CENTER FOR TEXTUAL STUDIES AND
DIGITAL HUMANITIES

DIGH 400 - Introduction to Digital Humanities Research

Fall Semester 2013

Week 15

Today's Class

- DH site
- After digitisation
- Classification

DIGH 400 - Introduction to Digital Humanities Research

www.digital-humanities.com

After Digitisation - Different uses of Metadata

- metadata can be used for different purposes
    - resource discovery metadata
    - descriptive metadata
    - provenance and rights metadata
    - technical metadata
    - administrative metadata
    - preservation metadata
    - structural metadata

After Digitisation - Origins of Metadata

- often derived from one of two locations
    - derived automatically from the digital resource itself
    - input and associated by digitisation staff, project members...
- intrinsic or implicit metadata
    - EXIF, file formats, resolution, bit depth...

Flickr Example

- extrinsic or explicit metadata
    - human created and processed
    - can also include user processed tags, vocab, annotations...

After Digitisation - Metadata and our digitised images

- essential for providing the means to
    - describe, share, search, manage and preserve our data
- metadata should be tailored to meet specific collection and user needs
- metadata should make data sharing possible with other collections, catalogues, systems...
- preparation of a set of specifications
- digital images can be complex to describe dependent upon project requirements
- digitised images can also contain many different layers
- judgement decisions will need to be made relative to a project and collection

DIGH 400 - Introduction to Digital Humanities Research

After Digitisation - Metadata and our digitised images

An example of Technical Metadata - XMP

- Extensible Metadata Platform
- open XML-based Adobe standard
- XMP can also incorporate metadata from other schemas
    - Dublin Core, IPTC...
- embeds the metadata within the image file itself
    - titles
    - author
    - author title
    - description
    - description writer
    - keywords
    - copyright status
    - copyright notice

After Digitisation - Image Optimisation

- optimise images for online publication, manipulation, and distribution
- consider how the images will be used within the website
- optimise images relative to tools and usage to gain best rendering
- tools can include
    - Adobe Photoshop
    - The Gimp
    - ImageMagick

After Digitisation - Image Optimisation

- we can often consider optimisation as two-tiered
    - generic
    - specific

Suggestions for generic optimisation
- create a working copy
- crop if necessary
- check and correct colour, contrast and density levels
- sharpen or soften as required (can also be deferred to the specific optimisation)
- scale image to required pixel resolution
- save new output to a lossy compressed file format such as JPEG or lossless compression format such as PNG

Suggestions for specific optimisation
- further colour correction
- image repairs
- specific crop issues or concerns
- sharpen or soften as required
- any other specific image adjustments...
- save and save again...

- document project workflow

After Digitisation - Storage

- many different considerations often dependent upon project specifics
- simple server based solutions to full media management solutions
    - filenames and folder names
    - database solutions
    - online image storage such as Picasa or Flickr (APIs available)
    - proprietary image management (Canto, Extensis, Luna...)
    - open source image management (Gallery, Coppermine...)
    - media management solutions (Greenstone, Fedora, DSpace, ExLibris, Koha, Omeka...)

After Digitisation - Rendering and user manipulation

- rendering will be dependent upon chosen project medium
    - thumbnails
    - galleries and gallery images
    - catalogue, taxonomy, and search
    - contextual linking and relative data
- images can be viewed and manipulated by tools such as
    - zoom (Zoomify, Google Maps, custom zoom)
    - magnify
    - transparent
    - ehinman
    - OCR (Tesseract and OCRopus)
    - OCR and transparent (example 1)
    - Woolf examples (example 1, example 2)

After Digitisation - continued

Jules Verne and H.G.Wells

After Digitisation - continued

- metadata has been specified and added to our digitised material
- classification is required for overarching project material
    - assignment of material to a specified class
    - class is a group of material (or objects) that share a common property
    - brings common objects together in a class
    - conversely divides disparate objects into different classes
    - views of these differences and similarities can be expressed and manipulated

Classification - scope

- loosely tied classification of data based on any set of properties
    - what is in and what is out
    - conforming and non-conforming
    - standards and schemas such as XML 1, TEI, Unicode...
- pre-existing classification systems in databases, XML...
- ad hoc categorisation of data
- subject based categorisation of material
- identification of an object's properties
- perfect classification presents perceived perfect knowledge of the object
    - objects near related topics
    - objects distant from unrelated topics
    - n-dimensional space
    - map of intellectual terrain

Classification - 1D

- semantically weak or nominal classifications
- often defined using simple class labels
- each object may take any one of a number of possible discrete values
- initially classes of objects not ordered relative to each other
- ordinal classification can now apply some loose ordering
- classify objects based on the value of a single characteristic
- classification becomes harder with borderline cases

Classification - multi-dimensional

- multiple characteristics may be applied
- Dewey Decimal Classification or System
- tree like classification
- context within tree is important to semantic value of the classification
- tree classifications often referred to as hierarchical classification systems
- pattern as follows
    - most general to most specific
    - biological classification system

Classification - an intro to faceted classifications

- Ranganathan (Indian mathematician and librarian)
- Colon Classification System
- Bliss Bibliographic Classification System
- Art and Architecture Thesaurus
- Flamenco Search Interface Project
- FAT or Faceted Analytical Theory

Classification - faceted

- separate defined subjects into constituent parts
- ideal for combining searching and browsing
- each resource classified by several separate hierarchical classifications called facets
- multiple classifications per item
- allows classification to be ordered in multiple ways
- subject into standard component parts
- consider material in the context of high level facets
- create hierarchy with narrower terms
- groupings as facets, individual terms as value or attribute
- should be based upon a controlled vocabulary

Classification - faceted online

- faceted particularly useful online
- both specific and general subject access
- general subject access can become time consuming
- combine searching and browsing and often providing breadcrumbs
- refine searches by drilling down a given hierarchy
- combine simple search with faceted options

Classification - Advantages and Disadvantages of faceted

- effective system because it divides subjects into component parts
- allows retrieval of data based upon a user's consideration of importance
- combination of hierarchical browsing and searching
    - switch between the two as needed
- number of results per option helps refinement of required result

- requires commitment to the creation and maintenance of the classification
- metadata particularly useful and important
- easiest to implement with well-organised, structured, and tagged data

eg: Mod Mags projects

Classification - a few rules...

- avoidance of cross-clarification

"it is written that animals are divided into: (a) those that belong to the Emperor, (b) embalmed ones, (c) those that are trained, (d) suckling pigs, (e) mermaids, (f) fabulous ones, (g) stray dogs, (h) those that are included in this classification, (i) those that tremble as if they were mad, (j) innumerable ones, (k) those drawn with a very fine camel's-hair brush, (1) others, (m) those that have just broken a flower vase, (n) those that resemble flies from a distance."

- distinguish some objects from each other
- be relevant to the purpose of the classification scheme
- be definite and ascertainable
- be permanent, in order to avoid the need for constant reclassification
- have an enumerable list of possible values which exhausts all possibilities

C. M. Sperberg-McQueen

Classification - a how-to for faceted

- faceted navigation for a site
- consider faceted relative to filters, options presented to the user
- user should be able to add and remove available filters as required to rearrange the available results
- faceted can also have deleterious impact on search engine crawlers

issues

- naive or lazy faceted
- noindex or nofollow
- robots.txt

Classification - a how-to for faceted

- provision of a dynamic filter solution for faceted based search results
- effective website crawling with top results and content correctly indexed
- index AJAX content
- AJAX can provide a user experience for easily adding and subtracting filters
- fallback static HTML navigation block
- faux facets that are links to deeper HTML pages
- build facets based upon search volume, weighted options...
- guide search engine crawlers to our top categories or facets

Classification - other options

taxonomy

- refers to the science of classifying objects
- traditionally it has come to specifically refer to the classification of plants and animals, eg: Linnaean classification system
- a popular reference for any hierarchical classification or categorisation system
- taxonomy as a kind of controlled vocabulary with hierarchy

eg: Woolf Online taxonomy

Classification - other options

ontology - overview

- a set of concepts with attributes and relationships
- define a domain of knowledge
- expressed in a format that is machine readable
- terms and relationships as the end goal
- customised relationship pairs that contain specific meaning
- differences in application relative to chosen field or discipline

Classification - other options

ontology - structural overview

- most ontologies describe
    - individuals or instances, classes or concepts, attributes, relations...
- individual = objects or instances
- classes = collections, sets, concepts, groupings...
- attributes = values, properties, features, characteristics...
- relations = relationships between classes and individuals
- function terms = complex structures formed from certain relations...
- plus
    - restrictions, grammars, events...

Classification - other options

ontology - domains...

- refers to a specific part of a defined something...
- define meanings for terms that apply specifically to that domain
- similar words considered relative to domain's ontology
- upper ontology or foundation ontology
- common core glossary
- Dublin Core upper ontology
- concepts specific and customised to the domain

OWL & RDF