NINES

a federated model for integrating digital scholarship

September 2005

Foreword

NINES is a scholarly initiative to establish a coordinated network of peer-reviewed content and useful tools (both organizational and interpretive) for pedagogical and research materials developed by educators and scholars working in 19th-century British and American literary and cultural studies. The goal is to establish this aggregated body of scholarly and educational materials within those existing professional frameworks and organizations that monitor and accredit professional publication.

NINES is designing a working model for a federated network akin to the Open Access Initiative but operating under coordinated professional oversight. Such a model decentralizes scholarly work, allowing individuals and groups to work and archive scholarly materials in their local "IT" environments and at the same time integrate that work into a widely distributed network.

A Federated Response

Initial designs for the technical implementation of NINES assumed that great emphasis should be placed on content and metadata standardization, with the further assumption that all NINES resources would be uniformly coded, hierarchically organized, and archived from a centrally served location. Subsequent research and

development have persuaded us that this top-down, centralized approach is not the best way to proceed.

We have been drawn to a federated model for a number of reasons: experiments in the UVA Library's "Sustaining Digital Scholarship" initiative; the emergence of RDF and other new protocols for information sharing; consultation with groups building related tools; and a deeper understanding of the use and collection of descriptive metadata. All of these factors expose the clear advantages of a federated, collaborative, and non-hierarchical approach to NINES. This document articulates that approach. We will therefore proceed to give: 1. the (d)evolution of our conception of NINES from earlier models; 2. the relation of a federated model of NINES to trends in data management; and 3. the model's fullest expression in Collex, a semantic research hub for NINES.

The Initial Approach

In March of 2004, the NINES technical development team presented preliminary work on a METS-based "wrapper" or metadata schema meant, in part, to ensure that some nonstandard and legacy digital objects could find a place in NINES. Most files, however, would be required to share a common, NINES-approved TEI-based markup system and interface — including all future contributions to what we imagined would be a centralized database operating in Fedora or some other controlled content management system. When new materials came into NINES, contributors or editors would be asked to position them explicitly within an interpretive hierarchy or stemma based on the concept of the "work." The schema expressing this hierarchy, formalized

as an extension to METS (the Library of Congress Metadata Encoding and Transmission Standard), evolved from the hierarchical organization of a NINES flagship resource, the Rossetti Archive, in which site functions are governed by parent-child relations among works and their documentary or pictorial expressions. Although we weren't aware of it, our theoretical assumptions in this regard paralleled the International Federation of Library Associations' findings on the Functional Requirements of Bibliographic Records (FRBR). The FRBR recommendations form a guideline for the professional cataloging of works (abstract intellectual or creative concepts), their expressions (particular mediations of works), manifestations (such as editions, issues, or states), and dependent items (or actual physical documents) in hierarchical relation (Plassard).

An initial problem loomed, however. In contrast to the Rossetti Archive and to those library collections administered by professional organizations endorsing FRBR, NINES would come into being without metadata specialists or cataloging staff solely dedicated to overseeing, supporting, and maintaining a uniform set of coding standards. The technical development team assembled at the University of Virginia was meant to serve as midwife to the NINES community, but that team could not be imagined to remain in place past the early development of the project, when contributions from individual scholars would begin to flesh out the data set. Maintenance of the intellectual structure of NINES would therefore fall on the community itself — nor even solely on the scholars who had been invited to serve as NINES board members, only a dedicated subset of whom are humanities

technologists. More critically, contributors to and users of NINES tools and materials would be required to operate within and uphold the standards we set up.

The development team's initial approach was to view this conception of the future operation of NINES as a problem to be managed. How could we ensure the continued operation of an evolving resource? Our first solution was an attempt to limit the technological evolution of NINES by providing tools and interfaces for markup and metadata collection in compliance with three rigid sets of standards:

1. for the internal organization of documents (a NINES extension of TEI),

2. for the way documents were to be described and linked (the METS-based NINES "wrapper"),

3. and for the protocols through which contributions could be submitted to and archived in a centralized server.

Working to these standards and operating under the guidance of its editorial boards, NINES could, we imagined, essentially run itself. It would still be an evolving resource in the sense that new tools and content could be incorporated and the whole body of material could be converted to alternate formats in the future, but it would, for some time, represent a monolithic view of the markup and metadata relevant to research in 19th-century literature and culture. We viewed the limitations such a system would impose on the critical and creative freedom of contributing scholars as a necessary evil, when balanced against the great advantages of bringing hitherto incompatible digital resources together in an integrated, peer-reviewed research environment.

A Working Model for NINES

Fast-forward to the spring-summer of 2005. Work on NINES-related tools such as Juxta (a bibliographical collation system) and IVANHOE (a multi-player game of interpretation) has continued. In addition, we've begun concrete development of Collex, a tool for collecting and annotating digital objects, and for publishing rich, interlinked online exhibits in collaborative research environments. We now recognize Collex as a natural pivot around which the community of NINES will revolve, producing new scholarship on NINES materials and feeding that work back into the system. Furthermore, the SIMILE project at MIT has demonstrated that RDF, the "resource description framework" that NINES will use to represent complex material and conceptual objects in its aggregated corpora, is a viable component of real semantic web tools.

Although our initial, centralized and standardized model for NINES would promote future tool-building and data migration, we are now persuaded that NINES must be understood as a *social system*, especially as it gains traction in a scholarly community still constrained by traditional paper-based communication conventions. We cannot set the bar for participation in NINES too high: perfect compliance with standards for markup, metadata, interface, and archiving will slow the growth of the resource at its most critical juncture.

This fact was brought home to us in the most powerful way during the past year's bi-weekly meetings with the UVA library and related conversations with its Fedora development team. The library is determined to develop a library-based "publishing"

model and is throwing its best resources to the practical development of such a model. We can now see, however, that it will be many years before a Fedora-type mechanism will be ready for even a preliminary practical implementation, in particular for a set of NINES-type aggregated materials. In addition, publishing models located elsewhere (for example, in the university press network) have been exposed as inadequate for a project like NINES — though it is entirely imaginable that NINES would enter into alliances with various publishing entities, including university presses that have transitioned to digital production. Finally, we learned from the legal counsel who participated in these library meetings that while copyright restrictions are crippling to the kinds of formal "publications" created by (say) university presses, they become far less constraining in the case of decentralized, fair-use scholarly resources. This fact is partly responsible for the remarkable recent growth of interest in institutional repositories (such as Fedora and DSpace) as a way of archiving scholarly and educational materials.

So a model for aggregating and disseminating online humanities scholarship that assumes a centralized organization and a more or less strict set of governing standards was beginning to present formidable practical difficulties for NINES. In addition, we were continually seeing that that kind of approach was running counter to the major trends in software development and the metadata cataloguing that would support a robust and flexible system of knowledge and information management, search, and analysis.

A federated model for NINES pursues a design that exploits and develops these very trends. Briefly, the model incorporates the advantages of an Open Archives Initiative (OAI) approach to a professionally-oriented institutional structure of NINES. A set of "institutional repositories" is thus put into online aggregation where new kinds of knowledge and information management tools can be shared to generate deep explorations across the set of aggregated materials. These tools — described more particularly below — include a powerful search mechanism based in Lucene; an RDF syntax that greatly facilitates the description and semantic integration of online material (in this case, within a trusted environment); and the Collex tool (being developed at UVA) to collect, organize, and display second-order research acts carried out by anyone working with, and ultimately adding to, the corpora of scholarship assembled in NINES. The Collex application allows students and scholars to **COLL**ect trusted NINES materials and refashion them into fresh research presentations or "**EX**hibits". These research projects, individual or collaborative, can then be fed back into NINES as work-in-progress or for inclusion within the peer-reviewed corpus of NINES materials.

The federated model permits individual contributors of scholarly resources (who are also its users) to:

1. integrate fully into their local college or university IT environments, including established institutional repository systems like DSpace or Fedora;

2. maintain complete creative and critical freedom over matters of interface and markup particular to the scholarly needs of their users and developers;

3. manage the publication of NINES peer-reviewed scholarship alongside new or experimental material without compromising either body of work;

4. expose several levels of detailed metadata to a central harvester, enabling full-text search (both locally and across the federation), faceted browsing (by genres, dates, etc.), and user collection, annotation, sharing, and re-purposing through Collex and other tools at a central NINES web site.

Although these advantages may appear to be purely technical, the real benefits of the federated model are institutional and scholarly. A federated NINES holds its greatest promise for the scholars, teachers, institutions, and students who collaborate in its resource and content development. (See Appendix 1 for more information about the participation requirements placed on different types of NINES contributors: scholars maintaining digital archives; scholars contributing to or editing online and print journals; curators of museums and libraries exposing objects for NINES harvesting; and end users of the Collex tool, whose work can feed back into emergent data structures in NINES.)

In order to test and refine our model for federated digital scholarship, and to make its implications clear to potential funders, we are undertaking a pilot implementation of NINES. The milestone date for its rollout is December 2005. This pilot includes two large XML-based archives (Rossetti and Walt Whitman), two smaller archives (the Poetess Project and the Swinburne Archive), one electronic journal (Romanticism on the Net), and one collection of legacy resources in transition (Romantic Circles). These conceptually disparate and geographically distributed resources will be integrated for

full-text search using Nutch, an open-source Lucene-based engine. Because these resources will be sharing metadata in standard formats such as Dublin Core and RDF, they will also be available for faceted browsing (in categories based on genres, authors, dates, etc.) and for the several Collex operations that enable scholars to collect and annotate NINES objects, to explore these materials in knowledge discovery operations, and to feed this work into the NINES environment in exhibits that can be shared with other NINES users. Finally, NINES will provide other interpretive applications — Juxta and IVANHOE — that can either be launched from Collex or deployed on their own for collation and text analysis, on the one hand (Juxta), and on the other (IVANHOE) for exploratory investigations of NINES materials within a collaborative playspace.

The Federated NINES: How Does It Work?

Humanists eager to develop new ways to integrate and explore digital works currently lack the institutional and technical resources to do this. Think about the models, even the best models, that scholars now can follow and imitate: The Whitman Archive, The Rossetti Archive, The William Blake Archive. These are stand-alone projects that can only be loosely integrated through web browsers, even when shared through OAI protocols. Romantic Circles is itself a closed system, and its different parts are largely disjunct from each other except in the most general way. As a consequence, what you see now on the web is what you get: an agglomeration of sites and projects whose content is atomized and whose scholarly and educational value is indeterminate. While

it is possible for tech-savvy scholars, using ad-hoc tools and methods, to produce and distribute annotated, re-organized, or selected versions of existing online resources, they presently lack coordination within a peer-reviewed digital publishing environment. Because of this, their productions — personal web pages and online course packets — are difficult to maintain, are not readily interoperable or standards-compliant, and are easily dismissed as heterogeneous grab-bags of links. NINES was founded to work against that debilitating situation.

Two broad requirements have to be met: 1. an information management system for submitting material to NINES, peer-reviewing it, and then certifying its integration within the official NINES environment; 2. a set of tools that take advantage of the special knowledge-discovery and interpretive capacities made possible through digitization, and which integrate an individual's work with others'.

To date, NINES development has been directed at the second of these requirements, and the remainder of this document will be a description and explanation of that development work. We have focused in this area because so little has been done to develop a coherent approach to these basic research needs. Information management systems, on the other hand, have been well-developed and we believe we can identify an open-source peer review application and modify it to the special needs of a project like NINES, or even roll out one of our own.

The scholarly and educational software being developed for NINES started with IVANHOE and then passed to designing and building Juxta and Collex. In the course of this work, we began to see that Collex was leading us toward a more coherent and

robust understanding of how to design and build what used to be called “A Scholar’s Work Station” (IATH). Our view of that idea is very different from the way it was being imagined in 1993. The individual scholar can and should do more than simply access the resources available through a large distributed online network. The digital environment should be so designed that individual projects and research acts will get reintegrated into the larger network at deep semantic and structural levels. We now see that this can best be accomplished through design frameworks that exploit a decentralized and “bottom up” approach.

The key to such a flexible design is RDF (Resource Description Framework). RDF is a shared, semantically rich means of describing the objects in a set of digital resources like NINES. A specialized Dublin Core “flavor” of RDF will provide a NINES-standard metadata framework for digital objects such as the transcriptions, page images, art work, and reproductions available from a multimedia digital resource like the Rossetti Archive. RDF puts distinct and distributed resources — The Walt Whitman Archive, Romantic Circles, the Poetess Project — in a framework of interoperability that is more flexible and user/contributor-malleable than an ontologically structured metadata system.

When a primary resource like The Rossetti Archive is introduced into NINES, a flat array of RDF metadata descriptors is attached to its various “objects”. Initially defined by the creator of the resource, the basic set consists of author, title, date, and genre. Content contributors may define as many or as few “objects” as they wish; there need not be a one-to-one relationship between web pages and RDF-defined objects. Each

poem in a digitized book, for example, may be represented as a discrete “object” in NINES. (See Appendix 1 for more discussion of RDF production by content contributors.) The RDF is then harvested and made available for search and browsing at the NINES aggregation site. Users can also employ NINES tools that act on provided RDF metadata to expose unapparent relationships among peer-reviewed objects housed in distributed archives and encoded using disparate schemas and DTDs.

The RDF framework drives a powerful search and indexing procedure that consists of a customized indexing system (Nutch) and the search engine, Lucene — both open-source software tools. The NINES system uses RDF to locate specific objects for collection and reuse. It also directs the content fetched by Nutch into a Lucene index, to which queries are addressed for faceted browsing and searching. Nutch includes a crawler, content parsers, an indexer, and a basic search interface, which we have customized for NINES by integrating faceted browsing with search. The crawler fetches content from web sites or portions of web sites which have been designated, through a process of peer review, as NINES affiliates. Content from these sites is parsed and any extracted links feed into the next round of crawling. (See http://lucene.apache.org/nutch/.)

Nutch leverages Lucene, a widely-acclaimed open source search engine, for indexing and searching. Lucene itself is solely a library, however, not a user interface or document parser. Developers embed Lucene in an application — in this case, in Collex — and map the application domain to the concept of "documents", the smallest

granularity of a search result. Documents are made up of developer- or user-definable metadata fields, such as the basic starter array for NINES “objects”: author, title, date, genre. Sophisticated querying capabilities include: boolean search (with “and”, “or”, and “not” operators); fuzzy queries using the Levenshtein distance algorithm; and wildcard, range, or phrase queries.

Collex is the NINES application that brings this indexing and search design framework into a collaborative research environment.

Collex

Collex is an open-source collections- and exhibits-builder to aid humanities scholars doing research in complex digital collections like the Rossetti Archive or within federated research environments like NINES. Such environments often stymie their users through the sheer quantity of information made available to them in top-level tables of contents, sitemaps, and idiosyncratic search engines. Collex operates under the assumption that the best paths through a complex digital resource are those forged by use and interpretation. A Collex approach works to assist scholars in recording, sharing, and building on the interpretive purposes to which they put their online teaching and research environments.

“Clio” is our codename for an extension of the Collex idea through RDF and faceted browsing. Clio leverages semantic search and current developments in collaborative tagging or “folksonomy” tools to perform data mining operations and enhance

knowledge discovery. In this approach, similarities and relationships are exposed among objects in user-collections through a series of automated "reveal more like this" queries — more pencil sketches from this decade, more poetry in this sub-genre, more commentary referencing these concepts. The scholar is led to see connections among objects and ideas based on the contexts into which they have been placed (implicitly or explicitly) by past scholarly activity within the system.

By integrating Collex with established digital archives, NINES can enable the expression of alternate interpretive visions within editorially-organized electronic environments. Software applications designed for more particular interpretive uses — the collation tool Juxta, for instance, and the interpretive playspace IVANHOE — can be launched from within the Collex environment and applied to objects users have "collected". For further information about these applications see Appendix 3 below.

The Collex application is designed so that scholars and students using NINES may:

1. collect, tag, analyze, and annotate trusted objects (digital texts and images vetted for scholarly integrity);

2. reorganize and publish NINES objects in fresh critical perspectives;

3. share these new collections with students and colleagues, in a variety of output formats;

4. and, without any special technical training, produce interlinked online and print "exhibits" using a set of professional design templates.

Faceted Browsing and Knowledge Discovery

Collex uses semantic web principles and technologies to explore and develop the research potential of the digital scholarship aggregated in NINES. Two critical concepts embodied in a NINES environment shaped by the Collex application fall under the rubrics widely known as "faceted classification" and "folksonomy." Facets and folksonomies structure an approach to descriptive metadata. They generate an evolving interface between the peer-reviewed electronic resources that constitute NINES and the user communities that reimagine NINES content through interpretation, contextualization, and critical and creative re-fashioning.

In this approach to the design of NINES, the content is built and deepened not only by ingesting NINES content objects (like texts or images from the Whitman Archive or the Poetess Project), but by the interpretive reuse and transformation of materials already aggregated within the NINES environment. Note that this design approach mirrors precisely the character of traditional scholarly work that is executed within the spaces of our pre-digital institutions.

Faceted navigation is a method for browsing resources or refining search results along multiple non-exclusive, non-hierarchical dimensions. It is based on a classification scheme designed by Indian librarian S. R. Ranganathan in the early decades of the twentieth century as an alternative to traditional, enumerative taxonomy.

Objects in a faceted collection are not limited to a single "location" in a top-down navigational hierarchy or enumerative taxonomy. Nor must a metadata specialist

understand complex cataloguing rules or make the futile attempt to predict all categorizations under which users may expect to find a given object. Faceted systems are predicated on the notion that "there is more than one way to view the world, and that even those classifications that are viewed as stable are in fact provisional and dynamic" (Kwasnick).

In the traditional enumerative taxonomy of the Library of Congress, a book like Ken Daley's "The Rescue of Romanticism: Walter Pater and John Ruskin" may be cross-listed with Ruskin and early 19th-century literature, but will always find itself shelved with Pater criticism at a particular spot in a rigid hierarchy. Cross-listing is an attempt to facilitate faceted browsing within pre-established hierarchies in a world limited by the physicality of the book. The complex subject headings through which books like Daley's are located, however, betray the complications of an unfaceted system:

> **Subject**: Pater, Walter, 1839-1894--Criticism and interpretation.
>
> **Subject**: Ruskin, John, 1819-1900--Criticism and interpretation.
>
> **Subject**: English literature--19th century--History and criticism.
>
> **Subject**: Art criticism--Great Britain--History--19th century.
>
> **Subject**: Romanticism--Great Britain.

Furthermore, note that this book would not be locatable through a number of its clearest additional facets — for instance, a set that would hang off a "Victorian" catalogue heading.

In a faceted system, subject headings could be composed on the fly by users engaged in an active winnowing of resources related to their interests. "Art criticism," "Great Britain," "History," and "19th century" would appear as combinatorial options within larger facet categories (time, location, genre, and so forth), rather than as one highly-specialized and absolutely specified node in a tree. Faceted classification is a bottom-up scheme, in which structure emerges on demand from the attributes applied to objects and the interests of end users of the collection. Facets become "dimensions in a Cartesian n-dimensional space, and the value of a facet is the position of the object in that dimension" (Quintarelli). Even the facet-values themselves need not be hierarchically structured (in general-to-specific or whole-to-part systems): the alphabet itself can serve as a useful and easily-generated facet for browsing titles and authors (KM).

Joseph Busch, past president of the American Society for Information Science, describes the advantages of faceted classification in orders of magnitude: "Four facets of 10 nodes each have the same discriminatory power as one taxonomy of 10,000 nodes." In other words, forty individual subject headings, organized in a manageable set of four facets, allow the precise retrieval of a set of objects that might — if, in an idealized case, they were sufficiently different as to warrant them — require ten thousand separate, complex hierarchical descriptors (Papa).

> Thus, a faceted structure relieves a classification scheme from the procrustean bed of rigid hierarchical and excessively enumerative subdivision that resulted in the assignment of fixed "pigeonholes" for

> subjects that happened to be known or were foreseen when a system was designed but often left no room for future developments and made no provision for the expression of complex relationships and their subsequent retrieval. (Taylor)

Faceted browsing systems, the interfaces giving access to objects so classified, are typically constructed so that users can begin with any facet, so narrowing the collection and the set of available facets to those that have valid results. The system only displays those facets under which objects meeting the requirements of "all" selected facets can be found. It is impossible to return an empty result or failed search to a user, and users are never required to play a guessing game with keywords. Ultimately, this process should deliver a subset of objects small enough for reading.

The advantages of a combinatorial set of facets are clear. It empowers users and content providers alike by remaining manageably small and not requiring overly-specialized subject knowledge for creation and navigation. Put simply, you don't need to know that "the history of nineteenth-century art criticism in Great Britain" is a sanctioned category in order to place or find an object on that "shelf." Because facets operate independently and don't require a single, over-arching taxonomy, their "schemaless" data model allows hetereogeneous collections to be combined and new facets to be added at any time (Papa). And the n-dimensional structures that result from user demands on faceted classification systems are particularly appropriate to and evocative in digital environments (Broughton). Played out across a rich research archive — in which other authors' relationships to works, genres, time periods,

subjects, and so forth become calculable — allusive RDF faceting will make NINES an important experimental playground for the semantic web.

Folksonomy and Research Communities

"Folksonomy" is a portmanteau word coined to describe recent developments in so-called "folk taxonomical" collaborative keywording or tagging. A folksonomy allows many users — all working in their own self-interest — to cooperate in assigning free-form metadata to digital objects. Like faceted classification, folksonomy responds to complications in practices of metadata assignment common to enumerative taxonomy. These complications have been introduced by the proliferation of online resources, both born-digital and digitized from print sources, hitherto requiring formal description for full exploitation.

While descriptive metadata can be produced by its traditional gatekeepers — professional cataloguers — to a very high standard of quality, it comes at a high cost in terms of time and money. In addition, the limitations of hierarchical classification systems (see the discussion of facets, above) and the explosion of ephemeral resources on the Web combine to make professionally-created metadata an unattractive or difficult option. In recent years, driven largely by commercial and blogging applications, the Web has begun to turn away from dependence on metadata professionals in two clear ways:

1. in favor of allowing content creators (using XML, Dublin Core standards, and Technorati tags) to classify their own information and

2. in gathering *implicit* information from end users (for example, through Google's PageRank technology or Amazon's "recommendations") in order to sort and group or "collaboratively filter" digital objects.

Faceted browsing in NINES is an exploitation of content-creator-generated metadata, while collaborative tagging is a growing extension of the latter impulse to gather data from real-world use. At this third deliberate remove, end users are prompted to offer *explicit* information about the ways in which they think about and apply resources on the Web.

Such explicit information is saved and shared in the form of "tags," typically one-word or short-phrase identifiers composed in natural language and suited to the exigencies of each individual tagger's research project. Tagging is useful as a flexible, non-hierarchical and personal organizing system for objects like web pages (del.icio.us) and digital photographs (Flickr). Connotea and CiteULike have also brought tagging to the sciences, where academic research is published and consumed quickly online, and where personal classification systems work well in the context of quickly-evolving studies and projects (Hammond et al). All of these systems make it easy for users to develop and refine categorizations that make local sense. But folksonomic tagging or "social bookmarking" really comes into its own as a group endeavor, when users' tags are shared across a large body of material and where facilitating software systems suggest relevant tags and resources to participants:

> The power of folksonomy is connected to the act of aggregating, not simply to the creation of tags. Without a social distributed environment that suggests aggregation, tags are just flat keywords, only meaningful for the user that has chosen them... The term-significance relationship emerges by means of an implicit contract between the users. (Quintarelli)

Folksonomies that reach a certain critical mass can function as sensitive aggregation and concept-matching tools, despite variations in terminology introduced by the uncontrolled nature of tag vocabularies. Aggregation can be facilitated in the practice of tag stemming, in which tags sharing a common root are automatically matched. Objects tagged “sonnet” and “sonnets,” therefore, can be linked by the computer without requiring taggers to agree to a single term.

Tag bundling is another algorithmic operation that helps folksonomies function as emergent ontological or concept-matching tools. In this practice, the computer monitors all the tags applied to a set of objects and draws relations among them, noting which tags tend to overlap. Those overlapping tags can then be presented to the user as “bundles” related to a particular tag or object of interest, thereby perhaps alerting the user to unexpected ways in which others have classified NINES material. Tag bundles are also useful as an alternate means of grouping search results. Instead of presenting results as an ordered list of matching hits, a NINES interface might cluster them into groups emerging from user practice and application. For an excellent example of disambiguation and knowledge discovery through tag bundles, see cluster

pages for the tags “turkey,” “sole,” and “twister” on the popular photo sharing site, Flickr (http://www.flickr.com/photos/tags).

Flickr also groups search results by levels of “interestingness,” an algorithmic measure that takes into account the number of times and the ways a given object has been tagged or otherwise repurposed. NINES could easily implement a similar system, which would then serve as an aid to researchers, alerting them to objects that have received either a great deal of critical attention or very little notice, and assisting them in characterizing that attention. A simpler variation on this idea would allow researchers to “subscribe” to an RSS feed, automatically produced for each tag used in the NINES folksonomy. This subscription would serve as a customized research alert, notifying users whenever new objects fitting their criteria have been added to Collex collections, or when existing objects are newly classified in interesting ways.

Folksonomies lend themselves to graphic presentation. The most common visualization is a weighted list, or “tag cloud,” in which tags are coded by size and/or intensity of color to correspond to their frequency of use. (The Flickr link above presents a typical tag cloud.) Tag clouds offer viewers the Zeitgeist of a folksonomic community at a glance, or can be keyed to the tags of an individual user and mapped over time to trace classification patterns. They also serve as navigational devices. Clicking text in a weighted list opens the selected tag as a browseable facet and gives access to all objects so tagged.

The swell of enthusiasm for collaborative tagging or social bookmarking that is so notable an influence on Internet development in 2005 leads to the announcement of

new folksonomy visualization methods and tools almost every week. These include: Venn diagrams for picturing clusters or bundling; undirected graphs in which related tags are represented as connected nodes in a network; and temporal plots of the evolution in use of particular tags. Tom Coates of BBC Interactive suggests applications for temporal analysis of tag use:

> But what do changes in a tag-cloud mean? Probably the most obvious underlying cause for a change in the words used to describe a site would be that *the site itself has changed*. You could probably use an analysis of the changing tag-cloud to get a handle on what's happening to the site. That's quite interesting.
>
> After that - or alongside that - another underlying cause could be *a change in the vocabulary* around a subject. At a really grand level, if you can imagine a one hundred year tag-cloud around a gay novel, then it might start with lots of people using the tag invert, with this gradually giving way to homosexual, then gay and potentially after that, queer….
>
> But there's also a third potential cause for changes in a tag-cloud over time - that people might approach the very act of tagging differently - that their understanding of what they're doing might develop. (Coates)

Clearly, this sort of data mining and analysis only comes at a second remove, when the NINES user community is well enough established to *have* a history of critical practice. Folksonomy tools, as implemented in Collex, will be of supreme importance both in tracing critical trends in the NINES user community and — through the added research and collaboration benefits they bring in the short term — in ensuring that that community thrives.

The federated model for NINES, in eschewing one-size-fits-all digital encoding and top-down taxonomy, clearly favors the kind of emergent structure native to collaborative tagging. Indeed, we think it likely that new organizing structures will emerge from the uses to which students and scholars put NINES materials, and will one day be codified into alternate interfaces for NINES. Architectural theorist Christopher Alexander describes, in *A Pattern Language*, a method employed by landscape designers who share our faith in praxis. After foundations are laid, buildings are built, and occupants take their places in new structures, these designers seed the spaces in-between for grass. They wait to pave, until "desire lines" emerge in the form of walkways, well beaten in receptive dirt.

Research and Publication in NINES

What has just been described is how students and scholars could access research and pedagogical resources distributed across a broad range of institutional repositories and servers, but integrated and accessible through a central NINES interface. Scholars would use those resources to develop content in their local IT environments for whatever purposes they have in mind. A scholar might be working on a particular research project or developing teaching materials for a specific course. If the work were part of an ongoing project (research or teaching), it could be exported as "Work in Progress" for comment and discussion by the NINES community. A collaborative project could also be facilitated through the NINES environment. And of course any

kind of finished work, regardless of where it is housed, could be vetted in NINES peer review and, ultimately, admitted to the sanctioned corpus of NINES materials.

Furthermore, the federated design is flexible enough to enable a good deal of interoperability at different scales. NINES resources can integrate with Fedora, DSpace, or other institutional repositories, nor would there be, in principle, any obstacle to prevent collaboration with resources like JSTOR, and least of all with other initiatives like NINES should they be created. In addition, NINES is specifically conceived to help paper-based periodicals like, for instance, *Nineteenth Century Studies*, to make a digital migration that would allow them to maintain their local identity while at the same time becoming aggregated in a developing corpus of related scholarly materials.

Appendix 1: Making Digital Resources NINES-Ready

All indexed contributions to NINES — that is to say, all material available to users through search and faceted browsing at http://www.nines.org — will be peer reviewed by the NINES editorial boards and (if necessary) vetted for basic technical integrity by the technical advisory board. We've written software specification for the NINES peer review system, and anticipate that it will be online by December 2005. The peer review system will help contributors input basic metadata about their projects and will, when applicable, translate that metadata into NINES-ready RDF to facilitate search, browsing, and other acts of knowledge discovery.

This should be a simple process for contributors of single objects, such as journal articles or scholarly monographs. Complex projects like digital archives, however, will require more detailed, object-specific RDF. Primary responsibility for the production of NINES-ready RDF lies with contributors.

NINES contributions fall into four basic categories:

1. User-created material in Collex:

Collections and exhibits created in Collex are already NINES-ready. NINES RDF is generated and stored automatically as you create and modify your own Collex

collections. It is also produced and embedded in user-created Collex exhibits. Think of the RDF associated with your Collex exhibit as an automatic "works cited," generated on the fly as you reference NINES objects in your exhibit commentary.

Most Collex collections and exhibits will be housed in a special section of the NINES site, reserved for community-created objects not subject to peer review. You may, however, wish to submit an annotated collection or Collex exhibit for peer review and inclusion in the sanctioned set of NINES materials. In either case, the RDF process is automated and any additions you wish to make to the generated RDF can be done through web forms.

2. Federated scholarly archives:

Maintainers of digital archives and other scholarly resource collections housed in institutional repositories or on private servers must produce NINES-ready RDF for each "object" they wish to make collectable in Collex and discoverable through Clio and NINES faceted browsing. For digital objects expressed in XML, the normal method of RDF production will involve XSL transforms. NINES will provide a basic set of XML-to-RDF stylesheets for TEI-encoded documents by way of example, and will make available a schema for verifying that the RDF you generate from your data meets our standards. Once generated, this RDF may be embedded directly in your rendered HTML documents or stored separately and linked in via META tags.

3. Resource records from NINES-approved institutions:

The federated system makes it possible for approved institutions such as museums, galleries, libraries, and archives to contribute RDF records expressing their holdings for integration into the NINES search interface. This is an especially practicable option for small collections of 19th-century objects that may not be networked into other finding aids, or which have been overwhelmed by the enormity of databases like OCLC. Bringing RDF records into NINES will not only open them to search and browsing; it will also allow curators or other NINES users to feature them in Collex exhibits for greater exposure to the community of scholars.

In most cases, institutions wishing to contribute to NINES will be able to generate RDF easily through XSL transforms or other database queries. Unlike the electronic scholarly archives described above, NINES-approved institutions may not offer discrete web pages describing each artifact in their holdings. In this case, RDF can be made available for search and browsing without any link to a specific web resource. Instead, users may be offered a link to the institution's home page and accession numbers for relevant objects.

4. Contributions to NINES-affiliated journals:

NINES-affiliated journals may be published online or in print formats — or in some transitional or hybrid state between the two. The metadata or RDF production process is identical from the point of view of journal contributors. Contributors will submit

material to the specifications of the journal via a NINES peer review interface, which will also assist them in assigning metadata categories corresponding to NINES-ready RDF fields. An added feature of this process for journal articles (print and digital) may be the automated assignment, on publication, of a Digital Object Identifier (see http://www.doi.org), which can be used as a unique reference number, allowing users to collect and catalog resources not only in Collex, but in bibliography software such as EndNote and Connotea. (The Nature Publishing Group, for instance, has begun listing DOIs on every page of their print publications.)

In each of these cases, the fundamental procedure for bringing materials into NINES involves submission for peer review and collection of relevant metadata, which is then transformed into RDF. Once resources are vetted and approved, they can be crawled for full-text search and their RDF records can be incorporated in the NINES database for faceted browsing and knowledge-discovery operations through Collex and Clio. We hope to build user-friendly interfaces and provide excellent documentation and schema materials for making this necessary work as easy as possible.

Appendix 2: Issues for consideration by the NINES board

We will require authority lists for valid fields in NINES-approved RDF. This does not mean that contributors must use these terms and formats in their own markup, but rather that their metadata must be mapped to NINES standards through XSL transformation into RDF. Fields needing standardization include: names, dates, and genres.

We can use the Library of Congress naming authority for persons. Is a single "agent" field sufficient for initial implementation of faceted browsing? (Or must we break this into roles for authors, editors, models, scribes, etc?)

We will require a simple and standardized format for dates. We propose to limit date records to years and ranges of years in the following formats: "1861" and "1861-1865." Is this sufficient for the initial implementation?

Objects are typed in the RDF as "primary" or "secondary" resources and then assigned one or more genre facets. For genre listings, we offer the following set for discussion, derived in part from the *Cambridge Bibliography of English Literature*:

TEXTUAL	**VISUAL**
Poetry	Oil Painting
Novel	Watercolour
Short Story	Drawing
Drama	Sketch
Non-fiction Prose	Photograph
Art Criticism	Engraving
History	Lithograph
Philosophy	Reproduction
Religion	Sculpture
Newspaper	Stained Glass
Magazine	Furniture
Translation	Architecture
Travel	Clothing and Jewelry
Education	Other
Leisure	
Other	

Another topic of discussion is the issue of peer review in the context of NINES-affiliated journals. We assume that contributions would continue to be made directly to these journals (through their web sites, etc.). Given that a robust peer review system

will be in place for editorial contributions to NINES, how can it be modified to assist journals and encourage the migration of print journals to digital formats?

Appendix 3: Related Tools: IVANHOE and Juxta

IVANHOE is an online collaborative playspace designed to promote imaginative reinvestigations of cultural materials. It is an especially useful pedagogical tool for students at all levels from middle school through graduate school, but it can also be used by groups of scholars who want to work together in a joint research project. It has been successfully deployed in both modes and at all those educational levels.

Descriptions and discussions of the software are widely available. (See the most recent discussion: Jerome McGann, "Like Living on the Nile. IVANHOE, A User's Manual", *Literature Compass* 2 (2005): VI 149: 1-27. For the IVANHOE hompage, which has a QuickTime demo plus other demo versions (including a hands-on demo), see http://patacriticism.org/ivanhoe.)

Juxta is a software application that carries out collation and text comparison operations on any number of textual witnesses. It is a cross-platform application that can be used to collate either poetry or prose documents and to arbitrarily choose as the basis for the collation any of the textual witnesses. The tool allows the user to annotate the collations or any of the text comparison operations, and to export the results of the analyses in a number of formats that include raw data, selected transcriptions, and traditional calculi of variants. The tool displays the results of its collations in several analytical visualizations, including a histogram of textual variance

and a “heat-map“ of variance keyed to an arbitrarily chosen base text. This color-coded overlay can be clicked at any point to reveal schedules of variants in the collated witnesses.

The basic Juxta interface can be toggled for a focused comparison of two texts only. In addition, the tool keys its transcriptions to any digital images that are the basis for the transcriptions.

Juxta’s functionality can be usefully presented in the following workflow chart:

Collect — This operation involves collecting the set of comparands for the analytic operation. Witnesses brought into Juxta for collation and analysis will bring with them a basic set of metadata identifying the witness. The collecting operation might be accomplished via a web interface or by selecting documents from the local file system. In a NINES environment, Juxta and Collex would be integrated so that a Collex operation could be output to Juxta, or an operation in Juxta could be exported to Collex.

Normalize — Once the textual witnesses are marshaled, if they are not already in a uniform format (marked or unmarked), they should be normalized for comparison. The default normalization would preserve capitalization and italics and identify non-standard textual forms (like gothic type).

Collate — Collation is an iterative step. Over the course of a single Juxta session the scholar may perform many collations. The collation may be executed with any witness

as the base text and witnesses may be added or removed from the comparison set. The collations can be filtered for different types and levels of collation (e.g., excluding accidentals or collating only accidentals).

Analyze — The software facilitates interpretation of the results through an interactive user interface. The analyses may be of various kinds: a two text comparison; a study of variance by manipulating a distance algorithm; a study through the histogram visualization; analysis through a search controlled by a dictionary of word forms.

Annotate — This functionality allows the scholar to mark up the collated text, thus adding new metadata to the individual witnesses and/or creating annotations that relate to the acts of collation and analysis themselves.

Export — Saving the operations executed in Juxta and moving them elsewhere for reception and repurposing, for instance as stand-alone web interfaces, Collex exhibits, digital or print edition apparatus, a pure data dump, etc.

Receive — In the NINES context, this step would probably involve the Collex function, in which new users receive, interpret, and repurpose or revise the results of a Juxta collation.

Bibliography

Alexander, C., et al. A Pattern Language: Towns, Buildings, Construction. Oxford UP: 1977.

Broughton, V. “Faceted Classification as a Basis for Knowledge Organization in a Digital Environment; the Bliss Bibliographic Classification and the Creation of Multi-Dimensional Knowledge Structures.” New Review of Hypermedia and Multimedia, no. 7 (2001): 67-102.

Coates, Tom. “Two Cultures of Fauxonomies Collide.” Plasticbag.org (4 June 2005). online: http://www.plasticbag.org/archives/2005/06/two_cultures_of_fauxonomies_collide.shtml

CiteULike. online: http://www.citeulike.org

Connotea. online: http://www.connotea.org

Dspace. online: http://www.dspace.org/

Del.icio.us. online: http://del.icio.us

Denton, William. “How to Make a Faceted Classification and Put It on the Web.” November 2003. online: http://www.miskatonic.org/library/facet-web-howto.html

Flickr. online: http://www.flickr.com

Fedora. online: http://www.fedora.info

Golder, Scott and Bernardo Huberman. “The Structure of Collaborative Tagging Systems.” Information Dynamics Lab, HP Labs: 2005. online: http://www.hpl.hp.com/research/idl/papers/tags/tags.pdf

Hammond, Tony et al. “Social Bookmarking Tools (I): a General

Overview." D-Lib Magazine, April 2005. online: http://www.dlib.org/dlib/april05/hammond/04hammond.html

He, Ping et al. "Faceted Classification." School of Library, Archival and Information Studies, University of British Columbia, 2 April 2003. online: http://www.slais.ubc.ca/courses/libr517/02-03-wt2/projects/faceted/

IATH. "1992-93 Annual Report of the Institute for Advanced Technology in the Humanities." online: http://web.archive.org/web/20000711144156/http://www.iath.virginia.edu/iath/annrep.93.html

KM (Knowledge Management Connection). "Faceted Classification of Information." online: http://www.kmconnection.com/DOC100100.htm

Kwasnick, Barbara H. 1999. "The role of classification in knowledge representation and discovery." Library Trends 48 (1): 22-47.

Mathes, Adam. "Folksonomies - Cooperative Classification and Communication Through Shared Metadata." Graduate School of Library and Information Science, University of Illinois Urbana-Champaign. December 2004. online: http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

Papa, Steve. "A Primer on Faceted Navigation and Guided Navigation." Best Practices in Enterprise Knowledge Management, Vol. IV, A Supplement to KMWorld, November/December 2004, Volume 13, Issue 10.

Plassard, Marie-France, ed. "Functional Requirements for Bibliographic Records: Final Report." IFLA Study Group on the Functional Requirements for Bibliographic Records. UBCIM Publications: New Series Vol 19. Munich, 1998.

Quintarelli, Emanuele. "Folksonomies: Power to the People." ISKO Italy-UniMIB: Milan. 24 June 2005. online:

http://www.iskoi.org/doc/folksonomies.htm

SIMILE. online: http://simile.mit.edu/

Taylor, A. G. “Classification of Library Materials.” In Wynar's Introduction to Cataloging and Classification. 9th ed. Library and Information Science Text Series. Englewood, CO: Libraries Unlimited, 2000.

Veen, Jeffrey. “Faucet Facets: A Few Best Practices for Designing Multifaceted Navigation Systems.” Adaptive Path, 4 June 2002. online: http://www.adaptivepath.com/publications/essays/archives/000034.php