Please note, this report has been compiled from my notes, and so I remain solely responsible for any inaccuracies that might appear. For authoritative information about NINES visit http://www.nines.org/

# **NINES Report**

#### Contents

1.0	NINES: Projects, software, methodologies
1.1	NINES overview
1.2	projects underway
1.3	projected projects
1.4	NINES applications
2.0	Rossetti Archive, IATH / UVA
2.1	Fedora
2.2	Rossetti Archive
3.0	Useful Aspects of Humanities Computing
3.1	metadata
3.2	folksonomy
3.3	imaging / mapping
3.4	data extraction / language tools
3.5	XML / XSL / XPATH / TEI
3.6	OCR
3.7	Romantic Circles and the Book of NINES
3.8	project management
3.9	User Interfaces
3.10	sustaining digital scholarship
3.11	Intellectual Property
4.0	My conclusions
4.1	folksonomy principles
4.2	databases

This presentation is supported by a del.icio.us account (http://del.icio.us/Jim\_ncse) which contains all the main websites, tagged with labels that correspond to the sections in which I am using them. This is what I am using to provide the websites.

Dr Jim Mussell

Nineteenth-Century Serials Edition Faculty of Continuing Education Birkbeck College 26 Russell Square London WC1B 5DQ

# Part One: NINES: Projects, Software, Methodologies

# 1.1 NINES Overview

- www.nines.org
- NINES is a federated consortium, in which peer-reviewed digital projects will operate with each other through metadata and crawlers.
- Individual projects retain their own UIs, responsibility for their projects, and all property.
- Part of a political, curatorial, scholarly enterprise:
  - C19th objects are physically sick, dying in libraries. Equally, digital objects will not last for ever. But scholarship is about care – for cultural heritage.
  - XML and TEI, if they are well used, will be incorporated into the next technological phase. But individual uses of TEI / XML must move with the times. NINES represents a consortium of projects with an investment in digital conservation. The distributive model of NINES ensures ownership is with the community.
  - It works through the following suite of software:
    - o JUXTA
    - IVANHOE
    - COLLEX
    - Paracritical Demon
- And a search engine powered by Lucene

# **1.2 Projects that exist**

Note NINES intends to go live in December with Romantic Circles, Rossetti Archive, and Whitman Archive. It has three editorial boards, but no publishing mechanism as yet.

# **1.2.1** Stephanie Browner (Berea College): The Charles Chesnutt Digital Archive: Vision, Shape and Update

- Already exists http://www.berea.edu/faculty/browners/chesnutt/intro.html
- Author centred, but aims to capture the diverse activities of the author (i.e. works about him as well as by him)
- Currently transcription but is exploring facsimiles.
- Project has been integrated with teaching.

# 1.2.2 Amy Earhart (Texas A&M University): 19th Century Concord: A Historical and Literary Center Website

- http://www-english.tamu.edu/concord/
- Ambitious attempt to map a place in historical time ongoing project.
- Has access to local archives and significant resources site already exists, and she is developing XML.

- Plans to incorporate a flash 3d map, with 360° VR and hotspots.
- See Emerson house: http://www-english.tamu.edu/concord/emersonhouse/emhouse.html

# 1.2.3 Neil Fraistat & Steve Jones (University of Maryland & Loyola University): Romantic Circles. Also Carl Stahmer.

- http://www.rc.umd.edu/
- In the process of shifting to XML.
- Repository for editions, original work, and discussion. Contains scholarly editions: http://www.rc.umd.edu/editions/
- Has facsimiles, transcripts, comparison software, extensive bibliographies and a school's page.

# **1.2.4** Laura Mandell (Miami University): Early-Nineteenth-Century Popular Literature in Britain and America

- http://www.orgs.muohio.edu/womenpoets/poetess/
- She has already converted an html website into XML.
- Poetess archive contains both primary and secondary works.
- She has opened the processes she has gone through (made a feature of displaying the markup etc...).

# **1.2.5** Julia Thomas & Tim Killick (Cardiff University): A Web-Mounted Database of Mid-Victorian Wood-Engraved Illustrations

- http://www.cf.ac.uk/encap/ceir/dmvi/index.html
- AHRC project which is designed to facilitate access to Victorian woodengravings which are otherwise lost in periodicals.
- Theorized the extraction of images from context, and are trying to make the database searchable by thematic content by keywords in the illustrations with no reliance on the text.
- They are using access databases, and so not in XML.

# **1.3. Projected Projects**

# **1.3.1** Gavin Budge (University of Central England): The Romantic Periodicals Project

- Early stages: has digitized volumes 1 and 2 of *Quarterly Review* as part of a pilot to be incorporated into Romantic Circles.
- Has resources onsite at UCE, and has digital text of vols 3 and 4.

# **1.3.2 Dennis Denisoff & Lorraine Janzen (Ryerson University & Nipissing University): The Yellow Book Project**

• Still in planning stage. Good local resources but dependent on scholars time.

- They have innovative approaches to the *Yellow Book*, and are interested in capturing things like paper quality, covers, page detail etc.
- They are also building concept maps to link the community of figures together.

# 1.3.3 Susan Garfinkel & Jurretta Jordan Heckscher (Library of Congress): "American Memory" Re-Imagined: Alternative Interfaces for Online Historical Collections

- They are planning an XML encoded version of a *The Diary of Horatio Nelson Taft*. They intend this as a pilot for a wider re-evaluation of LOC's vast existing digital archives *American Memory*.
- Diary is already there but has no metadata. (http://lcweb2.loc.gov/ammem/tafthtml/tafthome.html)
- They plan ambitious deep markup on three levels: cultural (using the *Outline of Cultural Materials*); literary (according to docs own internal categories); and library (LOC subject classifiers).

# 1.3.4 David Hanson (Southeastern Louisiana University): The Early Ruskin Manuscripts, 1826-1842

• An XML edition of the early Ruskin MS with facsimiles and transcripts. This will eventually be incorporated into the Ruskin Project at Lancaster.

# **1.3.5** David Latané (Virginia Commonwealth University): Digitization of Victorian Institute Journal

- The only exclusively contemporary work.
- *Victorians Institute Journal*, allowing its content to add new context to the scholarly editions in NINES.
- place where newly discovered MS, or edited works, could be published quickly in a twice peer-reviewed environment.

# **1.3.6** Edward Whitley (Lehigh University): The Vault at Pfaff's: An Online Archive of Works by American Bohemian Writers and Artists of the Nineteenth Century

- Is beginning project to recreate the network of artists who frequented Pfaffs beer cellar in NY.
- Will present works, and allow users to map the connections between figures.

# **1.4 NINES applications**

- As mentioned, NINES will:
  - Gather digital scholarship and criticism
  - Set up technical framework for producing digital scholarship
  - Develop suite of easily accessible tools.

# **1.4.1 NINES applications: harvesting:**

- NINES will harvest individual collections in COLLEX through screen scraping (create rdf through the html) or from the xsl.
- The rdf is Resource Description Framework: it is a language for representing resources *identified* on the web particularly metadata about them. Machine readable, and operates with triples (subject, predicate, object).
  - Represents an object, not a page 40,000 in Rossetti.
  - Each project is responsible for creating rdf records that are referenced in the META heads of html pages. This could be done through the XSLT (which we have) or screen scraping available metadata in html.
  - NINES server would harvest both the rdf and the archive content using Nutch (which I describe below). The rdf basically refines the metadata to improve searching.
  - COLLEX uses the rdf records to identify and recommend objects for collection / exhibition. They see rdf as a NINES peer reviewed tag.
  - They are going to use an open source rdf database like Kowari: http://kowari.sourceforge.net/
  - As rdf uses triples made up of the URI (uniform resource identifier http etc), these must be stable.

# 1.4.2 NINES applications: searching

- Searching can be achieved at a number of levels:
  - Traditional relational databases.
  - Structured searching with xpath and xquery (see below).
  - Full-text searching.
  - Latent semantic indexing recognizes combinations of words: implies a document is more than its vocabulary.

# Nutch

- http://lucene.apache.org/nutch/
- Nutch will search both the full text and browse the faceted metadata of projects
- It is an open source web crawler, indexer and search engine
- Java based, and powered by Lucene (coming up below) it is extremely scalable
- It respects web crawling standards, and is plugin capable it can be integrated into databases like kowari.

# Lucene

- Lucene http://lucene.apache.org
- Erik Hatcher (literally) wrote the book on Lucene: *Lucene in Action*.

- As well as Nutch, it also powers: Guardian jobs, Flamenco (see below, under metadata), Wikipedia, and certain US government agencies.
- It needs to be worked upon as its API (Application Program Interface) is in Java.
- It indexes documents into 4 levels: an Index (sequence of docs); documents (sequence of fields); fields (named sequence of terms); and terms (strings or words).
- All documents go into the parser before being indexed and it works on text.
- The parser turns text into tokens, with each character being attributed a value and located by position increments. It runs a range of different ways of doing this, and their order can be customized. They are:
  - WhitespaceAnalyzer strips out whitespace
  - SimpleAnalyzer lowercase, no whitespace or special characters
  - StopAnalyzer removes stopwords (why Google returns "not" for a "to be or not to be search")
  - StandardAnalyzer recognizes when certain variants are part of a token (e.g. initials); also sequences of letters and digits, interior apostrophes, acronyms, companies, email and hostnames, serial numbers, and CJK characters.
  - SnowballAnalyzer employs stemming
  - SynonymAnalyzer employs synonyms that can be incorporated from anywhere (currently wordnet).
- The parser also works on a query: it can
  - o do Boolean operators
  - $\circ$  field selection
  - wildcard searches
  - o fuzzy terms ("kountry" will return "country")
  - o proximity (for authors who occasionally use middle names)
  - grouping (OR and AND)
  - and ranges (spans etc)
- Lucene has been incorporated into the Rossetti Archive:
- http://www.rossettiarchive.org/rose/

# 1.4.3 JUXTA

- Allows comparisons without the need to posit a base text.
- It collects; normalizes; collates; analyzes; annotates; publishes; and receives.
- It is fluid, and very easy to manage complex differences: e.g. it can do multiple docs, compare words, punctuation, whitespace, and can output results as a histogram.
- They are integrating Lucene into Juxta to make it more powerful.

# 1.4.4 Ivanhoe

- Playspace where users, playing roles, can amend texts
- It has a timeline, so you can see moves. Currently it is represented in a circle that shows characters' revisions as slices of a pie (highlights the

roles), but can also be presented as a text centred version (a branching tree). There is a move centered one, which would be dynamic.

- I see it as a valuable pedagogical tool:
  - Has a journal to encourage reflexivity
  - Could allow users to "correct" classic texts (explore absent motivations for instance)
  - Moves could be theoretical positions
  - o Or could model exchanges of correspondence, editing etc...

#### 1.4.5 COLLEX

- A way to browse and identify objects, and then annotate them (with tags and notes) and exhibit them.
- Works like folders (it derives from social bookmarking see below), but they are exposed and cross-referenced so others can benefit from suggestions.
- **Collection:** It works through a tag and collect bookmarklet, which then imports objects into COLLEX proper. The rdf is semantically linked to provide a "more like this window"
- Because you need an identity to get the bookmarklet, your tags are linked to you. You write them in, and they are linked to the object with a hyperlink. Currently limited to one word tags, but these can be yours or other peoples.
- When you "COLLEX this page," all objects are greyed out until you supply your tags.
- **Exhibition:** side is presentation tool and bibliography.
- Exhibits (and their metadata) can be re-published back into NINES environment.
- At the moment limited to NINES, but they will make it open source.

# Part Two: Rossetti Archive, IATH / UVA

#### 2.1 Fedora

- http://www.fedora.info/
- Open source digital archive designed by Cornell and developed by UVA.
- Produce views of objects, and relationships between them
- UVA use it to house born-local digital collections curative.
- Cross collection search restricted to image metadata and TEI heads.
- Images can be collected, saved, displayed as multiple slide shows etc.
- Fedora being incorporated into Rossetti.

# 2.2 Rossetti Archive

- Rossetti Archive http://www.rossettiarchive.org/
- The archive's structure is through what they call RAC (Rossetti Archive Commentary) and RAW (Rossetti Archive Work either image or text)
- Filenames are derived from the existing archives they stress the importance of good, regular, user-friendly filenaming (i.e. no Rossetti1.tif)
- They've paid fees per image where required for copyright sometimes for one year sometimes for five. Some galleries insist on having their name imprinted as a digimark. Thumbnails are exempt.
- For the Rossetti Archive they have three levels of proof-reading: copy; content; code (although the latter is done by their xml editor).
- It has just had a massive redesign: previously it was in SGML using Dynaweb. This meant their images and texts were stuck in links at the bottom of pages. Now in xml, they use xsl and cascading style sheets to re-order: now they are at the top. :
  - Use Blessed Damozel e.g.
  - Old: http://www.rossettiarchive.org/newdesign/old/RAWs/damozelraw.html
  - New: http://www.rossettiarchive.org/newdesign/RAWs/damozelraw.html
- Archive says if pictures aren't available. For docs they provide desc, page image (but not always full facsimiles) and transcripts.

# Part Three: Useful Aspects of Humanities Computing

#### 3.1. Metadata

- Means to create relationships between objects. And identify objects when out of context.
- Standards less important than well-encoded data consistency. Also, document decisions for creating standards.
- Three types of metadata:
  - Descriptive (aids to identification) human
    - Administrative (file types etc) automated
    - Structural
- There exist library resources for data values e.g.LOC lists and Getty's *Art and Architecture Thesaurus*.
- Consider the future of metadata: see OAI allow data harvesting across archives.

#### 3.1.1 metadata examples

# • History Wired at the Smithsonian

- http://historywired.si.edu
- An attractive two dimensional representation of metadata categories
- Is limited in that it cannot show too many items, but could be developed for more complex concepts with another axis.
- It has a further level of meaning in the size of the spaces this might be combined with folksonomy components such as social bookmarking.

# • Flamenco

- Stands for Flexible information Access using Metadata in Novel Combinations.
- http://bailando.sims.berkeley.edu/flamenco.html
- uses hierarchical facted metadata in an extremely flexible way users can see the integrity of the collection change as they delimit it, and they can expand and contract as they wish.

#### 3.2 Folksonomy

• The idea that relationships between digital objects are dynamic and social. User activity brings relationships into being (emergent ontologies). Users can pick any tags they want, and see what other people have tagged things as – tagging is consensual rather than prescriptive.

#### 3.2.1. Del.icio.us

• http://del.icio.us/

- What this presentation has created
  - Username Jim\_ncse.
- Provides a bookmarklet that is incorporated into your browser. Add sites (similar to folders) but also to see what other people have called your sites.
- Allows you to search others' activity, see what sites are most popular etc...

# 3.2.2 connotea

- http://www.connotea.org/
- Based on delicious
- Allows scientists to follow each others work, and discover cross references

# 3.2.3 citeulike

- http://www.citeulike.org/
- Similar to the others but more general academic papers
- Lets you export refs into endnote etc...

# 3.2.3 flickr

- http://flickr.com
- people place their images in a communal space, and then invite annotation, tags etc...

# 3.3 Imaging / mapping

# 3.3.1 Zoomify

- http://www.zoomify.com/
- Very good way to zoom in without refreshing etc (uses tiles)
- E.g. BL literary landscapes use Hardy's Dorset.
- Better e.g. is National Gallery (US) small French Paintings shows brushstrokes.

# 3.3.2 Formats

- Jpegs pixellate too readily
- Note the importance of filenames

# 3.3.3 Annotation

- Iconclass is too unwieldy
- Interesting discussion lead by Julia Thomas and Tim Killick: they are attempting to derive their own keywords, but some delegates felt this introduced an unwelcome claim for objectivity.

- There is a *Thesaurus for Graphic Materials* (download from http://www.loc.gov/rr/print/tgm1/) and *Outline of Cultural Materials* (which is available on microfilm).
- Worthy Martin, Associate Professor of Computer Science at UVA, recommends James Wang's research group at upenn. They have a 10 year project on annotating images. http://wang.ist.psu.edu/docs/projects.shtml
- TEI forms have space for commentary, but any annotation needs to be lengthy to co-operate with statistical features of search engines.

# 3.3.4 mapping

- Maps show (spatial) change over time can be easily achieved by layering maps.
- Geographical data is available from GIS (Geographic Information Systems), but it is difficult to move information. It is important to develop databases and presentation methods simultaneously.
- Vectors are a good way to show changes through time (bitmaps need to be drawn, pixel by pixel, each time i.e. frames.). Vectors draw lines and curves, and can be modified easily (as the size of the graphic isn't the limiting factor). Flash uses vectors, but was designed to animate bitmaps.

#### 3.3.5 examples

# • BL online gallery turning the pages

- o http://www.bl.uk/onlinegallery/ttp/digitisation3.html
- zooms, can be read, transcripts see Austen.

# • UVAs Salem Witchcraft accusations maps:

- o http://jefferson.village.virginia.edu/salem/maps/
- o towns, events, and time linked in database, flash front end.
- **BBC Backstage** they've opened up their data and are encouraging people to use at as they wish.
  - http://backstage.bbc.co.uk/
  - could be combined with things like:
- Google Earth the whole planet.
- Second Story. Produced these 5 which are worth a look:

# • John Paul Getty animation

 http://www.secondstory.com/search.php?ID=58&SubjectMas k=0&CategoryMask=0

# • Monticello Explorer

- http://explorer.monticello.org/
- use the plantation map to get into the house.

- Move into the house, and turn around.
- Demonstrate the photograph (the camera)
- Demonstrate animation (the cross)
- Demonstrate the objects (the icons at the bottom).

# • Peabody Essex Museum

- http://www.pem.org/flash.php
- Go into collections
- Show them artscape and mention the integrated pdas.

# • America on the Move

- http://americanhistory.si.edu/onthemove/
- Three portals in:
  - Through time which brings you to curated examples
  - Through space (plus metadata and keywords)
  - Or with concept maps which here are curated exhibitions which follow certain stories.

# • Design Archives

- http://designarchives.aiga.org/
- Another spatial arrangement that can be filtered
- Has a collection / exhibition space below (the lightbox) that can be shared with the community.

# 3.4 data extraction

• The various language software applications are important as they allow tags, metadata etc to be structured – e.g. "me" and "self portrait" as Flickr tags.

# 3.4.1 wordnet

- http://wordnet.princeton.edu/
- cog sci project groups lexical items into synonym sets linked by a single underlying relationship.
- Works semantically and formally with all parts of speech.

# 3.4.2 simile

- http://www.simile.mit.edu
- uses rdf and semantic web to extend Dspace a place where data is exchanged – in order to make various systems interoperable. Has a piggy bank which allows the creation of local collections (via rdf) like NINES, it is powered by Lucene, and works with Googlemaps etc.

# 3.4.3 GATE

- http://gate.ac.uk based at Sheffield
- They can extract data itself, rather than just keyword hits.

# 3.5 xml / xsl / xpath / rdf / TEI

# 3.5.1 xml

- Xml creates flexible self-contained databases
- Useful suggestion was to use comments <!...>in the code in order to explain why things were done and how. Also blocks out bits that don't parse.
- DTD provides the rules xml follows.
- Its simplicity has ensured its proliferation. Not necessarily universal for instance Proquest are in SGML.
- It is not presentational, a second process must present things marked in xml.
- It is fairly sustainable: files should be kept in xml, and then generated into other forms like pdfs using xsl.

# 3.5.2 xsl

- A family of recommendations for defining xml document transformation and presentation. Abbreviation of xslfo (formatting objects). Part of three with xslt (transforming) and xpath (navigating)
- XSL is powerful because it can hide or show parts of XML.
- Can be used for exporting into pdf or txt
- Addresses the problem of xml extensibility xml docs can be made compatible.

# 3.5.3 XSLT

- is a stylesheet that converts xml into a form that is presentable by browsers. It is built into xml editors: i.e. Saxon id in <oXygen/>
- XSLT 1.0 more widely used than 2.0

# 3.5.4 Xpath

• provides a format for describing element structure

# 3.5.5 rdf

• rdf allows an object to come with a description saying what it is.

# 3.5.6 TEI

- consensus at the workshop was TEI (for instance it doesn't cover pseudonyms) was limited and if something needed to be added, we should, and then let them know.
- Discussion about keeping TEI lite: a lite markup ensures a base level of interoperability

• UVA library found TEI inadequate for serials. They introduced an element that corresponded to a local thesaurus to declare what each item was. They had TEI headers for numbers and for runs as a whole, and the relationship between them is in the header. See: http://www.lib.virginia.edu/digital/metadata/bestpractices.html

# 3.6 OCR

- decision whether to go for OCR generated text depends on scope.
- Recommends keyboarding vendors, who can be 99.95% accurate, as they do it twice and check against digital images. Some even do the first stages of TEI.
- Matt Gibson, Associate Director of their etext centre, warns against presenting dirty OCR.
- Prime recognition software can work to correct OCR.

# 3.7 Markup

- A solution to converting html sites into NINES compatible xml.
- The Book of NINES is a Romantic Circles product which automates the NINES wrapper. It uses a database to categorize XML. A simple SQL (Structured Query Language) database is effective as demonstrated by its wide use in the corporate world. Data is held in table, which are structured in a database. They put their poems (words, punctuation, white spaces) in a database which then encodes in XML and arranges content non-hierarchically. The data can then be extracted and run according to TEI. The databases can be searched, even though users are using an HTML interface. The databases are devised from the NINES wrapper, and this will be made open source when someone funds it. This means that it could be used to make XML databases from other people's projects.
- This is similar to stand-alone markup or Just In Time Markup. Gavin Budge strongly advocated it, on the basis of a 2000 paper by Philip William Berrie

(http://idun.itsc.adfa.edu.au/ASEC/JITM/Wollongong200004PWB.pdf).

- Makes the following points:
  - Suggests that electronic editions are transcriptions, and so embedded markup affects their accuracy.
  - And embedded markup privileges some users over others.
  - So advocates separating metadata markup from transcription files.
  - Speeds proofreading. And metadata can be added incrementally becomes a dynamic process.
  - However, it requires a static reference file (the transcription).
  - As the markup has a built in authentication system, it can be swapped between members of the community while also ensuring the integrity of the transcription.

- The different JITM tag sets could be viewed simultaneously through multiple browser windows in order to make comparisons.
- Copyright issues: the transcription copyright rests with the owners of the original, metadata may not be selected therefore possibly depriving transcriber of revenue.
- It also demotes the role of the scholarly edition.
- Metadata files can be stored in a database.
- But, although offering a solution to overlapping hierarchy problems in xml, it seems to rest on a very confident notion of the original.

# 3.8 Project Management

- Collaboration is increasingly important and the power of paper helps to cohere groups. There are software packages such as CVS (Concurrent Versions System) that help to control processes in collaborative work.
  http://www.nongnu.org/cvs/
- Five useful suggestions (Melissa White, UVA):
  - Everyone does everything: project leaders mucking in; new people showing your presumptions; staff meetings are important to build teams and delegate responsibility (ownership)
  - Training: allows you to grasp vision of big picture; practice skills; experienced people can learn from new people.
  - Standards: dating; spelling; author's names; citation; imaging specs etc
  - Procedures: numbered stages; centralized documents; weigh carefully implications of changing standards (back-dating)
  - Archiving: automatic and human; beware of duplicates. They have two servers, a working one, and an archive one (when docs move to archive the working files are deleted).

# 3.9 User Interfaces

- Need to identify who is coming to the site web statistic are available from free to a fee.
- Every archive makes an argument
- Search is context free, and only works if you have a question: good UI will show users what is there.
- Linking to other projects (and vice versa) will raise our google ranking however, locating suitable sources is a professional issue.

# 3.10 Sustaining Digital Scholarship

- Digital scholarship needs to be supported by institutions: both while created and to sustain them
- UVA treat them as gifts, just like special collections and see them as camera-ready copy. There may be parts they do not preserve, such as some software applications.

# 3.11 Intellectual Property

- Bridgeman vs Corel settlement in the US, which ruled photographic reproductions are not copyright. You can photograph works of art.
- All institutions have intellectual property policies, but they vary (e.g. UVA cedes ownership to authors unless significant university resources are used but many don't, treating academics as employees).
- We are publishers, so have to check our content now and for the future.
- If prevented from publishing an image by copyright, files should still be kept as the law may change.
- LOC only charge for labour in making copies.

# **Part Four: My Conclusions**

# 4.0 These are some final things I think nose might consider.

NINES is a project worth pursuing, and think its goals are laudable. Context is all, and its meanings are largely structured by the mobility of objects.

# 4.1 Folksonomy.

# 4.2 Database for advanced metadata.