org

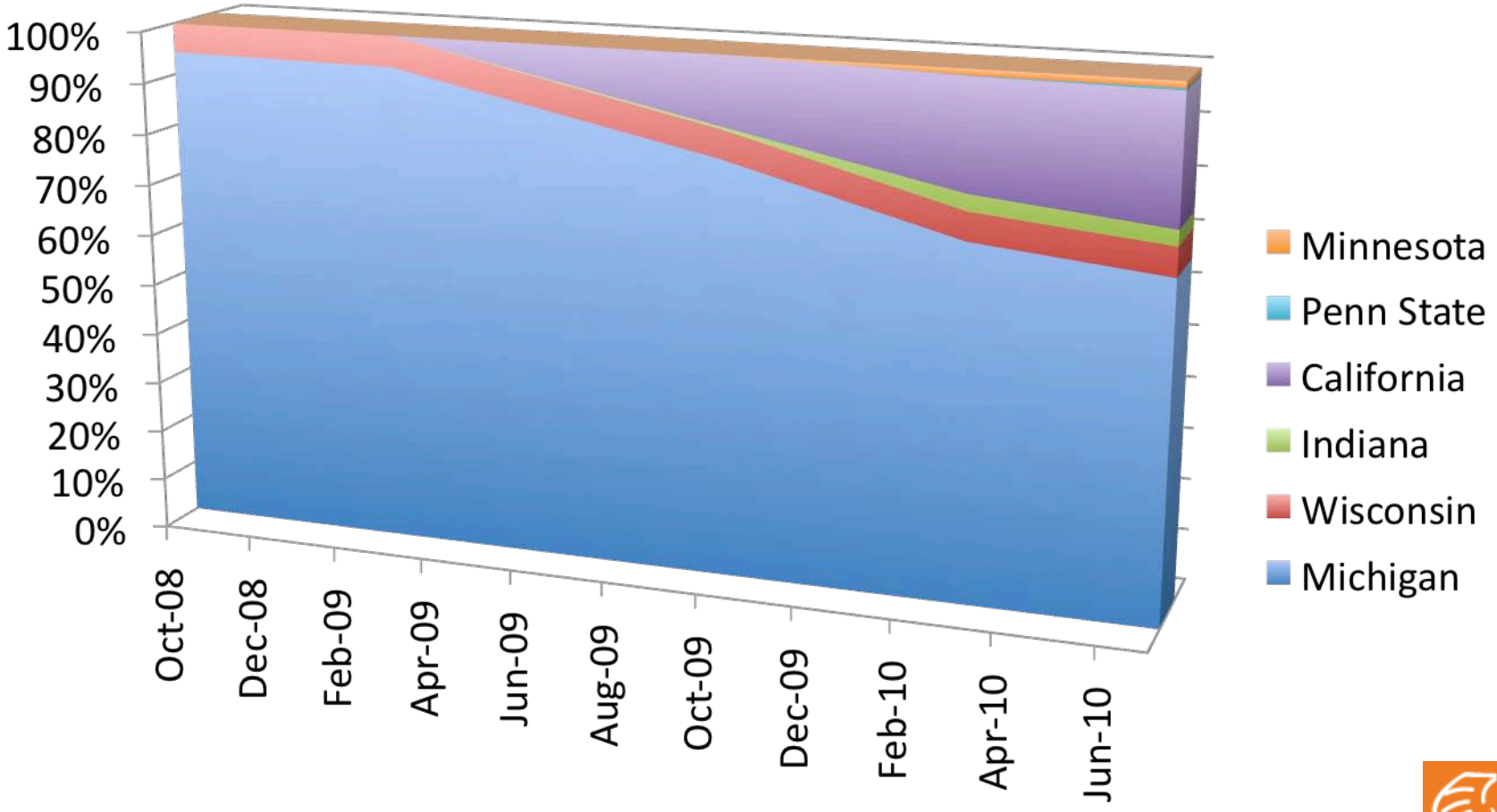
Material and Data Flow



Content Growth



Content Distribution Over Time



* As of July 25, 2010

What do I worry about?

Yesterday's worry	…is a non-issue due to…	…but today's worry is
Managing too many separate devices	Block/file virtualization	Storage system software reliability and change management.
What if I have to fsck this hulking beast?	Non-volatile journals and online integrity checks	
Bit rot, misdirected writes, …	Online error detection and repair	

- Trend is obvious, but not necessarily bad
- External error detection may be impossible

What's the Data Integrity Roadmap?

- Not all systems provide integrity features
- It's time for the data integrity model of systems to be a *primary* purchase criterion
- SNIA Data Integrity and Long Term Retention Technical Working Groups may help to surface minimum standards or common approaches; can anyone speak to progress?

HATHI TRUST
a shared digital repository

Questions?

Cory Snavely
csnavely@umich.edu

www.hathitrust.org