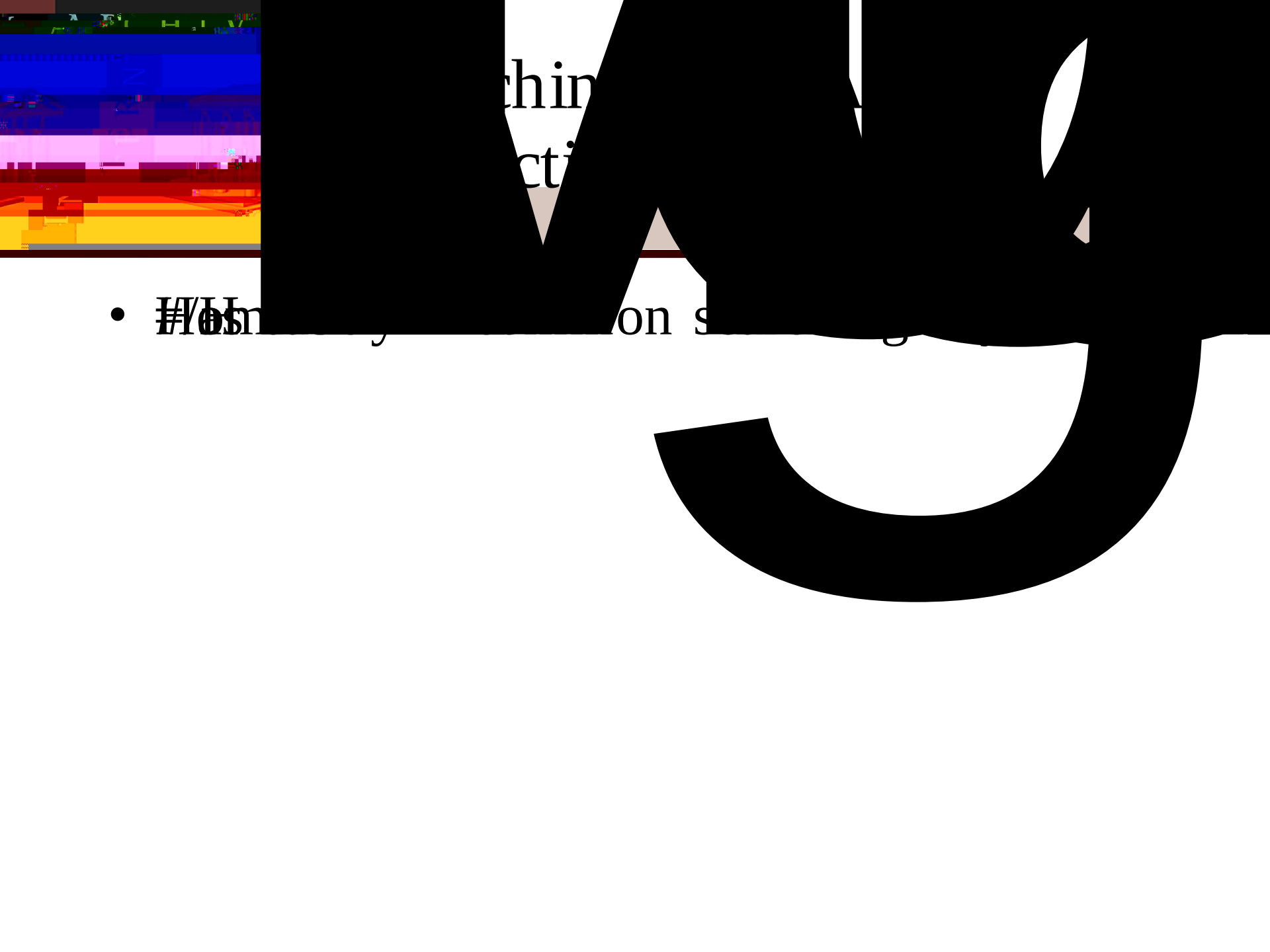
IWA

Internet Archive

- IA is a 501(c)(3)non-profit
- Mission is to build a publia

Public Access Limitation

- But Wayb



What is a Web Archive Collection?

- Web Archioq

WAC Attributes

- Key attributes

Nutch Overview

- Popure

Nutch Overview: Indexing 1

- Run

Nutch Overview: Indexing 2

- Steps:
  1. Ask Nutch DB to generate URLs to fetch.
  2. Fetch and Parse the downloaded pages.
  3. Update Nutch DB, run analysis on Document content, page rankings.

Nutch Overview: Querying

- Start the Nutch search Web application.
  - Run multiple to distribute query processing.
    - Distributes by remotely invoking queries against all query cluster participants.
      - Each query cluster participant is responsible for some subset of all “segments”.
- Queries return ranked lists of results

Adaptingi' iFãT

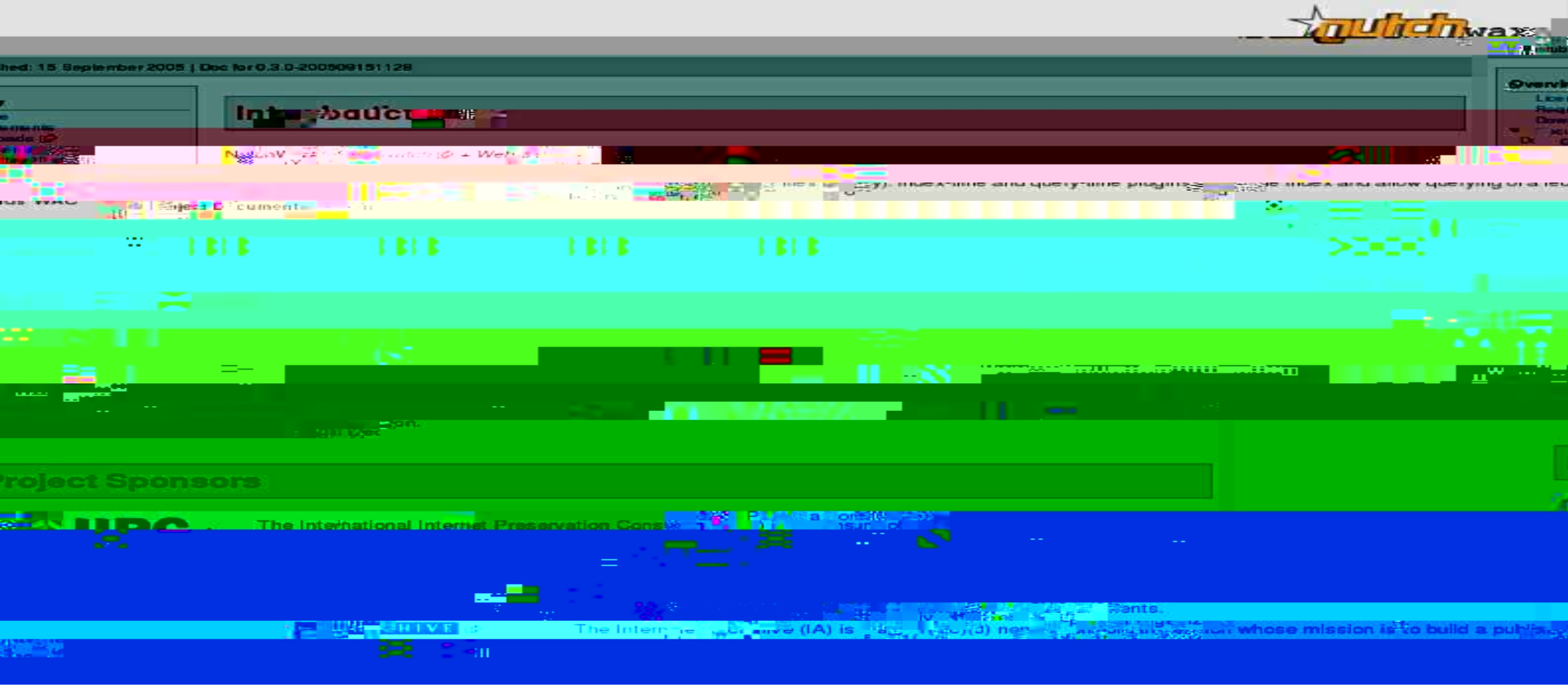
Adapting Nutch: Mode 1

- Nutch fetcher step recast to pull content from a

Adapt

Nutchwax

- All Nutch WAC plugin extensions, documentation, and scripts are open source, hosted on Sourceforge under the *Nutchwax* -- Nutch with Web Archive eXtensions – project: http://archive-access/projects/nutch/



Running WAC Search: Indexing Stats

- Indexing Machine Profile
  - Single processor 2.80GHz Pentium 4s with 1GB RAM and 4x400GB IDE disks running Debian GNU/Linux.
  - Indexing, CPU-bound with light I/O loading.
  - RAM sufficient (no swapping).
  - All source ARC data NFS mounted.
- Only documents of type text/* or application/* and HTTP status code 200 were indexed.

Indexing Stats: Small Collection

- Collection
  - Three crawls

Obse

Observations 2

- Inclusion of

Future 2

- Viewer applications
  - NWA WEP PÖVpÖ•