Human Rights Web Archive Portal – Technical Summary

Columbia University Libraries

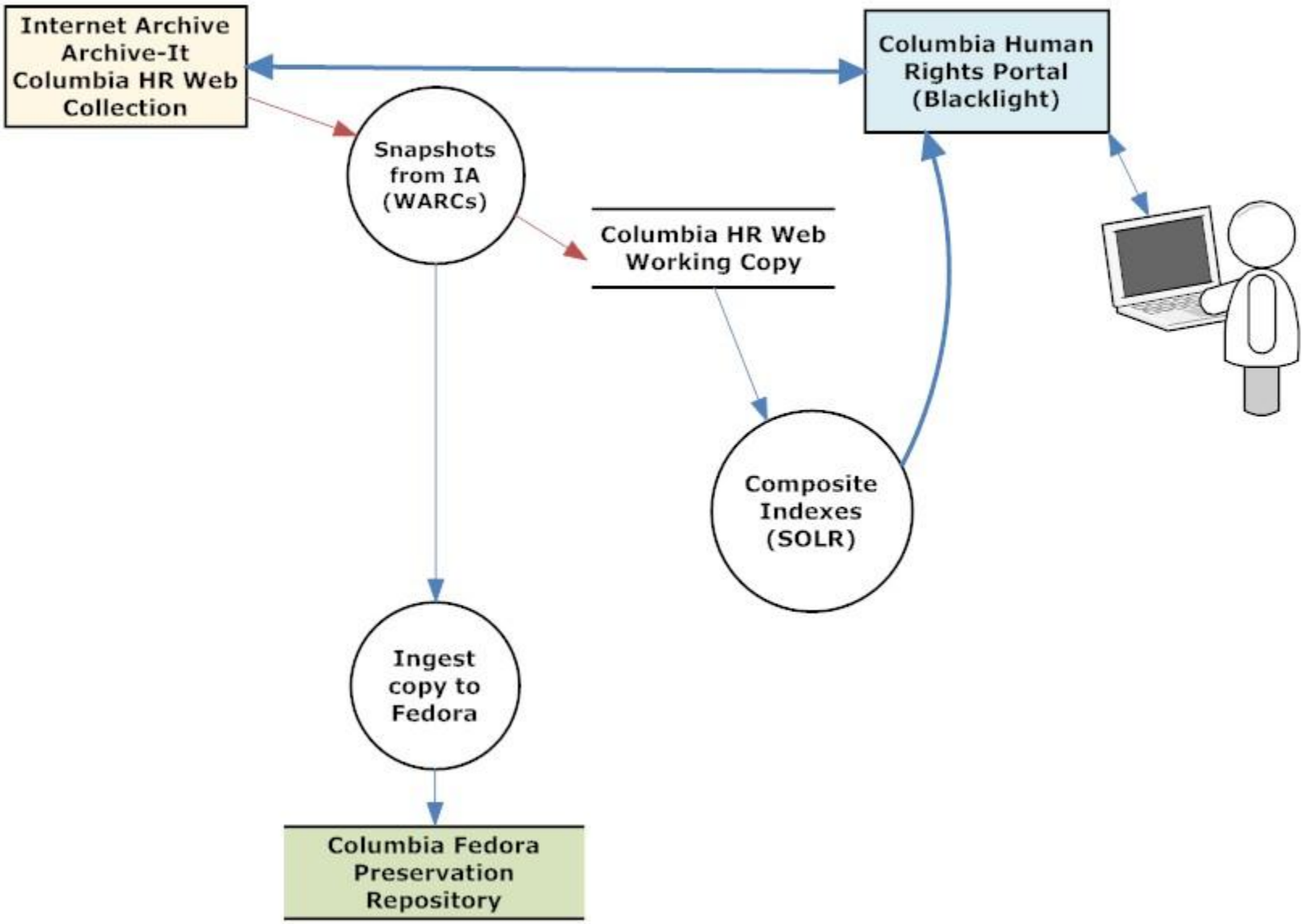
HRWA
Statistics, through July 31, 2012

- ca. 500 web sites
- 26 million pages / documents
  - HTML pages = 24.5 million
  - Document files (e.g., doc) = .5 million
  - PDFs = . 5 million
  - XML = 100,000
  - Presentations (e.g., ppt) = ca. 1,800
  - Spreadsheets (e.g., xls) = ca. 700
- ca. 65 languages

HRWA
Relevant Tech Terms

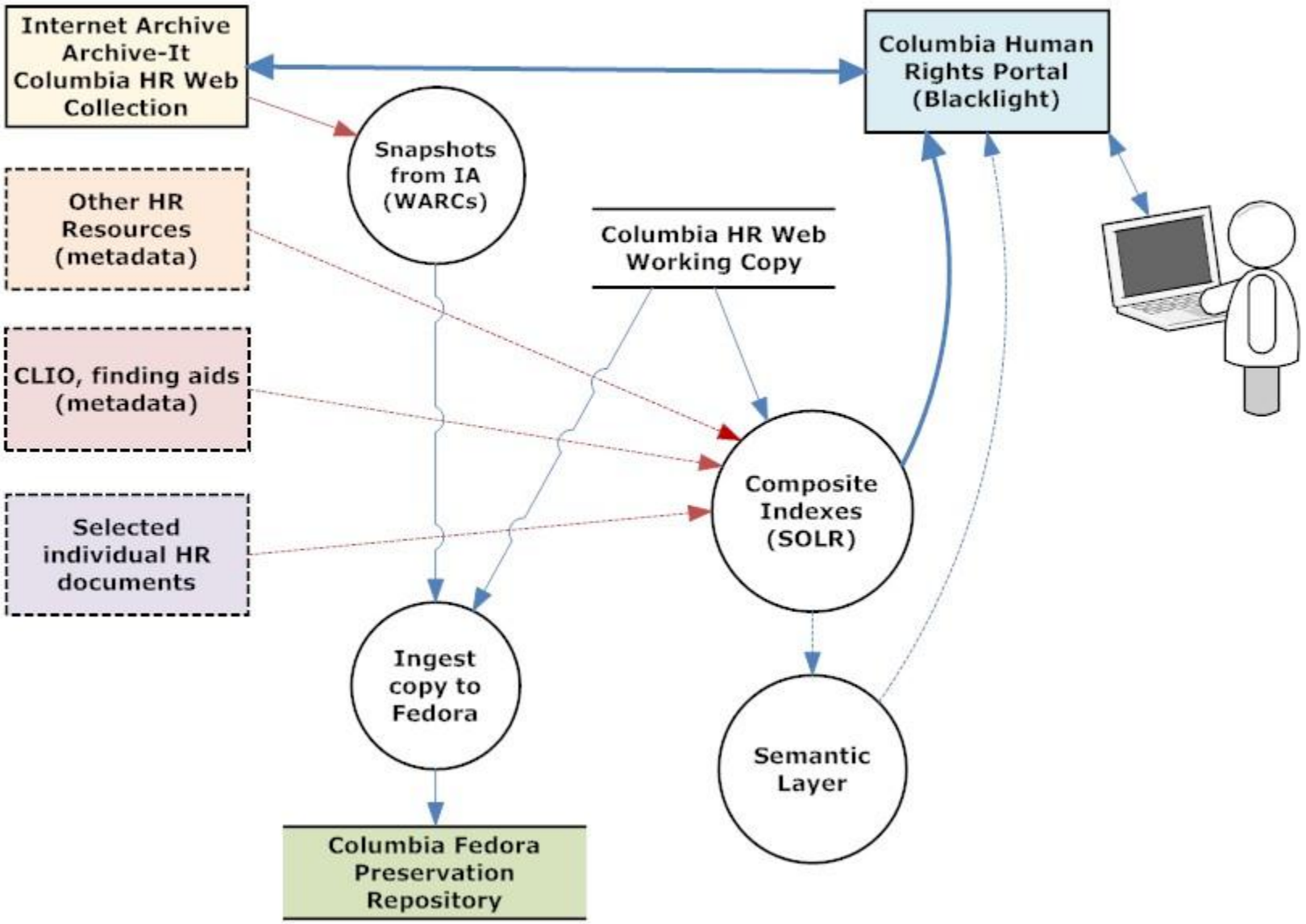
- Archive-It – IA's web archiving service
- SOLR (Lucene) – indexing tool
- Blacklight – Discovery Interface for SOLR
- MySQL – used as an intermediate index db
- WARC (Web Archive Format) – web storage
- Fedora – Columbia's preservation repository

Columbia Human Rights Web Archive Portal (HRWA)
Schematic Overview



11/12/2012 spd

Columbia Human Rights Web Archive Portal (HRWA)
Schematic Overview



11/12/2012 spd

HRWA
Challenges

Most challenging and innovative LDPD project to date.

- Most data in single project (ca. 2 TB)
- Largest indexes
- Greatest number of servers for indexing / production
- Most complex data (WARC / Web)
- Most challenging end-user design requirements
- Most uncharted in terms of users, possible uses, possible value added features, scoping, etc.
- Most cutting edge, most unanswered tech questions

HRWA
More Information

- CUL/IS *Behind the Scenes* page
- CUL/IS Mellon Web Resources Wiki
- Archive-It: Columbia's Web Archive Collections
- Columbia's Human Rights Web Archive Portal