1 Objectives

Digital libraries have already begun to serve humanists in two major ways. First, they enhance the reach of traditional researchers: once a text is digitized, even the simplest of search facilities allows users to interact with and study texts in entirely new ways. Second, where the technologies of industrialized print made possible a network of research libraries centered primarily in Europe and North America, electronic collections, linked by high-speed networks, are becoming digital libraries that serve a global audience, reaching far beyond universities into schools, public libraries, and private homes. The cultural heritage of humanity can play new roles in the lives of professional scholars and the general public alike. A properly designed global information infrastructure can strengthen our own cultural identities while opening up for us cultures radically different from our own.

But humanists cannot develop their information technologies in isolation. They must build upon and help shape a common information infrastructure which they must share with scientific enterprises on the one hand and businesses on the other. The range of technologies is impressive and growing rapidly. In the United States, for example, the National Science Foundation is creating a National Science Digital Library (NSDL) aimed at distributing information to an audience ranging from elementary school through adult practitioners. The Digitally Literate Europe Action Plan shares similar goals. The US Defense Department, meanwhile, is supporting an initiative entitled *Translingual Information Detection, Extraction, and Summarization* (TIDES). TIDES is developing ways to apply a range of language technologies such as machine translation and information extraction to dozens of languages. Programs such as these provide a technological foundation that could revolutionize the position of cultural heritage languages in scholarship and in society as a whole.

Programs such as the NSDL and TIDES, though of potential benefit to humanists, are aimed at very different audiences. Scientists do not read journal articles in the same ways that humanists read Shakespeare or Dante. Intelligence analysts interpreting Albanian background materials during a crisis have needs that differ from those of the scholar reflecting intensively and at length over tenth century Icelandic culture. We in the humanities need to extend and generalize technologies designed for scientists, scholars, and diplomats if we are to enjoy their benefits.

We propose to concentrate in this project on the problems of integrating electronic text corpora and scholarly resources for cultural heritage languages. The overall problem is immense: digital libraries not only facilitate work on individual questions within existing discipline (e.g., a study of "guest friendship" in archaic Greece) but can also place on a new level our ability to compare cultures (e.g., Homeric Epics and Old Norse Sagas). For the purposes of this grant, we have elected to work with three languages: classical Greek, early modern Latin, and Old Norse. First, Classical Greek is an intensively studied language of world historical importance. One of our collaborative projects (The Perseus Project) has, over the past fourteen years, created a suite of tools for the study of classical Greek, including a corpus of major texts, two

lexica, grammars, an encyclopedia, dozens of commentaries, a morphological analyzer, and various document analysis tools. Classical Greek thus provides us with one well-established dataset on which we can build. Second, two of our collaborators (Istituto di Linguistica Computazionale del CNR and The Perseus Project) have created similar foundations for Classical Latin. In this grant, we choose to concentrate on the problems of early modern Latin — an immense corpus, far too large for conventional translation, unmanageable for those who are not fluent in idiosyncratic forms of Latin, but essential to the study of European literature, philosophy, science, and culture. Third, we have chosen to work on Old Norse, a language with an immense and rich literature to study the problems of boot-strapping a cultural heritage language into an already existing system.

Our focus for all three of these languages will be the creation of advanced digital library applications that (1) adapt computational linguistic and data mining techniques for the needs of students and scholars in the humanities, (2) establish an international framework for the long-term preservation of data, the sharing of metadata, and interoperability between affiliated digital libraries, and (3) lower the barriers to reading these texts.

The final result of this collaboration will be a suite of applications that includes multi-lingual information retrieval facilities; concept identification and visualization tools; vocabulary profiles; and a syntactic parsing toolbox with facilities for word sense and morphological disambiguation, and the resolution of attachment ambiguity. It will also include infrastructure-level programs that share data, metadata, and tools among affiliated digital libraries. This infrastructure will allow partner libraries to generate automatic hypertexts that link similar resources in different collections, federate their search facilities, and share resource-intensive programs. Finally, we will create or integrate new corpora of texts as testbeds for these applications. These corpora include approximately 300 MB (more than 60,000 printed pages) of literary and scientific early modern Latin texts, including many of Isaac Newton's papers, and 12 MB of Old Norse literature with many texts linked to manuscript images.

We have chosen the particular languages to work on based on the research interests and existing corpora of the collaborators. Because our project includes both humanists (literary scholars, linguists, historians of science) and computer scientists, and because some of our collaborators are also researchers with significant experience in the theory and practice of digital library systems, we are well placed to assess the needs of humanities scholars and the potential applications of information technology to these materials. We believe this collaboration among humanists, programmers, and humanist-programmers will result in a richer, more usable system than either humanists or digital library developers could create alone.

2 Contribution To Key Action Objectives

The objectives of our international consortium fulfill many of the goals of the Next Generation Digital Collections Action Line (IST 2001 III.1.3). The objectives that we address are:

- **Development of Advanced Digital Library Applications**: Digital libraries need to manage full content as well as metadata. We therefore proposal to integrate separate language technologies into a coherent Digital Library architecture. Thus we propose to create a syntactic parsing toolbox, vocabulary profiling, keyword and concept identification and visualization tools, and multi-lingual information retrieval facilities that all contribute to this objective.

- **Construction of New Visualization and Navigation Tools**: Several members of our consortium will be developing tools to work with page images of manuscripts and early modern printed books. We will also integrate many of our tools with clustering and visualization facilities.

- **International Collaboration and Metadata Sharing** International collaboration is one of the linchpins of our research group. We have members from five countries of the European Union and from the United States. The system that we propose with this grant will facilitate collaboration among scholars in many countries. It will also create a network of interoperable affiliated digital libraries in both the United States and Europe.

- **Creation of Thematically Coherent Collections**: The corpora that we will create as our testbeds consist of three thematically coherent collections of documents: Post-classical Latin literature, Renaissance and Early Modern scientific texts, and the corpus of Old Norse literature.

- **Delivery of Cultural and Scientific Content**: Our infrastructure is designed for wide dissemination and free access to any person with access to the internet. The collections that we propose to build are important both because of their cultural significance and because of their importance for the history of science.

- **Archiving and Preserving Data**: Our proposal for a collaborative, international digital library infrastructure involves the creation of texts that conform to published archival standards, a facility for the open exchange of metadata, and the storage of multiple copies of the data in many locations.

3 Innovations

We will bring advances in computer science and information technology to bear on the problems of studying documents that are important for understanding the European cultural tradition, but that are written in languages that few people fully understand. Our team consists of computer scientists, scholars in the humanities, and experts in the theory and practice of digital libraries. We are thus in a unique position to explore new technologies, assess their relevance, and deploy tools based on these technologies widely and for a general audience.

Many aspects of our proposed research expand on the state of the art, including the application of computational linguistics to the problems of cultural heritage languages and humanities computing; the creation of a working infrastructure for citation linking, searching, and resource discovery across digital libraries on an international scale; and the deployment of an infrastructure allowing wide dissemination and free public access to central documents of the European intellectual tradition in an integrated reading environment that lowers the barriers to reading these documents.

3.1 Computational Linguistics and the Digital Library

One of our key areas of innovation is our focus on integrating computational linguistics tools into the digital library. While a great deal of computer science work has been done in this area, very little of this work is ever reincorporated into digital libraries. If these tools are applied to texts in the digital library and made widely available to students and scholars, they can fundamentally transform the sorts of questions that people ask about these texts (see, for example, [4]).

The extension or development of morphological analysis facilities for early modern Latin and Old Norse is fundamental for these tools. It is not practical to use simple stemming techniques such as Porter's stemming algorithm for a highly inflected language such as Latin; a full parser is required if we wish to perform syntactic analysis and not just information retrieval. Two of our research groups (Istituto di Linguistica Computazionale del CNR and The Perseus Project) will modify an existing morphological analysis system for Classical Latin so that it can parse words from early modern Latin texts, while two other research groups (Arnamagnaean Institute and University of California at Los Angeles) will create similar morphological tools for Old Norse [2].

A parsing system will allow us to build search indexes so that users can search for all inflected instantiations of any given lexical form. This tool, along with a multi-lingual dictionary, can also serve as the basis for multi-lingual information retrieval facilities. One of our collaborators (University of Missouri at Kansas City) has done some preliminary work in this area, allowing users to enter queries in English to search Latin or Greek texts and automatically associating words with similar definitions in Greek and Latin. With this grant, we will develop these tools much more extensively to aid non-specialist users who want to work with the texts in our collections [18, 17, 21]. We will also work on innovative tools to cluster and visualize the

results of these searches. Traditionally, search results are largely unstructured, perhaps presented to the user simply as a ranked list. One of our collaborators (Imperial College) will develop a system that automatically identifies keywords, uses clustering algorithms to sort the repository subset into groups, labels those groups accordingly, visualizes them, and enables the user to focus on sub-clusters in the task of narrowing down a search (cf. [14, 24, 9, 6]). This strategy shifts the user's mental load from the slow thought-intensive process of reading lists to the faster perceptual process of pattern recognition in a visual display. With these techniques, one can study, and draw conclusions from, the change in use of a particular word within a context over time, and even hope to detect the emergence of new concepts [1, 3].

University of Missouri at Kansas City will do complementary work on applications to help users understand the vocabulary of the documents in the digital library. For example, a reader of a text written in an unfamiliar language such as Old Norse or Renaissance Latin could benefit from a display of all of the words in that text ranked according to frequency or another weighted score such as an inverted document frequency. Word usage profiles that integrate key words, the relative frequency of a word, and other information will also help users to understand more fully how a word is being used. In scientific documents, identifying and finding relationships among different technical terms can be particularly helpful [20, 12]. We will also develop tools to display aggregate information about the vocabulary in a document or a corpus by calculating — and clearly explaining — measures of lexical richness such as Yule's constant [23, 22].

Three research groups in the grant (Cambridge University, University of Missouri at Kansas City, and The Perseus Project) will work to develop an open syntactic parsing system for Greek and Latin, building a syntactic parsing toolbox with facilities for word sense and morphological disambiguation, resolution of attachment ambiguity, and the generation of parse trees. We will also work on programs that will allow us to understand the selectional preferences and subcategorization frames of the verbs in the texts our corpora. Scholars might find it very interesting, for example, to know the most common direct objects of very common verbs such as the Greek *tithêmi* (to establish or place) or the Latin *cano* (to sing). While algorithms for these functions are widely known for English and other modern languages, the differing word order and morphological conventions of Greek and Latin require modification of these algorithms as noted by [15, p. 381]. Collaborator CR 6 has had a great deal of success with these adaptations for collocation analysis and information retrieval programs and we are well positioned to build a toolset that has the potential to transform the ways that scholars work with documents written in these languages [17].

While these sorts of systems can produce a great deal of interesting Scholars studying a specific text or topic will always want to correct, annotate, or extend the results of automated analysis. Although digital libraries often allow simple annotation, their usual practice is to require users to make complex corrections outside of the digital library system. Every time a scholar does this, the new data thus created are lost to the system because there are no mechanisms for reintegrating these data. This situation is untenable; these data should be used to improve the performance of the system. Our workplan, therefore, allows for detailed evaluation and assessment of the results of these programs. Researchers at several of our collaborating in-

stitutions (Imperial College, Cambridge University, Arnamagnaean Institute, University of California at Los Angeles) will work intensively with these tools with the aim of creating or refining dictionaries and other linguistic reference works. This process will help the collaborators focusing on infrastructure (Imperial College, Cambridge University, University of Missouri at Kansas City, The Perseus Project, the Stoa Consortium) to improve these tools as they are being developed.

3.2 Collaborative Digital Libraries

A second area of innovation is collaborative digital libraries. A collaborative digital library environment involves co-operation among scholars and also among their software systems. We propose the creation of an infrastructure for collaboration based on well-known metadata sharing techniques and protocols to make disparate collections from various digital libraries work together as easily as the texts within a single library [16, 10].

A rudimentary version of this infrastructure is already taking shape. Two partner institutions (The Perseus Project and the Stoa Consortium) have deployed the basic technical infrastructure for federated searching, resource discovery, and linking. With support for this proposal, we will be able to make this platform more robust and allow it to link now disparate research projects in both Europe and the United States.

The technological platform for metadata sharing is the Open Archives protocol, developed by the Open Archives Initiative (www.openarchives.org) [13]. We propose to use the OAI protocol to share not only basic Dublin Core metadata, but also the more detailed metadata used in our digital library systems. For each text in a digital library, we must store catalog-level facts — title, author, and so on — as Dublin Core fields whose values are drawn from the header of the XML or SGML text. In addition to these fields, we add cataloging information about what abstract bibliographic object (ABO) this text instantiates or comments on (e.g., is it an edition of Homer's *Iliad*?) Thus, when a reader requests "Homer's *Iliad*," we can offer a choice among all available editions and translations of the text.

We are also able to supply notes from all available commentaries in all of the libraries in our consortium. These notes and citations are converted to hyperlinks. When the referring text is displayed, and the cited text is in the same digital library, we simply make a direct link. When the cited text is displayed, we offer the reverse citation as a comment or footnote. (This system is described in [18] and [21].) We propose to extend this mechanism to texts in co-operating digital libraries. To do this, we will augment the existing citation metadata with "location" or "provenance" information, then distribute the metadata using the standard OAI protocol. When we make a link to a remote text, we will use the location information to construct the hyperlink in the correct form.

Citations are the most obvious interconnections among texts, but other metadata also permit texts to be grouped or analyzed within a digital library group. Our metadata tables include all the dates and all

the toponyms referenced in each text, allowing for basic visualizations of the information in the text, in the form of timelines and maps. We can already, within a single digital library (The Perseus Project), plot all the dates or places in a group of texts. The same technique can be extended to a group of texts distributed throughout the library consortium.

3.3 An Electronic Reading Environment

Through our consortium, we will have several large testbeds of material available for these tools. Most of our partner institutions in this grant are already engaged in the process of building electronic corpora of texts. Some of our partners are entering and tagging texts, others will be adapting already existing electronic editions, and still others will be working on strategies for integrating page images of early modern books and manuscripts with the digitized text. These corpora include approximately 300 MB (more than 60,000 printed pages) of literary and scientific early modern Latin texts, including many of Isaac Newton's papers; and 12 MB of Old Norse texts with many texts linked to manuscript images.

All of these texts will be placed in an integrated reading environment that uses a morphological analysis program (developed by Istituto di Linguistica Computazionale del CNR, The Perseus Project, Arnamagnaean Institute, and University of California at Los Angeles) to generate hypertexts linking inflected forms in the text to lexical resources such as dictionaries and grammars. While developing a morphological analysis system is not technically innovative, very few existing projects deploy them in an integrated reading environment of this sort. The net result of this environment is a substantial reduction of the barriers to reading documents that are important for understanding the European cultural heritage, whether it be the literary heritage of the Old Norse Sagas or the early publications, working papers, letters, and notes of major scientific writers.

An essential feature of texts in this sort of integrated reading environment is the documentation of text encoding practices according to consistent standards that are open and available to any other scholarly group who might want to work with these texts. It is also essential to have stable archives for any texts themselves and their encoding standard so that they can be used in the future. (CR6 and CR7 have done work in editorial theory [5, 19]. We also follow and respect the models of [7], [8], and [11]). The collaborative infrastructure that we envision will provide exactly this sort of stable archive by replicating data at multiple institutions in many countries. At the same time, we anticipate taking advantage of already existing archives for our texts such as the Oxford Text Archive.

4 Workpackages and Deliverables

4.1 Introduction

Our consortium involves a large number of participants in many countries. Each of our participants has developed a workpackage that contributes to the larger whole. Each of the workpackages is intended to integrate with and extend an existing digital library infrastructure developed by one of our collaborators (The Perseus Project). The first deliverable of this grant will involve making this infrastructure available to all of the collaborators in this consortium. Once this infrastructure has been delivered, each workpackage can move forward in a parallel and complementary development process.

The greatest danger in our parallel development process is the creation of tools that work well on their own but are not interoperable with each other or the larger digital library system. To avoid this problem, constant communication between the participants is essential for our work; we plan annual collaborators meetings with the first meeting taking place as soon as possible after our work begins. At this meeting, we can deliver the infrastructure, review our workplans, and lay the foundations for successfully completing this project. Subsequent meetings will provide an opportunity for all collaborators to assess their work, describe their progress, and evaluate their results.

Our workpackages are organized around a series of advanced digital library applications. Each of these applications will be developed in parallel by different workgroups in our consortium but ultimately integrated back into a single integrated digital library system. Our tool creation methodology relies on the development of a robust indexing architecture that easily scales both across systems and to other languages. This will allow our programming teams to focus on developing interesting tools instead of dealing with different tagging systems on an ad hoc basis. These tools can function as 'modules' within the federated digital library system rather than stand-alone programs; this means we can apply these tools to every text that is added to the library without custom programming for every set of texts in the system.

Several of our workpackages involve the creation of corpora as testbeds for our digital library applications. Although collaborators at Imperial College, the Istituto di Linguistica Computazionale del CNR, the University of Missouri at Kansas City, The Perseus Project, and the Stoa Consortium all have access to existing corpora, the infrastructure that we are developing also allows us easily to integrate other texts at a very low cost per megabyte. In the end, the corpora that we create or integrate into our system will in themselves be substantial contributions to the digitization of the European cultural heritage. At the end of three years, we will have added to our system approximately 300 MB (more than 60,000 printed pages) of literary and scientific early modern Latin texts, including many of Isaac Newton's papers, and 12 MB of Old Norse literature with many texts linked to manuscript images.

Proper preservation of these materials is of great concern to all of the collaborators in our group. Accordingly, each collaborator in our group encodes both the layout and the content of their texts in SGML or XML following the standards of the Text Encoding Initiative. Our distributed library system will allow

for easy international duplication of our materials and include archiving facilities. We will plan to deposit many of the texts that we produce with an established European data repository such as the Oxford Text Archive.

As noted at the outset, our consortium intends to create advanced digital library applications that (1) adapt computational linguistic and data mining techniques for the needs of students and scholars in the humanities, (2) establish an international framework for the long-term preservation of data, the sharing of metadata, and interoperability between affiliated digital libraries, and (3) lower the barriers to reading these texts.

We have developed seven workpackages for our consortium that directly relate to these goals. Five of these workpackages are directed towards the creation of advanced applications and an integrated reading environment for these texts, one is intended to digitize new texts and integrate existing texts from outside collaborators, and one allows time for coordination of effort and reporting our progress.

4.2 Collaborative Infrastructure and Metadata Sharing

Workpackage WP 3: Imperial College, The Perseus Project, and the Stoa Consortium will work to deliver the basic infrastructure for metadata sharing and collaborative resource discovery including the establishment of naming rules for objects in the digital library group, generalization of an existing OAI data provider for use by all the co-operating projects, implementation of a metadata harvester, and incorporation of metadata from remote systems into a local DL repository. To leverage the resources collaboratively created in other parts of the grant, they also will implement cross-searching of the full content of our various repositories.

This workpackage provides the core infrastructure to which all other partners in the consortium will contribute. A functional version of this infrastructure already exists and it will be made immediately available to the other members of the consortium at the outset of our work. As the developments envisioned in this grant are implemented, the improvements will be propagated to each member of the group.

4.3 Advanced Digital Library Applications and an Integrated Reading Environment

Workpackage WP 1: In this workpackage, Imperial College will focus on tools for extracting key terms and phrases from document collections, clustering these phrases into related groups, and developing tools for visualizing these clusters. They will also focus on problems of determining when concepts emerge and change over time in a collection of thematically related documents.

Workpackage WP 2: In this workpackage, Cambridge University and the University of Missouri at Kansas City will work to develop multi-lingual information retrieval facilities for the digital library in addition to a series of tools that provide vocabulary profiles for texts and corpora within the digital library system and also tools for syntactic parsing of Greek and Latin texts.. These tools will also incorporate extensive

user feedback mechanisms to evaluate and improve the results of these programs.

Workpackage WP 5: The Istituto di Linguistica Computazionale del CNR will work to create a morphological analyzer for Renaissance and later Latin that can be used in the integrated reading environment of our digital library.

Workpackage WP 4: The Arnamagnaean Institute and the University of California at Los Angeles will similarly digitize lexical resources and write a morphological analyzer for Old Norse texts that will also be used to help students and scholars read and understand Old Norse texts. They will also edit a corpus of Old Norse literature and link the tagged texts to digitized manuscript images as a testbed for the integrated reading environment.

Both of these morphological analyzers will be built around existing standards for both input and output of the results so that they can serve as 'plug in' modules for the existing digital library infrastructure and participate in the data sharing models that we will develop with this grant.

4.4 Co-ordination and Assessment

Workpackage WP 7: Imperial College in conjunction with the University of Missouri at Kansas City will organize the efforts of each of the institutions participating in this consortium. These two institutions will be responsible for organizing annual meetings of the collaborators, submitting reports, and publishing accounts of our progress in scientific journals.

References

[1] Peter Au, Matthew Carey, Shalini Sewraz, Yike Guo, and Stefan M. Rüger. New paradigms in information visualization. In *Proceedings of the 23rd International ACM SIGIR Conference*, pages 307–309, 2000.

[2] A. Bozzi. Character recognition and the linguistic spelling checker: An integrated example. In A. Bozzi, editor, *Computer Aided Recovery and Analysis of Damaged Text Documents*, pages 161–186. CLUEB, Bologna, 2000.

[3] M. Carey, F. Kriwaczek, and S. M. Rüger. A visualization interface for document searching and browsing. In *Proceedings of NPIVM 2000*, Washington, D.C., 24-28 July 200. ACM Press.

[4] Gregory Crane. *The Blinded Eye: Thucydides and the New Written Word*. Rowman & Littlefield Publishers, Inc., 1996.

[5] Gregory Crane and Jeffrey A. Rydberg-Cox. New technology and new roles: the need for "corpus editors". In *Proceedings of the Fifth ACM Conference on Digital Libraries*, pages 252–253, San Antonio, Texas, June 2000.

[6] Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th Annual ACM SIGIR Conference*, pages 318–329, Copenhagen, 1992. ACM Press.

[7] LeeEllen Friedland, Nancy Kushigian, Christina Powell, David Seaman, Natalia Smith, and Perry Willett. *TEI Text Encoding in Libraries: Draft Guidelines for Best Encoding Practices*, 30 July 1999. See `http://www.indiana.edu/~letrs/tei/`.

[8] Charles Harvey and Jon Press. *Databases in Historical Research: Theory, Methods and Applications*. Macmillan Press, London, 1996.

[9] Marti Hearst and Jan O. Pedersen. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Proceedings of the 19th Annual ACM SIGIR Conference*, pages 76–84, Zurich, 1996. ACM Press.

[10] Steve Hitchcock, Les Carr, Zhuoan Jiao, Donna Bergmark, Wendy Hall, Carl Lagoze, and Stevan Harnad. Developing services for open eprint archives: Globalisation, integration and the impact of links. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, pages 143–151, San Antonio, Texas, June 2000.

[11] Susan M. Hockey. *Electronic Texts in the Humanities: Principles and Practice*. Oxford University Press, Oxford and New York, 2001.

[12] John S. Justeson and Slava M. Katz. Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Lanuage Engineering*, 1(1):9–27, 1995.

[13] Carl Lagoze and Herbert Van de Sompel. The Open Archives Initiative: Building a low-barrier interoperability framework. In *Proceedings of the First ACM+IEEE Joint Conference on Digital Libraries*, pages 54–62, 2001.

[14] Bjornar Larsen and Chinatsu Aone. Fast and effective text mining using linear-time document clustering. In *Proceedings of the 5th ACM SIGKDD Conference*, pages 16–22, San Diego, CA, 1999. ACM Press.

[15] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.

[16] Andreas Paepcke, Chen-Chuan K. Chang, Terry Winograd, and Héctor García-Molina. Interoperability for digital libraries worldwide. *Communications of the ACM*, 41(4):33–42, April 1998.

[17] Jeffrey A. Rydberg-Cox. Word co-occurrence and lexical acquisition in Ancient Greek texts. *Literary and Linguistic Computing*, 15(2):121–129, 2000.

[18] Jeffrey A. Rydberg-Cox, Robert F. Chavez, Anne Mahoney, David A. Smith, and Gregory R. Crane. Knowledge management in the Perseus digital library. *Ariadne*, 25, 2000. `http://www.ariadne.ac.uk/issue25/rydberg-cox/`.

[19] Jeffrey A. Rydberg-Cox, Anne Mahoney, and Gregory Crane. Document quality indicators and corpus editions. In *Proceedings of the First ACM + IEEE Joint Conference on Digital Libraries*, pages 435–436, Roanoke, VA, 24-28 June 2001.

[20] Bruce Schatz, William Mischo, Timothy Cole, Ann Bishop, Susan Harum, Eric Johnson, Laura Neumann, Hsinchun Chen, and Dorbin Ng. Federated search of scientific literature. *Computer*, 32(2):51–59, February 1999.

[21] David A. Smith, Anne Mahoney, and Jeffrey A. Rydberg-Cox. Management of XML documents in an integrated digital library. In *Proceedings of Extreme Markup Languages 2000*, pages 219–224, Montreal, August 2000.

[22] G. U. Yule. On sentence length as a statistical characteristic of style in prose with application to two cases of disputed authorship. *Biometrika*, 30:363–390, 1938.

[23] G. U. Yule. *The Statistical Study of Literary Vocabulary*. Cambridge University Press, Cambridge, 1944.

[24] Oren Zamir, Oren Etzioni, Omid Madani, and Richard M. Karp. Fast and intuitive clustering of web documents. In *Proceedings of the 3rd ACM SIGKDD Conference*, 1997.