Project G

USC Libraries

Presentation to the 2006 DLF Fall Forum

Why

- Search large datasets
  - HPCC
- Multitude of custom interfaces
    - *Monastic Matrix* (http://monasticmatrix.org)
- ILS Limitations
- Gandhara

Gandhara Platform

- Open Source
  - Struts
- Flexibility
  - Customization
    - Analyzers
    - Indexers
    - Interfaces
  - Components

Ingest and Transformation

- Catalog Data
  - Standard library catalogs
  - XML Parser and Indexer
  - 2 Stage Process
    - MARC XML
    - OAI-PMH XML (DC)
- OAI Data Ingest
  - Descriptive Metadata Elements

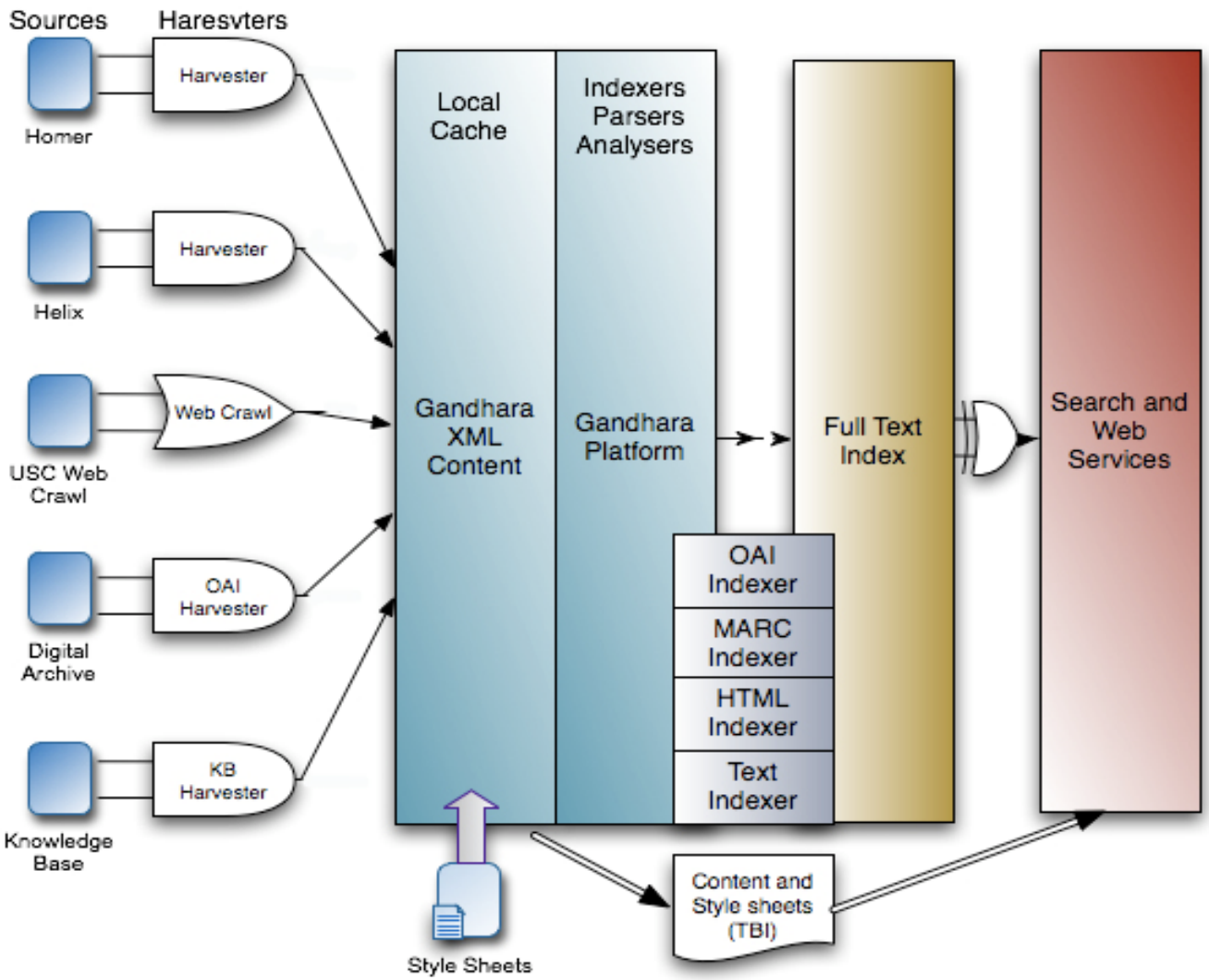
Pooling and Extraction

- Local Indexing
  - Heritrix to crawl
  - Lucene to index
- Remote Indexing
  - Google et al.

Next Steps

- Semantics
  - Taxonomies, terminologies, and controlled vocabularies (LCSH, MESH)
- Faceted Browsing
- Context Searching
- Results Clustering
- Recommender and Subscription Services

Architecture



Local Cache

- Gandhara XML
- Content
  - Text
  - XML
  - Html
- Style Sheets
- Rendering rules
- Schema

Wrapper (Gandhara) XML

- <?xml version="1.0"?>
- <!ELEMENT page-list (page*)>
- <!ELEMENT page(name,title,subject*,author*,description,source,archive,offset,sourceurl,electronicurl,thumbnailurl,collection,content,date)
- <!ELEMENT name (#PCDATA)>
- <!ELEMENT title (#PCDATA)>
- <!ELEMENT subject (#PCDATA)>
- <!ELEMENT author (#PCDATA)>
- <!ELEMENT description (#PCDATA)>
- <!ELEMENT source (#PCDATA)>
- <!ELEMENT archive (#PCDATA)>
- <!ELEMENT offset (#PCDATA)>
- <!ELEMENT sourceurl (#PCDATA)>
- <!ELEMENT electronicurl (#PCDATA)>
- <!ELEMENT thumbnailurl (#PCDATA)>
- <!ELEMENT collection (#PCDATA)>
- <!ELEMENT content (#PCDATA)>
- <!ELEMENT date (#PCDATA)>

Demo

- URL:
  - http://gandhara.usc.edu/

Todd Grappone
Associate Executive Director
Information Development and Management
em:grappone@usc.edu

Zahid Rafique
Director Digital Information Development
USC Libraries
em:zrafique@usc.edu