Linked Data for researchers, YCBA Pilot Project

Lec Maj
Head of Information Technology & Innovation
Yale Center for British Art (YCBA)
lec.maj@yale.edu
Who and what are we working on?

• Began with access to online collections via website [2010]
• Extended XML data provider, [2011] data used by:
  
  Discover Yale Digital Content  
  Yale  
  ARTSTOR  
  Art Project  
  YOUR PAINTINGS  
  s[edition]  
  ARTSY  
  Art Finder  

• In 2012 began LOD work, connecting with other repository data sets in a pilot project.

“Linked Data for researchers, YCBA Pilot Project” by Lec Maj @ LDBC2012
What is going on with this data?

“Linked Data for researchers, YCBA Pilot Project” by Lec Maj @ LDBC2012
Answer: it requires domain expert

- Image from Tate
- Data from YCBA
- Different objects
- BM, V&A, etc. have related objects as well
Data for Research

• Answer in-depth questions
• Refine and collect better data
• Current research
  – Spreadsheets
  – Paper print outs
  – Google search
  – Image search
• Research Questions
  – What related types of objects are in existence
  – What drawings were created and owned by Paul Sandby (analysis of image markings to identify marks on them)
  – History map of ownership migration for Sandby’s drawings
  – Compare compositions of water colors to other objects to help speculate artistic models (how was the work influenced, how it is treated by various artists)
• Tools
  – Need visual tools for domain experts to analyze data
  – Combination of maps, timelines, etc. without technical expertise and complicated tools

http://collection.britishart.yale.edu/id/object/178

“Linked Data for researchers, YCBA Pilot Project” by Lec Maj @ LDBC2012
Architecture Integration

Metadata Providers
- YCBA - Art Collections Management (TMS)
- Library Collections Management (Voyager / ORBIS)
- Other Yale Data Providers

Aggregators and Indexing
- YCBA OAI Data Provider
- Library OAI Data Provider
- OAI-PMH Metadata Harvester
- RDFer
- Cross Collection Discovery (Lucene index)
- Solr Service
- Triple Store OWLIM

Online Collections
- Federated Global Data
- E-cards
- Website
- VuFind

Media (Image, Video, 3D) Providers
- MediaTools
- LadyBird
- Drop folder

Digital Asset Management
- Yale Media Manager and Preservation

Content Delivery Service

“Linked Data for researchers, YCBA Pilot Project” by Lec Maj @ LDBC2012
Data

• Schema and Data Complexity:
  – XML all kinds of schemas / vocabularies (12 mapped over 150 fields)
  – In pilot project CIDOC-CRM Ontology (Conceptual Reference, no limit to describing data)

• Scale (at the moment metadata only):
  – **Pilot Project: YCBA** (100K XML objects ~= 20M RDF triples)
    • Assuming current mapping of 1 XML object ~= 200 RDF triples
  – 3 Museums (2-3M XML objects ~= 400-600M RDF triples)
    • Painting, Prints, Drawings, Books, Rare Books, Sculptures, Bones, Cake...
  – Professional Schools (1M XML objects ~= 200M RDF triples)
    • Law, SOM, Drama, Music...
  – 21 Libraries (16M XML objects ~= 3.2Billion RDF triples)
    • Book, Rare Books, Maps, Sketchbooks, Journals...

• Type of queries we expect:
  – NLP, multi-join, federated (in/out of cloud)
  – in-depth research oriented, visualization of data for domains (maps, timelines,...)
Key Challenges

• Loading data without running out of time or computing resources (memory)
  – RDF triples updated (serialized) slow
  – Parallel asynchronous updates (multi-threaded)

• Need to know what is happening under the hood
  – Indexing / distributed processing
  – Commits

• Queries
  – Federation across global repositories
  – Dynamic Ontology Matching

• Analyzing natural language questions for translation to SPARQL, voice Q/A

• 3rd Party / Partner real-time data integration

http://collection.britishart.yale.edu/id/object/236

“Linked Data for researchers, YCBA Pilot Project” by Lec Maj @ LDBC2012
Limitations (computing vs. time)

• Performance Expectation
  – Query: Google like performance (3 seconds at most to process queries)
  – Loading data, cannot take longer then next load cycle
  – Offload whatever takes to long (multi-thread, cluster, cloud)

• Functionality
  – Multi-threaded and parallel data loading from cloud (ie: 100 servers)
  – Local indexing vs. in-cloud distributed indexing (ran into problems with CCD)

• Interoperability
  – Federation support across product and other products
  – URI mapping / lookups
  – Various ontology mapping (CRM, EDM,...)

http://collection.britishart.yale.edu/id/object/10386

“Linked Data for researchers, YCBA Pilot Project” by Lec Maj @ LDBC2012
What to expect from benchmarks

- Configuration recommendations
- Processing benchmarks
  - Load: triples per sec
  - Inference
  - Indexing
  - Poor query examples and timings (# of joins)

- Scale and hardware
  - Physical vs. VM vs. Cloud

- At what point will system start breaking down

- Performance measurement tool to diagnose problems in real time

"Linked Data for researchers, YCBA Pilot Project" by Lec Maj @ LDBC2012