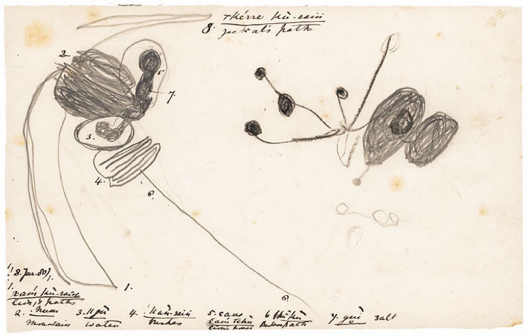
South African Digitization Initiative

on Building Large Scale Aggregations



Hussein Suleman
hussein@cs.uct.ac.za

University of Cape Town
Department of Computer Science
Centre for ICT for Development & Digital Libraries Laboratory

December 2014

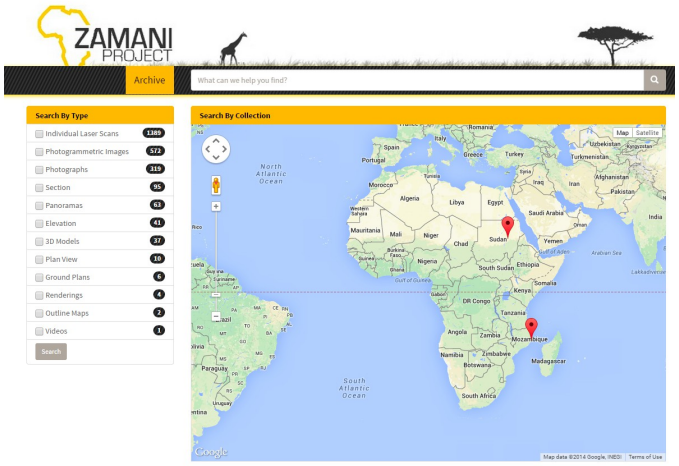
Why a National Aggregator?

- Provide discovery services to researchers, students, the general public
- Raise the profile of SA heritage internationally
- Encourage good practices in building archives through participation
- Fulfil a government mandate with a transformation imperative of nation building
- Develop skills and a base from which we can support other neighbouring countries

Principles

- A showcase for the public
  - It has to be designed and built for the public
- A focal point for development of heritage archives
  - Capacity development
- Access via source archives
  - No aggregation of actual resources

Services



- Search
- Browse
  - Thumbnail-oriented views
- Trails
- User Contributions and Enhancements
- Facet-based discovery
- Machine interfaces

Example: NSDL

- Aggregator of teaching resources across USA
- All levels: K12-university-adult learning
- Best effort principle: used many data gathering approaches and metadata formats
- Multiple directed portals over a metadata archive

Example: NDLTD

- International aggregator of ETD metadata
- A small set of metadata formats: DC/ETDMS
- OAI data harvesting only
- 3.5 million records at present
- Updates twice daily from about 200 sites
- Sites can be individual institution (MIT) or province (OhioLINK) or country (South Africa)
- Separated archive from portals

Example: Europeana

- European heritage resource archive and portal
- Very large scale ...
- EDM Metadata based on linked data and Semantic Web (RDF) principles
- OAI harvesting or static FTP pull
- Machine interfaces available for developing services over the dataset
- Both items and collections supported

Metadata: Simple DC

- Dublin Core that everyone knows
- Title, creator, date, ...
- Adv: simple
- Disad: vague; difficult to build services beyond simple search; no default controlled vocabularies; many concepts cannot be expressed

Metadata: Europeana Data Model

- RDF / Semantic Web model
- Extensible
- Graph-based metadata, relating concepts
- Large amount of information can be encoded precisely
- Fairly complex for new archive managers to deal with

Metadata: VRA Core

- Visual Resources Association Core metadata
- For multimedia objects
- Like EDM (but long before it), clearly differentiates between work and representation
- Works well for, say, paintings … but maybe not for entire caves

Metadata Principles

- Avoid reinventing the wheel!

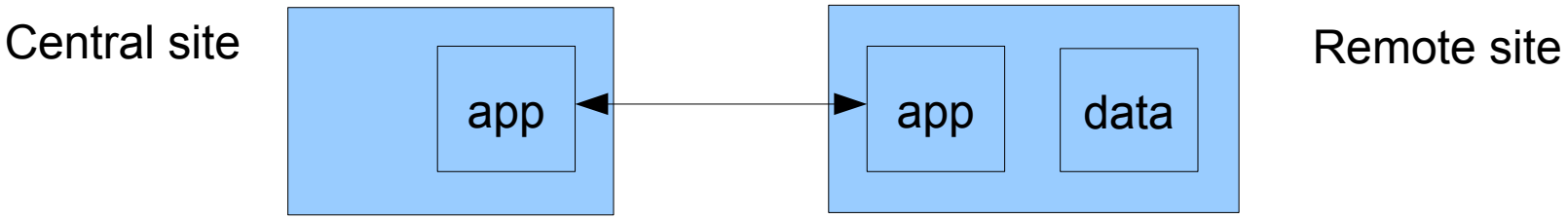
- Descriptive metadata
- Controlled vocabularies wherever possible
- Visual elements, such as thumbnails

Metadata Issue: Named Entities

- How do we handle these?
- Organizations?
- Persons?
- Collection Ids?

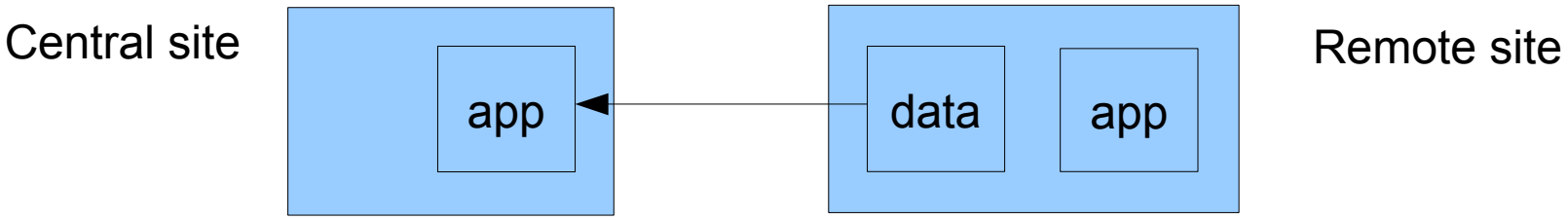
Gathering: OAI Harvesting

- Remote site runs a Web application to provide chunks of metadata on demand
- Central site initiates transfer periodically
- Handles updates and increments to collections, as well as efficient and robust transfer
- XML and REST



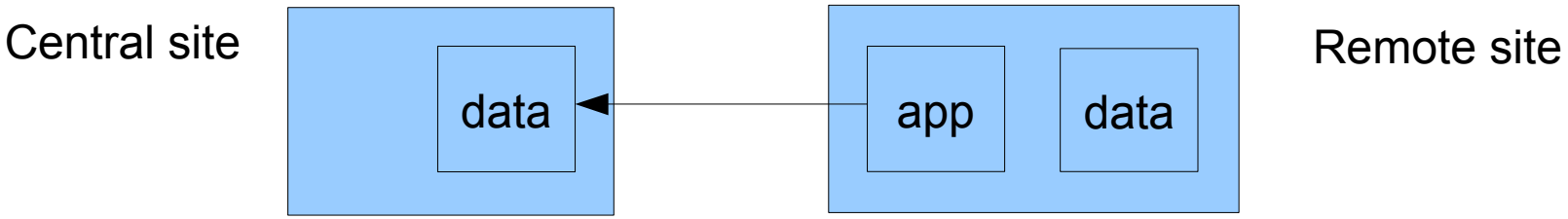
Gathering: FTP pull

- Remote site runs FTP server and periodically dumps metadata there
- Central site periodically fetches metadata using FTP client
- Not efficient – entire site transferred each time



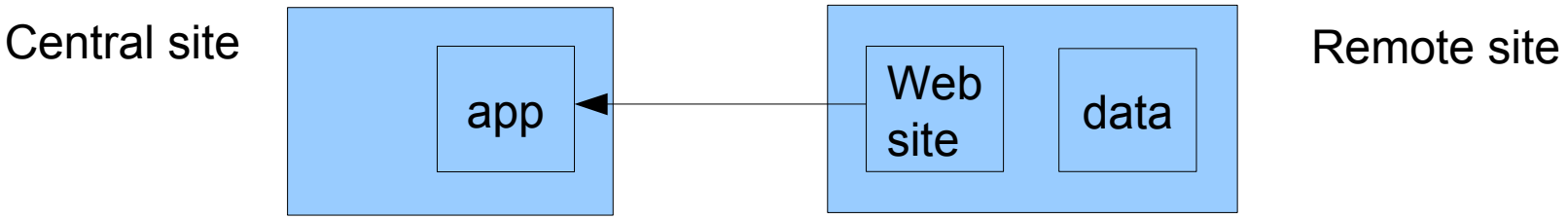
Gathering: FTP push

- Remote site runs FTP client and periodically uploads metadata dump to central site
- Central site runs FTP server and waits for uploaded data, then processes
- Remote sites are in control but more work
- Difficult to scale to multiple central sites



Gathering: Crawling

- No work for remote sites, except to be on WWW
- Central site runs focused crawler to gather Web pages and auto-generate metadata
- Poor quality and unreliable
- Used when all else fails :)



Issue: Aggregations and Collections

- Do we collect item-level or collection-level metadata or both?
- How do we merge metadata about one theme from multiple places?
  - e.g., Bleek and Lloyd collection has items in at least 6 physical archives
- Duplicates?

Issue: Languages

- How many to support? All!
- Translations needed for portal
- Multilingual services
- Must provide language-specific resources in language of communities

Issue: Legalities

- What is the agreement between remote sites and the central aggregator?
- Can this be an enabler to make sharing easier?

Issue: Testing and Automation

- Need automatic archive validation
- Automation can reduce costs
- Remotes sites need to take responsibility for quality of data and machine interfaces

Issue: Infrastructure

- Server infrastructure
- Internet bandwidth
- Replicas for reliability/preservation?
- Ongoing maintenance planning

Issue: Staffing

- What staff will we need?
- To manage technical infrastructure
- To liaise with remote sites
- To organize training

Issue: Local Constraints

- No natural home for project
- Little funding
- Limited bandwidth
- Unstable network and power!
- Skill levels are low
- Many archives have barely started scanning!
- Many potential end users are not educated

Issue: Toolsets

- Can we create skills in common tools for the community?
- Tools with basic configuration out of the box
- Easy to build community and peer support groups
- Kickstart approach
  - Worked very well for ETD community using DSPace

Issue: Evaluation

- Evaluation as a key part of the project
- Independence of monitoring team

A Proposal

- Start with a pilot to test the basic idea in 2015. One portal + small number of existing stable archives
- Then scope and seek support and funding for the full-blown project
- First host at NRF; until DAC takes ownership

questions, comments, … ?

Google "hussein suleman"

Facebook/slumou

Twitter@slumou

hussein@cs.uct.ac.za